Springer Proceedings in Mathematics & Statistics

Hervé Abdi
Vincenzo Esposito Vinzi
Giorgio Russolillo
Gilbert Saporta
Laura Trinchera *Editors*

The Multiple Facets of Partial Least Squares and Related Methods

PLS, Paris, France, 2014

Springer

Springer Proceedings in Mathematics & Statistics

Volume 173

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Hervé Abdi • Vincenzo Esposito Vinzi
Giorgio Russolillo • Gilbert Saporta
Laura Trinchera
Editors

The Multiple Facets of Partial Least Squares and Related Methods

PLS, Paris, France, 2014

Springer

Editors
Hervé Abdi
School of Behavioral and Brain Sciences
The University of Texas at Dallas
Richardson, TX, USA

Giorgio Russolillo
CNAM
Paris, USA

Laura Trinchera
NEOMA Business School
Rouen, France

Vincenzo Esposito Vinzi
ESSEC Business School
Cergy Pontoise CX, France

Gilbert Saporta
CNAM
Paris Cedex 03, France

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-40641-1 ISBN 978-3-319-40643-5 (eBook)
DOI 10.1007/978-3-319-40643-5

Library of Congress Control Number: 2016950729

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

In 1999, the first meeting dedicated to partial least squares methods (abbreviated as PLS and also, sometimes, expanded as *projection to latent structures*) took place in Paris. Other meetings in this series took place in various cities all over the world, and in 2014, from the 26th to the 28th of May, the eighth meeting of the partial least squares (PLS) series returned to Paris to be hosted in the beautiful building of the Conservatoire National des Arts et Métiers (CNAM) under the double patronage of the Conservatoire National des Arts et Métiers and the ESSEC Paris Business School. This venue was again a superb success with more than 250 authors presenting more than one hundred papers during these 3 days. These contributions were all very impressive by their quality and by their breadth. They covered the multiple dimensions and facets of partial least squares-based methods, ranging from partial least squares regression and correlation to component-based path modeling, regularized regression, and subspace visualization. In addition, several of these papers presented exciting new theoretical developments. This diversity was also expressed in the large number of domains of application presented in these papers such as brain imaging, genomics, chemometrics, marketing, management, and information systems to name only a few.

After the conference, we decided that a large number of the papers presented in the meeting were of such an impressive high quality and originality that they deserved to be made available to a wider audience, and we asked the authors of the best papers if they would like to prepare a revised version of their paper. Most of the authors contacted shared our enthusiasm, and the papers that they submitted were then read and commented on by anonymous reviewers, revised, and finally edited for inclusion in this volume; in addition, Professor Takane (who could not join us for the meeting) accepted to contribute a chapter for this volume. These papers included in *The Multiple Facets of Partial Least Squares and Related Methods* provide a comprehensive overview of the current state of the most advanced research related to PLS and cover all domains of PLS and related domains.

Each paper was overviewed by one editor who took charge of having the paper reviewed and edited (Hervé was in charge of the papers of Beaton et al., Churchill et al., Cunningham et al., El Hadri and Hanafi, Eslami et al., Löfstedt et al., Takane

and Loisel, and Zhou et al.; Vincenzo was in charge of the paper of Kessous et al.; Giorgio was in charge of the papers of Boulesteix, Bry et al., Davino et al., and Cantaluppi and Boari; Gilbert was in charge of the papers of Blazère et al., Bühlmann, Lechuga et al., Magnanensi et al., and Wang and Huang; Laura was in charge of the papers of Aluja et al., Chin et al., Davino et al., Dolce et al., and Romano and Palumbo). The final production of the LaTeXversion of the book was mostly the work of Hervé, Giorgio, and Laura. We are also particularly grateful to our (anonymous) reviewers for their help and dedication.

Finally, this meeting would not have been possible without the generosity, help, and dedication of several persons, and we would like to specifically thank the members of the scientific committee: Michel Béra, Wynne Chin, Christian Derquenne, Alfred Hero, Heungsung Hwang, Nicole Kraemer, George Marcoulides, Tormod Næs, Mostafa Qannari, Michel Tenenhaus, and Huiwen Wang. We would like also to thank the members of the local organizing committee: Jean-Pierre Choulet, Anatoli Colicev, Christiane Guinot, Anne-Laure Hecquet, Emmanuel Jakobowicz, Ndeye Niang Keita, Béatrice Richard, Arthur Tenenhaus, and Samuel Vinet.

Dallas/Paris
April 2016

Hervé Abdi
Vincenzo Esposito Vinzi
Giorgio Russolillo
Gilbert Saporta
Laura Trinchera

Contents

Part I Keynotes

1 Partial Least Squares for Heterogeneous Data 3
Peter Bühlmann

2 On the PLS Algorithm for Multiple Regression (PLS1) 17
Yoshio Takane and Sébastien Loisel

3 Extending the Finite Iterative Method for Computing the Covariance Matrix Implied by a Recursive Path Model 29
Zouhair El Hadri and Mohamed Hanafi

4 Which Resampling-Based Error Estimator for Benchmark Studies? A Power Analysis with Application to PLS-LDA 45
Anne-Laure Boulesteix

5 Path Directions Incoherence in PLS Path Modeling: A Prediction-Oriented Solution 59
Pasquale Dolce, Vincenzo Esposito Vinzi, and Carlo Lauro

Part II New Developments in Genomics and Brain Imaging

6 Imaging Genetics with Partial Least Squares for Mixed-Data Types (MiMoPLS) 73
Derek Beaton, Michael Kriegsman, ADNI, Joseph Dunlop, Francesca M. Filbey, and Hervé Abdi

7 PLS and Functional Neuroimaging: Bias and Detection Power Across Different Resampling Schemes 93
Nathan Churchill, Babak Afshin-Pour, and Stephen Strother

8 Estimating and Correcting Optimism Bias in Multivariate PLS Regression: Application to the Study of the Association Between Single Nucleotide Polymorphisms and Multivariate Traits in Attention Deficit Hyperactivity Disorder 103
Erica Cunningham, Antonio Ciampi, Ridha Joober, and Aurélie Labbe

9 Discriminant Analysis for Multiway Data 115
Gisela Lechuga, Laurent Le Brusquet, Vincent Perlbarg, Louis Puybasset, Damien Galanaud, and Arthur Tenenhaus

Part III New and Alternative Methods for Multitable and Path Analysis

10 Structured Variable Selection for Regularized Generalized Canonical Correlation Analysis 129
Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe, Edouard Duchesnay, Vincent Frouin, and Arthur Tenenhaus

11 Supervised Component Generalized Linear Regression with Multiple Explanatory Blocks: THEME-SCGLR 141
Xavier Bry, Catherine Trottier, Fréderic Mortier, Guillaume Cornu, and Thomas Verron

12 Partial Possibilistic Regression Path Modeling 155
Rosaria Romano and Francesco Palumbo

13 Assessment and Validation in Quantile Composite-Based Path Modeling 169
Cristina Davino, Vincenzo Esposito Vinzi, and Pasquale Dolce

Part IV Advances in Partial Least Square Regression

14 PLS-Frailty Model for Cancer Survival Analysis Based on Gene Expression Profiles 189
Yi Zhou, Yanan Zhu, and Siu-wai Leung

15 Functional Linear Regression Analysis Based on Partial Least Squares and Its Application 201
Huiwen Wang and Lele Huang

16 Multiblock and Multigroup PLS: Application to Study Cannabis Consumption in Thirteen European Countries 213
Aida Eslami, El Mostafa Qannari, Stéphane Legleye, and Stéphanie Bougeard

17 A Unified Framework to Study the Properties of the PLS Vector of Regression Coefficients 227
Mélanie Blazère, Fabrice Gamboa, and Jean-Michel Loubes

18 A New Bootstrap-Based Stopping Criterion in PLS Components Construction 239
Jérémy Magnanensi, Myriam Maumy-Bertrand, Nicolas Meyer, and Frédéric Bertrand

Part V PLS Path Modeling: Breakthroughs and Applications

19 Extension to the PATHMOX Approach to Detect Which Constructs Differentiate Segments and to Test Factor Invariance: Application to Mental Health Data 253
Tomas Aluja-Banet, Giuseppe Lamberti, and Antonio Ciampi

20 Multi-group Invariance Testing: An Illustrative Comparison of PLS Permutation and Covariance-Based SEM Invariance Analysis 267
Wynne W. Chin, Annette M. Mills, Douglas J. Steel, and Andrew Schwarz

21 Brand Nostalgia and Consumers' Relationships to Luxury Brands: A Continuous and Categorical Moderated Mediation Approach 285
Aurélie Kessous, Fanny Magnoni, and Pierre Valette-Florence

22 A Partial Least Squares Algorithm Handling Ordinal Variables 295
Gabriele Cantaluppi and Giuseppe Boari

Author Index 307

Subject Index 313

List of Contributors

Hervé Abdi School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Babak Afshin-Pour Rotman Research Institute, Baycrest Hospital, Toronto, ON, Canada

Tomas Aluja-Banet Universitat Politecnica de Catalunya, Barcelona Tech, Barcelona, Spain

Derek Beaton School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Frédéric Bertrand Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, Strasbourg Cedex, France

Mélanie Blazère Institut de mathématiques de Toulouse, Toulouse, France

Giuseppe Boari Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy

Stéphanie Bougeard Department of Epidemiology, French agency for food, environmental and occupational health safety (Anses), Ploufragan, France

Anne-Laure Boulesteix Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Munich, Germany

Laurent Le Brusquet Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506), CentraleSupélec-CNRS-Université Paris-Sud, Paris, France

Xavier Bry Institut Montpelliérain Alexander Grothendieck, UM2, Place Eugène, Bataillon CC 051 - 34095 Montpellier, France

Peter Bühlmann Seminar for Statistics, ETH Zurich, Zürich, Switzerland

Gabriele Cantaluppi Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy

Wynne W. Chin Department of Decision and Information Systems, C.T. Bauer College of Business, University of Houston, Houston, TX, USA

Nathan Churchill Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada

Antonio Ciampi Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, QC, Canada

Guillaume Cornu Cirad, UR Biens et Services des Ecosystèmes Forestiers tropicaux, Campus International de Baillarguet, Montpellier, France

Erica Cunningham Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

Cristina Davino University of Macerata, Macerata, Italy

Pasquale Dolce University of Naples "Federico II", Naples, Italy

Edouard Duchesnay NeuroSpin, CEA Saclay, Gif-sur-Yvette, France

Joseph Dunlop SAS Institute Inc, Cary, NC, USA

Zouhair El Hadri Faculté des Sciences, Département de Mathématiques, Université Ibn Tofail, Equipe de Cryptographie et de Traitement de l'Information, Kénitra, Maroc

Aida Eslami LUNAM University, ONIRIS, USC Sensometrics and Chemometrics Laboratory, Rue de la Géraudière, Nantes, France

Vincenzo Esposito Vinzi ESSEC Business School, Cergy Pontoise Cedex, France

Francesca M. Filbey School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Vincent Frouin NeuroSpin, CEA Saclay, Gif-sur-Yvette, France

Damien Galanaud Department of Neuroradiology, AP-HP, Pitié-Salpêtrière Hospital, Paris, France

Fabrice Gamboa Institut de mathématiques de Toulouse, Toulouse, France

Vincent Guillemot Bioinformatics/Biostatistics Core Facility, IHU-A-ICM, Brain and Spine Institute, Paris, France

Fouad Hadj-Selem NeuroSpin, CEA Saclay, Gif-sur-Yvette, France

Mohamed Hanafi Oniris, Unité de Sensométrie et Chimiométrie, Sensometrics and Chemometrics Laboratory, Nantes, France

Lele Huang School of Economics and Management, Beihang University, Beijing, China

Ridha Joober Douglas Mental Health University Institute, Verdun, QC, Canada

Aurélie Kessous CERGAM, Faculté d'Economie et de Gestion, Aix-Marseille Université, Marseille, France

Michael Kriegsman School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Giuseppe Lamberti Universitat Politecnica de Catalunya, Barcelona Tech, Barcelona, Spain

Aurélie Labbe Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada

Carlo Lauro University of Naples "Federico II", Naples, Italy

Gisela Lechuga Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506), CentraleSupélec-CNRS-Université Paris-Sud, Paris, France

Siu-wai Leung State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

School of Informatics, University of Edinburgh, Edinburgh, UK

Tommy Löfstedt Computational Life Science Cluster (CLiC), Department of Chemistry, Umeå University, Umeå, Sweden

Sébastien Loisel Heriot-Watt University, Edinburgh, UK

Jean-Michel Loubes Institut de mathématiques de Toulouse, Toulouse, France

Jérémy Magnanensi Institut de Recherche Mathématique Avancée, UMR 7501, LabEx IRMIA, Université de Strasbourg et CNRS, Strasbourg Cedex, France

Fanny Magnoni CERAG, IAE Grenoble Pierre Mendès France University, Grenoble, France

Myriam Maumy-Bertrand Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, Strasbourg Cedex, France

Nicolas Meyer Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430, Université de Strasbourg, Strasbourg Cedex, France

Annette M. Mills Department of Accounting and Information Systems, College of Business and Economics, University of Canterbury, Ilam Christchurch, New Zealand

Fréderic Mortier Cirad – UR Biens et Services des Ecosystèmes Forestiers tropicaux, Montpellier, France

Francesco Palumbo University of Naples Federico II, Naples, Italy

Vincent Perlbarg Bioinformatics/Biostatistics Platform IHU-A-ICM, Brain and Spine Institute, Paris, France

Cathy Philippe Gustave Roussy, Villejuif, France

Louis Puybasset AP-HP, Surgical Neuro-Intensive Care Unit, Pitié-Salpêtrière Hospital, Paris, France

El Mostafa Qannari LUNAM University, ONIRIS, USC Sensometrics and Chemometrics Laboratory, Rue de la Géraudière, Nantes, France

Rosaria Romano University of Calabria, Cosenza, Italy

Giorgio Russolillo Conservatoire National des Arts et Métiers, Paris, France

Gilbert Saporta Conservatoire National des Arts et Métiers, Paris, France

Andrew Schwarz Louisiana State University, Baton Rouge LA, USA

Douglas J. Steel School of Business, Department of Management Information Systems, University of Houston-Clear Lake, Houston, TX, USA

Stephen Strother Rotman Research Institute, Baycrest Hospital, Toronto, ON, Canada

Yoshio Takane University of Victoria, Victoria, BC, Canada

Arthur Tenenhaus Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506), CentraleSupélec-CNRS-Université Paris-Sud and Bioinformatics/ Biostatistics Platform IHU-A-ICM, Brain and Spine Institute, Paris, France

The Alzheimer's Disease Neuroimaging Initiative (ADNI) http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Laura Trinchera NEOMA Business School, Rouen, France

Catherine Trottier Université Montpellier 3, Montpellier, France

Pierre Valette-Florence CERAG, IAE Grenoble, Université Grenoble Alpes, Grenoble, France

Thomas Verron ITG-SEITA, Centre de recherche SCR, Fleury-les-Aubrais, France

Huiwen Wang School of Economics and Management, Beihang University, Beijing, China

Yi Zhou State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

Yanan Zhu State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

Part I
Keynotes

Chapter 1
Partial Least Squares for Heterogeneous Data

Peter Bühlmann

Abstract Large-scale data, where the sample size and the dimension are high, often exhibits heterogeneity. This can arise for example in the form of unknown subgroups or clusters, batch effects or contaminated samples. Ignoring these issues would often lead to poor prediction and estimation. We advocate the maximin effects framework (Meinshausen and Bühlmann, Maximin effects in inhomogeneous large-scale data. Preprint arXiv:1406.0596, 2014) to address the problem of heterogeneous data. In combination with partial least squares (PLS) regression, we obtain a new PLS procedure which is robust and tailored for large-scale heterogeneous data. A small empirical study complements our exposition of new PLS methodology.

Keywords Partial least square regression (PLSR) • Heterogeneous data • Big data • Minimax • Maximin

1.1 Introduction

Large-scale complex data, where the the total sample size n and the number of variables p (i.e., the "dimension") are large, arise in many areas in science. For the case with high dimensions, regularized estimation schemes have become popular and are well-established (cf. Hastie et al. 2009; Bühlmann and van de Geer 2011). Partial least squares (PLS) (Wold 1966) is an interesting procedure and is widely used in many applications: besides good prediction performance, with its "vague similarity" to Ridge regression (Frank and Friedman 1993), and usefulness for dimensionality reduction, it is computationally attractive for large-scale problems as it operates in an iterative fashion based on empirical covariances only (Geladi and Kowalski 1986; Esposito Vinzi et al. 2010).

When the total sample size n is large, as in "big data" problems, we typically expect that the observations are heterogeneous and not i.i.d. or stationary realizations from a single probability distribution. Ignoring such heterogeneity

P. Bühlmann (✉)
Seminar for Statistics, ETH Zurich, Zürich, Switzerland
e-mail: buhlmann@stat.math.ethz.ch

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_1

(e.g., unknown subpopulations, batch and clustering effects, or outliers) is likely to produce poor predictions and estimation. Classical approaches to address these issues include robust methods (Huber 2011), varying coefficient models (Hastie and Tibshirani 1993), mixed effects models (Pinheiro and Bates 2000) or mixture models (McLachlan and Peel 2004). Mostly for computational reasons with large-scale data, we aim for methods which are computationally efficient with a structure allowing for simple parallel processing. This can be achieved with a so-called maximin effects approach (Meinshausen and Bühlmann 2015) and its corresponding subsampling and aggregation "magging" procedure (Bühlmann and Meinshausen 2016). As we will discuss, the computational efficiency of partial least squares together with the recently proposed maximin effects framework leads to a new and robust PLS scheme for regression which is appropriate for heterogeneous data.

To get a more concrete idea about (some form of) inhomogeneity in the data, we focus next on a specific model.

1.1.1 A Mixture Regression Model for Heterogeneous Data

In the sequel we focus on the setting of regression but allowing for inhomogeneous data. We consider the framework of a mixture regression model

$$Y_i = X_i^T B_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where Y_i is a univariate response variable, X_i is a p-dimensional covariable, B_i is a p-dimensional regression parameter, and ε_i is a stochastic noise term with mean zero and which is independent of the (fixed or random) covariable. Some inhomogeneity occurs because, in principle, every observation with index i can have its own and different regression parameter B_i, arising from a different mixture component. The model in (1.1) is often too general: we make the assumption that the regression parameters $B_1, \ldots, B_n$ are realizations from a distribution F_B:

$$B_i \sim F_B,\ i = 1, \ldots, n, \tag{1.2}$$

where the B_i's do not need to be independent of each other. However, we assume that the B_i's are independent from the X_i's and ε_i's.

Example 1 (known groups). Consider the case where there are G known groups $\mathscr{G}_g$ ($g = 1, \ldots, G$) with $B_i \equiv b_g$ for all $i \in \mathscr{G}_g$. Thus, this is a clusterwise regression problem (with *known* clusters) where every group $\mathscr{G}_g$ has the same (unknown) regression parameter vector b_g.

Example 2 (smoothly changing structure). Consider the situation where there is a changing behavior of the B_i's with respect to the sample indices i: this can be achieved by positive correlation among the B_i's. In practice, the sample index often corresponds to time.

Example 3 (unknown groups). This is the same setting as in Example 1 but the groups $\mathscr{G}_g$ are unknown. From an estimation point of view, there is a substantial difference to Example 1 (Meinshausen and Bühlmann 2015).

1.2 Magging: Maximin Aggregation

We consider the framework of grouping or subsampling the entire data-set, followed by an aggregation of subsampled regression estimators. A prominent example is Breiman's bagging method (Breiman 1996) which has been theoretically shown to be powerful with homogeneous data (Bühlmann and Yu 2002; Hall and Samworth 2005). We denote the subsamples or subgroups by

$$\mathscr{G}_g \subset \{1, \ldots, n\},\ g = 1, \ldots, G, \tag{1.3}$$

where $\{1, \ldots, n\}$ are the indices of the observations in the sample. We implicitly assume that they are "approximately homogeneous" subsamples of the data. Constructions of such subsamples are described in Sect. 1.2.2.

Magging (Bühlmann and Meinshausen 2016) is an aggregation scheme of subsampled estimators, designed for heterogeneous data. The wording stands for **m**aximin **agg**regat**ing**, and the maximin framework is described below in Sect. 1.2.1. We compute a regression estimator $\hat{\theta}_g$ for each subsample $\mathscr{G}_g,\ g = 1, \ldots, G$:

$$\hat{b}_1, \ldots, \hat{b}_G.$$

The choice of the estimator is not important for the moment. Concrete examples include ordinary least squares or regularized versions thereof such as Ridge regression (Hoerl and Kennard 1970) or the LASSO (Tibshirani 1996), and we will consider partial least squares regression in Sect. 1.3. We aggregate these estimates to a single p-dimensional parameter estimate. More precisely, we build a convex combination

$$\hat{b}_{\text{magging}} = \sum_{g=1}^{G} \hat{w}_g \hat{b}_g,\ \hat{w}_g \geq 0,\ \sum_{g=1}^{G} \hat{w}_g = 1, \tag{1.4}$$

where the convex combination weights are given from the following quadratic optimization. Denote by $H = [\hat{b}_1, \ldots, \hat{b}_G]^T \hat{\Sigma} [\hat{b}_1, \ldots, \hat{b}_G]$ the $G \times G$ matrix, where $\hat{\Sigma} = X^T X/n$ is the empirical Gram- or covariance (if the mean equals zero) matrix of the entire $n \times p$ design matrix X containing the covariates. Then:

$$\begin{aligned} &\hat{w}_1, \ldots, \hat{w}_G = \text{argmin}_w w^T (H + \gamma I_{G \times G}) w, \\ &\text{subject to } w_g \geq 0,\ \sum_{g=1}^{G} w_g = 1, \end{aligned} \tag{1.5}$$

where $\gamma = 0$ if H is positive definite which is typically the case if $G < n$; and otherwise, $\gamma > 0$ is chosen small such as 0.05, making $(H + \gamma I_{G\times G})$ positive definite (and in the limit for $\gamma \searrow 0^+$, we obtain the solution $\hat{w}$ with minimal squared error norm $\|.\|^2$).

Computational implementation. Magging is computationally feasible for large-scale data. The computation of $\hat{b}_g$ can be done in parallel, and the convex aggregation step involves a typically low-dimensional (as G is typically small) quadratic program only. An implementation in the R-software environment (R Core Team 2014) looks as follows.

```
library(quadprog)
hatb <- cbind(hatb1,...,hatbG)
 #matrix with G columns:
 #each column is a regression estimate
hatS <- t(X) %*% X/n
 #empirical covariance matrix of X
H <- t(hatb) %*% hatS %*% hatb
 #assume that it is positive definite
 #(use H + xi * I, xi > 0 small, otherwise)
A <- rbind(rep(1,G),diag(1,G))
 #constraints
b <- c(1,rep(0,G))
d <- rep(0,G)
 #linear term is zero
w <- solve.QP(H,d,t(A),b, meq = 1)
 #quadratic programming solution to
 #argmin(x^T H x) such that Ax >= b and
 #first inequality is an equality
```

1.2.1 The Maximin Effects Parameter

The magging aggregation scheme in (1.4) is estimating the so-called maximin parameter. To explain the concept, consider a linear model as in (1.1) but now with the fixed p-dimensional regression parameter b which can take values in the support of F_B from (1.2):

$$Y_i = X_i^T b + \varepsilon_i,\ i = 1, \ldots, n, \tag{1.6}$$

where X_i and ε_i are as in (1.1) and assumed to be i.i.d. The variance which is explained by choosing a parameter vector β in the linear model (1.6) is

$$V_{\beta,b} := \mathbb{E}|Y|^2 - \mathbb{E}|Y - X^T\beta|^2 = 2\beta^T \Sigma b - \beta^T \Sigma \beta,$$

where Σ denotes the covariance matrix of X_i. We aim for maximizing the explained variance in the worst (or most adversarial) situation. This leads to the definition of the maximin effects.

Definition (Meinshausen and Bühlmann 2015). The maximin effects parameter is

$$b_{\text{maximin}} = \arg\max_{\beta} \min_{b \in \text{supp}(F_B)} V_{\beta,b}.$$

The name "maximin" comes from the fact that we consider "maximization" of a "minimum", that is, optimizing on the worst case.[1]

The maximin effects can be interpreted as an aggregation among the support points of F_B to a single parameter vector (i.e., among all the B_i's, as, e.g., in Example 2 in Sect. 1.1.1) or among all the clustered values b_g (e.g., in Examples 1 and 3 in Sect. 1.1.1), see also Fact 1 below. The maximin effects parameter is different from the pooled effects

$$b_{\text{pool}} = \arg\min_{\beta} \mathbb{E}_B[-V_{\beta,B}]$$

which is the population analogue when considering the data as homogeneous. Maybe surprisingly, the maximin effects are also different from the prediction analogue

$$b_{\text{pred-maximin}} = \arg\min_{\beta} \max_{b \in \text{supp}(F_B)} \mathbb{E}[(X^T b - X^T \beta)^2].$$

In particular, the value zero has a special status for the maximin effects parameter b_{maximin}, unlike for $b_{\text{pred-maximin}}$ or b_{pool}, (see Meinshausen and Bühlmann 2015). The following is an important "geometric" characterization which indicates the special status of the value zero.

Fact 1. *Meinshausen and Bühlmann (2015) Let $\mathscr{H}$ be the convex hull of the support of F_B. Then*

$$b_{\text{maximin}} = \arg\min_{\gamma \in \mathscr{H}} \gamma^T \Sigma \gamma.$$

That is, the maximin effects parameter b_{maximin} is the point in the convex hull $\mathscr{H}$ which is closest to zero with respect to the distance $d(u, v) = (u - v)^T \Sigma (u - v)$. In particular, if the value zero is in $\mathscr{H}$, the maximin effects parameter equals $b_{\text{maximin}} \equiv 0$.

The characterization in Fact 1 leads to an interesting robustness property. If the support of F_B is enlarged, e.g. by adding additional heterogeneity to the model, there are two possibilities: either, (i) the maximin effects parameter b_{maximin} does not change; or (ii) if it changes, it moves closer to the value zero because the convex

[1] In game theory and mathematical statistics, the terminology "minimax" is more common. To distinguish, and to avoid confusion from statistical minimax theory, Meinshausen and Bühlmann (2015) have used the reverse terminology "maximin".

hull is enlarged and invoking Fact 1. Therefore, the maximin effects parameter and its estimation exhibit an excellent robustness feature with respect to breakdown properties: an arbitrary new support point in F_B (i.e., a new sample point with a new value of the regression parameter) cannot shift b_{maximin} away from zero. We will exploit this robustness property in an empirical simulation study in Sect. 1.3.3.

Magging as described above in (1.4)–(1.5) turns out to be a reasonably good estimator for the maximin effects parameter b_{maximin}. This is not immediately obvious but a plausible explanation is given by Fact 1 as follows. For the setting of Example 1 in Sect. 1.1.1, that is with known groups $\mathcal{G}_g$ each having its regression parameter b_g, the maximin effects parameter is the point in the convex hull which is closest to zero. This can be characterized by

$$b_{\text{maximin}} = \sum_{g=1}^{G} w_g^0 b_g, \ \sum_g w_g^0 = 1$$

where the weights w_g^0 are the population analogue of the optimal weights in (1.5) (i.e., with b_g instead of $\hat{b}_g$ and Σ instead of $\hat{\Sigma}$). Thus, the magging estimator is of the same form as b_{maximin} but plugging in the estimated quantities instead of the true underlying parameters b_g $(g = 1, \ldots, G)$ and Σ.

1.2.1.1 Interpretation of the Maximin Effects

An estimate of the maximin effects b_{maximin} should be interpreted according to the parameter's meaning. The definition of the parameter implies that b_{maximin} is optimizing the explained variance under the worst case scenario among all possible values from the support of the distribution F_B in the mixture model (1.1). Furthermore, Fact 1 provides an interesting geometric characterization of the parameter.

Loosely speaking, the maximin effects parameter b_{maximin} describes the "common" effects of the covariates to the response variable in the following sense. If a covariable has a strong influence among all possible regression values from the support of F_B in model (1.1), then the corresponding component of b_{maximin} is large in absolute value; vice-versa, if the effect of a covariable is not common to all the possible values in the support of F_B, then the corresponding component of b_{maximin} is zero or close to zero.

In terms of prediction, the maximin effects parameter is typically leading to enhanced prediction of future observations in comparison to the pooled effect b_{pool}, whenever the future observations are generated from a regression model with parameter from the support of F_B. In particular, the prediction is "robust" and not mis-guided by a few or a group of outliers. Some illustrations of this behavior on real financial data are given in Meinshausen and Bühlmann (2015).

1.2.2 Construction of Groups and Sampling Schemes for Maximin Aggregation

The magging scheme relies on groups or subsamples $\mathscr{G}_g$ $(g = 1, \ldots, G)$. Their construction is discussed next.

1.2.2.1 Known Groups

As in Example 1 in Sect. 1.1.1, consider the situation where we have J known groups of homogeneous data. That is, the sample index space has a partition

$$\begin{aligned} &\mathscr{J}_1, \ldots, \mathscr{J}_J, \quad \mathscr{J}_j \subset \{1, \ldots, n\}, \\ &\cup_{j=1}^{J} \mathscr{J}_j = \{1, \ldots, n\}, \quad \mathscr{J}_j \cup \mathscr{J}_k = \emptyset \ (j \neq k) \end{aligned}$$

where

$$B_i \equiv b_j \text{ for all } i \in \mathscr{J}_j.$$

For such a scenario, we deterministically subsample the data corresponding to the known groups[2]:

$$\begin{aligned} &\mathscr{G}_1, \ldots, \mathscr{G}_G, \\ &\text{where } G = J \text{ and } \mathscr{G}_g = \mathscr{J}_g \text{ for all } g = 1, \ldots, G. \end{aligned} \tag{1.7}$$

1.2.2.2 Smoothly Changing Structure

As in Example 2 in Sect. 1.1.1, consider the situation where there is a smoothly changing behavior of the B_i's with respect to the sample indices i. This can be achieved by positive correlation among the B_i's. In practice, the sample index often corresponds to time. There are no true (unknown) groups in this setting.

In some applications, the samples are collected over time, as mentioned in Example 2. For such situations, we construct:

$$\begin{aligned} &\text{disjoint groups } \mathscr{G}_g \ (g = 1, \ldots, G), \text{ where each } \mathscr{G}_g \text{ is a} \\ &\text{block of } \textit{consecutive} \text{ observations of (usually) the same size } m. \end{aligned} \tag{1.8}$$

[2] We distinguish notationally the true (known) groups $\mathscr{J}_j$ from the sampled groups $\mathscr{G}_g$, although here for this case, they coincide exactly. For other cases though, the sampled groups do not necessarily correspond to true groups.

The group size m is a tuning parameter which needs to be chosen: a reasonable guidance is to choose m as a fraction of n such that the resulting $G = n/m$ is rather small (e.g. in the range of $G \in [3, 10]$). From a theoretical perspective, Meinshausen and Bühlmann (2015) provide some arguments leading to asymptotic consistency for b_{maximin}. Note that the true underlying structure has no strictly defined groups while the estimator does.

1.2.2.3 Without Structural Knowledge

Corresponding to Example 3 in Sect. 1.1.1, consider the case where the groups are unknown. We then construct G groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ where each $\mathcal{G}_g \subset \{1, \ldots, n\}$ encodes a subsample of the data, and these subsample do not need to be disjoint. A concrete subsampling scheme is as follows:

$$\begin{aligned} &\text{for each group } \mathcal{G}_g \ (g = 1, \ldots, G)\text{: subsample } m \text{ data points without replacement,} \\ &\text{while subsampling between groups is with replacement.} \end{aligned} \tag{1.9}$$

The number of groups G and the group size m are tuning parameters which need to be specified. A useful guideline is to choose m reasonably large (e.g., $m = f \cdot n$ with $f \in [0.2, 0.5]$) and G not too large (e.g. $G \in [3, 10]$). Some theoretical considerations leading to consistency for b_{maximin} are given in Meinshausen and Bühlmann (2015).

1.2.2.4 Contaminated Samples

An interesting special case occurs with outliers and associated robust inference. There are unknown groups of the entire sample: one large group with "clean" observations, all having the true regression parameter b_{true} and many singleton groups of size 1 each having its own contaminated regression parameter. We would then sample subgroups as in (1.9) and use these subsampled groups for the magging scheme (1.4)–(1.5). Interestingly, magging then becomes a robust method for estimating the true regression parameter b_{true} (Meinshausen and Bühlmann 2015; Bühlmann and Meinshausen 2016), and we will also illustrate this fact in Sect. 1.3.3.

1.3 A PLS Algorithm for Heterogeneous Data

The use of magging in (1.4) for PLS in a regression setting is straightforward. The subsampled estimators $\hat{b}_g = \hat{b}_{\text{PLS},g}$ are now from PLS regression with a specified number of components (and the number of components can vary for different subsamples $\mathcal{G}_g$); the construction of the groups used in magging is as in Sect. 1.2.2, depending on the situation whether we have known or unknown subpopulations, or whether there is an underlying smoothly changing trend. The obtained aggregated magging estimator is denoted by $\hat{b}_{\text{PLS}-\text{magging}}$.

1.3.1 PLS for Heterogeneous Data

The estimated parameter $\hat{b}_{\text{PLS}-\text{magging}}$ itself can serve as an appropriate value of the maximin effects regression parameter. In addition, we might want a more genuine PLS estimate with all its usual output. This can be easily obtained by running a standard PLS regression on the noise free entire data where we replace the response variable Y by the fitted values $X\hat{b}_{\text{PLS}-\text{magging}}$ and using the covariables from the entire original design matrix X. The output of such an additional standard PLS regression yields orthogonal linear combinations of the covariables and the corresponding obtained PLS regression coefficients are typically not too different from $\hat{b}_{\text{PLS}-\text{magging}}$, depending on the number of components we allow in the additional PLS regression.

1.3.2 A Small Empirical Experiment: Heterogeneous Data from Known Groups

Consider a linear model with changing regression coefficients as in (1.1). The total sample size is $n = 300$. There are $p = 500$ covariables which are generated as

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}_{500}(0, I), \tag{1.10}$$

and they are then centered and scaled to have empirical mean 0 and empirical variance 1, respectively. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, 1)$ are standard Gaussian.

We assume that there are six different known groups or clusters each with 30 observations such that

$$\begin{aligned} &B_1 = \ldots = B_{30} = b_1, \\ &B_{31} = \ldots = B_{60} = b_2, \\ &\ldots \\ &B_{271} = \ldots = B_{300} = b_6, \end{aligned}$$

that is, in every group $\mathcal{G}_g$ we have the same regression coefficient b_g for $g = 1, \ldots, 6$. These regression coefficients are realizations of

$$\begin{aligned} &b_1 \sim \mathcal{N}_p(2\mathbf{1}, I), \\ &b_g = \text{diag}(Z_1^g, \ldots, Z_p^g) b_{g-1} \ (g = 2, \ldots, 6), \end{aligned} \tag{1.11}$$

where the Z_j^g's are i.i.d. $\in \{-1, 1\}$ with $\mathbb{P}[Z_j^g = 1] = \xi$. Thus, for ξ close to 1, the coefficient vectors $b_1, \ldots, b_6$ are rather similar whereas for $\xi = 0.5$, the sign switches from b_{g-1} to b_g for each component independently with probability 0.5.

We also consider a sparse version of (1.11):

$$\begin{aligned} b_1 &= \mathcal{N}_5(2\mathbf{1}, I), \\ b_g &= \mathrm{diag}(Z_1^g, \ldots, Z_p^g) b_{g-1} \ (g = 2, \ldots, 6), \end{aligned} \tag{1.12}$$

where we use a short-hand notation for b_1, saying that the first 5 components are Gaussian and all others are zero. The variables $Z_j^{(g)}$ are as in (1.11).

We use magging in (1.4)–(1.5) with the PLS regression estimator $\hat{b}_g$ for the groups $\mathcal{G}_g$: thereby, the number of PLS components is set to 10. The groups are assumed as known and they are constructed as in (1.7). We report in Table 1.1 the out-of-sample squared error for a single representative training sample and for a test set of exactly the same structure and size as the training set described above:

$$300^{-1} \sum_{i \in \text{ test}} (Y_i - \hat{Y}_i)^2, \tag{1.13}$$

where $\hat{Y}_i$ is the prediction of Y_i based on the estimated parameters from the training data.

Table 1.1 Out-of-sample squared error (1.13) for magging with PLS regression $\hat{b}_{\mathrm{PLS-magging}}$, the pooled PLS regression estimator $\hat{b}_{\mathrm{PLS-pool}}$ (also with 10 components) based on the entire data-set, and using the mean $\overline{y}$ of the entire data-set: relative gain (+) or loss (−) over the pooled estimator. (By chance, we obtained exactly the same realized data-set for (1.12) with $\xi = 0.95$ and $\xi = 0.90$). Total sample size is $n = 300$, dimension equals $p = 500$ and there are 6 known groups each having their own regression parameter vector and each consisting of 50 homogeneous data

Model	$\hat{b}_{\mathrm{PLS-magging}}$ (%)	$\hat{b}_{\mathrm{PLS-pool}}$ (%)	$\overline{y}$ (%)
(1.11), $\xi = 0.95$	2.0	0	−14.9
(1.11), $\xi = 0.90$	32.5	0	26.9
(1.11), $\xi = 0.80$	46.3	0	43.6
(1.11), $\xi = 0.70$	44.6	0	41.4
(1.11), $\xi = 0.60$	56.3	0	55.3
(1.11), $\xi = 0.50$	52.4	0	51.3
(1.12), $\xi = 0.95$	−13.1	0	−27.5
(1.12), $\xi = 0.90$	−13.1	0	−27.5
(1.12), $\xi = 0.80$	−6.0	0	−22.6
(1.12), $\xi = 0.70$	57.4	0	58.5
(1.12), $\xi = 0.60$	55.5	0	56.2
(1.12), $\xi = 0.50$	46.7	0	44.0

We clearly see that if the degree of heterogeneity is becoming larger (smaller value of ξ), the magging estimator with PLS has superior prediction performance over the standard pooled PLS regression.

1.3.3 A Small Empirical Example: Contaminated Samples and Robustness

We consider the model in (1.1) but now with two different groups: one with clean data consisting of 285 observations, and one with 15 outlier datapoints, that is,

$$B_1 = \ldots = B_{185} = \beta^0,$$
$$B_{286} = \ldots = B_{300} = b_2.$$

Note that the outliers have all the same regression parameter b_2, but we believe that the findings below are also relevant for the case where each outlier would have its own regression parameter. The regression coefficients are realizations of

$$\beta^0 \sim \mathcal{N}_p(2\mathbf{1}, I),$$
$$b_2 \sim \mathcal{N}_p(\mu\mathbf{1}, I), \quad \mu \in \{-10, 10\}. \tag{1.14}$$

The covariates X_i are as in (1.10), and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ i.id. $\sim \mathcal{N}(0, 1)$ are standard Gaussian.

We use magging in (1.4)–(1.5) with PLS regression (with 10 components) for each subsample, and the random subsamples are constructed as in (1.9) with $G = 6$ and $m = 100$. The choice of G and m are rather ad-hoc. We report in Table 1.2 for a single representative training sample and for a test set of exactly the same structure

Table 1.2 Robustness with 5 % outliers having a different regression parameter vector than the target parameter β^0 in (1.14). Magging with PLS regression $\hat{b}_{\mathrm{PLS-magging}}$, the pooled PLS regression estimator $\hat{b}_{\mathrm{PLS-pool}}$ (also with 10 components) based on the entire data-set, and the overall mean $\overline{y}$ based on the entire data-set. Total sample size is $n = 300$ and the dimension equals $p = 500$. Out-of-sample squared error (1.13) and estimation errors (1.15) are given in the respective rows: relative gain (+) or loss (−) over the pooled estimator.

Model	Performance measure	$\hat{b}_{\mathrm{PLS-magging}}$ (%)	$\hat{b}_{\mathrm{PLS-pool}}$ (%)	$\overline{y}$
(1.14), $\mu = -10$	Squared out-sample error	50.1	0	40.4 %
(1.14), $\mu = -10$	ℓ_2-norm est. error	32.6	0	–
(1.14), $\mu = -10$	ℓ_1-norm est. error	27.0	0	–
(1.14), $\mu = 10$	squared out-sample error	18.1	0	3.6 %
(1.14), $\mu = 10$	ℓ_2-norm est. error	13.2	0	–
(1.14), $\mu = 10$	ℓ_1-norm est. error	4.4	0	–

and size as the training set the out-of-sample squared error (1.13) as well as the estimation error

$$\|\hat{b} - \beta^0\|_q \text{ for } q \in \{1, 2\}. \tag{1.15}$$

We conclude that magging is an effective and simple strategy to robustify PLS regression in presence of (at least some kind of) outliers.

1.4 Conclusions

Maximin effects estimation (Meinshausen and Bühlmann 2015) and the associated magging procedure (Bühlmann and Meinshausen 2016) are effective methods for addressing the issue of statistical estimation when the data are heterogeneous. They are computationally attractive, and especially the magging scheme is a very generic subsampling and aggregation scheme with a simple algorithmic implementation allowing for parallel processing. We present here magging for partial least squares regression: the method is appropriate and computationally feasible in presence of heterogeneity or outliers in large-scale data, and a small empirical study confirms its usefulness.

References

Breiman, L.: "Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
Bühlmann, P., Meinshausen, N.: Magging: maximin aggregation for inhomogeneous large-scale data. Proc. of the IEEE **104**, 126–135 (2016)
Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York (2011)
Bühlmann, P., Yu, B.: Analyzing bagging. Ann. Stat. **30**, 927–961 (2002)
Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H.: Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer, New York (2010)
Frank, L.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**(2), 109–135 (1993)
Geladi, P., Kowalski, B.R.: Partial least-squares regression: a tutorial. Analytica Chimica Acta **185**, 1–17 (1986)
Hall, P., Samworth, R.J.: Properties of bagged nearest neighbour classifiers. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**, 363–379 (2005)
Hastie, T., Tibshirani, R.: Varying-coefficient models. J. R. Stat. Soc. Ser. B (Statist. Methodol.) **55**, 757–796 (1993)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, Data Mining, Inference and Prediction, 2nd edn. Springer, New York (2009)
Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
Huber, P.J.: Robust Statistics, 2nd edn. Springer, New York (2011)

McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2004)
Meinshausen, N., Bühlmann, P.: Maximin effects in inhomogeneous large-scale data. Ann. Statist. **43**, 1801–1830 (2015)
Pinheiro, J., Bates, D.: Mixed-Effects Models in S and S-PLUS. Springer, New York (2000)
R Core Team: R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. http://www.R-project.org (2014)
Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B (Statist. Methodol.) **58**, 267–288 (1996)
Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P. (ed.) Multivariate Analysis, pp. 391–420. Academic, New York (1966)

Chapter 2
On the PLS Algorithm for Multiple Regression (PLS1)

Yoshio Takane and Sébastien Loisel

Abstract Partial least squares (PLS) was first introduced by Wold in the mid 1960s as a heuristic algorithm to solve linear least squares (LS) problems. No optimality property of the algorithm was known then. Since then, however, a number of interesting properties have been established about the PLS algorithm for regression analysis (called PLS1). This paper shows that the PLS estimator for a specific dimensionality S is a kind of constrained LS estimator confined to a Krylov subspace of dimensionality S. Links to the Lanczos bidiagonalization and conjugate gradient methods are also discussed from a somewhat different perspective from previous authors.

Keywords Krylov subspace • NIPALS • PLS1 algorithm • Lanczos bidiagonalization • Conjugate gradients • Constrained principal component analysis (CPCA)

2.1 Introduction

Partial least squares (PLS) was first introduced by Wold (1966) as a heuristic algorithm for estimating parameters in multiple regression. Since then, it has been elaborated in many directions, including extensions to multivariate cases (Abdi 2007; de Jong 1993) and structural equation modeling (Lohmöller 1989; Wold 1982). In this paper, we focus on the original PLS algorithm for univariate regression (called PLS1), and show its optimality given the subspace in which the vector of regression coefficients is supposed to lie. Links to state-of-the-art algorithms for solving a system of linear simultaneous equations, such as the Lanczos bidiagonalization and the conjugate gradient methods, are also discussed

Y. Takane (✉)
University of Victoria, Victoria, BC, Canada
e-mail: Yoshio.Takane@mcgill.ca

S. Loisel
Heriot-Watt University, Edinburgh, UK
e-mail: sloisel@gmail.com

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_2

from a somewhat different perspective from previous authors (Eldén 2004; Phatak and de Hoog 2002). We refer the reader to Rosipal and Krämer (2006) for more comprehensive accounts and reviews of new developments of PLS.

2.2 PLS1 as Constrained Least Squares Estimator

Consider a linear regression model

$$\mathbf{z} = \mathbf{G}\mathbf{b} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{z}$ is the N-component vector of observations on the criterion variable, $\mathbf{G}$ is the $N \times P$ matrix of predictor variables, $\mathbf{b}$ is the P-component vector of regression coefficients, and $\mathbf{e}$ is the N-component vector of disturbance terms. The ordinary LS (OLS) criterion is often used to estimate $\mathbf{b}$ under the *iid* (independent and identically distributed) normal assumption on $\mathbf{e}$. This is a reasonable practice if N is large compared to P, and columns of $\mathbf{G}$ are not highly collinear (i.e., as long as the matrix $\mathbf{G}'\mathbf{G}$ is well-conditioned). However, if this condition is not satisfied, the use of OLS estimators (OLSE) is not recommended, because then these estimators tend to have large variances. Principal component regression (PCR) is often employed in such situations. In PCR, principal component analysis (PCA) is first applied to $\mathbf{G}$ to find a low rank (say, rank S) approximation, which is subsequently used as the set of new predictor variables in a linear regression analysis. One potential problem with PCR is that the low rank approximation of $\mathbf{G}$ best accounts for $\mathbf{G}$ but is not necessarily optimal for predicting $\mathbf{z}$. By contrast, PLS extracts components of $\mathbf{G}$ that are good predictors of $\mathbf{z}$. For the case of univariate regression, the PLS algorithm (called PLS1) proceeds as follows:

PLS1 Algorithm

Step 1. Column-wise center $\mathbf{G}$ and $\mathbf{z}$, and set $\mathbf{G}_0 = \mathbf{G}$.
Step 2. Repeat the following substeps for $i = 1, \cdots, S$ ($S \leq \text{rank}(\mathbf{G})$):

Step 2.1. Set $\mathbf{w}_i = \mathbf{G}'_{i-1}\mathbf{z}/\|\mathbf{G}'_{i-1}\mathbf{z}\|$, where $\|\mathbf{G}'_{i-1}\mathbf{z}\| = (\mathbf{z}'\mathbf{G}_{i-1}\mathbf{G}'_{i-1}\mathbf{z})^{1/2}$.
Step 2.2. Set $\mathbf{t}_i = \mathbf{G}_{i-1}\mathbf{w}_i/\|\mathbf{G}_{i-1}\mathbf{w}_i\|$.
Step 2.3. Set $\mathbf{v}_i = \mathbf{G}'_{i-1}\mathbf{t}_i$.
Step 2.4. Set $\mathbf{G}_i = \mathbf{G}_{i-1} - \mathbf{t}_i\mathbf{v}'_i = \mathbf{Q}_{G_{i-1}w_i}\mathbf{G}_{i-1}$ (deflation),

where $\mathbf{Q}_{G_{i-1}w_i} = \mathbf{I} - \mathbf{G}_{i-1}\mathbf{w}_i(\mathbf{w}'_i\mathbf{G}'_{i-1}\mathbf{G}_{i-1}\mathbf{w}_i)^{-1}\mathbf{w}'_i\mathbf{G}'_{i-1}$, and where $'$ denotes the transpose operation, and $||.||$ denotes the L_2 norm of a vector (i.e., $||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}}$, see, e.g., Takane (2014), for details); vectors $\mathbf{w}_i$, $\mathbf{t}_i$, and $\mathbf{v}_i$ are called (respectively) weights, scores, and loadings, and are collected in matrices $\mathbf{W}_S$, $\mathbf{T}_S$, and $\mathbf{V}_S$. For a given S, the PLS estimator (PLSE) of $\mathbf{b}$ is given by

$$\hat{\mathbf{b}}^{(S)}_{PLSE} = \mathbf{W}_S(\mathbf{V}'_S\mathbf{W}_S)^{-1}\mathbf{T}'_S\mathbf{z} \tag{2.2}$$

(see, e.g., Abdi 2007). The algorithm above assumes that S is known and, actually, the choice of its value is crucial for good performance of PLSE (a cross validation method is often used to choose the best value of S). It has been demonstrated (Phatak and de Hoog 2002) that for a given value of S, the PLSE of $\mathbf{b}$ has better predictability than the corresponding PCR estimator.

The PLSE of $\mathbf{b}$ can be regarded as a special kind of constrained LS estimator (CLSE), in which $\mathbf{b}$ is constrained to lie in the Krylov subspace of dimensionality S defined by

$$\mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) = \mathrm{Sp}(\mathbf{K}_S), \tag{2.3}$$

where $\mathrm{Sp}(\mathbf{K}_S)$ is the space spanned by the column vectors of $\mathbf{K}_S$, and

$$\mathbf{K}_s = [\mathbf{G}'\mathbf{z}, (\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{z}, \cdots, (\mathbf{G}'\mathbf{G})^{S-1}\mathbf{G}'\mathbf{z}] \tag{2.4}$$

is called the Krylov matrix of order S. Because $\mathrm{Sp}(\mathbf{W}_S) = \mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ (see Eldén 2004, proposition 3.1; Phatak and de Hoog 2002) $\mathbf{b}$ can be re-parameterized as $\mathbf{b} = \mathbf{W}_S\mathbf{a}$ for some $\mathbf{a}$. Then Eq. (2.1) can be rewritten as

$$\mathbf{z} = \mathbf{G}\mathbf{W}_S\mathbf{a} + \mathbf{e}. \tag{2.5}$$

The OLSE of $\mathbf{a}$ is given by

$$\hat{\mathbf{a}} = (\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}, \tag{2.6}$$

from which the CLSE of $\mathbf{b}$ is found as

$$\hat{\mathbf{b}}_{CLSE}^{(S)} = \mathbf{W}_S\hat{\mathbf{a}} = \mathbf{W}_S(\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}. \tag{2.7}$$

To show that (2.7) is indeed equivalent to (2.2), we need several well-known results in the PLS literature (Bro and Eldén 2009; de Jong 1993; Eldén 2004; Phatak and de Hoog 2002). First of all, $\mathbf{W}_S$ is column-wise orthogonal, that is,

$$\mathbf{W}_S'\mathbf{W}_S = \mathbf{I}_S. \tag{2.8}$$

Secondly, $\mathbf{T}_S$ is also column-wise orthogonal,

$$\mathbf{T}_S'\mathbf{T}_S = \mathbf{I}_S, \tag{2.9}$$

and

$$\mathbf{T}_S\mathbf{L}_S = \mathbf{G}\mathbf{W}_S, \tag{2.10}$$

where $\mathbf{L}_S$ is an upper bidiagonal matrix. Relations (2.8), (2.9) and (2.10) imply that

$$\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S = \mathbf{L}_S'\mathbf{L}_S = \mathbf{H}_S, \tag{2.11}$$

where $\mathbf{H}_S$ is tridiagonal. Thirdly,

$$\mathbf{V}_S' = \mathbf{T}_S'\mathbf{G}, \tag{2.12}$$

so that

$$\mathbf{L}_S = \mathbf{T}_S'\mathbf{G}\mathbf{W}_S = \mathbf{V}_S'\mathbf{W}_S. \tag{2.13}$$

Now it is straightforward to show that

$$\begin{aligned}
\hat{\mathbf{b}}_{CLSE}^{(S)} &= \mathbf{W}_S(\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z} \\
&= \mathbf{W}_S\mathbf{H}_S^{-1}\mathbf{L}_S'\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S(\mathbf{L}_S'\mathbf{L}_S)^{-1}\mathbf{L}_S'\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S\mathbf{L}_S^{-1}\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S(\mathbf{V}_S'\mathbf{W}_S)^{-1}\mathbf{T}_S'\mathbf{z} \\
&= \hat{\mathbf{b}}_{PLSE}^{(S)},
\end{aligned} \tag{2.14}$$

and this establishes the equivalence between Eqs. (2.7) and (2.2).

The PLSE of regression parameters reduces to the OLSE if $S = \text{rank}(\mathbf{G})$ (when $\text{rank}(\mathbf{G}) < P$, we use $\mathbf{G}^+$, which is the Moore-Penrose inverse of $\mathbf{G}$, in lieu of $(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}$ in the OLSE for regression coefficients).

2.3 Relations to the Lanczos Bidiagonalization Method

It has been pointed out (Eldén 2004) that PLS1 described above is equivalent to the following Lanczos bidiagonalization algorithm:

The Lanczos Bidiagonalization (LBD) Algorithm

Step 1. Column-wise center $\mathbf{G}$, and compute $\mathbf{u}_1 = \mathbf{G}'\mathbf{z}/||\mathbf{G}'\mathbf{z}||$ and $\mathbf{q}_1 = \mathbf{G}\mathbf{u}_1/\delta_1$, where $\delta_1 = ||\mathbf{G}\mathbf{u}_1||$.

Step 2. For $i = 2, \cdots, S$ (this is the same S as in PLS1),

(a) Compute $\gamma_{i-1}\mathbf{u}_i = \mathbf{G}'\mathbf{q}_{i-1} - \delta_{i-1}\mathbf{u}_{i-1}$.
(b) Compute $\delta_i\mathbf{q}_i = \mathbf{G}\mathbf{u}_i - \gamma_{i-1}\mathbf{q}_{i-1}$.

Scalars γ_{i-1} and δ_i ($i = 2, \cdots, S$) are the normalization factors to make $||\mathbf{u}_i|| = 1$ and $||\mathbf{q}_{i-1}|| = 1$, respectively.

Let $\mathbf{U}_S$ and $\mathbf{Q}_S$ represent the collections of $\mathbf{u}_i$ and $\mathbf{q}_i$ for $i = 1, \cdots, S$. It has been shown (Eldén 2004, Proposition 3.1) that these two matrices are essentially the same as $\mathbf{W}_S$ and $\mathbf{T}_S$, respectively, obtained in PLS1. Here "essentially" means that these

two matrices are identical to $\mathbf{W}_S$ and $\mathbf{T}_S$ except that the even columns of $\mathbf{U}_S$ and $\mathbf{Q}_S$ are reflected (i.e., have their sign reversed). We show this explicitly for $\mathbf{u}_2$ and $\mathbf{q}_2$ (i.e., $\mathbf{u}_2 = -\mathbf{w}_2$ and $\mathbf{q}_2 = -\mathbf{t}_2$). It is obvious from Step 1 of the two algorithms that

$$\mathbf{w}_1 = \mathbf{u}_1 \quad \text{and} \quad \mathbf{t}_1 = \mathbf{q}_1. \tag{2.15}$$

Let $\alpha_1 = ||\mathbf{G}'\mathbf{z}||$. Then

$$\begin{aligned}
\mathbf{w}_2 &\propto \mathbf{G}'\mathbf{Q}_{Gw_1}\mathbf{z} \quad \text{(from Step 2.4 of the PLS1 algorithm)} \\
&= \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{w}_1(\mathbf{w}_1'\mathbf{G}'\mathbf{G}\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{G}'\mathbf{z} \\
&= \alpha_1(\mathbf{w}_1 - \mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2) \qquad (2.16) \\
&\propto -\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1 + \delta_1\mathbf{w}_1, \qquad (2.17)
\end{aligned}$$

where $\propto$ means "proportional." To obtain the last expression, we multiplied Eq. (2.16) by δ_1/α_1 (> 0). This last expression is proportional to $-\mathbf{u}_2$, where $\mathbf{u}_2 \propto \mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1$ from Step 2(a) of the Lanczos algorithm. This implies $\mathbf{u}_2 = -\mathbf{w}_2$, because both $\mathbf{u}_2$ and $\mathbf{w}_2$ are normalized.

Similarly, define $\beta_1^2 = \mathbf{w}_1'(\mathbf{G}'\mathbf{G})^2\mathbf{w}_1$. Then

$$\begin{aligned}
\mathbf{t}_2 &\propto \mathbf{Q}_{Gw_1}\mathbf{G}\mathbf{G}'\mathbf{Q}_{Gw_1}\mathbf{z} \quad \text{(from Step 2.2 of the PLS1 algorithm)} \\
&= \alpha_1(\mathbf{G}\mathbf{w}_1 - \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2 - \mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^4}\mathbf{G}\mathbf{w}_1) \qquad (2.18) \\
&\propto -\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^2}\mathbf{G}\mathbf{w}_1. \qquad (2.19)
\end{aligned}$$

To obtain Eq. (2.19), we multiplied (2.18) by δ_1^2/α_1 (> 0). On the other hand, we have

$$\begin{aligned}
\mathbf{q}_2 &\propto \frac{1}{\delta_1\gamma_1}(\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - \delta_1^2\mathbf{G}\mathbf{u}_1 - \gamma_1^2\mathbf{G}\mathbf{u}_1) \quad \text{(from Step 2(b) of the Lanczos algorithm)} \\
&\propto \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - (\delta_1^2 + \gamma_1^2)\mathbf{G}\mathbf{u}_1. \qquad (2.20)
\end{aligned}$$

To show that $\mathbf{q}_2 \propto -\mathbf{t}_2$, it remains to show that

$$\gamma^2 + \delta^2 = \beta_1^2/\delta_1^2. \tag{2.21}$$

From Step 2(a) of the Lanczos algorithm,

$$\begin{aligned}
\gamma^2 &= (\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1)'(\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1) \\
&= \beta^2/\delta^2 - \delta^2, \qquad (2.22)
\end{aligned}$$

and so indeed (2.21) holds. Again, we have $\mathbf{q}_2 = -\mathbf{t}_2$, because both $\mathbf{q}_2$ and $\mathbf{t}_2$ are normalized.

The sign reversals of $\mathbf{u}_2$ and $\mathbf{q}_2$ yield $\mathbf{u}_3$ and $\mathbf{q}_3$ identical to $\mathbf{w}_3$ and $\mathbf{t}_3$, respectively, by similar sign reversals, and $\mathbf{u}_4$ and $\mathbf{q}_4$ which are sign reversals of $\mathbf{w}_4$ and $\mathbf{t}_4$, and so on. Thus, only even columns of $\mathbf{U}_s$ and $\mathbf{Q}_s$ are affected (i.e., have their sign reversed) relative to the corresponding columns of $\mathbf{W}_S$ and $\mathbf{T}_S$, respectively. Of course, these sign reversals have no effect on estimates of regression parameters. The estimate of regression parameters by the Lanczos bidiagonaliation method is given by

$$\hat{\mathbf{b}}_{LBD}^{(S)} = \mathbf{U}_s(\mathbf{L}_S^*)^{-1}\mathbf{Q}_S'\mathbf{z}, \tag{2.23}$$

where

$$\mathbf{L}_S^* = \mathbf{Q}_S'\mathbf{G}\mathbf{U}_S, \tag{2.24}$$

which is upper bidiagonal, as is $\mathbf{L}_S$ (defined in Eq. (2.13)). matrix $\mathbf{L}_S^*$ differs from matrix $\mathbf{L}_S$ only in the sign of its super-diagonal elements. The matrices $\mathbf{L}_S^{-1}$ and $(\mathbf{L}_S^*)^{-1}$ are also upper bidiagonal, for which the super-diagonal elements are opposite in sign, while their diagonal elements remain the same. Thus

$$\begin{aligned}\mathbf{W}_S\mathbf{L}_S^{-1}\mathbf{T}_S' &= \sum_{i=1}^{s}(\ell_{i,i}\mathbf{w}_i\mathbf{t}_i' + \ell_{i,i+1}\mathbf{w}_i\mathbf{t}_{i+1}') \\ &= \sum_{i=1}^{s}(\ell_{i,i}^*\mathbf{u}_i\mathbf{q}_i' + \ell_{i,i+1}^*\mathbf{u}_i\mathbf{q}_{i+1}') \\ &= \mathbf{U}_s(\mathbf{L}_s^*)^{-1}\mathbf{Q}_s',\end{aligned} \tag{2.25}$$

where $\ell_{i,j}$ and $\ell_{i,j}^*$ are the ij-th element of (respectively) $\mathbf{L}_S$ and $\mathbf{L}_S^*$. Note that

$$\ell_{i,i} = \ell_{i,i}^*, \quad \mathbf{w}_i\mathbf{t}_i' = \mathbf{u}_i\mathbf{q}_i', \quad \ell_{i,i+1} = -\ell_{i,i+1}^*, \quad \text{and} \quad \mathbf{w}_i\mathbf{t}_{i+1}' = -\mathbf{u}_i\mathbf{q}_{i+1}' \tag{2.26}$$

It is widely known (see, e.g., Saad 2003) that the matrix of orthogonal basis vectors generated by the Arnoldi orthogonalization of $\mathbf{K}_S$ (Arnoldi 1951) is identical to $\mathbf{U}_S$ obtained in the Lanczos algorithm. Starting from $\mathbf{u}_1 = \mathbf{G}'\mathbf{z}/\|\mathbf{G}'\mathbf{z}\|$, this orthogonalization method finds $\mathbf{u}_{i+1}$ $(i = 1, \cdots, S-1)$ by successively orthogonalizing $\mathbf{G}'\mathbf{G}\mathbf{u}_i$ $(i = 1, \cdots, S-1)$ to all previous $\mathbf{u}_i$'s by a procedure similar to the Gram-Schmidt orthogonalization method. This yields $\mathbf{U}_S$ such that $\mathbf{G}'\mathbf{G}\mathbf{U}_S = \mathbf{U}_S\mathbf{H}_S^*$, or

$$\mathbf{U}_S'\mathbf{G}'\mathbf{G}\mathbf{U}_S = \mathbf{L}_S^{*'}\mathbf{L}_S^* = \mathbf{H}_S^*, \tag{2.27}$$

where $\mathbf{H}_S^*$ is tridiagonal as is $\mathbf{H}_S$ defined in Eq. (2.11). The diagonal elements of this matrix are identical to those of $\mathbf{H}_S$ while its sub- and super-diagonal elements have their sign reversed. Matrix $\mathbf{H}_S^*$ is called the Lanczos tridiagonal matrix and it is useful to obtain eigenvalues of $\mathbf{G}'\mathbf{G}$.

2.4 Relations to the Conjugate Gradient Method

It has been pointed out (Phatak and de Hoog 2002) that the conjugate gradient (CG) algorithm (Hestenes and Stiefel 1951) for solving a system of linear simultaneous equations $\mathbf{G}'\mathbf{G}\mathbf{b} = \mathbf{G}'\mathbf{y}$ gives solutions identical to $\hat{\mathbf{b}}^{(s)}_{PLSE}$ $[s = 1, \cdots, \mathrm{rank}(\mathbf{G})]$, if the CG iteration starts from the initial solution $\hat{\mathbf{b}}^{(0)}_{CG} \equiv \mathbf{b}_0 = \mathbf{0}$. To verify their assertion, we look into the CG algorithm stated as follows:

The Conjugate Gradient (CG) Algorithm

Step 1. Initialize $\mathbf{b}_0 = \mathbf{0}$. Then, $\mathbf{r}_0 = \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{b}_0 = \mathbf{G}'\mathbf{z} = \mathbf{d}_0$. (Vectors $\mathbf{r}_0$ and $\mathbf{d}_0$ are called initial residual and initial direction vectors, respectively.)

Step 2. For $i = 0, \cdots, s-1$, compute:

(a) $a_i = \mathbf{d}_i'\mathbf{r}_i / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i = ||\mathbf{r}_i||^2 / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i$.

(b) $\mathbf{b}_{i+1} = \mathbf{b}_i + a_i\mathbf{d}_i$.

(c) $\mathbf{r}_{i+1} = \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{b}_{i+1} = \mathbf{r}_i - a_i\mathbf{G}'\mathbf{G}\mathbf{d}_i = \mathbf{Q}'_{d_i/G'G}\mathbf{r}_i$, where $\mathbf{Q}_{d_i/G'G} = \mathbf{I} - \mathbf{d}_i(\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i)^{-1}\mathbf{d}_i'\mathbf{G}'\mathbf{G}$ is the projector onto the space orthogonal to $\mathrm{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$ along $\mathrm{Sp}(\mathbf{d}_i)$ [its transpose, on the other hand, is the projector onto the space orthogonal $\mathrm{Sp}(\mathbf{d}_i)$ along $\mathrm{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$].

(d) $b_i = -\mathbf{r}'_{i+1}\mathbf{G}'\mathbf{G}\mathbf{d}_i / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i = ||\mathbf{r}_{i+1}||^2 / ||\mathbf{r}_i||^2$.

(e) $\mathbf{d}_{i+1} = \mathbf{r}_{i+1} + b_i\mathbf{d}_i = \mathbf{Q}_{d_i/G'G}\mathbf{r}_{i+1}$.

Let $\mathbf{R}_j = [\mathbf{r}_0, \cdots, \mathbf{r}_{j-1}]$ and $\mathbf{D}_j = [\mathbf{d}_0, \cdots, \mathbf{d}_{j-1}]$ for $j \le S$. We first show that

$$\mathrm{Sp}(\mathbf{R}_j) = \mathrm{Sp}(\mathbf{D}_j) = \mathscr{K}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) \tag{2.28}$$

by induction, where, as before, $\mathrm{Sp}(\mathbf{A})$ indicates the space spanned by the column vectors of matrix $\mathbf{A}$. It is obvious that $\mathbf{r}_0 = \mathbf{d}_0 = \mathbf{G}'\mathbf{z}$, so that $\mathrm{Sp}(\mathbf{R}_1) = \mathrm{Sp}(\mathbf{D}_1) = \mathscr{K}_1(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. From Step 2(c) of the CG algorithm, we have

$$\mathbf{r}_1 = \mathbf{Q}'_{d_i/G'G}\mathbf{r}_0 = \mathbf{r}_0 - \mathbf{G}'\mathbf{G}\mathbf{d}_0 c_0 \tag{2.29}$$

for some scalar c_0, so that $\mathbf{r}_1 \in \mathscr{K}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ because $\mathbf{G}'\mathbf{G}\mathbf{d}_0 \in \mathscr{K}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. From Step 2(e), we also have

$$\mathbf{d}_1 = \mathbf{Q}_{d_0/G'G}\mathbf{r}_1 = \mathbf{r}_1 - \mathbf{d}_0 c_0^* \tag{2.30}$$

for some c_0^*, so that $\mathbf{d}_1 \in \mathscr{K}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. This shows that $\mathrm{Sp}(\mathbf{R}_2) = \mathrm{Sp}(\mathbf{D}_2) = \mathscr{K}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. Similarly, we have $\mathbf{r}_2 \in \mathscr{K}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ and $\mathbf{d}_2 \in \mathscr{K}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$, so that $\mathrm{Sp}(\mathbf{R}_3) = \mathrm{Sp}(\mathbf{D}_3) = \mathscr{K}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$, and so on.

The property of $\mathbf{D}_j$ above implies that $\mathrm{Sp}(\mathbf{W}_S)$ is identical to $\mathrm{Sp}(\mathbf{D}_S)$, which in turn implies that

$$\hat{\mathbf{b}}^{(S)}_{CG} = \mathbf{D}_S(\mathbf{D}_S'\mathbf{G}\mathbf{G}\mathbf{D}_S)^{-1}\mathbf{D}_S'\mathbf{G}\mathbf{z} \tag{2.31}$$

is identical to $\hat{\mathbf{b}}^{(S)}_{CLSE}$ as defined in Eq. (2.7), which in turn is equal to $\hat{\mathbf{b}}^{(S)}_{PLSE}$ defined in Eq. (2.2) (Phatak and de Hoog 2002) by virtue of Eq. (2.14). It remains to show that $\hat{\mathbf{b}}^{(S)}_{CG}$ defined in (2.31) coincides with $\mathbf{b}_S$ generated by the CG algorithm. By the $\mathbf{G}'\mathbf{G}$-conjugacy of $\mathbf{d}_j$'s (the orthogonality of $\mathbf{d}_j$'s with respect to $\mathbf{G}'\mathbf{G}$, i.e., $\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0$ for any $i \neq j$, as will be shown later), Eq. (2.31) can be rewritten as

$$\hat{\mathbf{b}}^{(S)}_{CG} = \sum_{i=0}^{S-1} \mathbf{d}_i(\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i)^{-1}\mathbf{d}_i'\mathbf{G}'\mathbf{z}. \tag{2.32}$$

From Step 2(b) of the CG algorithm, on the other hand, we have

$$\mathbf{b}_1 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{r}_0 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{G}\mathbf{z} = \hat{\mathbf{b}}^{(1)}_{CG}, \tag{2.33}$$

and

$$\begin{aligned}\mathbf{b}_3 &= \hat{\mathbf{b}}^{(1)}_{CG} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{r}_1, \\ &= \hat{\mathbf{b}}^{(1)}_{CG} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}^{(2)}_{CG},\end{aligned} \tag{2.34}$$

since $\mathbf{d}_1'\mathbf{r}_1 = \mathbf{d}_1'\mathbf{Q}'_{d_0/G'G}\mathbf{r}_0 = \mathbf{d}_1'\mathbf{r}_0 = \mathbf{d}_1'\mathbf{G}\mathbf{z}$ (the second equality in the preceding equation holds again due to the $\mathbf{G}'\mathbf{G}$-conjugacy of $\mathbf{d}_1$ and $\mathbf{d}_0$). Similarly, we obtain

$$\begin{aligned}\mathbf{b}_3 &= \hat{\mathbf{b}}^{(2)}_{CG} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{r}_2, \\ &= \hat{\mathbf{b}}^{(2)}_{CG} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}^{(3)}_{CG},\end{aligned} \tag{2.35}$$

since $\mathbf{d}_2'\mathbf{r}_2 = \mathbf{d}_2'\mathbf{Q}'_{d_1/G'G}\mathbf{r}_1 = \mathbf{d}_2'\mathbf{r}_1 = \mathbf{d}_2'\mathbf{Q}'_{d_0/G'G}\mathbf{r}_0 = \mathbf{d}_2'\mathbf{r}_0 = \mathbf{d}_2'\mathbf{G}\mathbf{z}$. This extends to S larger than 3. This proves the claim made above that (2.31) is indeed identical to $\mathbf{b}_S$ obtained from the CG iteration.

It is rather intricate to show the $\mathbf{G}'\mathbf{G}$-conjugacy of direction vectors (i.e., $\mathbf{d}_j'\mathbf{G}'\mathbf{G}\mathbf{d}_i = 0$ for $j \neq i$), although it is widely known in the numerical linear algebra literature (Golub and van Loan 1989). The proofs given in Golub and van Loan (1989) are not very easy to follow, however. In what follows, we attempt to provide a step-by-step proof of this fact. Let $\mathbf{R}_j$ and $\mathbf{D}_j$ be as defined above. We temporarily assume that the columns of $\mathbf{D}_j$ are already $\mathbf{G}'\mathbf{G}$-conjugate (i.e., $\mathbf{D}_j'\mathbf{G}'\mathbf{G}\mathbf{D}_j$ is diagonal). Later we show that such construction of $\mathbf{D}_j$ is possible.

We first show that

$$\mathbf{d}_{j-1}'\mathbf{r}_j = 0. \tag{2.36}$$

From Step 2(c) of the CG algorithm, we have

$$\mathbf{d}_{j-1}'\mathbf{r}_j = \mathbf{d}_{j-1}'\mathbf{Q}'_{d_{j-1}/G'G}\mathbf{r}_{j-1} = \mathbf{d}_{j-1}'(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}_{j-1}'\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}_{j-1}')\mathbf{r}_{j-1} = 0, \tag{2.37}$$

as claimed above. We next show that

$$\mathbf{d}'_{j-2}\mathbf{r}_j = 0, \tag{2.38}$$

based on (2.36). From Step 2(c) of the algorithm, we have

$$\begin{aligned}\mathbf{d}'_{j-2}\mathbf{r}_j &= \mathbf{d}'_{j-2}\mathbf{Q}'_{d_{j-1}/G'G}\mathbf{r}_{j-1} \\ &= \mathbf{d}'_{j-2}(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1})\mathbf{r}_{j-1} \\ &= \mathbf{d}'_{j-2}\mathbf{r}_{j-1} = 0, \end{aligned} \tag{2.39}$$

as claimed. Note that $\mathbf{d}'_{j-2}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1} = 0$ by the assumption of the $\mathbf{G}'\mathbf{G}$-conjugacy (among the column vectors) of $\mathbf{D}_j$. The last equality in (2.39) holds due to (2.36). By repeating essentially the same process, we can prove that $\mathbf{d}'_{j-k}\mathbf{r}_j = 0$ for $k = 3, \cdots, j$, which implies

$$\mathbf{D}'_j\mathbf{r}_j = \mathbf{0}, \tag{2.40}$$

and

$$\mathbf{R}'_j\mathbf{r}_j = \mathbf{0}, \tag{2.41}$$

since $\mathrm{Sp}(\mathbf{D}_j) = \mathrm{Sp}(\mathbf{R}_j) = \mathscr{K}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. These relations indicate that in the CG method, the residual vector $\mathbf{r}_j$ is orthogonal to all previous search directions as well as all previous residual vectors.

We are now in a position to prove that

$$\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0. \tag{2.42}$$

To do so, we first need to show that

$$\mathbf{d}'_j\mathbf{r}_j = ||\mathbf{r}_j||^2, \tag{2.43}$$

and also that

$$\mathbf{d}'_j\mathbf{r}_{j-1} = ||\mathbf{r}_j||^2. \tag{2.44}$$

For Eq. (2.43), we note that

$$\begin{aligned}\mathbf{d}'_j\mathbf{r}_j &= \mathbf{r}'_j\mathbf{Q}'_{d_{j-1}/G'G}\mathbf{r}_j \quad \text{(by Step 2(e))} \\ &= ||\mathbf{r}_j||^2 - \mathbf{r}'_j\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1})\mathbf{r}_j = ||\mathbf{r}_j||^2, \end{aligned} \tag{2.45}$$

due to Eq. (2.36). For Eq. (2.44), we have

$$\begin{aligned}\mathbf{d}_j'\mathbf{r}_{j-1} &= \mathbf{r}_j'\mathbf{r}_{j-1} + b_{j-1}\mathbf{d}_{j-1}'\mathbf{r}_{j-1} \quad \text{(by Step 2(e))} \\ &= 0 + (||\mathbf{r}_j||^2/||\mathbf{r}_{j-1}||^2)||\mathbf{r}_{j-1}||^2 = ||\mathbf{r}_j||^2. \end{aligned} \tag{2.46}$$

To show that (2.42) holds is now straightforward. We note that

$$\mathbf{r}_j'\mathbf{d}_j = \mathbf{r}_{j-1}'\mathbf{d}_j - a_{j-1}\mathbf{d}_{j-1}'\mathbf{G}'\mathbf{G}\mathbf{d}_j \tag{2.47}$$

by Step 2(c), and that $\mathbf{r}_j'\mathbf{d}_j = \mathbf{r}_{j-1}'\mathbf{d}_j = ||\mathbf{r}_j||^2$ by Eqs. (2.43) and (2.44). Since $a_{j-1} \neq 0$, this implies that $\mathbf{d}_{j-1}'\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0$. That is, $\mathbf{d}_j$ is $\mathbf{G}'\mathbf{G}$-conjugate to the previous direction vector $\mathbf{d}_{j-1}$.

We can also show that $\mathbf{d}_j$ is $\mathbf{G}'\mathbf{G}$-conjugate to all previous direction vectors despite the fact that at any specific iteration, $\mathbf{d}_j$ is taken to be $\mathbf{G}'\mathbf{G}$-conjugate to only $\mathbf{d}_{j-1}$. We begin with

$$\mathbf{d}_{j-2}'\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0. \tag{2.48}$$

We first note that

$$\begin{aligned}\mathbf{r}_{j-2}'\mathbf{d}_j &= \mathbf{r}_{j-2}'\mathbf{r}_j + b_{j-1}\mathbf{r}_{j-2}'\mathbf{d}_{j-1} \quad \text{(by Step 2(e))} \\ &= 0 + (||\mathbf{r}_j||^2/||\mathbf{r}_{j-1}||^2)||\mathbf{r}_{j-1}||^2 \quad \text{(by (2.44))} \\ &= ||\mathbf{r}_j||^2. \end{aligned} \tag{2.49}$$

We also have

$$\mathbf{r}_{j-1}'\mathbf{d}_j = \mathbf{r}_{j-2}'\mathbf{d}_j - a_{j-2}\mathbf{d}_{j-2}'\mathbf{G}'\mathbf{G}\mathbf{d}_j \tag{2.50}$$

by Step 2(c). Since $\mathbf{r}_{j-1}'\mathbf{d}_j = \mathbf{r}_{j-2}'\mathbf{d}_j = ||\mathbf{r}_j||^2$ and $a_{j-2} \neq 0$, this implies (2.48). We may follow a similar line of argument as above, and show that $\mathbf{d}_{j-k}'\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0$ for $k = 3, \cdots, j$. This shows that $\mathbf{D}_j'\mathbf{G}'\mathbf{G}\mathbf{d}_j = \mathbf{0}$, as claimed.

In the proof above, it was assumed that the column vectors of $\mathbf{D}_j$ were $\mathbf{G}'\mathbf{G}$-conjugate. It remains to show that such construction of $\mathbf{D}_j$ is possible. We have $\mathbf{D}_1'\mathbf{r}_1 = \mathbf{d}_0'\mathbf{r}_1 = 0$ by (2.36). This implies that $\mathbf{R}_1'\mathbf{r}_1 = 0$ (since $\mathrm{Sp}(\mathbf{D}_1) = \mathrm{Sp}(R_1)$), which in turn implies that $\mathbf{D}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1 = \mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_1 = 0$. The columns of $\mathbf{D}_2 = [\mathbf{d}_0, \mathbf{d}_1]$ are now shown to be $\mathbf{G}'\mathbf{G}$-conjugate. We repeat this process until we reach $\mathbf{D}_j$ whose column vectors are all $\mathbf{G}'\mathbf{G}$-conjugate. This process also generates $\mathbf{R}_j$ whose columns are mutually orthogonal. This means that all residual vectors are orthogonal in the CG method. The CG algorithm is also equivalent to the GMRES (Generalized Minimum Residual) method (Saad and Schultz 1986), when the latter is applied to the symmetric positive definite (*pd*) matrix $\mathbf{G}'\mathbf{G}$.

It may also be pointed out that $\mathbf{R}_S$ is an un-normalized version of $\mathbf{W}_S$ obtained in PLS1. This can be seen from the fact that the column vectors of both of these matrices are orthogonal to each other, and that $\mathrm{Sp}(\mathbf{W}_S) = \mathrm{Sp}(\mathbf{R}_S) = \mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. Although some columns of $\mathbf{R}_S$ may be sign-reversed as are some columns of $\mathbf{U}_s$ in the Lanczos method, it can be directly verified that this does not happen to $\mathbf{r}_2$ (i.e., $\mathbf{r}_2/||\mathbf{r}_2|| = \mathbf{w}_2$). So it is not likely to happen to other columns of $\mathbf{R}_S$.

2.5 Concluding Remarks

The PLS1 algorithm was initially invented as a heuristic technique to solve LS problems (Wold 1966). No optimality properties of the algorithm were known at that time, and for a long time it had been criticized for being somewhat ad-hoc. It was later shown, however, that it is equivalent to some of the most sophisticated numerical algorithms to date for solving systems of linear simultaneous equations, such as the Lanczos bidiagonalization and the conjugate gradient methods. It is amazing, and indeed admirable, that Herman Wold almost single-handedly reinvented the "wheel" in a totally different context.

References

Abdi, H.: Partial least squares regression. In: Salkind, N.J. (ed.) Encyclopedia of Measurement and Statistics, pp. 740–54. Sage, Thousand Oaks (2007)

Arnoldi, W.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. Q. Appl. Math. **9**, 17–29 (1951)

Bro, R., Eldén, L.: PLS works. J. Chemom. **23**, 69–71 (2009)

de Jong, S.: SIMPLS: an alternative approach to partial least squares regression. J. Chemom. **18**, 251–263 (1993)

Eldén, L.: Partial least-squares vs Lanczos bidiagonalization–I: analysis of a projection method for multiple regression. Comput. Stat. Data Anal. **46**, 11–31 (2004)

Golub, G.H., van Loan, C.F.: Matrix Computations, 2nd edn. The Johns Hopkins University Press, Baltimore (1989)

Hestenes, M., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur Stand. **49**, 409–436 (1951)

Lohmöller, J.B.: Latent Variables Path-Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)

Phatak, A., de Hoog, F.: Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. J. Chemom. **16**, 361–367 (2002)

Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., et al. (eds.) SLSFS 2005. LNCS 3940, pp. 34–51. Springer, Berlin (2006)

Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. Society of Industrial and Applied Mathematics, Philadelphia (2003)

Saad, Y., Schultz, M.H.: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Comput. **7**, 856–869 (1986)

Takane, Y.: Constrained Principal Component Analysis and Related Techniques. CRC Press, Boca Raton (2014)

Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (ed.) Multivariate Analysis, pp. 391–420. Academic, New York (1966)
Wold, H. (1982) Soft modeling: the basic design and some extensions. In: Jöreskog, K.G., Wold, H. (eds.) Systems Under Indirect Observations, Part 2, pp. 1–54. North-Holland, Amsterdam (1982)

Chapter 3
Extending the Finite Iterative Method for Computing the Covariance Matrix Implied by a Recursive Path Model

Zouhair El Hadri and Mohamed Hanafi

Abstract Given $q + p$ variables (q endogenous variables and p exogenous variables) and the covariance matrix among exogenous variables, how to compute the covariance matrix implied by a given recursive path model connecting these $q + p$ variables? The finite iterative method (FIM) was recently introduced by El Hadri and Hanafi (Electron J Appl Stat Anal 8:84–99, 2015) to perform this task but only when all the variables are standardized (and so the covariance matrix is actually a *correlation* matrix). In this paper, the extension of FIM to the general case of a covariance matrix case is introduced. Moreover, the computational efficiency of FIM and the well-known Jöreskog's method is discussed and illustrated.

Keywords Finite iterative method (FIM) • Jöreskog's method • Covariance matrix • Correlation matrix • Endogenous variable • Exogenous variable

3.1 Introduction

Path analysis (Boudon 1965; Duncan 1966; Heise 1969; Hauser and Sewall 1975) is a set of statistical techniques used to examine cause and effect between observed (measured) variables on the same set of observations. Path analysis—which originated in the early twentieth century mostly from the work of Sewall Wright (1921) and Kline (2016)—generalizes multiple regression models because, in path analysis, all variables can be both dependent and independent variables. Today, path analysis—the simplest form of structural equation models with latent

Z. El Hadri (✉)
Faculté des Sciences, Département de Mathématiques, Université Ibn Tofail, Equipe de Cryptographie et de Traitement de l'Information, Kénitra, Maroc
e-mail: z.elhadri@yahoo.fr

M. Hanafi
Oniris, Unité de Sensométrie et Chimiométrie, Sensometrics and Chemometrics Laboratory, Nantes, France
e-mail: mohamed.hanafi@oniris-nantes.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_3

variables (Boudon 1965; Duncan 1966)—is applied in many fields going from ecological studies (Sanchez-Pinero and Polis 2000; Shine 1996; Shipley 1999) to social sciences (Blalock 1971; Wolfle 2003).

In the vocabulary of path analysis there are two types of variables: *exogenous* and *endogenous* variables. An exogenous variable is always an independent variable; its role is to explain other variables, which implies that it can never be an effect in the considered model, whereas an endogenous variable can be the cause (predictor) of one or more other variables and can also, at the same time, be caused by other variables.

The starting point of path analysis is a conceptual diagram, considered as a schematic representation of the model; This diagram should be specified by the modeler. In addition to the causal relationships between variables, the disturbances associated to endogenous variables as well as direct effects (parameters) between variables must also be specified. When two variables are correlated and there is no causal relationship between them, this is represented by a curved arc. This diagram is represented algebraically by a set of regression equations in which at least one variable is both explanatory and explained. Figure 3.1 shows a path analysis model with four variables ξ_1, η_1, η_2 and η_3, together with its system of structural equations.

When a variable $\boldsymbol{\xi}$ is the cause of the variable $\boldsymbol{\eta}$, much of the variance of $\boldsymbol{\eta}$ could be explained by $\boldsymbol{\xi}$. The disturbance $\boldsymbol{\zeta}$ associated to $\boldsymbol{\eta}$ represents the source of variability of $\boldsymbol{\eta}$ which is not explained by $\boldsymbol{\xi}$. The disturbance is then an exogenous variable not directly measured. The magnitude of an effect is measured by a numerical quantity called the path coefficient or parameter. It is a statistical estimate of the direct effect of an independent variable on the dependent variable taking into account other variables.

In path analysis, there are two basic kinds of models, the recursive model and the non-recursive model. In a recursive model, the causal effects are unidirectional. In other words, no variable is both a cause and an effect of another variable, directly or indirectly. In contrast, in a nonrecursive model there is a mutual causal influence among variables.

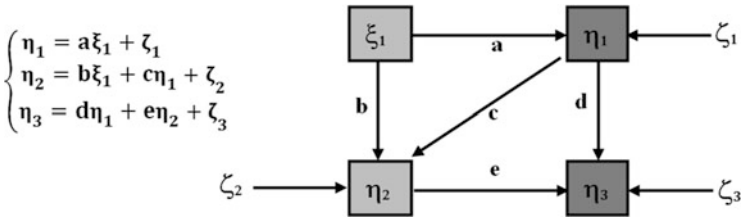

Fig. 3.1 Representation of a path analysis model with four variables and its system of structural equations. $\boldsymbol{\xi}_1$ is an exogenous variable (predictor), $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ and $\boldsymbol{\eta}_3$ are endogenous variables (explained), $\boldsymbol{\eta}_1$ is explained by $\boldsymbol{\xi}_1$ with direct effect noted by a, $\boldsymbol{\eta}_2$ is explained by $\boldsymbol{\xi}_1$ and $\boldsymbol{\eta}_1$ with direct effects noted respectively by b and c, $\boldsymbol{\eta}_3$ is explained by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ with direct effects noted respectively by d and e. Disturbances are associated to endogenous variables $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ and $\boldsymbol{\eta}_3$. These disturbances are noted respectively by $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ and $\boldsymbol{\zeta}_3$

The present paper focuses on the computation of the covariance matrix implied by a recursive path model noted $\widehat{\boldsymbol{\Sigma}}$. The computation of this matrix is needed at least at three different levels of the process of the analysis: (i) to test the validity of the model, (ii) to evaluate the total effect of exogenous variables, and (iii) to estimate the parameters where several criteria are minimized, such as, for example,

$$\mathscr{F} = \frac{1}{2}\text{trace}\left[\left(\mathbf{S} - \widehat{\boldsymbol{\Sigma}}\right)\right]^2$$

with $\mathbf{S}$ being the empirical covariance matrix.

The topic of the present paper can be summarized as follows. Given $q + p$ variables (q endogenous variables and p exogenous variables) and the covariance matrix of the exogenous variables, how to compute the covariance matrix implied by a given recursive path model (i.e., whose parameters are known) connecting these $q + p$ variables? For example, the model given in Fig. 3.1 corresponds to the following covariance matrix.

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} s_{11} & as_{11} & bs_{11} + acs_{11} & ads_{11} + bes_{11} + aces_{11} \\ as_{11} & s_{22} & abs_{11} + cs_{22} & ds_{22} + abes_{11} + ces_{22} \\ bs_{11} + acs_{11} & abs_{11} + cs_{22} & s_{33} & abds_{11} + cds_{22} + es_{33} \\ ads_{11} + bes_{11} + aces_{11} & ds_{22} + abes_{11} + es_{22} & abds_{11} + cds_{22} + es_{33} & s_{44} \end{bmatrix} \tag{3.1}$$

The well-known method proposed by Jöreskog (1977) (see also Sect. 3.4.2 in the present paper) can be considered as a solution for the computation of the covariance matrix implied by a given recursive model; but Jöreskog's method has two major drawbacks: (1) it requires matrix inversion and (2) the variances of disturbances should be computed based on model parameters. However the expression of these variances in terms of parameters is known only for simple models and this constitutes the main limitation of Jöreskog's method.

Alternatively, El Hadri and Hanafi (2015) have recently proposed a new method called the Finite Iterative Method (FIM) that can overcome the limitations of Jöreskog's method but, so far, this new method is limited to correlation matrices (i.e., all variables must be standardized). The present paper generalizes FIM to the case of a covariance matrix.

The paper is organized as follows. The second section presents the notations used in the rest of the paper. The third section presents the extension of the finite iterative method to the covariance matrix. The fourth section illustrates with examples the efficiency of FIM compared to Jöreskog's method. Finally, some conclusions are presented.

3.2 Notations

This section introduces the basic notations following Jöreskog (1977), Jöreskog and Wold (1982), and Hoyle (1995). The translation of the diagram of a recursive path model to equations is given by the following generic form:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_q \end{bmatrix} = \begin{bmatrix} 0 & \dots & \dots & 0 \\ \beta_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \beta_{q1} & \dots & \beta_{q,(q-1)} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_q \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \ddots & \gamma_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \gamma_{q1} & \gamma_{q2} & \cdots & \gamma_{qp} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_q \end{bmatrix} \tag{3.2}$$

System (3.2) can be also written in compact form as:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{3.3}$$

where :

(i) $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_q)$ is the $(q \times 1)$ vector of all endogenous variables,
(ii) $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_p)$ is the $(p \times 1)$ vector of all exogenous variables,
(iii) $\mathbf{B}$ is the $(q \times q)$ lower triangular matrix of structural coefficients relating endogenous variables. Matrix $\mathbf{B}$ is always lower triangular for a recursive model.
(iv) $\boldsymbol{\Gamma}$ is the $(q \times p)$ matrix of structural coefficients relating endogenous variables to exogenous variables,
(v) $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \dots, \zeta_q)$ is the $(q \times 1)$ vector of disturbances.

In addition, the following assumptions are made:

(i) the vector of disturbances $\boldsymbol{\zeta}$ is not correlated to the vector of exogenous variables $\boldsymbol{\xi}$:

$$\begin{cases} \mathrm{E}(\boldsymbol{\xi}\boldsymbol{\zeta}^t) = 0 \\ \mathrm{E}(\boldsymbol{\zeta}\boldsymbol{\xi}^t) = 0 \end{cases} \tag{3.4}$$

where E denotes the expected value and ^{t} the transpose operation.
(ii) disturbances are not correlated, which implies that $\mathrm{E}(\boldsymbol{\zeta}\boldsymbol{\zeta}^t)$ is a $q \times q$ diagonal matrix.

The covariance matrix implied by the model described by System (3.2) is the $(p + q) \times (p + q)$ symmetric matrix $\widehat{\boldsymbol{\Sigma}}$ whose elements are the covariance between each pair of variables in the model. This matrix can be defined as:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathrm{E}(\boldsymbol{\xi}\boldsymbol{\xi}^t) & \mathrm{E}(\boldsymbol{\xi}\boldsymbol{\eta}^t) \\ \mathrm{E}(\boldsymbol{\eta}\boldsymbol{\xi}^t) & \mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}^t) \end{bmatrix}. \tag{3.5}$$

3.3 Extention of the Finite Iterative Method to the Covariance Matrix

This part presents FIM—a new method to compute the matrix $\widehat{\boldsymbol{\Sigma}}$. Its principle, shown in Fig. 3.2, is to build $\widehat{\boldsymbol{\Sigma}}$ iteratively. Starting with the first square block of order p (corresponding to the p exogenous variables $(\xi_1, \xi_2, \ldots, \xi_p)$), then the $(p+1)$th row and the $(p+1)$th column (corresponding to the first endogenous variable η_1) and ending with the $(p+q)$th row and $(p+q)$th column (corresponding to the last endogenous variable η_q).

This building process is possible by using System (3.2) directly without recourse to Eq. (3.3). This system can be written as follows:

$$\begin{aligned}
\eta_1 &= \gamma_{11}\xi_1 + \cdots + \gamma_{1p}\xi_p + \zeta_1 \\
\eta_2 &= \gamma_{21}\xi_1 + \cdots + \gamma_{2p}\xi_p + \beta_{21}\eta_1 + \zeta_2 \\
&\vdots \\
\eta_j &= \gamma_{j1}\xi_1 + \cdots + \gamma_{j,p}\xi_p + \beta_{j1}\eta_1 + \cdots + \beta_{j,(j-1)}\eta_{j-1} + \zeta_j \\
&\vdots \\
\eta_q &= \gamma_{q1}\xi_1 + \cdots + \gamma_{q,p}\xi_p + \beta_{q1}\eta_1 + \cdots + \beta_{q,(q-1)}\eta_{q-1} + \zeta_q
\end{aligned} \tag{3.6}$$

Thereafter, we denote by $\mathbf{A}$ the $q \times (p+q)$ matrix of *model parameters* defined as:

$$\mathbf{A} = \left[\, \boldsymbol{\Gamma} \;\; \mathbf{B} \,\right] = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1p} & 0 & \cdots & \cdots & 0 \\ \gamma_{21} & \cdots & \gamma_{2p} & \beta_{21} & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \vdots \\ \gamma_{q1} & \cdots & \gamma_{q,p} & \beta_{q1} & \cdots & \beta_{q,(q-1)} & 0 \end{bmatrix} \tag{3.7}$$

For fixed k between 1 and q, and m between 1 and $(p+q)$, the following notations are used:

$$\mathbf{A}_{1:k,1:m} = \left[a_{i,j}\right]_{1 \le i \le k, 1 \le j \le m}. \tag{3.8}$$

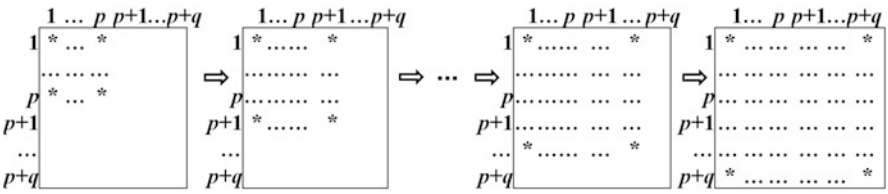

Fig. 3.2 Construction of the implied covariance matrix using the finite iterative method (FIM)

The Finite iterative method is defined by the following q (finite) iterations (Algorithm 3.1):

Algorithm 3.1: Finite iterative method

Repeat for $j = 1, 2, \ldots, q$

1. $\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{p+j,1:p+j-1} = \mathbf{A}_{j,1:p+j-1}\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:p+j-1,1:p+j-1}$
2. $\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:p+j-1,p+j} = \left(\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{p+j,1:p+j-1}\right)^t$
3. $\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{p+j,p+j} = s_{p+j,p+j}$

These iterations are initialized by $\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}{}_{1:p,1:p} = \boldsymbol{\Phi}$, the covariance matrix among exogenous variables

Theorem 3.1. *The matrix $\widehat{\boldsymbol{\Sigma}}$ can be computed by the algorithm above: $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\text{FIM}}$*

Proof. Variables $\xi_1, \ldots, \xi_p$ are all exogenous, therefore the block $\hat{\Sigma}^{\text{FIM}}_{1:p,1:p}$ corresponding to these variables is naturally identical to $\boldsymbol{\Phi}$, the covariance matrix among exogenous variables. Therefore

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}{}_{1:p,1:p} = \boldsymbol{\Phi}$$

1. The following notations are used:

   (i)
   $$\hat{\sigma}_{\xi_i\xi_{i'}} = \text{cov}\,(\xi_i, \xi_{i'}) = \text{E}\,(\xi_{i'}\xi_i) \qquad \forall (i, i') \in \{1, \ldots p\}$$

   (ii)
   $$\hat{\sigma}_{\eta_j\xi_i} = \text{cov}\left(\eta_j, \xi_i\right) = \text{E}\left(\eta_j\xi_i\right) \qquad \forall j \in \{1, \ldots q\} \text{ and } \forall i \in \{1, \ldots p\}$$

   (iii)
   $$\hat{\sigma}_{\eta_j\eta_{j'}} = \text{cov}\left(\eta_j, \eta_{j'}\right) = \text{E}\left(\eta_{j'}\eta_j\right) \qquad \forall (j, j') \in \{1, q\}$$

   Since the first equation in system (3.6) contains in the right term only the exogenous variables, we can separate it from the other equations. This first equation is,

   $$\eta_1 = \gamma_{11}\xi_1 + \cdots + \gamma_{1p}\xi_p + \zeta_1, \tag{3.9}$$

   thus, for $i \in \{1, p\}$ the multiplication of this equation on the right side by ξ_i gives:

   $$\eta_1\xi_i = \gamma_{11}\xi_1\xi_i + \cdots + \gamma_{1,p}\xi_p\xi_i + \zeta_1\xi_i \tag{3.10}$$

however, the model assumptions from Eq. (3.4) imply that ζ_1 is uncorrelated with all exogenous variables $\xi_1, \ldots \xi_p$. This in turn, implies that

$$\mathrm{E}\left(\zeta_1 \xi_i\right) = 0 \qquad \forall i \in \{1, \ldots p\}. \tag{3.11}$$

So, taking the respective mathematical expectation we obtain,

$$\hat{\sigma}_{\eta_1 \xi_i} = \gamma_{1,1} \hat{\sigma}_{\xi_1 \xi_i} + \cdots + \gamma_{1,p} \hat{\sigma}_{\xi_p \xi_i} \tag{3.12}$$

or equivalently,

$$\hat{\sigma}_{\eta_1 \xi_i} = \mathbf{A}_{1,1:p} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p,i}, \tag{3.13}$$

thus, for i in the set of integers between 1 and p we obtain,

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{p+1,1:p} = \left[\widehat{\boldsymbol{\sigma}}_{\eta_1 \xi_1}, \ldots, \widehat{\boldsymbol{\sigma}}_{\eta_1 \xi_p}\right] = \left[\mathbf{A}_{1,1:p} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p,1}, \ldots, \mathbf{A}_{1,1:p} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p,p}\right] = \mathbf{A}_{1,1:p} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p,1:p} \tag{3.14}$$

We now consider equations for the other endogenous variables $\eta_2, \ldots \eta_q$. Let $j \in \{2, q\}$, the structural equation for the jth endogenous variable η_j is:

$$\eta_j = \gamma_{j1} \xi_1 + \cdots + \gamma_{j,p} \xi_p + \beta_{j1} \eta_1 + \cdots + \beta_{j,(j-1)} \eta_{j-1} + \zeta_j. \tag{3.15}$$

Let $i \in \{1, p\}$ and $k \in \{1, j-1\}$, if we multiply this equation on the right side successively by ξ_i and η_k we obtain,

$$\eta_j \xi_i = \gamma_{j1} \xi_1 \xi_i + \cdots + \gamma_{j,p} \xi_p \xi_i + \beta_{j1} \eta_1 \xi_i + \cdots + \beta_{j,(j-1)} \eta_{j-1} \xi_i + \zeta_j \xi_i \tag{3.16}$$

and

$$\eta_j \eta_k = \gamma_{j1} \xi_1 \eta_k + \cdots + \gamma_{j,p} \xi_p \eta_k + \beta_{j1} \eta_1 \eta_k + \cdots + \beta_{j,(j-1)} \eta_{j-1} \eta_k + \zeta_j \eta_k \tag{3.17}$$

From Eq. (3.4), ζ_j is uncorrelated with all exogenous variables $\xi_1, \ldots$and ξ_p. Moreover, since η_j is explained by $\eta_1, \ldots$and η_{j-1}, ζ_j is uncorrelated with all the endogenous variables $\eta_1, \ldots$and η_{j-1}. Thus,

$$\begin{cases} \mathrm{E}\left(\zeta_j \xi_i\right) = 0 & \forall i \in \{1, p\} \\ \mathrm{E}\left(\zeta_j \eta_k\right) = 0 & \forall k \in \{1, j-1\} \end{cases} \tag{3.18}$$

So, taking the respective mathematical expectation, we obtain,

$$\hat{\sigma}_{\eta_j \xi_i} = \gamma_{j,1} \hat{\sigma}_{\xi_1 \xi_i} + \ldots + \gamma_{j,p} \hat{\sigma}_{\xi_p \xi_i} + \beta_{j,1} \hat{\sigma}_{\eta_1 \xi_i} + \ldots + \beta_{j,j-1} \hat{\sigma}_{\eta_{j-1} \xi_i}$$

and

$$\hat{\sigma}_{\eta_j \eta_k} = \gamma_{j,1} \hat{\sigma}_{\xi_1 \eta_k} + \ldots + \gamma_{j,p} \hat{\sigma}_{\xi_p \eta_k} + \beta_{j,1} \hat{\sigma}_{\eta_1 \eta_k} + \ldots + \beta_{j,j-1} \hat{\sigma}_{\eta_{j-1} \eta_k}$$

or, equivalently:

$$\hat{\sigma}_{\eta_j \xi_i} = \mathbf{A}_{j,1:p+j-1}\ \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,i}$$

and

$$\hat{\sigma}_{\eta_j \eta_k} = \mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,p+k},$$

thus for i in the set of integers between 1 and p, and k in the set of integers between 1 and $j-1$:

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{p+j,1:p} = \left[\hat{\sigma}_{\eta_j \xi_1}, \ldots, \hat{\sigma}_{\eta_j \xi_p}\right] = \left[\mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{FIM}{}_{1:p+j-1,1}, \ldots, \mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,p}\right]$$

and

$$\begin{aligned} \widehat{\Sigma}^{\mathrm{FIM}}_{p+j,p+1:p+j-1} &= \left[\hat{\sigma}_{\eta_j \eta_1}, \ldots, \hat{\sigma}_{\eta_j \eta_{j-1}}\right] \\ &= \left[\mathbf{A}_{j,1:p+j-1} \hat{\sigma}^{\mathrm{FIM}}_{1:p+j-1,p+1}, \ldots, \mathbf{A}_{j,1:p+j-1} \hat{\sigma}^{\mathrm{FIM}}_{1:p+j-1,p+j-1}\right] \end{aligned} \tag{3.19}$$

thus

$$\widehat{\Sigma}^{\mathrm{FIM}}{}_{p+j,1:p} = \mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,1:p} \tag{3.20}$$

and

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}_{p+j,p+1:p+j-1} = \mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}_{1:p+j-1,p+1:p+j-1}. \tag{3.21}$$

Together Eqs. (3.20) and (3.21) give

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{p+j,1:p+j-1} = \mathbf{A}_{j,1:p+j-1} \widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,1:p+j-1} \tag{3.22}$$

2. The matrix $\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}$ being a covariance matrix is symmetric and therefore:

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{1:p+j-1,p+j} = \left(\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{p+j,1:p+j-1}\right)^t \tag{3.23}$$

3. Since $\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}$ is also a covariance matrix then by definition:

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{FIM}}{}_{p+j,p+j} = s_{p+j,p+j} \tag{3.24}$$

□

3.4 Efficiency of FIM and Its Complementarity with Jöreskog's Method

The aim of this section is to analyze the efficiency of FIM, to compare it with Jöreskog's method, and discuss the relationships between these two methods. To do this, the model described in Fig. 3.1 is considered. The three structural equations in this model can be written as System (3.2) with,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ c & 0 & 0 \\ d & e & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} a \\ b \\ 0 \end{bmatrix}, \quad \boldsymbol{\Phi} = [s_{11}].$$

3.4.1 Jöreskog's Method

The expression given by Jöreskog to compute the covariance matrix implied by the model is as follows:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{\text{JOR}} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\Gamma}^t \left[(\mathbf{I}-\mathbf{B})^{-1}\right]^t \\ (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} & (\mathbf{I}-\mathbf{B})^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^t + \boldsymbol{\Psi}\right)\left[(\mathbf{I}-\mathbf{B})^{-1}\right]^t \end{bmatrix} \tag{3.25}$$

where $\boldsymbol{\Phi} = \mathrm{E}\left(\boldsymbol{\xi}\boldsymbol{\xi}^t\right) = \left[\mathrm{cov}\left(\xi_j, \xi_i\right)\right]_{1\leq j,i\leq p}$ is the $(p \times p)$ covariance matrix among exogenous variables and $\boldsymbol{\Psi} = \mathrm{E}\left(\boldsymbol{\zeta}\boldsymbol{\zeta}^t\right) = \left[\mathrm{cov}\left(\zeta_j, \zeta_i\right)\right]_{1\leq j,i\leq}$ is the $(q \times q)$ covariance matrix among disturbances.

3.4.2 Computation of $\widehat{\boldsymbol{\Sigma}}$ by Jöreskog's Method

For Jöreskog's method we start with

$$(\mathbf{I}-\mathbf{B})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ d+ce & e & 1 \end{bmatrix},$$

and thus

$$\mathrm{E}\left(\eta\xi^t\right) = \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ d+ce & e & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \times s_{11} = \begin{bmatrix} as_{11} \\ (ac+b)s_{11} \\ (ad+ace+be)s_{11} \end{bmatrix}, \tag{3.26}$$

and also

$$\mathrm{E}\left(\boldsymbol{\xi}\boldsymbol{\eta}^t\right) = \left(\mathrm{E}\left(\boldsymbol{\eta}\boldsymbol{\xi}^t\right)\right)^t = \left[\, as_{11} \;\; (ac+b)s_{11} \;\; (ad+ace+be)s_{11} \,\right].$$

However, the matrix $\boldsymbol{\Psi}$ is diagonal with elements:

$$\begin{cases} \theta_1^2 = \mathrm{var}\left(\boldsymbol{\zeta}_1\right) \\ \theta_2^2 = \mathrm{var}\left(\boldsymbol{\zeta}_2\right) \\ \theta_3^2 = \mathrm{var}\left(\boldsymbol{\zeta}_3\right) \end{cases}$$

Then

$$\boldsymbol{\Psi} = \begin{bmatrix} \theta_1^2 & 0 & 0 \\ 0 & \theta_2^2 & 0 \\ 0 & 0 & \theta_3^2 \end{bmatrix}.$$

Thus

$$\mathrm{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}^t\right) = \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ d+ce & e & 1 \end{bmatrix} \left(\begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \times s_{11} \begin{bmatrix} a & b & 0 \end{bmatrix} + \begin{bmatrix} \theta_1^2 & 0 & 0 \\ 0 & \theta_2^2 & 0 \\ 0 & 0 & \theta_3^2 \end{bmatrix} \right) \begin{bmatrix} 1 & c & d+ce \\ 0 & 1 & e \\ 0 & 0 & 1 \end{bmatrix},$$

thus

$$\mathrm{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}^t\right) = \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ d+ce & e & 1 \end{bmatrix} \begin{bmatrix} a^2 s_{11} + \theta_1^2 & abs_{11} & 0 \\ abs_{11} & b^2 s_{11} + \theta_2^2 & 0 \\ 0 & 0 & \theta_3^2 \end{bmatrix} \begin{bmatrix} 1 & c & d+ce \\ 0 & 1 & e \\ 0 & 0 & 1 \end{bmatrix},$$

and therefore, the covariance matrix implied by the model described by System (3.25) and provided by Jöreskog's method is written as:

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{JOR}} = \begin{bmatrix} [s_{11}] & \left[\, as_{11} \;\; (ac+b)s_{11} \;\; (ad+ace+be)s_{11} \,\right] \\ \begin{bmatrix} as_{11} \\ (ac+b)s_{11} \\ (ad+ace+be)s_{11} \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ d+ce & e & 1 \end{bmatrix} \begin{bmatrix} a^2 s_{11} + \theta_1^2 & abs_{11} & 0 \\ abs_{11} & b^2 s_{11} + \theta_2^2 & 0 \\ 0 & 0 & \theta_3^2 \end{bmatrix} \begin{bmatrix} 1 & c & d+ce \\ 0 & 1 & e \\ 0 & 0 & 1 \end{bmatrix} \end{bmatrix} \tag{3.27}$$

3.4.3 *Computation of $\widehat{\boldsymbol{\Sigma}}$ by the Finite Iterative Method*

For the finite iterative method we consider

$$\mathbf{A} = \begin{bmatrix} a & 0 & 0 & 0 \\ b & c & 0 & 0 \\ 0 & d & e & 0 \end{bmatrix}.$$

The covariance matrix implied by the model $\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}$ provided by FIM is built in three iterations ($q = 3$) and Algorithm 3.1 is initialized by

$$\widehat{\boldsymbol{\Sigma}}^{FIM}_{1:1,1:1} = \boldsymbol{\Phi} = [s_{11}]\,.$$

First iteration ($q = 1$): Computation for the first endogenous variable.

$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{2,1:1} = \mathbf{A}_{1,1:1}\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:1,1:1} = a \times s_{11} = as_{11}$
$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:1,2} = \left(\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{2,1:1}\right)^{t} = as_{11}$
$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{2,2} = s_{22}$, thus

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:2,1:2} = \begin{bmatrix} s_{11} & as_{11} \\ as_{11} & s_{22} \end{bmatrix}$$

Second iteration ($q = 2$): Computation for the second endogenous variable.

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{3,1:2} = \mathbf{A}_{2,1:2}\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:2,1:2} = \begin{bmatrix} b & c \end{bmatrix}\begin{bmatrix} s_{11} & as_{11} \\ as_{11} & s_{22} \end{bmatrix} = \begin{bmatrix} (b+ac)s_{11} & abs_{11} + cs_{22} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:2,3} = \left(\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{3,1:2}\right)^{t} = \begin{bmatrix} (b+ac)s_{11} \\ abs_{11} + cs_{22} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{3,3} = s_{33},$$

and thus

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:3,1:3} = \begin{bmatrix} s_{11} & as_{11} & (b+ac)s_{11} \\ as_{11} & s_{22} & abs_{11} + cs_{22} \\ (b+ac)s_{11} & abs_{11} + cs_{22} & s_{33} \end{bmatrix}$$

Third iteration ($q = 3$): Computation for the third endogenous variable.

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{4,1:3} = \mathbf{A}_{3,1:3}\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:3,1:3} = \begin{bmatrix} 0 & d & e \end{bmatrix}\begin{bmatrix} s_{11} & as_{11} & (b+ac)s_{11} \\ as_{11} & s_{22} & abs_{11} + cs_{22} \\ (b+ac)s_{11} & abs_{11} + cs_{22} & s_{33} \end{bmatrix}$$

$$= \begin{bmatrix} ads_{11} + (b+ac)es_{11} & ds_{22} + e(abs_{11} + cs_{22}) & d(abs_{11} + cs_{22}) + es_{33} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{1:3,4} = \left(\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{4,1:3}\right)^{t} = \begin{bmatrix} ads_{11} + (b+ac)es_{11} \\ ds_{22} + e(abs_{11} + cs_{22}) \\ d(abs_{11} + cs_{22}) + es_{33} \end{bmatrix}$$

$\widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{4,4} = s_{44}$, and thus

Fig. 3.3 Comparison between Jöreskog's method and "FIM"

	Jöreskog's method	FIM
Φ	*	*
Γ	*	*
B	*	*
$(\mathrm{I}-\mathrm{B})^{-1}$	*	
Ψ	*	

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{FIM}} = \begin{bmatrix} s_{11} & as_{11} & (b+ac)s_{11} & ads_{11}+(b+ac)es_{11} \\ as_{11} & s_{22} & abs_{11}+cs_{22} & ds_{22}+e(abs_{11}+cs_{22}) \\ (b+ac)s_{11} & abs_{11}+cs_{22} & s_{33} & d(abs_{11}+cs_{22})+es_{33} \\ ads_{11}+(b+ac)es_{11} & ds_{22}+e(abs_{11}+cs_{22}) & d(abs_{11}+cs_{22})+es_{33} & s_{44} \end{bmatrix}$$

3.4.4 Which Method Is More Flexible?

The answer to this question is summarized in Fig. 3.3. This figure lists the matrices needed to compute the implied covariance matrix for a given recursive model.

As clearly shown in Eq. (3.25), Jöreskog's method requires prior knowledge of the inverse matrix $(\mathbf{I}-\mathbf{B})^{-1}$, and computes $\hat{\boldsymbol{\Sigma}}$ as a function of the variances of disturbances (see Eq. 3.27). By contrast, FIM requires neither the inverse matrix $(\mathbf{I}-\mathbf{B})^{-1}$ nor the matrix $\boldsymbol{\Psi}$ but provides directly the matrix $\hat{\boldsymbol{\Sigma}}$ (the specific computations depend only upon matrices $\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$, which are all given). As a consequence, FIM seems more flexible than Jöreskog's method in computing the covariance matrix implied by the model.

3.4.5 Is FIM More Efficient Than Jöreskog's Method?

The answer to this question depends on the matrix of variances of disturbance $\boldsymbol{\Psi}$. When this matrix can be exhibited from $\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$, the efficiency of both methods is identical. Unfortunately, no known method allows computation of the matrix $\boldsymbol{\Psi}$ from $\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$. This leads to the conclusion that FIM is more efficient than Jöreskog's method. In practice, however, Jöreskog's method does not provide exactly the matrix $\hat{\boldsymbol{\Sigma}}$, it provides an approximation of it, denoted $\hat{\boldsymbol{\Sigma}}^{\mathrm{app}}$ because an approximation of $\boldsymbol{\Psi}$ is used.

To illustrate this situation, 100 data sets were randomly generated. Four variables are considered for each data set. The model in Fig. 3.1 is considered for each data set, the matrix $\boldsymbol{\Phi}$ is taken as the empirical covariance matrix between exogenous

variables, and the matrices $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are estimated as solutions of the minimization criterion

$$\frac{1}{2}\text{trace}\left[\left(\mathbf{S}-\widehat{\boldsymbol{\Sigma}}\right)\right]^2$$

with $\mathbf{S}$ being the empirical covariance. The `lavaan` package (Rosseel 2002) was used to estimate the model for each simulation. Thereby, for each data set, the matrix $\hat{\boldsymbol{\Sigma}}^{\text{app}}$ provided by Jöreskog's method was obtained as the output of the `lavaan` package, and the matrix $\hat{\boldsymbol{\Sigma}}^{\text{FIM}}$ was computed following Algorithm 3.1. We measured the distance between these two matrices as:

$$\Delta = \frac{1}{2}\text{trace}\left[\hat{\boldsymbol{\Sigma}}^{\text{app}} - \hat{\boldsymbol{\Sigma}}^{\text{FIM}}\right]^2. \tag{3.28}$$

Figure 3.4 displays the quantity Δ for each simulation. This figure shows that this difference is negligible as it does not exceed 0.0002. The most important result of these simulations is judging the quality of this approximation.

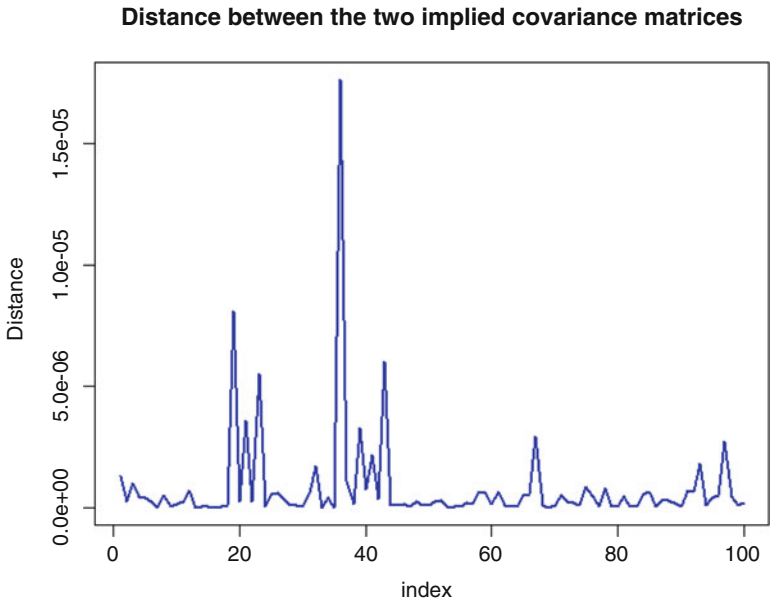

Fig. 3.4 Distance between the two implied matrices $\hat{\boldsymbol{\Sigma}}^{\text{App}}$ and $\hat{\boldsymbol{\Sigma}}^{\text{FIM}}$ for 100 simulations

3.4.6 Are Both Methods Complementary?

The simultaneous use of both methods has the advantage of determining exactly the matrix of the variances of disturbance $\boldsymbol{\Psi}$ as a function of parameters ($\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$, and $\mathbf{B}$). This saves the user from worrying about approximations. Indeed, the matrix $\boldsymbol{\Psi}$ can be calculated as follows:

(i) matrices $\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are given,
(ii) compute $\hat{\boldsymbol{\Sigma}}^{\text{FIM}}$,
(iii) compute $\boldsymbol{\Psi}$ as:

$$\boldsymbol{\Psi} = (\mathbf{I} - \mathbf{B})\, \widehat{\boldsymbol{\Sigma}}^{FIM}_{p+1:p+q,p+1:p+q} \left[(\mathbf{I} - \mathbf{B})^t - \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^t\right] \tag{3.29}$$

Indeed, from

$$\widehat{\boldsymbol{\Sigma}}^{\text{JOR}} = \widehat{\boldsymbol{\Sigma}}^{\text{FIM}}$$

and by identifying the blocks corresponding to the endogenous variables,

$$(\mathbf{I} - \mathbf{B})^{-1} \left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^t + \Psi\right) \left[(\mathbf{I} - \mathbf{B})^{-1}\right]^t = \widehat{\boldsymbol{\Sigma}}^{\text{FIM}}_{p+1:p+q,p+1:p+q}$$

And by applying (3.29) to the model of Fig. 3.1, the matrix Ψ is as follows :

$$\begin{aligned}
\boldsymbol{\Psi} = &\begin{bmatrix} 1 & 0 & 0 \\ -c & 1 & 0 \\ -d & -e & 1 \end{bmatrix} \\
&\times \begin{bmatrix} s_{22} & abs_{11} + cs_{22} & ds_{22} + e(abs_{11} + cs_{22}) \\ abs_{11} + cs_{22} & s_{33} & d(abs_{11} + cs_{22}) + es_{33} \\ ds_{22} + e(abs_{11} + cs_{22}) & d(abs_{11} + cs_{22}) + es_{33} & s_{44} \end{bmatrix} \\
&\times \left(\begin{bmatrix} 1 & -c & -d \\ 0 & 1 & -e \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \times [s_{11}] \times \begin{bmatrix} a & b & 0 \end{bmatrix} \right)
\end{aligned} \tag{3.30}$$

and thus,

$$\boldsymbol{\Psi} = \begin{bmatrix} s_{22} - as_{11} & 0 & 0 \\ 0 & s_{33} - b^2 s_{11} - c^2 s_{22} - 2abcs_{11} & 0 \\ 0 & 0 & s_{44} - d^2 s_{22} - e^2 s_{33} - 2abdes_{11} - 2dces_{22} \end{bmatrix}$$

To conclude, FIM does not determine the variances of disturbances, and Jöreskog's method does not allow the computation of $\widehat{\boldsymbol{\Sigma}}$. The simultaneous use of the two methods make it possible to compute both $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Psi}$.

3.5 Conclusion and Perspectives

The computation of the covariance matrix implied by a recursive path model is a crucial step in the global path analysis process. This paper proposes a new method called Finite Iterative Method (FIM) which will enrich and expand strategies for computing this matrix. On a conceptual level, FIM has two advantages compared to the well-known method of Jöreskog. First, it does not need matrix inversion and second, knowledge of variances of disturbances is not necessary. On a practical level, if we get good approximations of variances of disturbances, this new method is equivalent to Jöreskog's method.

The question that still remains to be answered and always open is to propose similar strategies of computation for non-recursive models, and for models with latent variables.

References

Blalock, H.: Causal Models in the Social Sciences. Aldine-Atherton, Chicago (1971)
Boudon, R.: Méthodes d'analyse causale. Revue Francaise de Sociologie **6**, 24–43 (1965)
Duncan, O.D.: Path analysis: sociological examples. Am. J. Soc. **72**, 1–16 (1966)
El Hadri, Z., Hanafi, M.: The finite iterative method for calculating the correlation matrix implied by a recursive path model. Electron. J. Appl. Stat. Anal. **8**, 84–99 (2015)
Hauser, R.M., Sewall, W.H.: Education, Occupation, and Earnings. Academic, New York (1975)
Heise, D.R.: Problems in path analysis and causal inference. In: Borgatta, E.F. (ed.) Sociological Methodology, pp. 38–73. Jossey-Bass, San Francisco (1969)
Hoyle, R.H.: Structural Equation Modeling: Concepts, Issues, and Applications. Sage, Thousand Oaks (1995)
Jöreskog, K.G.: Structural equation models in the social sciences: specification, estimation and testing. In: Krishnaiah, R. (ed.) Applications of Statistics, pp. 265–287. North-Holland, Amsterdam (1977)
Jöreskog, K.G., Wold, H.: Systems Under Indirect Observation: Causality, Structure, Prediction, Part I and Part II. North Holland, Amsterdam (1982)
Kline, R.B.: Practice of Principles of Structural Equation Modeling, 4th edn. Sage, Thousand Oaks (2016)
Rosseel, Y.: `lavaan`: an `R` package for structural equation modeling. J. Stat. Softw. **48**(2), 1–36 (2002)
Sanchez-Pinero, F., Polis, G.A.: Bottom-up dynamics of allochthonous input: direct and indirect effects of seabirds on islands. Ecology **81**, 3117–3132 (2000)
Shine, R.: Life-history evolution in Australian snakes: a path analysis. Oecologia **107**, 484–489 (1996)
Shipley, B.: Testing causal explanations in organismal biology: causation, correlation and structural equation modeling. Oikos **86**, 374–382 (1999)
Wolfle, L.M.: The introduction of path analysis to the social sciences, and some emergent themes: an annotated bibliography. Struct. Eq. Model. **10**, 1–34 (2003)
Wright, S.: Correlation and causation. J. Agric. Res. **20**, 557–585 (1921)

Chapter 4
Which Resampling-Based Error Estimator for Benchmark Studies? A Power Analysis with Application to PLS-LDA

Anne-Laure Boulesteix

Abstract Resampling-based methods such as *k*-fold cross-validation or repeated splitting into training and test sets are routinely used in the context of supervised statistical learning to assess the prediction performances of prediction methods using real data sets. In this paper, we consider methodological issues related to comparison studies of prediction methods which involve several real data sets and use resampling-based error estimators as the evaluation criteria. In the literature papers often claim that, say, "Method 1 performs better than Method 2 on real data" without applying any proper statistical inference approach to support their claims and without clearly explaining what they mean by "perform better." We recently proposed a new statistical testing framework which provides a statistically correct formulation of such paired tests—which are often performed in the machine learning community—to compare the performances of two methods on several real data sets. However, the behavior of the different available resampling-based error estimation procedures in this statistical framework is unknown. In this paper we empirically assess this behavior through an exemplary benchmark study based on 50 microarray data sets and formulate tentative recommendations regarding the choice of resampling-based error estimation procedures in light of the results.

Keywords Resampling • *K*-fold cross-validation • Supervised statistical learning • Statistical inference

4.1 Introduction

Resampling-based methods such as *k*-fold cross-validation or repeated splitting into training and test sets are routinely used in the context of supervised statistical learning to assess the prediction performance of prediction methods using real

A.-L. Boulesteix (✉)
Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany
e-mail: boulesteix@ibe.med.uni-muenchen.de

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_4

data sets. In this paper, we consider methodological issues related to comparison studies of prediction methods which involve several data sets and use resampling-based error estimators as evaluation criteria. Here such comparisons will be termed *benchmark studies* and the number of considered real data sets will be denoted as J.

In statistical literature most articles presenting new supervised learning methods include a sentence like "our method performed better than existing methods on real data sets"— usually in terms of prediction error (Boulesteix 2013; Boulesteix et al. 2013). However, these claims are often not based on proper statistical inference. To contrast, in machine learning literature it is common to compare the resampling-based error estimates of prediction methods using statistical tests such as paired t-tests or paired Wilcoxon tests (Demsar 2006). This approach consists (in the case of the t-test) of testing whether the means of the prediction errors are equal for the two considered supervised learning methods by considering the differences between the resampling-based estimates of the two methods obtained for each of the J real data sets as independent realizations. The tested hypothesis, however, is almost never clearly defined and poor attention is usually paid to the Type I and Type II error. To our knowledge, a proper statistical framework for these ad-hoc tests is, until our recent contribution (Boulesteix et al. 2015), missing from the literature completely.

To address this issue we proposed a new statistical testing framework which provides a statistically correct formulation of such paired tests (Boulesteix et al. 2015). With this framework in mind, we also examined benchmark studies in terms of their power to detect existing differences in performance in relation to the number J of data sets used in the study. However, we illustrated these ideas based on a single resampling procedure, namely repeated subsampling with a splitting ratio between training and test data sets of 4:1 (i.e., the training sets include 4/5 and the test sets 1/5 of the data).

The power of the considered paired test depends highly on the variance of the estimates over the data sets. The variance of resampling-based error estimates has been relatively well investigated in the context of simulation studies given a pre-specified underlying data distribution (Dougherty et al. 2010; Molinaro et al. 2005). In the context of real-data-based benchmark studies which include various data sets with different underlying distributions, however, this variance has to our knowledge never been examined systematically. Hence, the behavior of common resampling-based error estimators in our statistical testing framework is essentially unknown. This paper aims at filling this gap.

In Sect. 4.2 we give a brief overview of the considered resampling-based error estimation procedures: k-fold cross-validation (CV), leave-one-out cross-validation (LOOCV), repeated subsampling (SUB) and repeated bootstrapping (BOOT) and of our recently proposed statistical framework (Boulesteix et al. 2015) with particular emphasis on power issues. Section 4.3 presents the design of our empirical analyses and the obtained results.

4.2 Methods

In this paper we focus on binary class prediction problems, i.e. the response variable Y to be predicted is binary ($Y = 0, 1$). Prediction rules are constructed based on p predictor variables $X_1, \ldots, X_p$. The data set at hand is denoted by D and consists of n *i.i.d.* realizations of the random vector $(Y, X_1, \ldots, X_p)^\top$. If several data sets are considered they are denoted by $D_1, \ldots, D_J$, where J is the number of data sets in the benchmark study. Classification methods used to construct prediction rules are denoted by $M_1, \ldots, M_K$, where K is the number of methods compared in the study. Since the response Y is binary, the prediction error of a prediction rule is simply defined as the proportion of misclassified observations (or error rate) when this prediction rule is applied to make class predictions for test data. Note that other error criteria might be considered to assess classification accuracy. If Y is a continuous variable prediction error is typically assessed through the mean squared difference between predicted and true values. The testing framework and analyses presented in this paper are essentially generalizable to such other error criteria, though we focus on the error rate for simplicity.

4.2.1 Resampling-Based Methods for Prediction Error Estimation

In this paper we consider the following common resampling-based error estimation procedures: k-fold cross-validation (CV), leave-one-out cross-validation (LOOCV), repeated subsampling (SUB) and repeated bootstrapping (BOOT).

4.2.1.1 *K*-Fold Cross-Validation (CV)

In k-fold CV, the available data set is partitioned into k approximately equally sized folds. In each of the k CV iterations, the kth fold is considered the test data set and the $k - 1$ remaining folds form the training data set. The considered supervised learning method is used to fit a prediction rule from the training set, which is subsequently applied to the test set. Overall prediction error is assessed for all iterations by comparing the predicted and true value response variable and the average is built over the k iterations. To reduce the variability of this error estimate it is recommended (but not at all systematic in the literature) to repeat this procedure B_{CV} times for different random partitions of the data sets and to finally average the results over these B_{CV} repetitions. On the whole, if CV is repeated B_{CV} times, it implies that $B_{CV} \times k$ splittings into training and test sets are considered successively, which has to be kept in mind when comparing CV to other procedures such as repeated subsampling or repeated bootstrapping. The number of folds k is a parameter which is chosen by the user. Usual choices are $k = 3$, $k = 5$ or $k = 10$. The ratio between the sizes of the training and test sets is $(k - 1) : 1$.

4.2.1.2 Leave-One-Out Cross-Validation (LOOCV)

A special case of CV is when each fold consists of a single observation, corresponding to $k = n$. This variant of CV is known as leave-one-out cross-validation (LOOCV). The partition is then unique. Hence, LOOCV cannot be repeated, in contrast to k-fold CV with $k < n$. It yields a deterministic error estimate.

4.2.1.3 Repeated Subsampling (SUB)

Repeated subsampling (SUB) is similar to CV in the sense that it also considers splits of the data set D into training and test sets. But it differs from CV because these splits do not result from a single partition into k folds. Instead, at each subsampling iteration a new random partition into training and test sets is generated independently of the previous iterations. In other words, a training set is drawn without replacement out of the available data set (hence the term subsampling) and the rest of the data set forms the test set. As with CV, the ratio between the sizes of the training and test sets is a parameter which is chosen by the user. Usual choices are, for example, 2:1, 4:1 or 9:1. Note that these ratios correspond to the ratios of three-fold CV, five-fold CV and ten-fold CV, respectively. This procedure is repeated a large number B_{SUB} of times and the average error is taken over the B_{SUB} iterations.

4.2.1.4 Repeated Bootstrapping (BOOT)

The last resampling-based error estimation procedure which we will consider is bootstrapping. Repeated bootstrapping is very similar to repeated subsampling with the difference that at each iteration the training set is drawn from the available data set with replacement, i.e. observations are allowed to occur several times in the training set. Once the training set is drawn, the rest of the data set is taken as test set, as with the repeated subsampling procedure: note here however that the size of the test set varies at each iteration, depending on how many duplicates are included in the training set. It is usual to draw training sets of size n from the available data set, in which case an average of 63.2 % of the original observations are included in each training set. If this standard approach is adopted, the repeated bootstrap procedure does not involve any parameters. As with repeated subsampling, the procedure is repeated a large number B_{BOOT} of times and the average error is taken over the B_{BOOT} iterations.

4.2.2 Statistical Testing Framework for Real-Data Benchmark Studies

We recently proposed a proper statistical framework for hypothesis tests comparing the performances of supervised learning methods using several real data sets with unknown underlying distributions (Boulesteix et al. 2015). This statistical framework is briefly summarized here. For this purpose, we have to introduce the well-known concepts of *conditional* and *unconditional* errors.

Let us consider a data set $D_0 \sim P_0^{n_0}$, where n_0 denotes the number of observations in D_0 and P_0 the underlying distribution from which the data set is drawn. The *conditional* error $\varepsilon(M_k, D_0, P_0)$ of method M_k (for $k \in \{1, \ldots, K\}$) constructed from D_0 is defined as $\varepsilon(M_k, D_0, P_0) = \mathrm{E}_{P_0}(\hat{f}_{M_k,D_0}(X_1, \ldots, X_p) \neq Y)$, where $\hat{f}_{M_k,D_0}$ stands for the prediction rule constructed from D_0 with method M_k taking the predictors $X_1, \ldots, X_p$ as input and returning a prediction for Y.

If we consider D_0 as a random variable with distribution $P_0^{n_0}$, we can define the *unconditional* error $\varepsilon^*(M_k, n_0, P_0)$ of method M_k for distribution P_0 and size n_0 as $\varepsilon^*(M_k, n_0, P_0) = \mathrm{E}_{P_0^{n_0}}(\varepsilon(M_k, D_0, P_0))$, where the asterisk indicates that we are now considering the unconditional error.

Coming back to the problem of benchmarking based on J data sets $D_1, \ldots, D_J$, we denote by $P_1, \ldots, P_J$ their respective underlying distributions and $n_1, \ldots, n_J$ their respective numbers of observations. The data set D_j is thus a realization of $P_j^{n_j}$. The basic idea of our previously proposed statistical framework (Boulesteix et al. 2015) consists of considering P_j as the outcome of a random variable Φ taking values in the set of the possible underlying distributions and n_j as the outcome of a random variable N taking values in $\mathbb{N}$. The random variables $(\Phi_1, N_1), \ldots, (\Phi_J, N_J)$ are *i.i.d.*. Note that, for $j \in \{1, \ldots, J\}$ we cannot observe $\Phi_j = P_j$ but only a data set D_j of size n_j.

We now return to error estimation and denote $e_j(M_k)$ the error estimate for method M_k obtained using a chosen resampling-based error estimation procedure on data set D_j, for $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$. For simplicity, we will consider the case of the comparison of two methods M_1 and M_2. The difference between the two error estimates is $\Delta e_j = e_j(M_2) - e_j(M_1)$. The statistic of the paired t-test often performed in machine learning to compare the prediction errors of M_1 and M_2 based on data sets $D_1, \ldots, D_J$ can be formulated as

$$T = \frac{\overline{\Delta e}}{\sqrt{\frac{1}{J}\frac{1}{J-1}\sum_{j=1}^{J}(\Delta e_j - \overline{\Delta e})^2}},$$

where $\overline{\Delta e}$ stands for the empirical mean of $\Delta e_1, \ldots, \Delta e_J$. Researchers performing this t-test usually do not clearly state the tested null hypothesis.

With the above theoretical framework in mind, it becomes intuitive (and it can be shown Boulesteix et al. 2015) that the null hypothesis implicitly tested when

conducting (the one-sided version of) this paired t-test in a benchmark study is of the type:

$$H_0: \ \mathrm{E}(\varepsilon^*(M_2, N, \Phi) - \varepsilon^*(M_1, N, \Phi)) \geq 0 \text{ vs. } H_1: \ \mathrm{E}(\varepsilon^*(M_2, N, \Phi) - \varepsilon^*(M_1, N, \Phi)) < 0.$$

Readers are referred to the original paper Boulesteix et al. (2015) for more details.

4.2.3 Power Considerations

Considering the one-sided paired t-test outlined above, we can use standard formulae to derive the number of data sets J needed to detect a difference of δ at a power of $1 - \beta$ given that the standard deviation of the difference is σ (Bock 1998; Boulesteix et al. 2015):

$$J(\alpha, \beta, \delta, \sigma) \approx \frac{[t_{1-\alpha,J-1} + t_{1-\beta,J-1}]^2}{(\delta/\sigma)^2}, \tag{4.1}$$

with $t_{\alpha,df}$ standing for the α-quantile of Student's distribution with df degrees of freedom.

4.3 Analyses and Results

4.3.1 Aims and Design of Our Analyses

A question that remains unanswered, however, is whether the different common resampling-based approaches reviewed in Sect. 2.1 behave identically in terms of power within real-data benchmark studies. The variance and mean squared error of these methods as estimators of the unconditional prediction error have been extensively studied in the literature through simulations and theoretical considerations (Dougherty et al. 2010; Molinaro et al. 2005) for given underlying distributions of response and predictors. For example, LOOCV is known to have a small bias (since the considered data sets are of size $n - 1$, i.e. almost n) but a large variance because it is highly "data set dependent" and yields very different estimates depending on the data set at hand. In contrast, resampling procedures which use smaller training data sets have a larger positive bias (i.e., they overestimate the error) but a smaller variance. Existing literature mainly focuses on the properties of these error estimators for a given underlying distribution of response and predictors. Variability between data sets arises because the considered data sets are randomly sampled from this distribution.

It is unclear, however, how these properties impact the results of real-data-based benchmark studies. In real-data-based benchmark studies, the considered data sets

are drawn from different underlying distributions. Variability between data sets is not only due to random sampling but also to the variability of the underlying distributions. In this study, we aim at filling this gap by empirically investigating the behavior of the considered resampling-based error estimates in terms of variability and power in the context of a real-data-based benchmark study.

For this purpose, we consider an application of supervised classification methods based on Partial Least Squares (PLS) components to a collection of 50 high-dimensional microarray data sets used in two previous studies (Boulesteix et al. 2015; de Souza et al. 2010). More precisely, we consider the "PLS+LDA" classification method (Boulesteix 2004) as implemented in the R package CMA (Slawski et al. 2008): this consists of applying linear discriminant analysis (LDA) to the c first principal PLS components constructed by the SIMPLS algorithm considering the binary response as metric (with values 0 or 1). In our study the parameter c is set successively to the three values $c = 1, 2, 3$. Moreover, since this method is known to often yield better accuracy when applied to a subset of pre-selected predictor variables, we perform preliminary variable selection before applying the PLS+LDA algorithm. Variable selection is performed by (i) applying to each predictor variable successively a standard t-test to test the equality of the means of the two groups $Y = 0$ and $Y = 1$ and (ii) retaining the p^* variables yielding the smallest p-values. In our study the parameter p^* is successively set to the four values $p^* = 100, 200, 500, 1000$.

In total, since we have three values of c and four values of p^*, we consider $3 \times 4 = 12$ variants of the PLS + LDA classification method; we thus conduct $11 \times 12/2 = 66$ pairwise comparisons using the paired t-test described in Sect. 4.2.2, for each resampling-based estimation method. The used methods are: leave-one-out cross-validation, cross-validation with different numbers of folds (3, 5, and 10) and different numbers of repetitions, repeated subsampling with different training/test ratios (2:1, 4:1 and 9:1) and different numbers of iterations, and repeated bootstrapping with different numbers of iterations.

Note that the training/test ratio is identical in three-fold CV and 2:1 subsampling, in five-fold CV and 4:1 subsampling and in ten-fold CV and 9:1 subsampling. To make CV and subsampling completely comparable and ensure that observed differences are due to the partitioning scheme and not to the number of iterations, we set the number of CV repetitions and the number of subsampling iterations for each CV-subsampling pair in such a way that they yield the same total numbers of training/test iterations. For example, repeating five-fold CV 10 times corresponds to $B_{SUB} = 50$ iterations in repeated subsampling.

The real-data-based benchmark study is based on a collection of $J = 50$ high-dimensional (i.e., large p) microarray data sets of different moderate sizes with binary response variable (e.g., diseased versus healthy) previously used in the literature (Boulesteix et al. 2015; de Souza et al. 2010). These data sets as well as the R code reproducing all our results are publicly available from http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/plsproc.

4.3.2 Results

4.3.2.1 Results of the Benchmark Study

Although the purpose of this paper is to provide insights into the behavior of the testing procedure as influenced by the resampling scheme and not to discuss the results of the respective performances of the considered variants of PLS + LDA, we first briefly present key results in this section for three-fold CV repeated 333 times, since this procedure shows important advantages in our later analyses. It can be seen from Fig. 4.1 (left) which displays the histogram of the p-values obtained for the $12 \times 11/2 = 66$ pairs of variants with three-fold CV repeated 333 times, that many differences are significant. In particular, the smallest p-values are obtained for the comparisons $c = 1, p^* = 1000$ vs. $c = 2, p^* = 1000$ (p-value $= 8 \cdot 10^{-6}$), $c = 1, p^* = 500$ vs. $c = 2, p^* = 500$ (p-value $= 2 \cdot 10^{-5}$) and $c = 1, p^* = 1000$ vs. $c = 3, p^* = 1000$ (p-value $= 2 \cdot 10^{-5}$).

4.3.2.2 Ordering of the Methods and Splitting Ratio

One of the central questions addressed in this paper is the influence of the splitting ratio between training and test sets on the testing procedure (i.e., in the case of CV, the influence of the number of folds). The mean estimated error over the $J = 50$ data sets is displayed in Fig. 4.1 (right) for three-fold CV, five-fold CV and ten-fold CV repeated several times such that all estimates are based on ≈1000 splits into training and test data sets. A very similar picture is obtained if repeated subsampling (SUB) is considered in place of CV with splitting ratios 2:1, 4:1 and 9:1, respectively (data not shown). As expected from the theory of statistical learning, the mean estimated error consistently decreases with an increasing number of folds, i.e. with

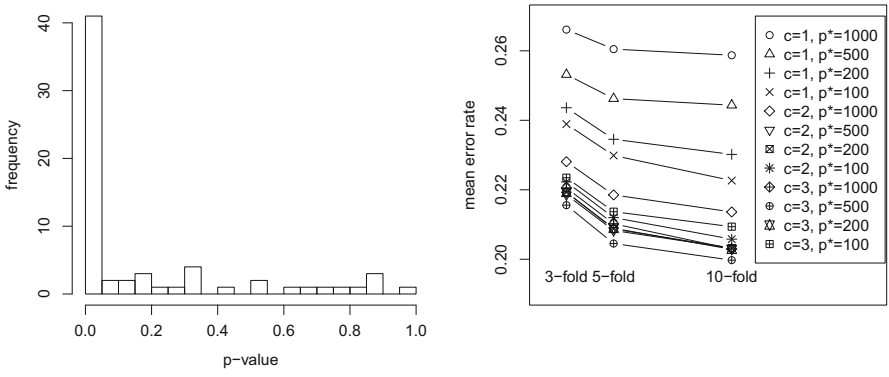

Fig. 4.1 *Left:* Histogram of the p-values obtained for the $12 \times 11/2 = 66$ pairs of variants with three-fold CV repeated 333 times. *Right:* Mean estimated error over the $J = 50$ data sets for the $K = 12$ considered variants with three-fold CV, five-fold CV and ten-fold CV

an increasing size of the training sets. This decrease, however, is moderate. A more interesting result is that, in this study, the curves of the $K = 12$ considered methods almost never cross and even remain approximately parallel. In other words, the variants ranking best with three-fold CV also rank best with five-fold CV and ten-fold CV.

The latter result is not obvious. We could also imagine scenarios where some variants for instance cope better with small training sets than others but perform worse when training sets get larger, in which case the curves might cross between three-fold CV and ten-fold CV. Such a scenario is not observed in the present study. This might seem comforting in the sense that the final ranking of methods resulting from the benchmark studies is not strongly influenced by arbitrary parameters like the number of CV folds or the splitting ratio—at least in our setting.

4.3.2.3 Absolute Value of the Mean Difference and Resampling Scheme

Figure 4.2 displays the absolute value $|\overline{\Delta e}|$ of the mean error difference over the $J = 50$ data sets for each of the $\frac{1}{2}(12 \times 11) = 66$ pairs of variants. Each boxplot summarizes 66 data points corresponding to the absolute difference for the 66 pairs of variants for a particular resampling-based error estimation method (SUB, CV, LOOCV, BOOT) for a given splitting ratios/number of folds (top: ratio 2:1, middle: ratio 4:1, bottom: ratio 9:1) and a given number of splittings into training and test data sets. Note that the results for BOOT are displayed on the top panel together with the ratio 2:1, since a bootstrap sample drawn with replacement includes an average of $63.2\,\% \approx 2/3$ of the original observations. It can be seen from Fig. 4.2 that the distribution of $|\overline{\Delta e}|$ is almost identical for all considered resampling-based error estimation procedures.

4.3.2.4 Standard Deviation of the Difference and Resampling Scheme

In the context of the testing framework for benchmark studies reviewed in Sect. 4.2, however, the variability over the $J = 50$ data sets also plays a major role. For Eq. (4.1)— which gives the required number $J(\alpha, \beta, \sigma, \delta)$ of data sets—the variance σ^2 of Δe_j over the data sets is of crucial importance since a small change in variance could lead to a large change in the number of required data sets. To visualize the variability over data sets, Fig. 4.3 displays similar boxplots to Fig. 4.2 but this time with the standard deviation of Δe_j over the $J = 50$ data sets rather than the absolute value of the mean. The right y-axis indicates the required number $J(\alpha, \beta, \sigma, \delta)$ of data sets according to Eq. (4.1) to achieve $\alpha = 0.05$ and $\beta = 0.2$ for $\delta = 0.05$.

It can be clearly seen from Fig. 4.3 that the distribution of the standard deviation over the $12 \times 11/2 = 66$ differences depends substantially on the considered resampling variant. The standard deviation is highest for LOOCV procedure, as suggested by theory (Dougherty et al. 2010). In the same vein, it is also higher for large training set sizes than for small training set sizes. While the median standard

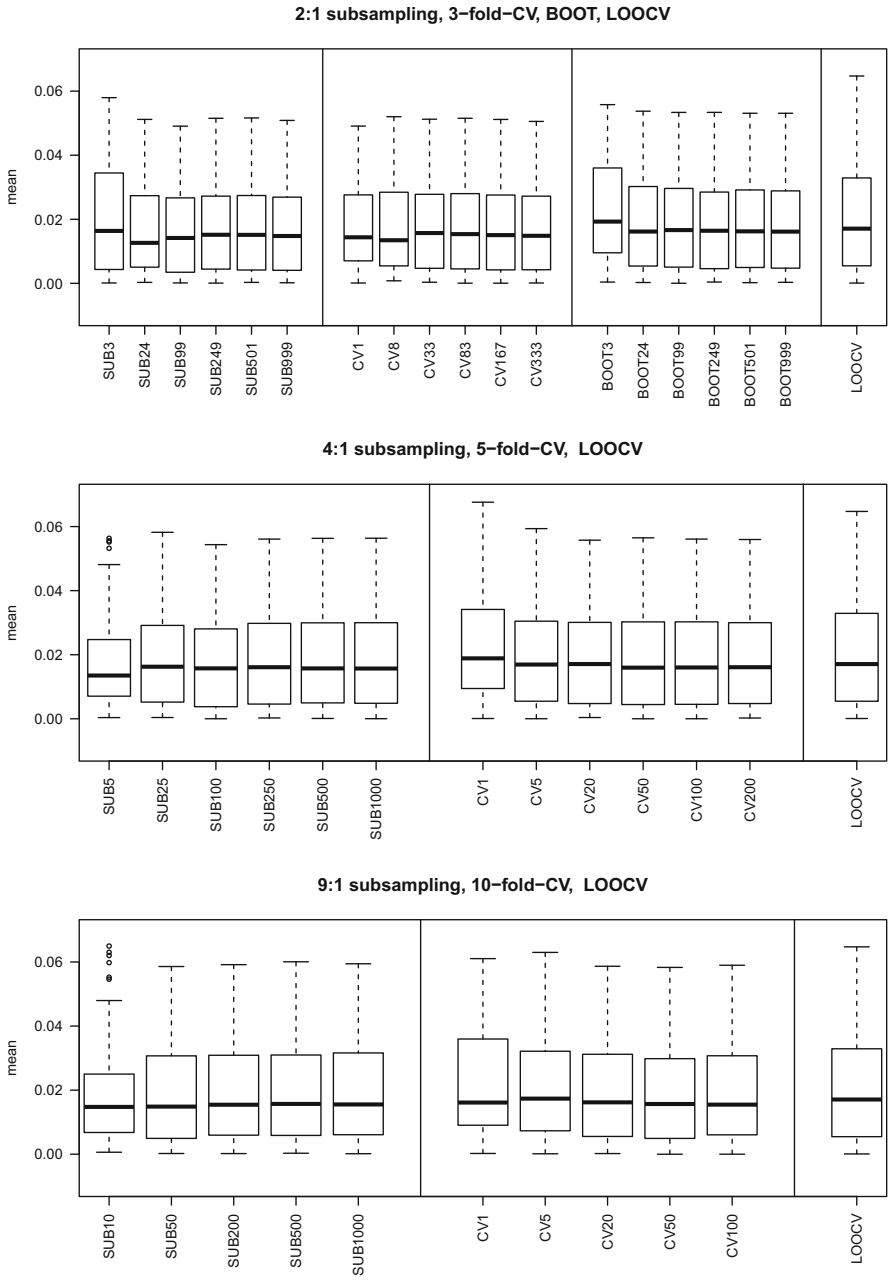

Fig. 4.2 Absolute value $|\overline{\Delta e}|$ of the mean error difference over the $J = 50$ data sets. Each boxplot depicts $11 \times 12/2 = 66$ data points corresponding the 66 pairwise differences between the $K = 12$ considered variants. SUB3,. . . ,SUB999,SUB5,. . . ,SUB1000: repeated subsampling with $B_{SUB} = 3, \ldots, 999, 5, \ldots, 1000$. CV1,. . . ,CV333: CV repeated $B_{CV} = 1, \ldots, 333$ times. BOOT3,. . . ,BOOT999: repeated bootstrapping with $B_{BOOT} = 3, \ldots, 999$

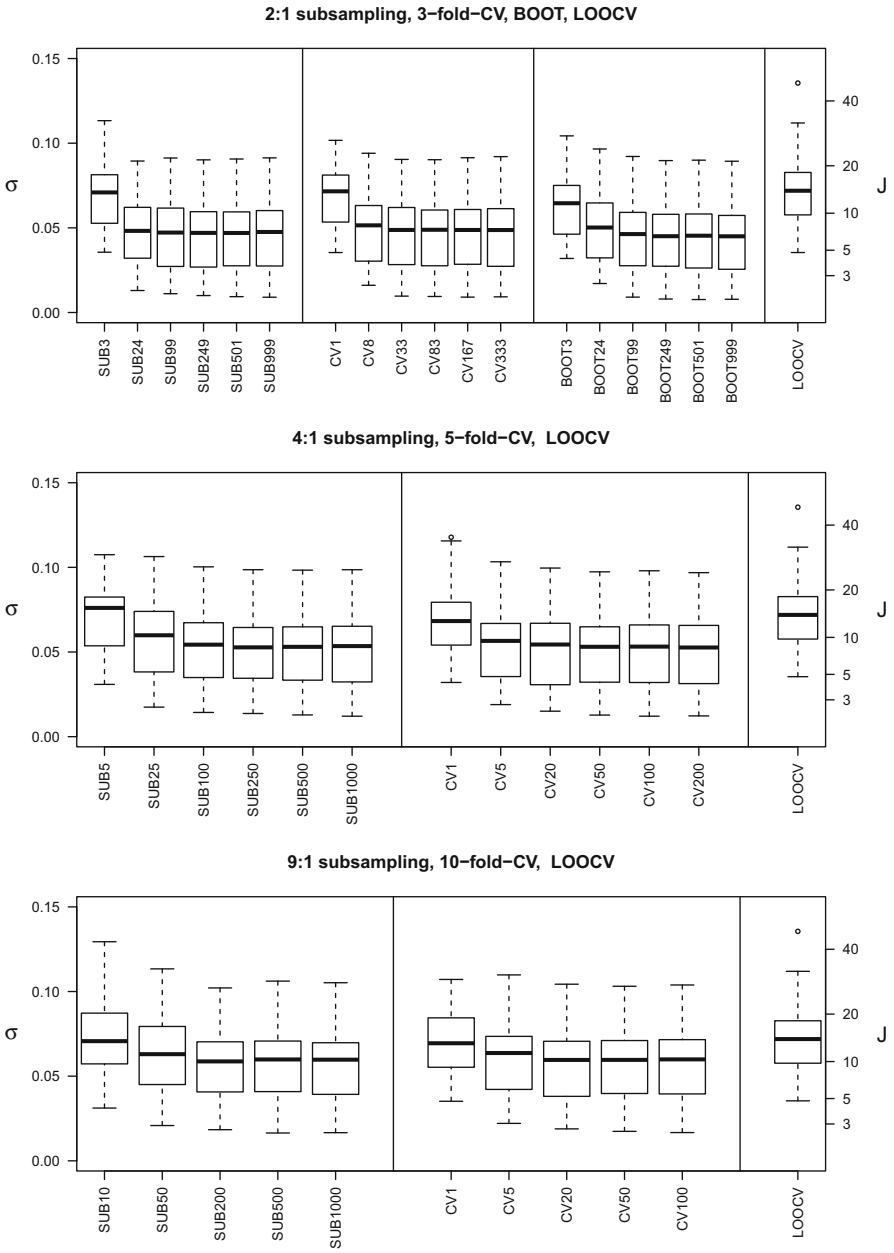

Fig. 4.3 Standard deviation of Δe_j over the $J = 50$ data sets. Each boxplot depicts $11 \times 12/2 = 66$ data points corresponding the 66 pairwise differences between the $K = 12$ considered variants. The right y-axis indicates the required number $J(\alpha, \beta, \sigma, \delta)$ of data sets according to Eq. (4.1) for $\alpha = 0.05$, $\beta = 0.2$ and $\delta = 0.05$. SUB3,. . . ,SUB999,SUB5,. . . ,SUB1000: repeated subsampling with $B_{SUB} = 3, \ldots, 999, 5, \ldots, 1000$. CV1,. . . ,CV333: CV repeated $B_{CV} = 1, \ldots, 333$ times. BOOT3,. . . ,BOOT999: repeated bootstrapping with $B_{BOOT} = 3, \ldots, 999$

deviation is below 0.05 for the 2:1 ratio/three-fold CV, it increases to ≈0.065 for the 9:1 ratio/ten-fold CV (for a sufficiently large number of splittings into training and test sets). These differences may appear minimal at first glance, but yield large differences in term of the required number of data sets. The right y-axis shows that the median number needed in three-fold CV, approximately 7, increases to almost $J = 15$ if LOOCV is used.

It can also be seen from Fig. 4.3 that very small numbers of splittings between training and test sets lead to higher standard deviations, as expected. However, the standard deviation rapidly stabilizes with increasing number of splittings: the standard deviations obtained for 100 splittings have the same distributions as those obtained for 1000 splittings. Note that increasing the number of splittings is in general always recommended, because it yields more precise error estimates for each considered data set. Our results merely suggest that further increasing the number of splittings *does not reduce the standard deviation over data sets*—at least in the investigated settings. So we do not aim to dissuade readers from performing many splittings in their analyses: we just claim that this would not reduce the variability across data sets. This is because, roughly speaking, the considered standard deviation is dominated by (i) the variability induced by the fact that each data set can be seen as a sample (of moderate size) randomly drawn from an underlying distribution and (ii) the variability of the underlying distributions, i.e., the fact that the real data sets are very different from each other. The variability induced by the randomness of the splitting procedure only accounts for a small part of the total variability, at least as soon as the number of splittings becomes "large enough" (≈100 in our analysis).

Finally, we observe that SUB and CV do not differ in terms of variance if they are based on the same total number of splitting into training and test data sets and use the same splitting ratio. Similarly, the bootstrap procedure (BOOT, top panel) shows the same pattern as SUB with ratio 2:1 and three-fold CV.

4.4 Conclusion

We conducted a benchmark study based on a large collection of $J = 50$ high-dimensional microarray data sets with binary response variable for comparing the performances of different variants of the PLS + LDA classification method. The aim of the study was to examine the impact of the resampling-based error estimation procedure (type of procedure, number of splittings, ratio between training and test set sizes) in the context of a statistical testing framework based on a paired t-test for the comparison of two methods. In particular, we focused on the variability of the estimates of the error difference over the $J = 50$ considered data sets and on the resulting power of the paired t-test. In our analyses:

- the ordering of the PLS + LDA variants according to their performance did not depend on the splitting ratio;

- the absolute mean difference did not depend on the resampling scheme;
- the standard deviation of the difference decreased for increasing number of iterations until 100 iterations and then remained stable;
- the standard deviation of the difference was larger for splitting ratios with large training sets;
- there were no substantial differences in terms of standard deviation between three-fold CV, repeated subsampling with ratio 2:1 and bootstrapping.

Based on these results (and considering the disadvantages of bootstrapping with replacement documented elsewhere Binder and Schumacher 2008), we recommend using three-fold CV or 2:1 repeating subsampling with in total at least 100 splittings into training and test sets in the context of real-data-based benchmark studies in the considered settings.

Acknowledgements We thank Rory Wilson for helpful comments.

References

Binder, H., Schumacher, M.: Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. Stat. Appl. Genet. Mol. Biol. **7**, 12 (2008)

Bock, J.: Bestimmung des Stichprobenumfangs. Oldenburg Verlag, München Wien (1998)

Boulesteix, A.-L.: PLS dimension reduction for classification with microarray data. Stat. Appl. Genet. Mol. Biol. **3**, 33 (2004)

Boulesteix, A.-L.: On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by Smith et al. Bioinformatics **29**, 2664–2666 (2013)

Boulesteix, A.-L., Lauer, S., Eugster, M.: A plea for neutral comparison studies in computational sciences. PLOS ONE **8**, 61562 (2013)

Boulesteix, A.-L., Hable, R., Lauer, S., Eugster, M.: A statistical framework for hypothesis testing in real data comparison studies. Am. Stat. **69**, 201–212 (2015)

Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)

de Souza, B.F., de Carvalho, A., Soares, C.: A comprehensive comparison of ML algorithms for gene expression data classification. In: The 2010 International Joint Conference of Neural Networks (IJCNN), Barcelona, pp. 1–8 (2010)

Dougherty, E.R., Sima, C., Hanczar, B., Braga-Neto, U.M.: Performance of error estimators for classification. Curr. Bioinform. **5**, 53–67 (2010)

Molinaro, A., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics **21**, 3301–3307 (2005)

Slawski, M., Daumer, M., Boulesteix, A.-L.: CMA: a comprehensive bioconductor package for supervised classification with high dimensional data. BMC Bioinform. **9**, 439 (2008)

Chapter 5
Path Directions Incoherence in PLS Path Modeling: A Prediction-Oriented Solution

Pasquale Dolce, Vincenzo Esposito Vinzi, and Carlo Lauro

Abstract PLS-PM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram (i.e., the path directions). The PLS-PM iterative algorithm analyzes interdependence among blocks and misses to distinguish explicitly between dependent and explanatory blocks in the structural model. This inconsistency of PLS-PM is illustrated using the simple two-blocks model. For the case of more than two blocks of variables, it is necessary to have a close look at the different criteria optimized by PLS-PM to show this issue. In general, the role of latent variables in the structural model depends on the way the outer weights are calculated. A recently proposed method, called Non-Symmetrical Component-based Path Modeling, which is based on the optimization of a redundancy-related criterion in a multi-block framework, respects the direction of the relationships specified in the structural model. In order to assess the quality of the model, we provide a new goodness-of-fit index based on redundancy criterion and prediction capability. Furthermore, we provide a procedure to address the problem of multicollinearity within blocks of variables.

Keywords PLS Path Modeling • Predictive Direction • Redundancy Index

5.1 Introduction

Multivariate techniques can be categorized as either interdependence or dependence techniques. Interdependence techniques involve the simultaneous analysis of the relationships among variables in the data set, where variables are not classified as

P. Dolce (✉) • C. Lauro
University of Naples "Federico II", Naples, Italy
e-mail: pasquale.dolce@unina.it; clauro@unina.it

V.E. Vinzi
ESSEC Business School, Cergy Pontoise Cedex, France
e-mail: vinzi@essec.edu

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_5

either dependent or explanatory. Thus, distinction between predictors and criteria is discarded and the direction of the relationships between blocks is symmetrical. In this case, an appropriate multivariate method predictive in both directions.

Dependence techniques take into account a priori information on the different roles of the variables or sets of variables (Lauro and D'Ambra 1992). A single variable or a set of variables is identified as the dependent variable to be explained or predicted by other variables known as explanatory or independent variables, and the analysis focuses on deriving those combinations of predictors which explain most of the variation in the set of dependent variables. In this case, the predictive direction of the relationship between the blocks of variables is asymmetrical.

PLS-PM is a method aimed at modeling a network of linear dependence relationships among several blocks of manifest variables (MVs), where each block is summarized by a latent variable (LV) defined as a component or a composite (i.e., an exact linear combination of the MVs). Since LVs are defined as components which aim to explain the variances of their own set of MVs, PLS-PM is commonly referred to as a component-based (or variance-based) approach (Lohmöller 1989; Tenenhaus et al. 2005; Wold 1982).

In order to respect the predictive directions of the structural relationships specified in the path diagram (i.e., the path directions), the estimation process should implicitly analyze the dependence relationships among LVs asymmetrically. However, it is known that PLS-PM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram (Vittadini et al. 2007; Dolce 2015).

The directions of the links in the structural model do not play a role in the algorithm apart from the specific case of the so-called path weighting scheme for the inner estimation (Tenenhaus et al. 2005). In the inner step of the PLS-PM algorithm, each LV is defined as a linear combination of all the connected LVs. Two LVs are connected if there exists a link between the two blocks: an arrow goes from one LV to the other in the path diagram, independently of the direction. When the path weighting scheme is applied, the path direction is taken into account only in the way the inner weights are computed, but each LV is still defined in the inner step of the algorithm as a function of all the connected LVs irrespective of the path directions.

Depending on the utilized outer schemes, PLS-PM provides components that are either optimally correlated to each other or as much correlated as possible while being somehow representative of each corresponding block of MVs. In the search for optimally correlated components, the estimation process amplifies interdependence among blocks and misses to distinguish between dependent and explanatory blocks in the structural model. As a consequence, there is often a difference between what PLS-PM wants to model and what is actually computed by the PLS-PM algorithm.

We will first illustrate this inconsistency of PLS-PM by using a simple model, the case of two blocks of variables. For the case of more than two blocks of variables, we will look at the different criteria optimized by PLS-PM (Esposito Vinzi and Russolillo 2013) in order to show this issue.

In general, the role of the LVs in the structural model depends on the way the outer weights are computed. We show that the only way for giving an explanatory role to a LV is to apply *Mode B*, while applying *Mode A* gives it a role of dependent variable, whatever the path direction is. In the case of more than two blocks, we cannot apply this rule (i.e., *Mode B* to the exogenous block and *Mode A* to the endogenous block) because some endogenous LVs appear as both explanatory and dependent LVs (we define them as "bridge" LVs).

Dolce's dissertation (Dolce 2015) describes a non-symmetrical component-based estimation approach—called Non-Symmetrical Component-based Path Modeling (NSC-PM)—based on the optimization of a redundancy-related criterion. It aims at maximizing the explained variance of the MVs in the dependent blocks, and it is more suitable for prediction purposes. As PLS-PM, NSC-PM is applied in a multiblock framework, where relationships among blocks are specified by a path diagram.

The NSC-PM respects the direction of the relationships specified in the structural model: bridge blocks are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when they play a dependent role.

In order to assess the quality of the model, we provide a global goodness of prediction index based on redundancy criterion and prediction capability. Furthermore, since in the NSC-PM algorithm multiple regressions are applied when the outer weights are computed for the explanatory LVs, we provide a procedure to address the issue of multicollinearity within the blocks of variables.

5.2 PLS-PM Incoherence with Path Directions

PLS-PM does not rigidly adhere to an underlying theoretical model (Chin 1998), and there is often a difference between what PLS-PM wants to model (the hypothesized model depicted in the path diagram) and what is actually computed by the PLS-PM algorithm.

Generally, the directions of the links in the structural model do not play a role in the algorithm. As a consequence PLS-PM misses to distinguish between dependent and explanatory LVs.

As for the measurement model, the choice between using *Mode A* instead of *Mode B*, for the computation of the outer weights, depends mainly on the theoretical difference between the two schemes, based essentially on the hypothesized relationships between LVs and their own MVs. Under conditions of low theoretical knowledge on the nature of the LVs, a rule of thumb in PLS-PM is to apply *Mode B* to the exogenous block and *Mode A* to the endogenous block (Wold 1980). However, to the best of our knowledge, there are hardly any studies in the literature that give reasons for following this rule and that analyze this issue into details.

In general, beyond the theoretical differences between the two different measurement model schemes, depending on the way the outer weights are calculated,

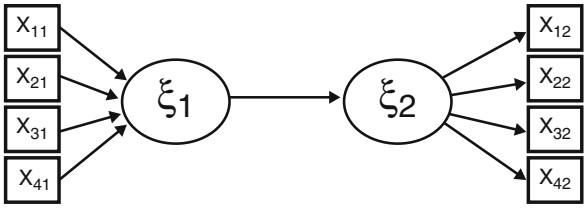

Fig. 5.1 Two-block model with inwards and outwards directed schemes: redundancy analysis

the role of the LV in the structural model changes. The only way for giving an explanatory role to an LV is to apply *Mode B*, while applying *Mode A* gives it a role of dependent variable, whatever the path direction. Thus, the predictive direction in the structural model is given by the utilized outer mode.

For two blocks of variables, the only case where PLS-PM adheres to the theoretical two-block model depicted in the path diagram is for the model in Fig. 5.1, that is, when the exogenous block is specified as formative (and the outer weights are computed by *Mode B*), and the endogenous block is specified as reflective (and the outer weights are computed by *Mode A*), which is equivalent to performing a Redundancy analysis (RA) of the endogenous block with respect to the exogenous one (Wollenberg 1977; Chin 1998; Tenenhaus et al. 2005).

When the weights are computed by using either *Mode A* for the two blocks of variables or *Mode B* for the two blocks of variables, predictive direction in the structural model (i.e., the direction of the relationship between the two LVs) is not explicitly considered in the algorithm. The procedure misses to distinguish between dependent and explanatory blocks in the model. Blocks are treated in the same way, that is the direction of the relationship between the two blocks of variables is symmetrical, that is, PLS-PM analyzes the interdependence relationship between the two blocks, instead of dependent relationship.

Recent works by Hanafi (2007), Krämer (2007) and Tenenhaus and Tenenhaus (2011), have shown that the PLS-PM iterative algorithm optimizes different statistical criteria according to the different options chosen for the computation of the outer and inner proxies of the components, also for the case of more than two blocks of variables.

Considering a network of dependence relationships between K blocks of MVs where each block, $\mathbf{X}_k$ $(k = 1, \ldots, K)$, is summarized by an LV, denoted by $\boldsymbol{\xi}_k$ $(k = 1, \ldots, K)$. A generic MV is denoted by $\mathbf{x}_{pk}$ $(p = 1, \ldots, P_k)$, $(k = 1, \ldots, K)$, where P_k is the number of MVs in the k-th block.

When all the outer weights are calculated by means of *Mode B*, Hanafi (2007) proved that the Wold's PLS-PM algorithm monotonically converges to the following criterion

$$\underset{||\mathbf{X}_k\mathbf{w}_k||^2=||\mathbf{X}_{k'}\mathbf{w}_{k'}||^2=1}{\arg\max} \quad \sum_{k\neq k'} c_{kk'}\, g\Big(\mathrm{cor}(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_{k'}\mathbf{w}_{k'})\Big) \tag{5.1}$$

where g is one of these two functions

$$g(x) = \begin{cases} x^2 & \text{if factorial} \\ |x| & \text{if centroid,} \end{cases} \tag{5.2}$$

while $c_{kk'}$ is the generic element of the Boolean square matrix $\mathbf{C}$ of order K, where $c_{kk'} = 1$ if $\boldsymbol{\xi}_k$ is connected to $\boldsymbol{\xi}'_k$ and $c_{kk'} = 0$ otherwise ($c_{kk} = 0$).

In 2007, Krämer (2007) showed that the PLS-PM algorithm was not based on a stationary equation related to the optimization of a twice differentiable function when *Mode A* was used for all the blocks in the model. In the same work, Kramer proposed a slightly modified version of the classical *Mode A* outer scheme in which a normalization constraint is put on outer weights rather than latent variable scores. If this new scheme—also called *New Mode A* by Tenenhaus and Tenenhaus (2011)—is used for all the blocks in the model, then the PLS-PM iterative algorithm monotonically converges to the criterion:

$$\underset{||\mathbf{w}_k||^2=||\mathbf{w}_{k'}||^2=1}{\arg\max} \quad \sum_{k \neq k'} c_{kk'} g\Big(\text{cov}(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_{k'}\mathbf{w}_{k'})\Big) \tag{5.3}$$

where g is defined as in Eq. 5.2.

Looking at the different optimized criteria, it is clear that the PLS-PM algorithm does not focus on directional analysis in terms of dependence relationships between blocks of variables. Depending on the chosen estimation modes (for the measurement model) and schemes (for the inner model), PLS-PM provides composite scores that are as much correlated as possible to each other while being somehow representative of each corresponding block of manifest variables. The PLS-PM estimation process analyzes symmetrical relationships between blocks, thus, it misses to distinguish between the role of dependent and explanatory blocks in the inner model.

Let us define $\tau_k = 1$ when Block k is estimated by *new Mode A* and $\tau_k = 0$ when Block k is estimated by *Mode B*. When both *new Mode A* and *Mode B* are used in the same model, Wold's procedure converges to the following criterion (Esposito Vinzi and Russolillo 2013; Tenenhaus and Tenenhaus 2011):

$$\begin{aligned} &\underset{\mathbf{w}_k}{\arg\max} \sum_{k \neq k'} c_{kk'} g\left(\text{cor}(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_{k'}\mathbf{w}_{k'}) \times \sqrt{\text{var}(\mathbf{X}_k\mathbf{w}_k)^{\tau_k}}\sqrt{\text{var}(\mathbf{X}_{k'}\mathbf{w}_{k'})^{\tau_{k'}}}\right) \\ &\text{subject to} \qquad \tau_k||\mathbf{w}_k||^2 + (1-\tau_k)||\mathbf{X}_k\mathbf{w}_k||^2 = 1, \quad k = 1, \ldots, K. \end{aligned} \tag{5.4}$$

Considering that, in the case of two blocks of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$, the redundancy analysis of $\mathbf{X}_2$ with respect to $\mathbf{X}_1$ maximizes the following criterion:

$$\begin{aligned} &\underset{\mathbf{w}_1,\mathbf{w}_2}{\arg\max}\, \text{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) \times \text{var}(\mathbf{X}_2\mathbf{w}_2)^{\frac{1}{2}} \\ &\text{subject to} \qquad ||\mathbf{X}_1\mathbf{w}_1||^2 = ||\mathbf{w}_2||^2 = 1 \end{aligned}. \tag{5.5}$$

Looking at Eqs. 5.5 and 5.4 it is clear that the role of the blocks in the structural model depends on the way the outer weights are calculated. The only way for giving an explanatory role to an LV is to apply *Mode B*, while applying *Mode A* to the dependent variable, whatever the path direction.

However, in PLS-PM, we cannot apply this rule (i.e., *Mode B* to the exogenous block and *Mode A* to the endogenous block), because some endogenous LVs appear only as dependent LVs, but others appear as both explanatory and dependent LVs.

5.3 Non-symmetrical Component-Based Path Modeling (NSC-PM)

The NSC-PM is a non-symmetrical component-based estimation approach for modeling a network of dependence relationships between K blocks of variables where each block, $\mathbf{X}_k$ $(k = 1, \ldots, K)$, is summarized by an LV, denoted by $\boldsymbol{\xi}_k$ $(k = 1, \ldots, K)$. Hence, NSC-PM is applied in a multiblock framework, where relationships among blocks are specified in a path diagram. A generic MV is denoted by $\mathbf{x}_{pk}$ $(p = 1, \ldots, P_k)$, $(k = 1, \ldots, K)$, where P_k is the number of MVs in the k-th block.

Similarly to the PLS-PM, the NSC-PM consists of two sub-models: the structural (or inner) model and the measurement (or outer) model. This method is based on the optimization of a redundancy-related criterion, and it is more suitable for prediction purposes. It aims at maximizing the explained variance of the MVs in one block given the others.

In this new approach, the distinction between reflective and formative measurement models is disregarded. The nature of LVs and the direction of relationships between LVs and MVs is not taken into account. On the contrary, great emphasis is placed on the dependence relationships between LVs in the structural model. We only make a distinction between explanatory blocks and dependent blocks in the structural model.

The NSC-PM respects the direction of the relationship specified in the structural model because the directions of the links in the structural model play a role in the algorithm. In particular, taking into account the two roles that bridge LVs play into the model (i.e., they appear as both explanatory and dependent LVs in the structural model), in NSC-PM algorithm Bridge LVs are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when they play a dependent role. When a block of variables plays an explanatory role in a specific step of the algorithm we apply *Mode B* for computing the outer weights, while we apply *Mode A* when a block of variables plays a dependent role (further details are provided in Dolce 2015).

5.3.1 Model Assessment

Since NSC-PM is based on the maximization of the explained variance of the MVs of the endogenous blocks, it is extremely important that the assessment of the quality of the model takes also into account appropriate measures of predictive ability. Generally, some measures commonly used in PLS-PM can be used as well. As in PLS-PM, the goodness of the structural model depends on the portion of variability of each endogenous LV explained by the corresponding exogenous predictors, that can be measured by the multiple linear determination coefficient (R^2).

As for the measurement model, the proportion of the variance of a generic MV $\mathbf{x}_{pk}$ reproduced by $\hat{\boldsymbol{\xi}}_k$ is equal to $\text{cor}^2(\mathbf{x}_{pk}, \hat{\boldsymbol{\xi}}_k)$ that, in the case of standardized MVs, corresponds to $\hat{\lambda}^2_{pk}$ (i.e., the so-called "communality").

If all the MVs are standardized, for each Block k, the average of the communalities is equal to the average variance extracted (AVE) that expresses the part of variance of the block explained by $\hat{\boldsymbol{\xi}}_k$:

$$\text{Com}_k = \frac{1}{P_k}\sum_{p=1}^{P_k} \text{cor}^2(\mathbf{x}_{pk}, \hat{\boldsymbol{\xi}}_k) = \frac{1}{P_k}\sum_{p=1}^{P_k} \hat{\lambda}^2_{pk} = \frac{\sum_{p=1}^{P_k} \hat{\lambda}^2_{pk}}{\sum_{p=1}^{P_k} \text{var}(\mathbf{x}_{pk})} = \text{AVE}_k \qquad (5.6)$$

In PLS-PM the weighted average of all the K blocks specific communality indexes, with weights equal to the number of MVs in each block, is used as a goodness of fit of the whole measurement model.

In NSC-PM communality index is conceptually appropriate just for dependent blocks. For LVs that appear at least in one equation of the structural model as predictors (i.e., exogenous and bridge LVs), the MVs do not necessarily measure the same underlying construct, (i.e., they are not supposed to be highly correlated). The components of the blocks that appear only as predictors (i.e., the exogenous blocks) are expected to maximize the explained MVs variance of the related dependent blocks. The components of the bridge blocks are expected to maximize the explained MVs variance of the related dependent blocks while being correlated with their own predictors LVs.

Moreover, since in the NSC-PM algorithm multiple regressions are applied when the outer weights are computed for explanatory LVs, excessive correlations among MVs are not desired. However, in order to avoid the multicollinearity problem, we propose a solution (see next Section).

The interpretation of exogenous and bridge LVs should be based on the weights. The weights provide information about the direct relation between the MVs and their own LV, which reflects the impact of the MVs on the LV (Bollen 1989), and a comparison among them gives information about which MV contributes most effectively to the LV. Loadings can also be used for interpretation, bearing in mind that while the outer weight is a measure of relative contribution of a MV to its LV, the loading can only be used to evaluate the absolute importance of a MV for its LV.

On the contrary, MVs of dependent blocks are expected to be unidimensional and to measure the same construct (i.e., the MVs in each block are supposed to be highly correlated among each others). In this case, multicollinearity is not an issue because only simple regressions are involved. The components of dependent blocks are expected to be as much correlated as possible to their predictor LVs, while being representative of their corresponding blocks of MVs. The interpretation of dependent LVs should be based on the loadings.

As a measure of the quality of the global model, the goodness-of-fit (GoF) index proposed by Amato et al. (2005) is not conceptually appropriate for measuring the global quality of NSC-PM. As a matter of fact, the Gof index—as described by Amato et al. (2005)—is computed as the geometric mean of the average communality and the average R^2 of the different inner models:

$$\text{GoF} = \sqrt{\overline{\text{Com}} \times \overline{R^2}} \tag{5.7}$$

Because the GoF index is partly based on average communality, it is conceptually appropriate only for dependent blocks. For this reason, we cannot use the Gof index in NSC-PM as a measure of the quality of the global model.

A way of assessing the global model in NSC-PM could be to measure the amount of variance in the sets of variables of the dependent blocks explained by their own latent predictors. In this direction, we can use the redundancy index which measures the portion of variability of dependent block of MVs explained by its own predictors.

Given two blocks of variables, $\mathbf{X}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{P_1 1})$ and $\mathbf{X}_2 = (\mathbf{x}_{12}, \ldots, \mathbf{x}_{P_2 2})$, the redundancy index as proposed by Stewart and Love (1968) measures the proportion of the variance in the dependent set $\mathbf{X}_2$ that is accounted for by the predictor set $\mathbf{X}_1$. The redundancy analysis model, proposed by Wollenberg (1977), searches for the linear combination, $\hat{\boldsymbol{\xi}}_1 = \mathbf{X}_1 \mathbf{w}_1$ (the so-called first redundancy variate), that maximizes the redundancy index, $R_{\mathbf{X}_2}$, defined as

$$R_{\mathbf{X}_2} = \sum_{p=1}^{P_2} \text{cor}(\hat{\boldsymbol{\xi}}_1, \mathbf{x}_{p2})^2 / P_2 \tag{5.8}$$

under the restriction that the variance of $\hat{\boldsymbol{\xi}}_1 = 1$.

In the context of canonical correlation analysis (Hotelling 1935, 1936), the redundancy index (Eq. 5.8) can be written as:

$$R_{\mathbf{X}_2} = \rho^2 \sum_{p=1}^{P_2} \text{cor}(\hat{\boldsymbol{\xi}}_2, \mathbf{x}_{p2})^2 / P_2 \tag{5.9}$$

where ρ is the canonical correlation coefficient and $\boldsymbol{\xi}_2 = \mathbf{X}_2 \tilde{\mathbf{w}}_2$ is the first canonical component of $\mathbf{X}_2$ (Rencher 1998).

For each endogenous block, in PLS-PM the redundancy index is computed as:

$$\text{Red}_k = \text{Com}_k \times R_k^2. \tag{5.10}$$

where Com_k is the average of the communalities in the kth block and R_k^2 is the multiple linear determination coefficient in the regression model of $\hat{\boldsymbol{\xi}}_q$ on its own predictor LVs. Looking at the redundancy index from the two different perspectives, it is clear that in PLS-PM the redundancy index is computed as in the context of CCA.

Because NSC-PM aims at maximizing the explained variance of the MVs in one block given the other (i.e., a redundancy-related criterion in a multi-block framework), as a redundancy measure in NSC-PM we propose to compute for each MV of endogenous blocks, the portion of its variability explained by its own predictors as:

$$\text{Red}_{\mathbf{x}_{pk}} = R^2 \left(\mathbf{x}_{pk}, \{ \hat{\boldsymbol{\xi}}'_{k'} s \; explaining \; \hat{\boldsymbol{\xi}}_k \} \right) \tag{5.11}$$

that is, as in the context of RA.

For Block k, the redundancy index is defined as

$$\text{Red}_k = \sum_{p=1}^{P_k} \text{Red}_{\mathbf{x}_{pk}} \tag{5.12}$$

In Lohmöller's dissertation (Lohmöller 1989) some advice about the evaluation of model quality is given. The author states that the fit of the global model (outer and inner model) can be judged as satisfactory if the average of the redundancy indexes is high enough. Thus, he considers the redundancy index as an index of Goodness of fit of the global model.

In this perspective, we consider the average of all the $\text{Red}_{\mathbf{x}_{pk}}$ as an index of global goodness of prediction, because it is based on redundancy criterion and prediction capability. Let us define the first J blocks $\mathbf{X}_k$ $(k = 1, \ldots, J)$ as exogenous blocks, and $\mathbf{X}_k$ $(k = J + 1, \ldots, K)$ as endogenous blocks; if we denote by $\tilde{P}$ the number of MVs of the endogenous blocks, the global goodness of prediction is defined as

$$\overline{\text{Red}} = \frac{1}{\tilde{P}} \sum_{k=J+1}^{K} P_k \times \text{Red}_k \tag{5.13}$$

Just as with canonical correlations, no generally accepted guidelines have been established for the minimum acceptable redundancy index needed to judge a fit of the model as satisfactory. The researcher must judge the specific research problem being investigated to determine whether the redundancy index is sufficient to justify interpretation.

Model validation regards also the way relations are modeled, in both the structural and the measurement model. In this respect, since NSC-PM does not require any distributional hypothesis on MVs, confidence intervals for model parameters can be obtained by resampling techniques, such as Jackknife and Bootstrap.

As said above, NSC-PM is a method for predictive purposes, and could be an important technique deserving a prominent place in research applications when the aims of the analysis is prediction. For these reasons, NSC-PM evaluation cannot focus only on parameter recovery and on the quality of the measurement model and the structural model—in terms of explained variance—indiscriminately. In order to evaluate the model in terms of predictive ability the so-called Blindfolding procedure, using the Stone-Geisser's approach to cross-validation, can be used (Stone 1974; Geisser 1975; Chin 1998).

5.3.2 A Solution to the Issue of Multicollinearity

In the NSC-PM algorithm, multiple regressions are applied when the outer weights are computed for explanatory LVs. As a consequence, the stability of the MV outer weights are affected by the strength of the MV intercorrelations. For this reason, multicollinearity is an important issue to take into account also in NSC-PM.

For LVs that appear only as dependent variables in the structural model, multicollinearity is not an issue because only simple regressions are involved, and theoretically it is desired. By contrast, excessive multicollinearity among MVs of explanatory LVs makes it difficult to separate the distinct influence of the individual MV on the LV or else the outer weights may be non-interpretable, having incoherent signs with the correlation with the corresponding LV.

A possible way to check for multicollinearity in a block of variables is to compute the "tolerance" of each MV as $1 - R^2$, where the R^2 is the coefficient of determination for the regression of the specific MV on the other MVs of the block. A measure related to the tolerance is the Variance Inflation Factor (VIF), computed as the inverse of the tolerance ($VIF = 1/TOL$) (Hair et al. 2010). A large VIF value indicates a high standard error of the specific weight due to multicollinearity among the MVs. As a rule of thumb, the VIF should not exceed a value of 10, but, particularly when samples size is small, the critical value may be smaller then 10 (Hair et al. 2010). In general, the critical value should be defined considering the specific analysis objectives.

As a preliminary analysis to NSC-PM, multicollinearity is checked in the blocks that appear as explanatory at least in one equation of the structural model. If excessive multicollinearity occurs in a block, we extract fewer components obtained by principal component analysis (PCA) on the specific block of variables, and then we use them instead of the original variables in the outer estimation step when the blocks play an explanatory role. In particular, a multiple regression can be performed to predict the instrumental inner composite from the extracted principal components and then the outer composite is computed as a weighted aggregate of the principal components.

A drawback of this procedure is that PCA creates components that explain the observed variability in the MVs, but do not consider the relationships of these variables with the MVs of the dependent blocks. An alternative approach could be

similar to the one proposed by Esposito Vinzi et al. (2009) in the PLS-PM algorithm, namely, providing PLS regression for estimating the outer weights as an alternative to OLS regression. As it is well known, PLS regression does take into account the relationships of the explanatory MVs with the response MVs.

5.4 Conclusions and Future Research

Generally speaking, there is a difference between what PLS-PM wants to model and what the iterative algorithm implicitly processes. As a matter of fact, PLS-PM analyzes relationships between LVs symmetrically, without taking into account the roles of dependent and explanatory LVs in the structural model. When theoretical knowledge about the nature of LVs is scarce, *Mode B* is suggested for exogenous blocks and *Mode A* for endogenous blocks (Wold 1980). However, in this paper we show that the only way for giving an explanatory role to an LV is to apply *Mode B* for its block, while applying *Mode A* gives it a role of a dependent variable, whatever the path direction. In the case of more than two blocks of variables, where some endogenous LVs appear as both explanatory and dependent LVs, this rule cannot be applied.

NSC-PM is a non-symmetrical approach that respects the direction of the relationship specified in the structural model, since the directions of the links play a role in the algorithm. In particular, LVs that appear as both explanatory and dependent LVs in the structural model are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when they play a dependent role. NSC-PM aims at maximizing the explained variance of the MVs of the endogenous blocks (i.e., an approach based on the optimization of a redundancy-related criterion in a multi-block framework), and seems to be a good compromise between favoring stability (high explained variance) in the blocks and correlation between components.

Further research is needed to study the properties and the performance of the method, and to find out if the NSC-PM algorithm optimizes a global criterion.

References

Amato, S., Esposito, V.V., Tenenhaus, M.: A global goodness-of-fit index for PLS structural equation modeling. Technical report, HEC School of Management, France (2005)

Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)

Chin, W.W.: The partial least squares approach for structural equation modeling. In: Marcoulides, G. (ed.) Modern Methods for Business Research, pp. 295–336. Lawrence Erlbaum Associates, London (1998)

Dolce, P.: Component-based path modeling: open issues and methodological contributions, PhD thesis, Università degli Studi di Napoli "Federico II", Italy (2015)

Esposito Vinzi, V., Russolillo, G.: Partial least squares algorithms and methods. WIREs Comput. Stat. **5**, 1–19 (2013)

Esposito Vinzi, V., Russolillo, G., Trinchera, L.: A joint use of PLS regression and PLS path modelling for a data analysis approach to latent variable modelling. In: 57th Session of the International Statistical Institute (ISI), International Statistical Institute (ISI), Durban (Invited Paper Session) (2009)

Geisser, S.: The predictive sample reuse method with applications. J. Am. Stat. Assoc. **70**, 320–328 (1975)

Hair, J., Black, W., Babin, B., Anderson, R.: Multivariate Data Analysis. A Global Perspective. Pearson Education, Inc., Upper Saddle River (2010)

Hanafi, M.: PLS path modeling: computation of latent variables with the estimation mode B. Comput. Stat. **22**, 275–292 (2007)

Hotelling, H.: The most predictable criterion. J. Educ. Psychol. **26**, 139–142 (1935)

Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)

Krämer, N.: Analysis of high-dimensional data with partial least squares and boosting. Phd thesis, Technische Universität Berlin, Berlin (2007)

Lauro, N., D'Ambra, L.: Non symmetrical exploratory data analysis. Statistica Applicata **4**(4), 511–529 (1992)

Lohmöller, J.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heildelberg (1989)

Rencher, A.: Multivariate Statistical Inference and Applications. Wiley Series in Probability and Statistics. Wiley, New York (1998)

Stewart, D., Love, W.: A general canonical correlation index. Psychol. Bull. **70**(3), 160–163 (1968)

Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. **36** 111–147 (1974)

Tenenhaus, A., Tenenhaus, M.: Regularized generalized canonical correlation analysis. Psychometrika **76**, 257–284 (2011)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)

Vittadini, G., Minotti, S.C., Fattore, M., Lovaglio, P.G.: On the relationships among latent variables and residuals in PLS path modeling: the formative-reflective scheme. Comput. Stat. Data Anal. **51**, 5828–5846 (2007)

Wold, H.: Model construction and evaluation when theoretical knowledge is scarce. In: Kmenta, J., Ramsey, J.B. (eds.) Evaluation of Econometric Models, pp. 47–74. NBER Books, Academic Press (1980)

Wold, H.: Soft modeling: the basic design and some extensions. In: Jöreskog, K., Wold, H. (eds.) Systems Under Indirect Observation, vol. 2, pp. 1–54. North-Holland, Amsterdam (1982)

Wollenberg, A.L.: Redundancy analysis an alternative for canonical correlation analysis. Psychometrika **42**, 207–219 (1977)

Part II
New Developments in Genomics and Brain Imaging

Chapter 6
Imaging Genetics with Partial Least Squares for Mixed-Data Types (MiMoPLS)

Derek Beaton, Michael Kriegsman, ADNI*, Joseph Dunlop, Francesca M. Filbey, and Hervé Abdi

Abstract "Imaging genetics" studies the genetic contributions to brain structure and function by finding correspondence between genetic data—such as single nucleotide polymorphisms (SNPs)—and neuroimaging data—such as diffusion tensor imaging (DTI). However, genetic and neuroimaging data are heterogenous data types, where neuroimaging data are quantitative and genetic data are (usually) categorical. So far, methods used in imaging genetics treat all data as quantitative, and this sometimes requires unrealistic assumptions about the nature of genetic data. In this article we present a new formulation of Partial Least Squares Correlation (PLSC)—called Mixed-modality Partial Least Squares (MiMoPLS)—specifically tailored for heterogeneous (mixed-) data types. MiMoPLS integrates features of PLSC and Correspondence Analysis (CA) by using special properties of quantitative

D. Beaton (✉) • M. Kriegsman • H. Abdi
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA
e-mail: derekbeaton@utdallas.edu; michael.kriegsman@utdallas.edu; herve@utdallas.edu

for the Alzheimer's Disease Neuroimaging Initiative

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

J. Dunlop
SAS Institute Inc., Cary, NC, USA
e-mail: joseph.dunlop@gmail.com

F.M. Filbey
Center for BrainHealth and School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA
e-mail: francesca.filbey@utdallas.edu

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_6

data and Multiple Correspondence Analysis (MCA). We illustrate MiMoPLS with an example data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) with DTI and SNPs.

Keywords Imaging genetics • Partial least squares • Alzheimer disease • (Multiple) Correspondence analysis • Burt's stripe • SNPs • Heterogenuous data

6.1 Introduction

Imaging genetics (and "imaging genomics") combines two scientific disciplines: neuroimaging—often from the cognitive neuroscience perspective—and genetics—often from the genomics perspective (Meyer-Lindenberg 2012; Thompson et al. 2010). Imaging genetics integrates neuroimaging and genetic data to understand how genetics contributes to brain structure and function—often with respect to diagnostic criteria or complex behavior and traits (such as personality). Usually, the data sets in imaging genetics are very large: neuroimaging data (measured in number of voxels) can comprise up to one million variables, whereas genetic data (often genome-wide with single nucleotide polymorphisms [SNPs]) can comprise more than three million variables. With such large data sets it is often impractical to use mass-univariate statistics, simply because the corrections for multiple comparisons become too drastic.

So, instead of using mass-univariate approaches, imaging genetics researchers often turn to multivariate methods (Liu and Calhoun 2014) such as sparse reduced rank regression (Vounou et al. 2010), distance matrix regression (Zapala and Schork 2006), independent components analysis (Meda et al. 2010; Liu et al. 2009), Canonical Correlation Analysis (CCA) (Sheng et al. 2014), or Partial Least Squares (PLS) (Le Floch et al. 2012). Because the goal of imaging genetics is to understand the relationships between imaging and genetics, researchers often turn to multivariate techniques designed to conjointly analyze two tables of data (e.g., imaging and genetics). However, nearly all implementations of CCA, PLS, and most other multivariate techniques are designed for quantitative data and this can be problematic because many types of genetic data—especially SNPs—are categorical data.

6.1.1 Ambiguity with Allelic Coding

With the advent of genome-wide technology, many biological, medical, and psychological disciplines conduct genome-wide association (GWA) studies. Typically, genome-wide data consist in single nuclear polymorphisms (SNPs) (Weiner and Hudson 2002). A SNP is expressed by the two nucleotide letters that exist at a particular genomic location. These two letters can be, for example, AA, AT,

or TT. For a given SNP, each letter can be a major allele—say A—or a minor allele—say T. For analyses, SNPs are often recoded into an allelic count; typically, SNPs emphasize the minor allele. Thus our example—AA, AT, and TT—would be recoded respectively as the numbers 0, 1, or 2 (because AA has 0 minor allele, and TT has 2 minor alleles). This {0,1,2} coding scheme is often called an "additive" model. In biological, medical, and psychological studies with SNPs, the minor allele is usually assumed to be associated with risk for diseases and disorders (Cantor et al. 2010; Visscher et al. 2012).

This allelic count makes several unrealistic assumptions. First, the {0,1,2} scheme is an implicit contrast—which, in GWA studies, emphasizes the minor allele for hundreds of thousands or even millions of SNPs. Second, this contrast is linear even though many risk factors are non-linear (e.g., risk of Alzheimer's Disease from ApoE) (Genin et al. 2012). Finally, because the minor allele frequency is usually computed *per study sample*, there is a possibility that a separate sample would detect a different minor allele, and so the "2" in one study would be a "0" in another study (and this could create problems with replication); thus the only unambiguous genotype—across different samples and populations—is the heterozygote marked as "1" (e.g., AT in our example).

To avoid these measurement assumptions, SNPs can be expressed in a purely categorical format that preserves exactly the alleles found without presuming a linear contrast effect. However, there exists only a few statistical methods (e.g., Multiple Factor Analysis, Bécue-Bertaut and Pagès 2008) designed to simultaneously analyze heterogeneous data such as SNPs (categorical) and neuroimaging (continuous). In this paper, we provide a new formulation of PLS designed for heterogeneous data types that allows both SNPs and imaging data to remain in their natural formats (categorical, and continuous, respectively). This approach—called"mixed-modality" PLS (MiMoPLS)—generalizes PLS for use with data sets that comprise both quantitative and categorical variables.

6.2 Notation and Prerequisites

This section presents the notations and a sketch of the main prerequisite methods: the singular value decomposition and its generalization, principal components analysis, (multiple) correspondence analysis, partial least squares correlation, and partial least squares correspondence analysis.

6.2.1 Notation

Uppercase bold letters denote matrices (e.g., $\mathbf{X}$) and lower case bold letters denote vectors (e.g., $\mathbf{x}$). The transpose operation is denoted $^{\mathrm{T}}$, the inverse operation $^{-1}$, and the diagonal operation—which turns a vector into a diagonal matrix, or extracts the

diagonal as a vector from a diagonal matrix—is denoted $\mathsf{diag}\,\{\}$. The identity matrix is denoted $\mathbf{I}$, an identity matrix of a specific size is denoted $\mathbf{I}_a$ where a indicates the size (i.e., the number of rows and columns) of $\mathbf{I}$; $\mathbf{1}_a$ is a vector of ones of length a. Matrices denoted as $\mathbf{Z}_*$ are centered and normalized (i.e., each column of $\mathbf{Z}_*$ has mean 0 and norm 1). Italic or bold subscripts of a matrix denote its relationship with an index or another matrix (e.g., matrix $\mathbf{Z}_\mathbf{Y}$ is centered and normalized $\mathbf{Y}$, matrix $\mathbf{W}_K$ denotes the "weights" matrix derived from the K set).

6.2.2 The Singular Value Decomposition

The singular value decomposition (SVD) of a $J \times K$ matrix $\mathbf{R}$ of rank L (with $L \leq \min(J, K)$) is expressed as

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}, \text{ where } \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\mathbf{V}, \tag{6.1}$$

where $\mathbf{U}$ is the $J \times L$ matrix of the left singular vectors, $\mathbf{V}$ the $K \times L$ matrix of the right singular vectors, and $\boldsymbol{\Delta}$ is an $L \times L$ diagonal matrix whose diagonal contains the singular values (ordered from the largest to the smallest). When squared, the singular values become eigenvalues and so $\boldsymbol{\Lambda} = \mathsf{diag}\,\{\boldsymbol{\Delta}\}^2$ is a diagonal matrix of eigenvalues. The first singular value and pair of first singular vectors are the solution to the following optimization problem:

$$\delta = \underset{\mathbf{u},\mathbf{v}}{\arg\max} \left(\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{v}\right) \quad \text{under the constraints } \mathbf{u}^{\mathrm{T}}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1. \tag{6.2}$$

The other pairs of singular vectors are solutions of the same optimization problem with the additional constraint that right (respectively, left) singular vectors are orthogonal to all other right (respectively, left) singular vectors associated with a larger singular value (see Greenacre (1984), Lebart et al. (1984), and Abdi (2007), for details).

6.2.3 The Generalized Singular Value Decomposition

The generalized singular value decomposition (GSVD) generalizes the SVD by imposing, on the left and right singular vectors, orthogonality constraints (also called "metrics") expressed by positive-definite matrices denoted $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}$ (Greenacre 1984; Lebart et al. 1984; Abdi 2007). The GSVD of a $J \times K$ matrix $\mathbf{R}$ of rank L (with $L \leq \min(J, K)$) is expressed as

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}, \text{ with } \mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{V}. \tag{6.3}$$

The first generalized singular value and pair of generalized singular vectors are solution of the following optimization problem (cf. 6.2):

$$\delta = \underset{\mathbf{u},\mathbf{v}}{\arg\max} \left(\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{v}\right) \quad \text{under the constraints } \mathbf{u}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{v} = 1. \tag{6.4}$$

The other pairs of singular vectors are solutions of the same optimization problem with the additional constraint that right (respectively left) singular vectors are $\boldsymbol{\Omega}$-orthogonal (respectively $\boldsymbol{\Phi}$–orthogonal) to all other right (respectively left) singular vectors associated with a larger singular value. Component (a.k.a. factor) scores are obtained as:

$$\mathbf{F}_J = \boldsymbol{\Omega}\mathbf{U}\boldsymbol{\Delta} \text{ and } \mathbf{F}_K = \boldsymbol{\Phi}\mathbf{V}\boldsymbol{\Delta}. \tag{6.5}$$

Often, the GSVD is expressed via the compact "triplet notation" (Escoufier 2006; Dray 2014; De la Cruz and Holmes 2010) and, for example, with this notation, the GSVD of Eq. 6.3 is presented as the analysis of the triplet $(\mathbf{R}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$.

6.2.3.1 Principal Components Analysis

PCA analyzes a quantitative data matrix $\mathbf{X}$ with I rows (observations) and J columns (variables) (Abdi and Williams 2010a). The matrix $\mathbf{X}$ is first pre-processed such that columns are centered and often normalized (i.e., the sum of squares of each column equals 1). With the centered and normed matrix denoted $\mathbf{Z}_{\mathbf{X}}$, PCA is then defined as the analysis of the triplet $(\mathbf{Z}_{\mathbf{X}}, \mathbf{I}_J, \mathbf{I}_I)$.

6.2.3.2 Correspondence Analysis

Correspondence Analysis (CA) is analogous to a PCA but for—typically—contingency tables (i.e., the cross product of two disjunctive data tables; see Table 6.1) (Greenacre 1984; Lebart et al. 1984; Abdi and Williams 2010b; Abdi and Béra 2014). CA requires specific pre-processing and constraints prior to the GSVD step. First, for a matrix $\mathbf{R}$ of size J by K we compute a matrix of *observed* values:

$$\mathbf{O}_{\mathbf{R}} = N^{-1}\mathbf{R} \tag{6.6}$$

where N is the total sum of $\mathbf{R}$. The row constraint matrix $\mathbf{M}$ and column constraint matrix $\mathbf{W}$ are defined as:

$$\mathbf{m} = \mathbf{O}_{\mathbf{R}}\mathbf{1}_J \text{ and } \mathbf{M} = \mathrm{diag}\left\{\mathbf{m}\right\}, \tag{6.7}$$

and as

$$\mathbf{w} = \mathbf{1}_K\mathbf{O}_{\mathbf{R}} \text{ and } \mathbf{W} = \mathrm{diag}\left\{\mathbf{w}\right\} \tag{6.8}$$

Table 6.1 Example of nominal data table, and its disjunctive counterpart

(a) Nominal

	Variable 1	...	Variable *J*
Subj.1	A	...	A
Subj.2	A	...	A
...	...	...	...
Subj.I-1	B	...	C
Subj.I	C	...	B

(b) Disjunctive

	Variable 1			...	Variable *J*		
	A	B	C		A	B	C
Subj.1	1	0	0	...	1	0	0
Subj.2	1	0	0	...	1	0	0
...	...	...	...	...	...	...	...
Subj.I-1	0	1	0	...	0	0	1
Subj.I	0	0	1	...	0	1	0

where $\mathbf{m}$ (respectively $\mathbf{w}$) is the vector of the row (respectively column) sums of $\mathbf{O_R}$. Next, we compute an *expected* matrix

$$\mathbf{E_R} = \mathbf{m}\mathbf{w}^{\mathrm{T}}. \tag{6.9}$$

Finally, we compute the matrix of deviations:

$$\mathbf{Z_R} = \mathbf{O_R} - \mathbf{E_R}. \tag{6.10}$$

The CA of $\mathbf{R}$ is performed from the analysis of the triplet $\left(\mathbf{Z_R}, \mathbf{W}^{-1}, \mathbf{M}^{-1}\right)$.

6.2.3.3 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is a specific version of CA applied to a single disjunctive data table (see Table 6.1). MCA can be carried out by following the steps of CA as outlined in Sect. 6.2.3.2. However, there are several ways to define MCA as a centered, non-normalized, and weighted PCA (Bécue-Bertaut and Pagès 2008). Here, we provide another alternative MCA formulation.

Given a matrix $\mathbf{N}$ with I rows as observations and N nominal columns (see Table 6.1a.), we then transform $\mathbf{N}$ into the disjunctive formated (see Table 6.1b.) matrix $\mathbf{R}$, which has I rows and J columns. First, we define the constraints, where R is the sum of $\mathbf{R}$:

$$\mathbf{M} = I\mathbf{I}_I \text{ and } \mathbf{m} = \mathsf{diag}\left\{\mathbf{M}\right\} \tag{6.11}$$

$$\mathbf{w} = R^{-1}\left(\mathbf{1}_J\mathbf{R}\right) \text{ and } \mathbf{W} = \mathsf{diag}\left\{\mathbf{w}\right\}. \tag{6.12}$$

MCA can now be performed as the analysis of triplet: $\left(R^{-1}\mathbf{Z_R}, \mathbf{W}^{-1}, \mathbf{M}\right)$, where $\mathbf{Z_R}$ is the centered non-normalized version of $\mathbf{R}$, and provides the same solution as standard MCA (within a constant scaling factor).

PCA and MCA are equivalent when all variables have *exactly* two levels (Greenacre 1984; Lebart et al. 1984). For example "yes" vs. "no" which would be coded as [1 0] and [0 1], respectively. The equivalence holds in the following case. Traditional MCA—as performed via CA—would be applied to the complete

disjunctive matrix (which represents all levels), whereas PCA would be applied to a *strictly binary* table where each variable is represented by only 1 column. In this case, for example, "yes" is denoted with a 1 whereas "no" is denoted with a 0 (essentially, just half of the usual table for MCA).

6.2.4 Partial Least Squares Correlation

Partial Least Squares Correlation (PLSC) (Abdi and Williams 2013; Krishnan et al. 2011; McIntosh et al. 1996; Bookstein 1994) exists under a wide varieties of other monikers such as: the SVD of two covariance fields (Bretherton et al. 1992), PLS-SVD (Wegelin 2000), canonical covariance analysis (Tishler et al. 1996), co-inertia analysis (Dray 2014), or the specifically named—though broadly applicable—"multivariate analysis of genotype-phenotype associations" (Mitteroecker et al. 2016); but PLSC probably best traced back to Tucker's inter-battery factor analysis (Tucker 1958)—a method that analyzes the information common to two data tables measured on the same set of observations. Given two matrices, $\mathbf{X}$ and $\mathbf{Y}$, each containing I rows (observations) with (respectively) J columns ($\mathbf{X}$'s variables) and K columns ($\mathbf{Y}$'s variables), the matrices $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are the centered and unitary normed versions of $\mathbf{X}$ and $\mathbf{Y}$. With $\mathbf{Z_R} = \mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}$, PLSC is then defined as the analysis of the triplet $(\mathbf{Z_R}, \mathbf{W_Y}, \mathbf{W_X})$ where (typically) $\mathbf{W_X} = \mathbf{I}_J$ and $\mathbf{W_Y} = \mathbf{I}_K$, PLSC extracts the information common to $\mathbf{X}$ and $\mathbf{Y}$ by computing two sets of latent variables defined as:

$$\mathbf{L_X} = \mathbf{Z_X}\mathbf{W_X}\mathbf{U} \text{ and } \mathbf{L_Y} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{V} \tag{6.13}$$

In PLSC, associated latent variables have maximal covariance. Specifically, call $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ the linear transformation coefficients for $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ respectively. A latent variable for each matrix is defined as $\mathbf{l_X} = \mathbf{Z_X}\mathbf{W_X}\mathbf{u}_\ell$ and $\mathbf{l_Y} = \mathbf{Z_Y}\mathbf{W_Y}\mathbf{v}_\ell$ where

$$\underset{\mathbf{u},\mathbf{v}}{\arg\max}\left(\mathbf{l_X}^{\mathrm{T}}\mathbf{l_Y}\right) = \underset{\mathbf{u},\mathbf{v}}{\arg\max}\,\mathrm{cov}\left(\mathbf{l_X}, \mathbf{l_Y}\right), \tag{6.14}$$

under the constraints that $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ have unit norm:

$$\mathbf{u_\ell}^{\mathrm{T}}\mathbf{W_X}\mathbf{u}_\ell = 1 = \mathbf{v_\ell}^{\mathrm{T}}\mathbf{W_Y}\mathbf{v}_\ell. \tag{6.15}$$

After the ℓ-th pair of latent variables are extracted, the subsequent ones are extracted under the additional constraint of orthogonality:

$$\mathbf{l_{X_\ell}}^{\mathrm{T}}\mathbf{l_{Y_{\ell'}}} = 0 \text{ when } \ell \neq \ell'. \tag{6.16}$$

Each successive $\mathbf{l_X}$ and $\mathbf{l_Y}$ is stored in $\mathbf{L_X}$ and $\mathbf{L_Y}$, respectively, where

$$\mathbf{L_X}^{\mathrm{T}}\mathbf{L_Y} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_X}^{\mathrm{T}}\mathbf{Z_Y}\mathbf{W_Y}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{Z_R}\mathbf{W_Y}\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{W_X}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}\mathbf{W_Y}\mathbf{V} = \boldsymbol{\Delta}, \tag{6.17}$$

because $\mathbf{U}^{\mathrm{T}}\mathbf{W}_{\mathbf{X}}\mathbf{U} = \mathbf{I}_L = \mathbf{V}^{\mathrm{T}}\mathbf{W}_{\mathbf{Y}}\mathbf{V}$ (where L is the rank of $\mathbf{Z}_{\mathbf{R}}$). The latent variables of PLSC maximize the covariance as expressed by the singular values (for proofs, see Bookstein 1994; Tucker 1958).

6.2.5 *Partial Least Squares-Correspondence Analysis*

Recently, we presented a PLSC method designed specifically for the analysis of two categorical data matrices: Partial Least Squares-Correspondence Analysis (PLSCA)—a technique that combines features of PLSC and CA (Beaton et al. 2013, 2016). PLSCA can be expressed as follows: $\mathbf{X}$ and $\mathbf{Y}$ are disjunctive matrices where $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is a contingency table. CA, as defined in Sect. 6.2.3.2, is applied to $\mathbf{R}$. The latent variables in PLSCA are computed according to Eq. 6.13, where:

$$\mathbf{Z}_{\mathbf{X}} = I^{\frac{1}{2}}X^{-1}\mathbf{X} \tag{6.18}$$

$$\mathbf{Z}_{\mathbf{Y}} = I^{\frac{1}{2}}Y^{-1}\mathbf{Y} \tag{6.19}$$

where X and Y are (respectively) the sums of $\mathbf{X}$ and $\mathbf{Y}$, and where $\mathbf{W}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{Y}}$ (cf. Eq. 6.13) are computed from Eqs. 6.7 and 6.8 (i.e., $\mathbf{M}$ and $\mathbf{W}$).

6.3 PLSC for Mixed Data Types

Though introduced in Beaton et al., here we establish a more efficient framework for PLSC that applies to mixed data types. We formalize this approach with respect to one table of continuous data and one table of categorical data. Categorical data can be treated as continuous data and analyzed with PCA to produce identical results to a MCA (see Sect. 6.2.3.3).

6.3.1 *Escofier-Style Transform for PCA*

In 1979, Brigitte Escofier presented a technique to analyze continuous data with CA to produce the same results as PCA (within a scaling factor) (Escofier 1979). Escofier showed that a quantitative variable, say $\mathbf{x}$ (i.e., a column from the matrix $\mathbf{X}$) that is centered with unitary norm, can be analyzed with CA if it is expressed as two vectors: $\frac{1-\mathbf{x}}{2}$ and $\frac{1+\mathbf{x}}{2}$ (see Table 6.2). Incidentally, dividing each set by 2 with this Escofier-style coding is superfluous when using the stochastic version of CA (see Sect. 6.2.3.2).

Call $\mathbf{Z}_{\mathbf{X}}$ the centered and unitary norm version of $\mathbf{X}$ with I rows and J observations, where $\mathbf{B}_- = 1 - \mathbf{Z}_{\mathbf{R}}$ and $\mathbf{B}_+ = 1 + \mathbf{Z}_{\mathbf{R}}$ where

$$\mathbf{B} = \left[\mathbf{B}_- \; \mathbf{B}_+\right], \tag{6.20}$$

Table 6.2 Example of Escofier's coding scheme of continuous data to perform a CA on continuous data. $\mathbf{x}_j$ denotes the j vector from a matrix $\mathbf{X}$ where $x_{i,j}$ denotes a specific value at row i and column j. This coding scheme is similar to the thermometer coding scheme often used for ordinal data in MCA

(a) Continuous data

	$\mathbf{x}_1$	...	$\mathbf{x}_J$
Subj.1	$x_{1,1}$	...	$x_{1,J}$
Subj.2	$x_{2,1}$	...	$x_{2,J}$
...	...	...	...
Subj.*I*-1	$x_{I-1,1}$	...	$x_{I-1,J}$
Subj.*I*	$x_{I,1}$	...	$x_{I,J}$

(b) Escofier-style transform

	$-\mathbf{x}_1$	$+\mathbf{x}_1$	...	$-\mathbf{x}_J$	$+\mathbf{x}_J$
Subj.1	$\frac{1-x_{1,1}}{2}$	$\frac{1+x_{1,1}}{2}$	...	$\frac{1-x_{1,J}}{2}$	$\frac{1+x_{1,J}}{2}$
Subj.2	$\frac{1-x_{2,1}}{2}$	$\frac{1+x_{2,1}}{2}$	...	$\frac{1-x_{2,J}}{2}$	$\frac{1+x_{2,J}}{2}$
...	...	...	...	...	...
Subj.*I*-1	$\frac{1-x_{I-1,1}}{2}$	$\frac{1+x_{I-1,1}}{2}$	...	$\frac{1-x_{I-1,J}}{2}$	$\frac{1+x_{I-1,J}}{2}$
Subj.*I*	$\frac{1-x_{I,1}}{2}$	$\frac{1+x_{I,1}}{2}$	...	$\frac{1-x_{I,J}}{2}$	$\frac{1+x_{I,J}}{2}$

The matrix $\mathbf{B}$—which has the properties of a disjunctive table, see Table 6.1—can then be analyzed with CA (as in Sect. 6.2.3.2) which is equivalent to a PCA via the analysis of the triplet: $\left(\frac{1}{IJ}\mathbf{Z}_\mathbf{X}, J\mathbf{I}_J, I\mathbf{I}_I\right)$.

There is one exception to the equivalence between these two methods: in the Escofier-style approach, the number of columns in $\mathbf{B}$ is $2J$, where J is the number of columns in $\mathbf{X}$. Each variable from $\mathbf{X}$ has essentially been duplicated in $\mathbf{B}$ much like "thermometer coding" (a.k.a. doubling or fuzzy coding Greenacre 2014) a là ordinal data analysis with MCA. Thermometer coding expresses each variable by two points that are equidistant from 0 (i.e., the mean).

6.3.2 *Escofier-Style Transform for PLSC*

To formalize PLSC for mixed data types, we first define PLSC approach for 2 continuous data matrices—$\mathbf{X}$ and $\mathbf{Y}$—but in the Escofier framework (Sect. 6.3.1 and also see Table 6.2). Let us call $\mathbf{B}_\mathbf{X}$ the Escofier-style transform of $\mathbf{X}$ and $\mathbf{B}_\mathbf{Y}$ the Escofier-style transform of $\mathbf{Y}$. If we use the standard form of PLSC, we decompose $\mathbf{B}_\mathbf{R} = \mathbf{B}_\mathbf{X}^\mathrm{T}\mathbf{B}_\mathbf{Y}$, where:

$$\mathbf{B}_\mathbf{R} = \begin{bmatrix} \left(\mathbf{B}_{\mathbf{X}-}^\mathrm{T}\mathbf{B}_{\mathbf{Y}-}\right) & \left(\mathbf{B}_{\mathbf{X}-}^\mathrm{T}\mathbf{B}_{\mathbf{Y}+}\right) \\ \left(\mathbf{B}_{\mathbf{X}+}^\mathrm{T}\mathbf{B}_{\mathbf{Y}-}\right) & \left(\mathbf{B}_{\mathbf{X}+}^\mathrm{T}\mathbf{B}_{\mathbf{Y}+}\right) \end{bmatrix}. \tag{6.21}$$

Because $\mathbf{B}_\mathbf{X}$ and $\mathbf{B}_\mathbf{Y}$ are each in the Escofier-style (i.e., pseudo-categorical), this problem can be treated as one tailored for PLSCA (i.e., PLSC for the two *categorical* matrices; see Sect. 6.2.5). The PLSCA of $\mathbf{B}_\mathbf{X}^\mathrm{T}\mathbf{B}_\mathbf{Y}$ is equivalent to the PLSC (see Sect. 6.2.4) of $\mathbf{Z}_\mathbf{X}$ and $\mathbf{Z}_\mathbf{Y}$ (within scaling factors). There are three items used to define equivalence between these approaches: (1) singular values, (2) component scores (for both rows and columns), and (3) latent variables.

Call (respectively) $\mathbf{\Delta}_{Z_R}$ and $\mathbf{\Delta}_{B_R}$ the singular values from a standard PLSC (of $\mathbf{Z_X}$ and $\mathbf{Z_Y}$) and the singular values from an Escofier-style PLSCA (of $\mathbf{B_X}$ and $\mathbf{B_Y}$). We use the Escofier style approach as the preferred method because, as an extension of CA, it provides a natural dual representation of the rows and columns. To transition between the two approaches, we do the following:

$$\mathbf{\Delta}_{B_R} = \frac{1}{I\sqrt{JK}}\mathbf{\Delta}_{Z_R}. \tag{6.22}$$

The transition between component scores is also defined as follows:

$$\mathbf{F}_{J\mathbf{B_R}} = \frac{1}{\frac{I}{J}\sqrt{J^2K}}\left[-\mathbf{F}_{J\mathbf{Z_R}} \quad \mathbf{F}_{J\mathbf{Z_R}}\right] \tag{6.23}$$

$$\mathbf{F}_{K\mathbf{B_R}} = \frac{1}{\frac{I}{K}\sqrt{K^2J}}\left[-\mathbf{F}_{K\mathbf{Z_R}} \quad \mathbf{F}_{K\mathbf{Z_R}}\right]. \tag{6.24}$$

And finally, the transition between latent variables are:

$$\mathbf{L}_{\mathbf{B_X}} = \sqrt{IJ}\,\mathbf{L}_{\mathbf{Z_X}} \tag{6.25}$$

$$\mathbf{L}_{\mathbf{B_Y}} = \sqrt{IK}\,\mathbf{L}_{\mathbf{Z_Y}}, \tag{6.26}$$

where the latent variables for the "standard" approach are defined as in Sect. 6.13, and the computation of latent variables for the Escofier-approach are defined as those for PLSCA in Sect. 6.2.5.

We have to duplicate the component scores from the standard PLSC and multiply by -1 because the Escofier-style transform is a "thermometer" style coding of the data (equidistant above and below 0; see Table 6.2). Given these properties, we can compute the standard PLSC with equivalence to the PLSCA via the GSVD as follows. First define $\mathbf{Z_{X*}}$ and $\mathbf{Z_{Y*}}$:

$$\mathbf{Z_{X*}} = J^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z_X} \tag{6.27}$$

$$\mathbf{Z_{Y*}} = K^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z_Y} \tag{6.28}$$

$$\mathbf{Z_{R*}} = \mathbf{Z}_{\mathbf{X*}}^{\mathrm{T}}\mathbf{Z_{Y*}}, \tag{6.29}$$

where (1) $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are centered and normed matrices of (respectively) $\mathbf{X}$ and $\mathbf{Y}$, (2) I are the number of rows (observations) in $\mathbf{X}$ and $\mathbf{Y}$, and (3) J and K are the number of columns (variables) for $\mathbf{X}$ and $\mathbf{Y}$. To produce the same results as the Escofier-style PLSC approach, the GSVD is described by the triplet: $(\mathbf{Z_{R*}}, K\mathbf{I}_K, J\mathbf{I}_J)$. Thus, for continuous data, we can transition between the standard approach to PLSC (see Sect. 6.2.4) and the Escofier-style approach to PLSCA (see Sect. 6.2.5).

6.3.3 Mixed Data and PLSC

The Escofier-style transformed matrix (see Table 6.2) is similar to a fully disjunctive matrix; and, because PLSC and PLSCA are equivalent when using Escofier-style pseudo-categorical matrices, we can use PLSCA to analyze mixed data types (i.e., one matrix of continuous data and one matrix of categorical data).

Call $\mathbf{B_Y}$ the Escofier-style transform of a continuous data matrix $\mathbf{Y}$ and call $\mathbf{X}$ a fully disjunctive data matrix (as in Table 6.1). Because both matrices are in a pseudo-categorical or categorical format we can define $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{B_Y}$ as a pseudo-contingency table since this $\mathbf{R}$ is the cross-product between a categorical matrix and a pseudo-categorical matrix. In fact, $\mathbf{R}$ expresses some of the properties we would expect from a contingency table but maintains the properties of $\mathbf{X}$ and $\mathbf{B_Y}$: the column sums of $\mathbf{R}$ are equal to one another—just as in $\mathbf{B_Y}$ and are also proportional to the column sums of $\mathbf{B_Y}$. This is also true for the row sums of $\mathbf{R}$ and the column sums of $\mathbf{X}$. Thus, the relationship between $\mathbf{X}$ and $\mathbf{B_Y}$ can be analyzed with PLSCA (see Sect. 6.2.5) and the properties that define PLSC still hold (see Sects. 6.2.4 and 6.2.5).

However, there is a minor drawback to this approach: The continuous data matrix, $\mathbf{Y}$, represents each variable twice in $\mathbf{B_Y}$ (see Sect. 6.2) and this could be problematic for very large data sets (e.g., neuroimaging, genomics). Thus, we now define a mixed data approach to PLS closer to PLSC, but that keeps key properties of CA (i.e., dual representation, distributional equivalence, emphasis on rare occurrences). Call $\mathbf{Y}$ a data matrix, with J columns, of continuous data where $\mathbf{Z_Y}$ is centered and normalized. Call $\mathbf{X}$ a fully disjunctive matrix, with K columns from N variables, where $\mathbf{Z_X}$ is centered but not normalized. Both $\mathbf{X}$ and $\mathbf{Y}$ have I rows (i.e., observations). First we define the data matrices derived from $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{Z_{X*}} = N^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z_X} \tag{6.30}$$

$$\mathbf{Z_{Y*}} = K^{-1}\sqrt{\frac{1}{I}}\,\mathbf{Z_Y}. \tag{6.31}$$

Next, we define weights associated to each set (where X is the sum of $\mathbf{X}$):

$$\mathbf{w_X} = X^{-1}\mathbf{1}_I\mathbf{X} \text{ and } \mathbf{W_X} = \mathsf{diag}\left\{\mathbf{w_X}\right\}, \tag{6.32}$$

and $\mathbf{W_Y} = K\mathbf{I}_K$. PLSC can then be performed on $\mathbf{Z_{Y*}}$ and $\mathbf{Z_{X*}}$ where $\mathbf{W_X}$ and $\mathbf{W_Y}$ are constraints for the GSVD. The GSVD step of PLSC in this case would analyze the triplet: $\left(\mathbf{R}, \mathbf{W_Y}, \mathbf{W_X^{-1}}\right)$ with $\mathbf{R} = \mathbf{Z}_{\mathbf{X*}}^{\mathrm{T}}\mathbf{Z_{Y*}}$. This approach is derived, in part, from MCA where MCA is treated as a centered, non-normalized, weighted PCA (see Sect. 6.2.3.3) and the standard approach to PLSC (see Sect. 6.2.4). We also imposed particular constraints on this formulation so that the results here would be equivalent to those done on $\mathbf{X}$ and $\mathbf{B_Y}$ obtained with PLSCA. However, there is also a drawback to this reformulation: supplemental projections are more difficult to

compute than in the CA approach. Therefore, we define MiMoPLS in one, final, way that combines the simplicity of the PLSCA approach with the minimally required data in the PLSC approach.

First, $\mathbf{X}$ is the complete disjunctive matrix where $\mathbf{B}_{\mathbf{Y}+} = \mathbf{Z}_{\mathbf{Y}} + 1$ (see Eq. 6.21 and Sect. 6.3.3), and $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{B}_{\mathbf{Y}+}$. The total sum of $\mathbf{B}_{\mathbf{Y}+}$ is equal to IK, where I is the number of observations and K is the number of columns in $\mathbf{Y}$ and we then use CA (see Sect. 6.2.3.2) where both $\mathbf{w}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{X}}$ are obtained from Eq. 6.32, and where $\mathbf{W}_{\mathbf{Y}} = K^{-1}\mathbf{I}_K$ where $\mathbf{w}_{\mathbf{Y}} = \mathsf{diag}\{\mathbf{W}_{\mathbf{Y}}\}$. Next we define the *observed*, *expected*, and *deviations* matrices (with R being the sum of all elements of $\mathbf{R}$):

$$\mathbf{O}_{\mathbf{R}} = R^{-1}\mathbf{R} \tag{6.33}$$

$$\mathbf{E}_{\mathbf{R}} = \mathbf{w}_{\mathbf{X}}\mathbf{w}_{\mathbf{Y}}^{\mathrm{T}} \tag{6.34}$$

$$\mathbf{Z}_{\mathbf{R}} = \mathbf{O}_{\mathbf{R}} - \mathbf{E}_{\mathbf{R}}. \tag{6.35}$$

The GSVD step then correspond to the analysis of the triplet $\left(\mathbf{Z}_{\mathbf{R}}, \mathbf{W}_{\mathbf{Y}}^{-1}, \mathbf{W}_{\mathbf{X}}^{-1}\right)$. Finally, the latent variables are computed as:

$$\mathbf{L}_{\mathbf{X}} = \left(I^{\frac{1}{2}}X^{-1}\mathbf{X}\right)\mathbf{W}_{\mathbf{X}}^{-1}\mathbf{U} \tag{6.36}$$

$$\mathbf{L}_{\mathbf{X}} = \left(I^{\frac{1}{2}}B^{-1}\mathbf{Z}_{\mathbf{Y}}\right)\mathbf{W}_{\mathbf{Y}}^{-1}\mathbf{V}, \tag{6.37}$$

where $\mathbf{Z}_{\mathbf{Y}}$ is the column centered and normalized version of $\mathbf{Y}$, and where X and B are (respectively) the sums of $\mathbf{X}$ and $\mathbf{B}_{\mathbf{Y}+}$. Recall that X is equal to IN, where N is the number of variables in $\mathbf{X}$, and B is equal to IK and this makes Eq. 6.36 analogous to the computation of the "observed" values in CA (see Sect. 6.2.3.2).

We now have an approach of analyzing mixed data types that (1) is in the PLSC fashion, (2) maintains the properties of PLSCA and CA (e.g., dual representation, simple supplemental projections), and (3) does not duplicate the representation of the continuous data matrix.

6.4 An Application to Alzheimer's Disease

We illustrate MiMoPLS with a data set—from the Alzheimer's Disease Neuroimaging Initiative (ADNI)—that contains brain imaging data obtained from diffusion tensor imaging (DTI)—as measured with fractional anisotropy (FA)—and genetic data obtained from single nuclear polymorphisms (SNPs). These data come from Phase 1 of the ADNI database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private funding partnership and includes public funding by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and the Food and Drug Administration. The primary goal

of ADNI has been to test a wide variety of measures to assess the progression of mild cognitive impairment and early AD. The ADNI project is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. Michael W. Weiner, MD (VA Medical Center and University of California San Francisco) is the ADNI PI. Subjects have been recruited from over 50 sites across the U.S. and Canada (for up-to-date information, see www.adni-info.org).

Participants include 29 individuals from the ADNI2 cohort classified into 4 groups: control ($N = 9$; CON), early mild cognitive impairment ($N = 11$; eMCI), late mild cognitive impairment ($N = 4$; ℓMCI), and Alzheimer's Disease ($N = 5$; AD). All participants were genotyped with genome-wide SNPs (Illumina HumanOmniExpress). SNPs underwent standard preprocessing (SNP & participant call rates were $\geq 90\,\%$, Hardy-Weinberg disequilibrium $\leq 1 \times 10^{-6}$, and minor allele frequency $\leq 5\,\%$). From the genome-wide data, we extracted 386 SNPs that, according the literature and aggregate sources (Bertram et al. 2007), should be associated with AD. We also extracted 35,062 voxels (of FA values) from 48 white matter tracts according to the JHU-ICBM-DTI-81 mask (see Fig. 6.1) (Oishi et al. 2008). We analyzed these data to identify the genetic contributions to white matter changes in an AD related population.

We present the analysis first with the descriptive component maps (Fig. 6.2). For illustrative purposes, we limit discussion to only the first two components. We can note that there is a higher variability of genotypes (top left; Fig. 6.2) than the FA values (top right; Fig. 6.2). Interpretation of these maps are done as they would be in CA: a genotype that is close to particular voxels is considered more related to those voxels than is the average genotype.

The latent variables suggest two interpretations of the components. First, Component 1 largely reflects the differences between ℓMCI (left side of Component 1) and AD (right side of Component 1), whereas Component 2 is characterized by {CON & eMCI} vs. {ℓMCI & AD}. This pattern suggests that Component 1 separates real AD pathology from possible misdiagnoses, whereas Component 2 appears to characterize non-pathological to pathological features. Further, we can interpret the

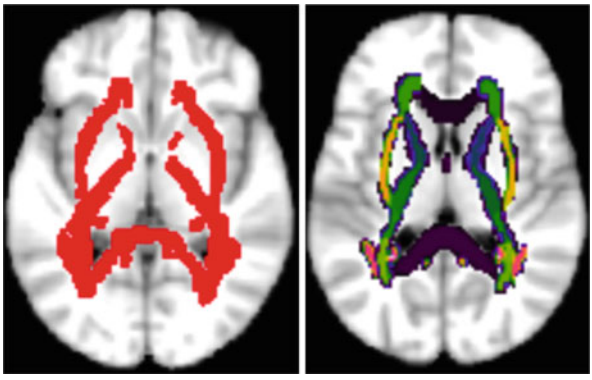

Fig. 6.1 Masks to identify white matter regions in a common (MNI) space. The *left figure* illustrates all the voxels included, whereas the *right figure* illustrates the separate tracts within this mask

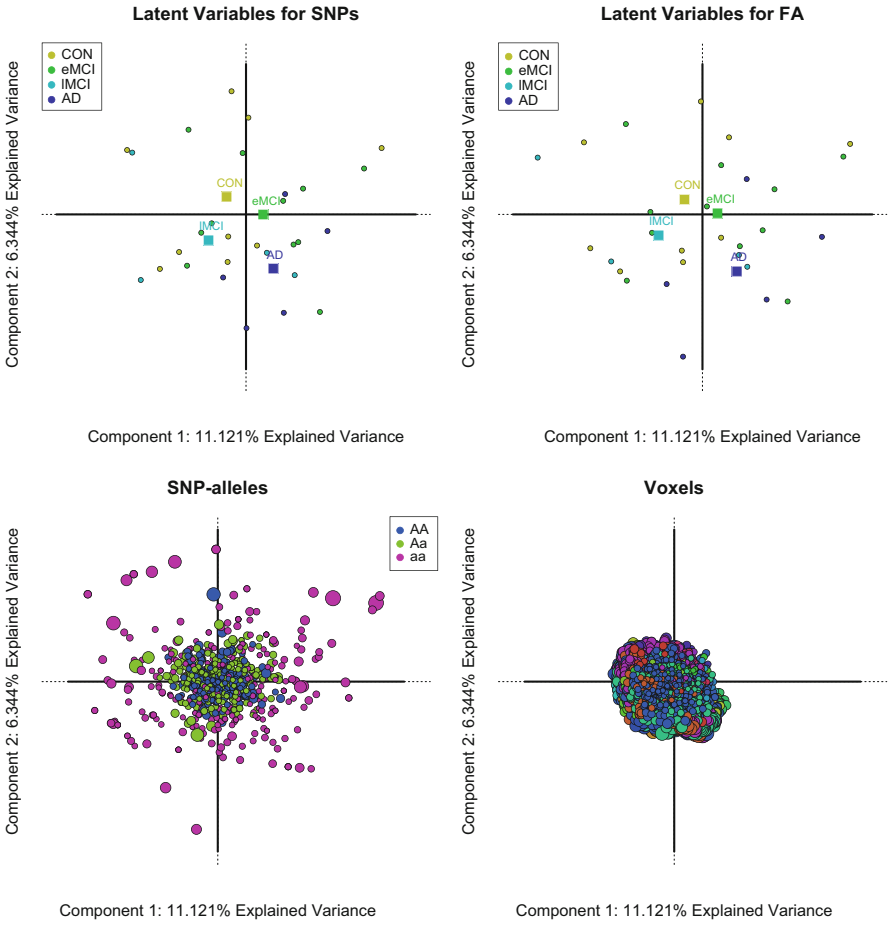

Fig. 6.2 *Top figures* show the individual participants' scores (latent variables) with respect to the genotypes (*left*) and FA values (*right*). The average of each participant group is labeled with a *large square*, whereas participants are labeled with *small circles*. *Bottom figures* show the component scores of the genotypes (*left*; colored by major homozygote, heterozygote, and minor homozygote), and the voxels (*right*; colored by tract)

latent variables (bottom; Fig. 6.2) as we would in both CA and in PLSC. Participants whose scores are closer to particular genotypes or FA values are more associated with those features than the average participant. Furthermore, we can include more meaningful information (e.g., group averages) to better understand the relationship between genotypes and white matter integrity. Doing so indicates that the CON group is associated with the upper left quadrant, the AD group is associated with the lower right quadrant, the eMCI group is associated with the upper right quadrant, and the ℓMCI group is associated with the lower left quadrant. Thus, we can infer that particular genotypes and voxels are more associated to these groups than others.

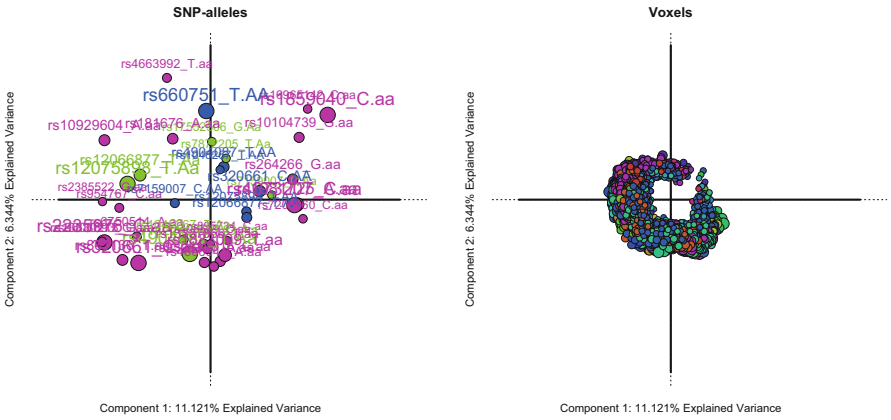

Fig. 6.3 Bootstrap ratios identify items that significantly contribute to the component structure

However, given that there are so many genotypes and voxels, we use inferential methods to eliminate non-significant genotypes and voxels.

One approach is to use the bootstrap (Efron 1979; Hesterberg 2011) and to compute bootstrap ratio values (BSRs) (Beaton et al. 2016; Krishnan et al. 2011) which are *t*-like statistics computed from the mean and standard deviation of the bootstrap distribution. With BSRs, we can reduce the number of items to interpret by selecting only the items that significantly contribute to the component structure (see Fig. 6.3): Here we only show items whose BSR magnitude is larger than 2.50. SNPs are labeled by the gene with which they are most associated, the voxels are plotted in standard MNI brain maps.

We first interpret the brain images (because more is known about white matter integrity than genetics in these populations); they provide a baseline from which a genetic relationship can be inferred. Component 1 (Fig. 6.4; lower left) shows small clusters in bilateral superior corona radiata and posterior internal capsule (blue colored voxels), whereas there are large clusters throughout anterior white matter tracts (i.e., genu and body of corpus callosum, internal and external capsule, and corona radiata; denoted with red voxels). Component 2, generally only has negative BSR values. The voxels (denoted in red) trace a path from lateral temporal lobe, to longitudinal tracts leading to frontal regions (i.e., internal and external capsuale, and corona radiata). Taken in context with the latent variables (Fig. 6.2), changes in white matter in anterior tracts are more associated with AD, whereas longitudinal tracts are more associated with ℓMCI. This pattern suggests that early biomarkers indicate the progression from ℓMCI to AD and, overall, as indicated by Fig. 6.4 that particular markers are associated with specific clinical groups: For example, UCK2 heterozygotes are more associated with ℓMCI whereas UCK2 major homozygotes are more associated with AD. Component 2 identifies fiber paths that interconnect temporal, parietal, and frontal regions—all regions often implicated in the progression of Alzheimer's pathology. Taken in context with the

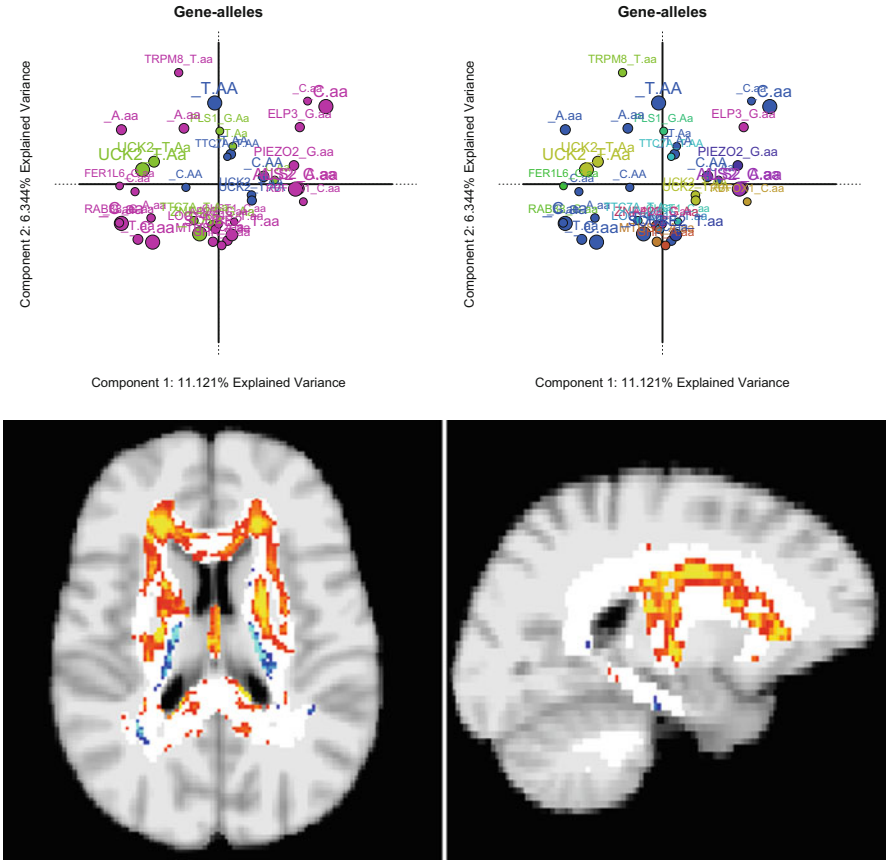

Fig. 6.4 All genotypes have been renamed to show which gene they are most associated with. genotypes colored by major homozygote, heterozygote, or minor homozygote (*top left*) and their respective genes (*top right*). Bootstrap ratio values are plotted in the voxels (*bottom*) to indicate their location and the strength of their contribution

latent variables (Fig. 6.2), this pattern suggest that there are substantial changes in these regions in late stage (ℓMCI) and pathological (AD) groups. Finally Fig. 6.4 shows that a heterozygote of a SNP associated with ZNF423 and the minor homozygote of a SNP associated with APOE—a pattern that confirms the importance of these two genes routinely associated with AD.

6.5 Conclusion

This article presents a new approach to PLS that integrates mixed-data types. Our presentation included continuous (brain imaging) and categorical (SNPs) data, but the method can be easily extended to ordinal data (via thermometer coding,

see Sect. 6.3.1). Though we present MiMoPLS via PLSC, MiMoPLS can easily be extended to other PLS approaches (e.g., regression, path-modeling). Future work includes regularization and sparsification designed specifically for block-wise categorical data (Takane and Hwang 2006) and two-way sparsification of the SVD (Allen 2013).

Acknowledgements **DB** is currently supported via training grant by the NIH and National Institute on Drug Abuse (F31DA035039).

FMF is currently supported by the NIH and National Institute on Drug Abuse (R01DA030344). **HA** would like to acknowledge the support of an EURIAS fellowship at the Paris Institute for Advanced Studies (France), with the support of the European Union's 7th Framework Program for research, and from a funding from the French State managed by the "Agence Nationale de la Recherche (program: Investissements d'avenir, ANR-11-LABX-0027-01 Labex RFIEA+)." **ADNI**: Data collection and sharing for this project was funded by the ADNI (NIH Grant U01 AG024904) and DOD ADNI (W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; ; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Abdi, H.: Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In: Salkind, N. (ed.) Encyclopedia of Measurement and Statistics, pp. 907–912. Sage, Thousand Oaks (2007)

Abdi, H., Béra, M.: Correspondence analysis. In: Alhajj, R., Rokne, J. (eds.) Encyclopedia of Social Networks and Mining, pp. 275–284. Springer, New York (2014)

Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdiscip. Rev.: Comput. Stat. **2**, 433–459 (2010a)

Abdi, H., Williams, L.J.: Correspondence analysis. In: Salkind, N. (ed.) Encyclopedia of Research Design, pp. 267–278. Sage, Thousand Oaks (2010b)

Abdi, H., Williams, L.J.: Partial least squares methods: partial least squares correlation and partial least square regression. In: Reisfeld, B., Mayeno, A. (eds.) Methods in Molecular Biology: Computational Toxicology, pp. 549–579. Springer, New York (2013)

Allen, G.I.: Sparse and Functional Principal Components Analysis (2013). arXiv preprint arXiv:1309.2895

Beaton, D., Filbey, F.M., Abdi, H.: Integrating partial least squares correlation and correspondence analysis for nominal data. In: Abdi, H., Chin, W.W., Esposito Vinzi, V., Russolillo, G., Trinchera, L. (eds.) New Perspectives in Partial Least Squares and Related Methods, pp. 81–94. Springer, New York (2013)

Beaton, D., Dunlop, J., ADNI, Abdi, H.: Partial least squares-correspondence analysis: a framework to simultaneously analyze behavioral and genetic data. Psychol. Methods **20** (2016, in press)

Bécue-Bertaut, M., Pagès, J.: Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computat. Stat. Data Anal. **52**, 3255–3268 (2008)

Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E.: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat. Genet. **39**, 17–23 (2007)

Bookstein, F.: Partial least squares: a dose–response model for measurement in the behavioral and brain sciences. Psycoloquy **5**(23), 1–10 (1994)

Bretherton, C.S., Smith, C., Wallace, J.M.: An intercomparison of methods for finding coupled patterns in climate data. J. Clim. **5**, 541–560 (1992)

Cantor, R.M., Lange, K., Sinsheimer, J.S.: Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet. **86**, 6–22 (2010)

De la Cruz, O., Holmes, S.P.: The duality diagram in data analysis: examples of modern applications. Ann. Appl. Stat. **5**, 2266–2277 (2010)

Dray, S.: Analyzing a pair of tables: co-inertia analysis and duality diagrams. In: Blasius, J., Greenacre, M. (eds.) Visualization and Verbalization of Data, pp. 289–300. CRC Press, London (2014)

Efron, B.: Bootstrap methods: another look at the Jackknife. Ann. Stat. **7**, 1–26 (1979)

Escofier, B.: Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Les Cahiers de l'Analyse Des Données **4**, 137–146 (1979)

Escoufier, Y.: Operators related to a data matrix: a survey. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT: 17th Symposium Proceedings in Computational Statistics, Rome, pp. 285–297. Physica Verlag, New York (2006)

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., ... Campion, D.: APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol. Psychiatry **16**, 903–907 (2012)

Greenacre, M.J.: Theory and Applications of Correspondence Analysis. Academic, London (1984)

Greenacre, M.: Data doubling and fuzzy coding. In: Blasius, J., Greenacre, M. (eds.) Visualization and Verbalization of Data, pp. 239–253. CRC Press, London (2014)

Hesterberg, T.: Bootstrap. Wiley Interdiscip. Rev.: Comput. Stat. **3**, 497–526 (2011)

Krishnan, A., Williams, L.J., McIntosh, A.R., Abdi, H.: Partial least squares (PLS) methods for neuroimaging: a tutorial and review. NeuroImage **56**, 455–475 (2011)

Lebart, L., Morineau, A., Warwick, K.M.: Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices. Wiley, New York (1984)

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., ... Duchesnay, É.: Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. NeuroImage **63**, 11–24 (2012)

Liu, J., Calhoun, V.D.: A review of multivariate analyses in imaging genetics. Front. Neuroinform. **8**, 29 (2014)

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N.I., Calhoun, V.: Combining *f*MRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Hum. Brain Mapp. **30**, 241–255 (2009)

McIntosh, A.R., Bookstein, F.S., Haxby, J., Grady, C.: Spatial pattern analysis of functional brain images using partial least squares. NeuroImage **3**, 143–157 (1996)

Meda, S.A., Jagannathan, K., Gelernter, J., Calhoun, V.D., Liu, J., Stevens, M.C., Pearlson, G.D.: A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. NeuroImage **53**, 1007–1015 (2010)

Meyer-Lindenberg, A.: The future of *f*MRI and genetics research. NeuroImage **62**, 1286–1292 (2012)

Mitteroecker, P., Cheverud, J.M., Pavlicev, M.: Multivariate analysis of genotype–phenotype association. Genetics **202**(4), 1345–1363 (2016)

Oishi, K., Zilles, K., Amunts, K., Faria, A., Jiang, H., Li, X., ... Mori, S.: Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter. NeuroImage **43**, 447–457 (2008)

Sheng, J., Kim, S., Yan, J., Moore, J., Saykin, A., Shen, L.: Data synthesis and method evaluation for brain imaging genetics. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, pp. 1202–1205 (2014)

Takane, Y., Hwang, H.: Regularized multiple correspondence analysis. In: Greenacre, M., Blasius, J. (eds.) Multiple Correspondence Analysis and Related Methods, pp. 259–279. Academic, London (2006)

Thompson, P.M., Martin, N.G., Wright, M.J.: Imaging genomics. Curr. Opin. Neurol. **23**, 368–373 (2010)

Tishler, A., Dvir, D., Shenhar, A., Lipovetsky, S.: Identifying critical success factors in defense development projects: a multivariate analysis. Technol. Forecast. Soc. Change **51**, 151–171 (1996)

Tucker, L.R.: An inter-battery method of factor analysis. Psychometrika **23**, 111–136 (1958)

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J.: Five years of GWAS discovery. Am. J. Hum. Genet. **90**, 7–24 (2012)

Vounou, M., Nichols, T.E., Montana, G.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. NeuroImage **53**, 1147–1159 (2010)

Wegelin, J.A.: A survey of partial least squares (PLS) methods, with emphasis on the two-block case. Technical report, University of Washington (2000)

Weiner, M.P., Hudson, T.J.: Introduction to SNPs: discovery of markers for disease. BioTechniques **10**(4–7), 12–13 (2002)

Zapala, M.A., Schork, N.J.: Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc. Natl. Acad. Sci. **103**, 19430–19435 (2006)

Chapter 7
PLS and Functional Neuroimaging: Bias and Detection Power Across Different Resampling Schemes

Nathan Churchill, Babak Afshin-Pour, and Stephen Strother

Abstract Correlation Partial-Least Squares (PLSC) provides a robust model for analyzing functional neuroimaging data, which is used to identify functional brain networks that show the largest covariance with task stimuli. However, neuroimaging data tend to be high-dimensional (i.e., there are far more variables *P* than samples *N*), with significant noise confounds and variability in brain response. It is therefore challenging to identify the significant, stable components of PLSC analysis. Empirical significance estimators are widely used, as they make minimal assumptions about data structure. The most common estimator in neuroimaging PLS is Bootstrapped Variance (BV), which tests whether bootstrap-stabilized mean component eigenvalues (i.e., covariance) are significantly different from a permuted null distribution; however, recent studies have highlighted issues with this model. Two alternatives were proposed that instead focus on reliability of the PLSC saliences (i.e., singular vectors): a Split-half Stability (SS) model that measures the consistency of reconstructed components for split-half data, and Split-half Reproducibility (SR) which measures the reliability across independent split-half analyses. We compare BV, SS, and SR estimators on functional Magnetic Resonance Imaging (*f*MRI) data, for both simulated and experimental datasets. The SS and SR methods have comparable sensitivity in detecting "brain" components for most simulated and experimental conditions. However, SR shows consistently greater sensitivity for "task" components. We demonstrate that this is due to relative bias in the SS model: both "brain" and "task" components have biased null distributions, but for the low-dimensional "task" vectors, this bias becomes sufficiently high that it is often impossible to distinguish a significant effect from the null distribution.

N. Churchill (✉)
Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada
e-mail: nchurchill.research@gmail.com

B. Afshin-Pour • S. Strother
Rotman Research Institute, Baycrest Hospital, Toronto, ON, Canada
e-mail: bafshinpour@research.baycrest.org; sstrother@research.baycrest.org

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_7

Keywords *f*MRI • Behavioral PLS • Bootstrap • Split-half resampling • Prediction • Reproducibility • PCA

7.1 Introduction

Functional neuroimaging techniques such as functional Magnetic Resonance Imaging (*f*MRI) provide information about how brain function is related to experimental stimuli. However, these data are extremely high-dimensional (i.e., the number of voxels P greatly exceeds the number of samples N), and significant noise confounds are often present. Moreover, brain responses exhibit both within- and between-subject variability. It is therefore a challenge to detect functional brain networks that have a significant, reliable relationship with experimental conditions.

Partial Least Squares Correlation (PLSC) provides a robust linear model (Wold 1985; Krishnan et al. 2011), which can be used to identify the brain networks that exhibit greatest covariance with the task conditions. The general PLSC problem, for data matrices $\mathbf{X}$ and $\mathbf{Y}$, is to find linear projections $\mathbf{u}$ and $\mathbf{v}$ such that:

$$\arg\max_{\mathbf{u},\mathbf{v}} \left(\text{cov}\left(\mathbf{Xu}, \mathbf{Yv}\right)\right)^2 \quad \text{under the constraints} \quad ||\mathbf{u}|| = ||\mathbf{v}|| = 1 \tag{7.1}$$

for which multiple solution algorithms have been developed (Rosipal and Krämer 2006). PLSC—the preferred method in brain imaging—applies the Singular Value Decomposition to the matrix product $\mathbf{R} = \mathbf{X}^T\mathbf{Y}$, decomposing into $\mathbf{R} = \mathbf{ULV}^T$ with sets of paired components $\mathbf{u}_k$ and $\mathbf{v}_k$ (left and right singular vectors of $\mathbf{R}$, respectively). For functional neuroimaging, the columns of $\mathbf{X}$ and $\mathbf{Y}$ are almost always centered and normalized to unit variance, and therefore $\mathbf{R}$ is almost always a cross-correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$. This approach estimates the shared information between variable sets (e.g., brain regions and experimental stimuli), rather than defining an explicit predictive relationship and is often used to identify brain networks that are modulated across different task conditions.

The PLSC approach provides a unique solution for a given data-set, but this is not guaranteed to reflect significant, generalizable covariance relationships. There are no straightforward analytic approaches to significance estimation in neuroimaging PLSC, partly due to the challenges of modeling unknown, data-dependent autcorrelations. Moreover, analytic estimators are often insensitive to true data dimensionality in multivariate network analysis (Yourganov et al. 2011). As an alternative, empirical resampling techniques are widely used to evaluate the significance of PLS components. Using this highly flexible approach, the test statistic of interest is compared against an empirical null distribution, obtained by randomly permuting variable labels.

The most well-established neuroimaging approach is to test the significance of Bootstrap-stabilized mean covariance estimates, (i.e., the diagonal elements, λ_{kk}, of the singular value matrix $\mathbf{L}$; see McIntosh et al. (1996), Krishnan et al. (2011), and Abdi et al. (2013) for details). This approach provides a simple, interpretable

statistic, but two recent publications have highlighted issues with the Bootstrapped variance (BV) model. Kovacevic et al. (2013) pointed out that this variance-driven approach is highly sensitive to outliers, and Churchill et al. (2013) demonstrated that Bootstrapped analysis of high-dimensional neuroimaging data can provide highly biased parameter estimates. Both Kovacevic et al., and Churchill et al., also noted that covariance magnitude provides no information about the underlying stability of brain and behavioral saliences (i.e., loadings on brain voxels and experimental conditions, stored in component matrices **U** and **V**), even though knowing this stability is crucial when interpreting results.

Both articles proposed alternatives to the Bootstrapped Variance (BV) model, based on split-half cross-validation. Kovacevic et al. (2013) proposed a test of split-half stability (SS), which measures how consistently PLS reconstructs **U** on data split-halves, using **V** estimated on the full dataset (and vice-versa). This test is based on PLS parameters estimated from the full data set, which maximizes detection power, but introduces an unknown estimation bias. Churchill et al. (2013) proposed a split-half reproducibility (SR) approach, which compares **U** and **V** estimated on independent data split-halves. It has lower detection power but minimizes bias in parameter estimates. It is unknown which model is best able to discriminate the test statistic from the null distribution for PLSC analysis of *f*MRI data. In this paper, we compare BV, SS, and SR testing approaches for (1) simulations over a range of different parameter settings, and (2) experimental task data.

7.2 Methods

7.2.1 PLSC in Task fMRI

We evaluate resampling methods for a standard task PLSC model, defined as follows. For $s = 1, \ldots, S$ subjects, we obtain a set of $V \times 1$ brain images, acquired during C different task conditions. We compute the average image per condition, and thereby obtain a $V \times C$ subject matrix $\mathbf{D}_s$. Finally, we compute the average matrix across subjects $\mathbf{D}_{avg}$, and perform singular value decomposition (SVD) of this matrix as $\mathbf{D}_{avg} = \mathbf{ULV}^T$. Matrices **U** and **V** provide a set of orthonormal basis vectors, where the kth pair of column vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ form the "brain" and "task" components respectively, and covariance scaling is given by diagonal element λ_{kk} of singular matrix **L**.

7.2.2 Resampling and Empirical Significance

Bootstrapped Variance (BV): In order to compute a stable estimate of the covariance statistic λ_{kk}, we resample on subject matrices $\mathbf{D}_s$ with replacement,

and perform PLSC on the bootstrapped average matrix $\mathbf{D}^*_{avg}$; this process produces eigenvalue estimates λ^*_{kk}. We compute the median λ^*_{kk} value over 1000 bootstrap samples, to obtain a stable estimate.

Split-half Stability (SS): for this model, we estimate the stability of $\mathbf{u}_k$ and $\mathbf{v}_k$ component vectors. Beginning with the full dataset $\mathbf{D}_{avg}$ we compute component matrices $\mathbf{U}$ and $\mathbf{V}$. Subjects are then randomly split into two equal-size groups and we obtain split-half matrices $\mathbf{D}_{sp1}$ and $\mathbf{D}_{sp2}$. For each split ($i = 1, 2$), we compute the "reconstruction" of each singular matrix, based on the other PLSC parameters estimated on the full-data matrix:

$$\mathbf{U}_{sp(i)} = \mathbf{D}_{sp(i)}\mathbf{V}\mathbf{L}^{-1} \tag{7.2}$$

and

$$\mathbf{V}_{sp(i)} = \mathbf{D}^T_{sp(i)}\mathbf{U}\mathbf{L}^{-1} \, . \tag{7.3}$$

We then compute the correlations between split-half estimates of the "brain" and"task" saliences $r_{brain,k} = \rho(\mathbf{u}_{sp1,k}, \mathbf{u}_{sp2,k})$ and $r_{task,k} = \rho(\mathbf{v}_{sp1,k}, \mathbf{v}_{sp2,k})$ respectively. These values reflect the sensitivity of brain and task saliences to subject heterogeneity. We obtain the median r_{brain} and r_{task} values over 100 resampling iterations to stabilize parameter estimates.

Split-half Reproducibility (SR): this model instead estimates the reproducibility of $\mathbf{u}_k$ and $\mathbf{v}_k$ vectors between independent data splits. As with SS, subjects are randomly split into two equal-size groups and we obtain split-half matrices $\mathbf{D}_{sp1}$ and $\mathbf{D}_{sp2}$, but for each split ($i = 1, 2$), we compute the independent split-half estimates:

$$\mathbf{D}_{sp(i)} = \mathbf{U}_{sp(i)}\mathbf{L}_{sp(i)}\mathbf{V}^T_{sp(i)} \, . \tag{7.4}$$

We then calculate the correlation between independent split-half estimates of the "brain" and "task" saliences $r_{brain,k} = \rho(\mathbf{u}_{sp1,k}, \mathbf{u}_{sp2,k})$ and $r_{task,k} = \rho(\mathbf{v}_{sp1,k}, \mathbf{v}_{sp2,k})$. These values quantify the reproducibility of salience patterns between independently analyzed data split-halves. We account for potential mismatch between components by performing constrained Procrustes matching of components in $\mathbf{V}_{sp(i)}$ relative to full-data $\mathbf{V}$, and components in $\mathbf{U}_{sp(i)}$ relative to full-data $\mathbf{U}$. This procedure permutes and sign-flips components of the split-half data in order to minimize a sum-of-squares cost function. We compute the median r_{brain} and r_{task} over 100 resampling iterations to stabilize parameter estimates.

For all three models, we compute p-values based on an empirical estimate of the null distribution. We randomly permute condition labels for all subjects' mean matrices $\mathbf{D}_s$, as we assume approximate independence between the mean conditions. We then calculate the test statistics of λ^*_k, $r^*_{brain,k}$, or $r^*_{task,k}$ on the randomized data. This process is repeated for 1000 resamples in order to generate an empirical null distribution. We compute empirical p-values based on the fraction of the null distribution that exceeds the test statistic (median λ^*_k, $r^*_{brain,k}$, or $r^*_{task,k}$ of the unpermuted data), and this provides a 1-tailed significance test.

7.2.3 Data

7.2.3.1 Simulated Dataset

We simulated *f*MRI data for a single-slice brain image, with additive Gaussian noise, based on the model developed by Lukic et al. (2002). This model consisted of a 100 × 100 pixel image, with a background structure of "grey matter" along the rim and in the center of the image, and "white matter" in between. The amplitude of background signal in "grey matter" was 4 times higher than in "white matter." The images were spatially smoothed by convolving with a 2D Gaussian kernel (FWHM = 2 pixels). After smoothing, the standard deviation of noise was 5 % of the background signal. Images contained 12 Gaussian activation loci in grey matter, which were added to the smoothed noisy background image. The FWHM of the activations varied between 2 and 4 pixels.

For each simulation run, we simulated 3 task salience vectors, by random sampling of a $K \times 1$ Gaussian vector, orthogonalized relative to all previously-sampled salience vectors. We simulated spatial salience vectors (i.e., a brain network associated with the task salience) by randomly assigning positive, negative, or zero expression of the task salience to each of the 12 activation loci. For this model, we varied four parameter settings: signal variance in "activated" regions, as the proportion of noise variance $V = 0$ to 4; number of subjects $S = 10$ to 80; and number of task conditions $C = 6$ to 20. We also simulated a confound of data heterogeneity, by randomly permuting condition labels on a subset of subjects, to create a percentage of "null" data $P_{null} = 0\,\%$ to $60\,\%$.

We generated 100 simulation datasets for each parameter setting, and computed empirical *p*-values on the BV, SS and SR resampling models for each simulation dataset. We then displayed the median *p*-value across simulation runs, for a single latent variable.

7.2.3.2 Experimental Dataset

Twenty-seven young normal subjects (20–33 years, 15 females) were scanned with *f*MRI while performing a block-design task with three different conditions: in Task-A, numbers 1–14 are pseudo-randomly displayed on a viewing screen and in Task-B, numbers 1–7 and letters A–G are displayed. Subjects used an MR-compatible tablet to draw a line connecting items in sequence (1-2-3-4-) or (1-A-2-B-), connecting as many as possible during a 20s interval, while maintaining accuracy (Tam et al. 2011). A Control stimulus was presented after each block, in which participants traced a line from the center of the screen to a dot (randomly placed at a fixed radius from the center of the screen) repeated 10 times. Subjects performed two repetitions of Task-A and Task-B, interleaved with a set of four Control blocks.

We used a 3 Tesla *f*MRI scanner to acquire axial, interleaved, multi-slice echo planar images of the whole brain during task performance (3.1 × 3.1 × 5 mm

voxels, TE/TR = 30/2000 ms). The resulting 4D *f*MRI time series were preprocessed using standard tools from the AFNI package, including rigid-body correction of head motion (`3dvolreg`), physiological noise correction with RETROICOR RETROICOR (`3dretroicor`), temporal detrending using Legendre polynomials and regressing out estimated rigid-body motion parameters (`3dDetrend`)

For each subject, we computed the average of each repetition of Task-A and Task-B, as well as the averages of Control blocks $1+2$ and $3+4$. This created a set of $P \times 6$ matrices with six different conditions, showing "early" and "late" responses to the three different task conditions. We displayed empirical resampling and null distributions for a single PLSC component, under the BV, SS and SR models.

7.3 Results

7.3.1 Simulated Results

Figure 7.1 depicts the median p-values for brain components r_{brain}, along with BV for comparison. Darker colours indicate lower p-values and thus greater sensitivity. All models show comparable performance as a function of signal variance V. However, the BV and SS models have lower sensitivity compared to SR, for fewer subjects S and conditions C, and a greater percentage of null data P_{null}. Figure 7.2 displays the median p-values for task components r_{task}, and BV for comparison. The empirical p-values for SS are considerably higher for "task" compared to "brain," whereas they are comparable for SR. Moreover, SS estimates are less sensitive than SR for a range of simulation parameters values. Figure 7.3 depicts sample empirical

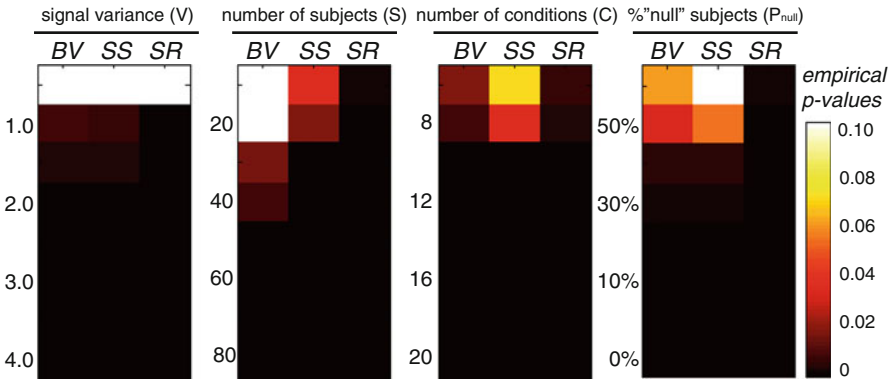

Fig. 7.1 Heat maps showing empirical p-values for the high-dimensional "brain" PLSC component of simulated data, for a range of simulated parameter settings, under split-half stability (SS) and split-half reproducibility (SR) models. We show bootstrap variance (BV) p-values of eigenvalue λ_k^* for comparison purposes. For each parameter setting, median p-value is computed across 100 simulated data-sets

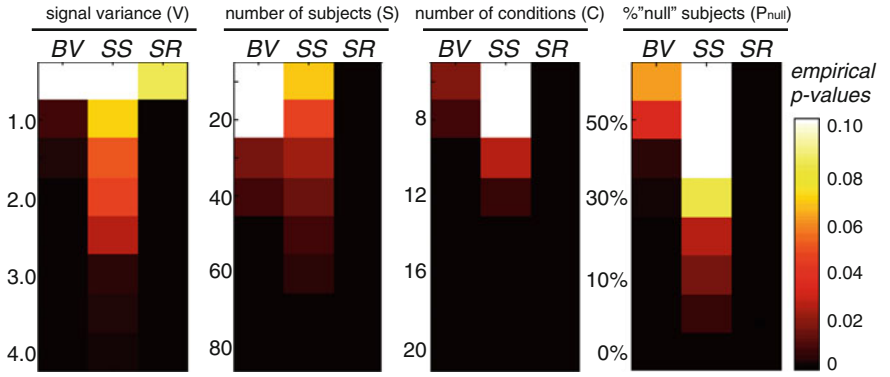

Fig. 7.2 Heat maps showing empirical p-values for the low-dimensional "task" PLSC component of simulated data, for a range of simulated parameter settings, under split-half stability (SS) and split-half reproducibility (SR) models. We show bootstrap variance (BV) p-values of eigenvalue λ_k^* for comparison purposes. For each parameter setting, median p-value is computed across 100 simulated data-sets

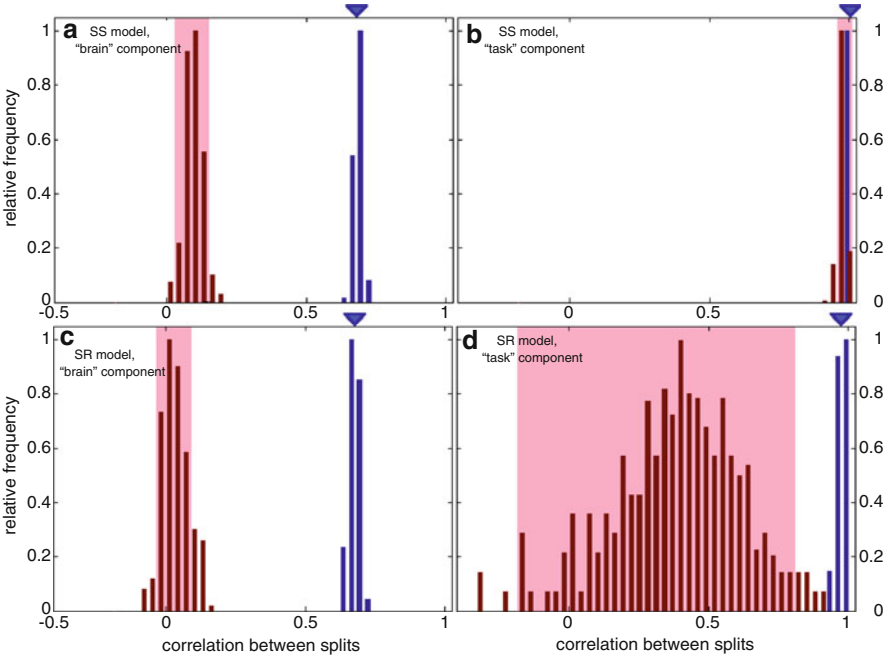

Fig. 7.3 Empirical histograms showing distributions of between-split correlations r_{brain} and r_{task} for a single simulated dataset with settings $V = 1.0, S = 50, C = 10, P_{null} = 0\,\%$. We show the distribution over 1000 splits (*blue*; sampling distribution) and the sampling distribution for permuted labels over 1000 splits (*red*; null distribution). A *blue arrow* denotes the median of the sampling distribution (i.e., our test statistic which is compared against the null). The *pink band* denotes the 95 % CI on the null. We show the distributions for "brain" and "task" components, for the split-half stability model (**a**–**b**), and the split-half reproducibility model (**c**–**d**)

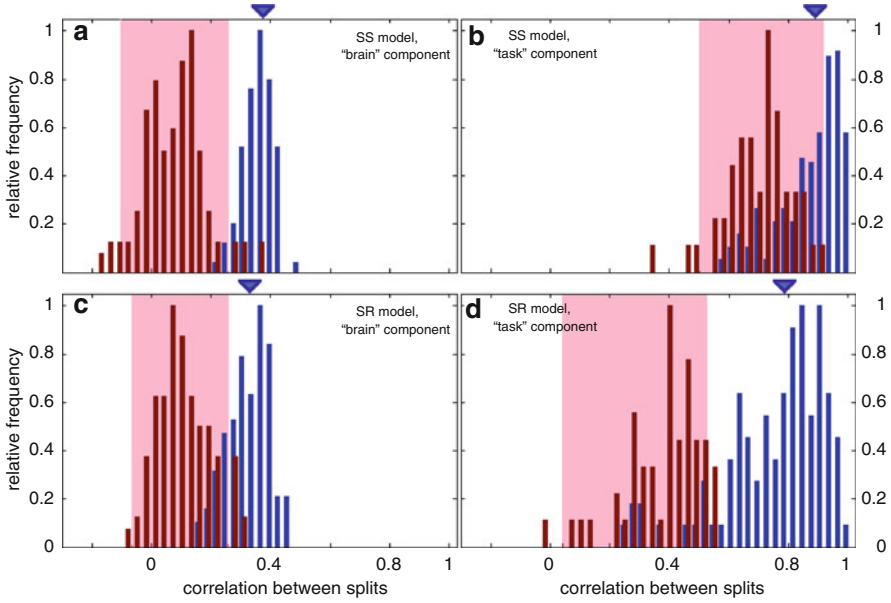

Fig. 7.4 Empirical histograms showing distributions of between-split correlations r_{brain} and r_{task} for the experimental data-set. We show the distribution over 1000 splits (*blue*; sampling distribution) and the sampling distribution for permuted labels over 1000 splits (*red*; null distribution). A *blue arrow* denotes the median of the sampling distribution (i.e. our test statistic which is compared against the null). The *pink band* denotes the 95 % CI on the null. We show the distributions for "brain" and "task" components, for the split-half stability (**a**–**b**), and the split-half reproducibility (**c**–**d**) models

distributions for SS and SR models. The SS model has upward bias in both sampling and null distributions. But for the "task" component, the bias is sufficiently high that we cannot distinguish null from alternative, as the upper bound on the null distribution's 95 % confidence interval (pink shading) is $r_{task} > 0.99$.

7.3.2 Experimental Results

Figure 7.4 depicts the empirical sampling distributions for "brain" and "task" components of the experimental data-set, for SS and SR models. As with the synthetic data, the high-dimensional "brain" distributions are comparable for SS and SR models. Whereas the low-dimensional "task" distribution is much more biased for SS, as the test statistic is not significantly different from the null.

7.4 Discussion and Conclusions

In this paper, we evaluated three different approaches to computing significance of PLSC components in functional neuroimaging data. These approaches included bootstrapped variance (BV), split-half stability (SS), and split-half reproducibility (SR). In general, the SS and SR models had comparable sensitivity to high-dimensional "brain" components, but SR was consistently more sensitive to low-dimensional "task" components, particularly in weaker-signal simulations. Unexpectedly, the BV model outperformed SS in some cases, including a low number of task conditions C and high percentage of null subjects P_{null}, whereas SS performed better for low sample sizes S. In addition, all of the resampling models showed relatively high tolerance to signal heterogeneity, as median p-values were below .10 for up to 30 % of the null data. This points towards the robustness of PLSC as an analytic tool.

From a detection standpoint, our results show that it is generally more important to ensure independence between split-halves (as in SR)—because doing so minimizes estimation bias—rather than to maximize sample power (as in SS). As a potential alternative, PLSC split-half resampling could be performed as part of a multi-level PCA decomposition, as implemented in Strother et al. (2002) for multivariate classification. This would allow for improved split-half stability with relatively small bias across splits.

We focused on estimating significance, based on the reliability of parameter estimates; in this case, the correlation of saliences between splits. This is appropriate, since we are using PLSC to estimate shared information between brain and task condition. A number of alternative metrics have been developed to assess goodness of fit under the predictive Regression PLS model, such as predicted sum of squares (PRESS) (Abdi 2010; Abdi and Williams 2012). In future work, it will be important to also compare different test statistics, to see which provide optimal sensitivity to covariance structure.

References

Abdi, H.: Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip. Rev.: Comput. Stat. **2**, 97–106 (2010)

Abdi, H., Williams, L.: Partial least square methods: partial least square correlation and partial least square regression. In: Reisfeld, B., Mayeno, A. (eds.) Methods in Molecular Biology: Computational Toxicology, pp. 549–579. Springer, New York (2012)

Abdi, H., Chin, W., Esposito Vinzi, V., Russolilo, G., Trinchera, L.: New Perspectives in PLS and Related Methods. Springer, New York (2013)

Churchill, N.W., Spring, R., Kovacevic, N., McIntosh, A.R., Strother, S.C.: The stability of behavioral PLS results in ill-posed neuroimaging problems. In: Abdi, H., Chin, W., Esposito Vinzi, V., Russolilo, G., Trinchera, L. (eds.) New Perspectives in PLS and Related Methods, pp. 171–183. Springer, New York (2013)

Kovacevic, N., et al.: Revisiting PLS resampling: comparing significance vs. reliability across range of simulations. In: Abdi, H., Chin, W., Esposito Vinzi, V., Russolilo, G., Trinchera, L. (eds.) New Perspectives in PLS and Related Methods, pp. 159–170. Springer, New York (2013)

Krishnan, A., Williams, L.J., McIntosh, A.R., Abdi, H.: Partial least squares (PLS) methods for neuroimaging: a tutorial and review. NeuroImage **56**, 455–475 (2011)

Lukic, A.S., Wernick, M.N., Strother, S.C.: An evaluation of methods for detecting brain activations from functional neuroimages. Artif Intell. Med. **25**, 69–88 (2002)

McIntosh, A.R., et al.: Spatial pattern analysis of functional brain images using partial least squares. NeuroImage **3**, 143–157 (1996)

Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. Subspace, latent structure and feature selection. In: Saunders, C. (ed.) Proceedings SLSFS, LNCS. Springer, Berlin (2006)

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D.: The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage **15**, 747–771 (2002)

Tam, F., Churchill, N.W., Strother, S.C., Graham, S.J.: A new tablet for writing and drawing during functional MRI. Hum. Brain Mapp. **32**, 240–248 (2011)

Wold, H.: Partial Least Squares. Encyclopedia of Statistical Sciences, vol. 6, pp. 581–591. Wiley, New York (1985)

Yourganov, G., et al.: Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data. NeuroImage **56**, 531–543 (2011)

Chapter 8
Estimating and Correcting Optimism Bias in Multivariate PLS Regression: Application to the Study of the Association Between Single Nucleotide Polymorphisms and Multivariate Traits in Attention Deficit Hyperactivity Disorder

Erica Cunningham, Antonio Ciampi, Ridha Joober, and Aurélie Labbe

Abstract In studies involving genetic data, the correlations between **X** and **Y** scores obtained from PLS regression models can be used as measures of association between genome-level measurements, **X**, and phenotype-level measurements, **Y**. These correlations may be overestimated due to potential overfitting (i.e., they may be vulnerable to optimism bias). We evaluate the optimism bias through simulations and examine the effect of increasing sample size and strength of correlation. We assess the effectiveness of bootstrap-based and permutation-based bias correction methods. We also investigate the selection of the appropriate number of components for PLS regression. We include an analysis of genetic data consisting of genotypes and phenotypes related to Attention Deficit Hyperactivity Disorder (ADHD).

E. Cunningham (✉)
Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada
e-mail: erica.cunningham@mail.mcgill.ca

A. Ciampi
Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada
e-mail: antonio.ciampi@mcgill.ca

R. Joober
Douglas Mental Health University Institute, Verdun, QC, Canada
e-mail: ridha.joober@douglas.mcgill.ca

A. Labbe
Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada
e-mail: aurelie.labbe@mcgill.ca

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_8

Keywords Partial least square regression (PLSR) • Optimism bias • Overfitting • SNPs • Bootstrap

8.1 Introduction

Partial least square (PLS) regression has become a useful and increasingly popular tool in genomics. Given two sets of measurements $\mathbf{Y}$ and $\mathbf{X}$, multivariate PLS determines two sets of orthogonal linear combinations of variables (called latent variables) $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots)$ and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots)$, such that the covariances between the latent variables $\text{cov}(\mathbf{t}_k, \mathbf{u}_k)$ are maximized (Abdi 2010). In genomic studies, the correlations between the latent variables (denoted ϱ_k's) are often used as measures of association between genome-level measurements, $\mathbf{X}$, and phenotype-level measurements, $\mathbf{Y}$. Because these correlations are parameters of central scientific interest, it is important to study the statistical properties of their PLS estimates because they are likely to be vulnerable to optimism bias (i.e., they may overestimate their actual values due to potential overfitting).

Our primary objective is to evaluate the optimism bias in the correlations $\varrho_k = \text{cor}(\mathbf{t}_k, \mathbf{u}_k)$ through simulations. We examine the effect of increasing sample size and the strength of the correlation, and we assess the effectiveness of bootstrap-based and permutation-based bias correction methods. We also investigate the choice of the number of components for PLS regression. The simulations are inspired by a real data analysis problem: the association between Single Nucleotide Polymorphisms (SNPs, genotype), and a number of behavioral and cognitive measurements (phenotype). We finish with a data analysis of genotypes at SNPs in two genes known to be associated with ADHD, NET/SLC6A2 (Kim et al. 2006) and TPH2 (Sheehan et al. 2005), and phenotypes consisting of behavioral and cognitive measurements related to ADHD.

8.2 Methods

The following describes the simulations and the real data analysis. All PLS models were fitted using the `pls` package in R (Mevik and Wehrens 2007).

8.2.1 Simulations to Evaluate the Bias

We consider a set of 16 traits which can be divided into two known components and 39 SNPs which are known to belong to two genes. As the basis for the simulations, we use the ADHD example where SNPs in one gene are associated with behavior and where SNPs in another gene are associated with cognition. A linear

combination of 30 SNP variables defines a genetic score $\mathbf{t}_1$ and a linear combination of 7 behavioral traits defines a behavior phenotype score $\mathbf{u}_1$; similarly, a linear combination of 9 (distinct) SNP variables defines another genetic score $\mathbf{t}_2$ and a linear combination of 9 cognitive traits defines a cognitive phenotype score $\mathbf{u}_2$. The coefficients of these linear relationships are called loadings.

The relationship between genotype and phenotype is embodied in ϱ_1 and ϱ_2, the correlations between $\mathbf{t}_1$ and $\mathbf{u}_1$, and $\mathbf{t}_2$ and $\mathbf{u}_2$, respectively. We have implemented several biologically motivated scenarios by varying the values of the correlations and of the loadings, as illustrated in Fig. 8.1. Each scenario was further divided into four sub-scenarios which vary the amount of information provided by the SNPs or traits:

1. All SNPs and all traits are informative (have a non-zero loading coefficient)
2. All SNPs and half of the traits from each category are informative

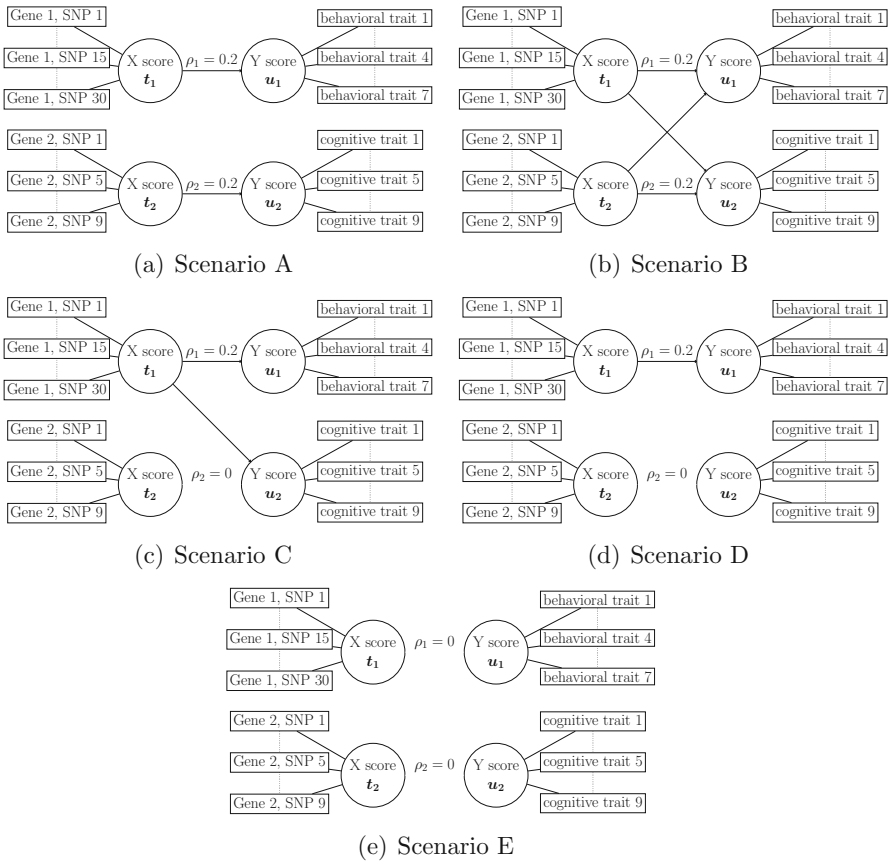

Fig. 8.1 Five main simulation scenarios. (**a**) Gene 1 is associated with behavioral traits and Gene 2 is associated with cognitive traits. (**b**) Genes 1 and 2 are associated with behavioral and cognitive traits. (**c**) Gene 1 is associated with behavioral and cognitive traits. (**d**) Gene 1 is associated with behavioral traits. (**e**) Neither gene is associated with behavioral and cognitive traits

3. Half of the SNPs from each gene and all traits are informative
4. Half of the SNPs from each gene and half of the traits from each category are informative.

The **X** and **Y** scores $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2)$ and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$ were generated as a sample of size 300 from a multivariate normal distribution. The variance–covariance matrix was constructed using the true correlations $\varrho_1 = \text{cor}(\mathbf{t}_1, \mathbf{u}_1)$ and $\varrho_2 = \text{cor}(\mathbf{t}_2, \mathbf{u}_2)$ as specified by each scenario. The **X** and **Y** matrices were then constructed as

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^\top + \mathbf{G}, \end{aligned} \tag{8.1}$$

where **P** and **Q** are matrices containing **X** and **Y** loadings, respectively (set to be 0.5 or 0 based on the scenario), and the columns of **E** and **F** are $\mathcal{N}(0, \sigma^2)$. The simulations were repeated for three values of σ: 0.1, 1.0, and 1.5.

Because the **X** variables represent SNP genotypes—which are often coded as counts of minor alleles taking values 0, 1, or 2—the simulated **X** variables were then converted to discrete variables. This conversion was performed by first obtaining the genotype frequencies at each SNP under study in the real data. These genotype frequencies were used as percentiles from the standard normal distribution to discretize **X**.

We applied multivariate PLS regression to each sample, and obtained the estimates of the correlations ϱ_1 and ϱ_2 (denoted, respectively $\hat{\varrho}_1$ and $\hat{\varrho}_2$). Some investigation revealed that the computed $\hat{\varrho}_i = \text{cor}(\hat{\mathbf{t}}_i, \hat{\mathbf{u}}_i)$ seemed to be constrained in some way to always be positive. We introduced an algorithm that allows for the possibility of a negative correlation and changes the sign of the correlation when appropriate. The algorithm is as follows: if $\text{cor}(\mathbf{t}_i, \hat{\mathbf{t}}_i)$ and $\text{cor}(\mathbf{u}_i, \hat{\mathbf{u}}_i)$ have different signs, then the sign of $\hat{\varrho}_i$ is switched, where $\mathbf{t}_i$ and $\mathbf{u}_i$ are the true scores used to generate the simulated data, $\hat{\mathbf{t}}_i$ and $\hat{\mathbf{u}}_i$ are the PLS estimates of these scores, and $i = 1, 2$ is the component number. After applying the sign change algorithm, the differences between $\hat{\varrho}_i$ and the true values were obtained and the bias was calculated by averaging these differences over 500 replications.

After completing the 20 simulation scenarios (5 main scenarios with 4 sub-scenarios each), further simulations were performed by increasing the number of subjects in each sample from 300 to 1000 and the strength of the correlation between scores from 0.2 to 0.5. These additional simulations were run for Scenario A, where one gene is associated with behavioral traits and the second gene is associated with cognitive traits, with all SNPs and all traits informative.

8.2.2 Simulations to Correct the Bias

In an attempt to correct for the optimism bias, simulations incorporating bootstrap-based and permutation-based bias correction methods were conducted based on a scenario that most closely represented the real dataset. In this scenario, SNPs in one

gene are associated with behavioral traits and SNPs in the other gene are associated with cognitive traits. We considered two situations: one where the true correlations ϱ_1 and ϱ_2 were both 0.2 and a second where both correlations were 0.

To further reflect the real dataset, the $\mathbf{P}$ matrix of $\mathbf{X}$ loadings was set up to represent a situation in which all SNPs in one gene contributed to the $\mathbf{X}$ score $\mathbf{t}_1$ and half of the SNPs in the other gene contributed to the $\mathbf{X}$ score $\mathbf{t}_2$. Similarly, the $\mathbf{Q}$ matrix of $\mathbf{Y}$ loadings was constructed to reflect a situation in which all behavioral traits contributed to the $\mathbf{Y}$ score $\mathbf{u}_1$ and half of the cognitive traits contributed to the $\mathbf{Y}$ score $\mathbf{u}_2$. The $\mathbf{X}$ and $\mathbf{Y}$ variables were simulated as described above using $\sigma = 1$ with 200 samples and multivariate PLS models were fitted to each sample.

8.2.2.1 Bootstrap-Based methods

We bootstrapped the residuals (Efron and Tibshirani 1993) from the fitted PLS models to obtain $\hat{\boldsymbol{\varepsilon}}_{\text{boot}}$ and applied PLS regression to $\mathbf{X}$ and $\mathbf{Y}_{\text{boot}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}_{\text{boot}}$, where $\hat{\boldsymbol{\beta}}$ is the matrix of coefficient estimates from the PLS model fitted on $\mathbf{X}$ and $\mathbf{Y}$. The correlation estimates for $\text{cor}(\mathbf{t}_1, \mathbf{u}_1)$ and $\text{cor}(\mathbf{t}_2, \mathbf{u}_2)$ were obtained, and denoted (respectively) $\hat{\varrho}_{1,\text{boot}}$ and $\hat{\varrho}_{2,\text{boot}}$. A similar approach for the sign change was applied here: Let $\hat{\mathbf{t}}_{\text{i,boot}}$ and $\hat{\mathbf{u}}_{\text{i,boot}}$ be the estimated scores obtained from a PLS model fitted on $\mathbf{X}$ and $\mathbf{Y}_{\text{boot}}$, where $i = 1, 2$ is the component number. If $\text{cor}(\hat{\mathbf{t}}_{\text{i,boot}}, \hat{\mathbf{t}}_{\text{i}})$ and $\text{cor}(\hat{\mathbf{u}}_{\text{i,boot}}, \hat{\mathbf{u}}_{\text{i}})$ had different signs, then the sign of $\hat{\varrho}_{\text{i,boot}}$ was changed.

The bootstrapped correlation estimates were averaged over the 200 bootstrap samples to get $\hat{\varrho}_{1,\text{boot}}$ and $\bar{\hat{\varrho}}_{2,\text{boot}}$ and the bias estimates for each repetition were calculated as:

$$\widehat{\text{Bias}}_{\text{a,boot}} = \bar{\hat{\varrho}}_{\text{a,boot}} - \hat{\varrho}_{\text{a,boot}}, \tag{8.2}$$

where $\text{a} = 1, 2$ is the component index and $\hat{\varrho}_{\text{a,boot}}$ is the correlation estimate obtained from the PLS model for $\mathbf{X}$ and $\mathbf{Y}$ simulated in that repetition. The bias estimates were used to get bias-corrected estimates of the correlations for each repetition:

$$\hat{\varrho}_{\text{C,a,boot}} = \hat{\varrho}_{\text{a,boot}} - \widehat{\text{Bias}}_{\text{a,boot}} \tag{8.3}$$

for $\text{a} = 1, 2$. We then recalculated the bias by subtracting the true correlations ϱ_1 and ϱ_2 from the bias-corrected estimates $\hat{\varrho}_{\text{C,1,boot}}$ and $\hat{\varrho}_{\text{C,2,boot}}$ and averaging them over the 200 repetitions.

8.2.2.2 Permutation-Based Methods

We permuted the observations of the $\mathbf{Y}$ variables to obtain $\mathbf{Y}_{\text{perm}}$ and applied PLS regression to $\mathbf{X}$ and $\mathbf{Y}_{\text{perm}}$. The correlation estimates were computed and denoted $\hat{\varrho}_{1,\text{perm}}$ and $\hat{\varrho}_{2,\text{perm}}$, respectively. We applied the sign change algorithm to the

estimates and averaged over 200 samples to get an estimate of the bias, $\widehat{\text{Bias}}_{\text{a,perm}}$. Bias-corrected correlation estimates were then obtained as:

$$\hat{\varrho}_{\text{C,a,perm}} = \hat{\varrho}_{\text{a,perm}} - \widehat{\text{Bias}}_{\text{a,perm}}, \tag{8.4}$$

where a $= 1, 2$.

8.2.3 Selecting the Number of Components

To examine the choice of the number of components used in fitting a PLS model, we simulated data as before, based on the model with one gene associated with behavioral traits and the other gene associated with cognitive traits, such that the true number of components was 2. We investigated three scenarios: (1) $N = 300$, $\varrho_1 = \varrho_2 = 0.2$, (2) $N = 1000$, $\varrho_1 = \varrho_2 = 0.2$, and (3) $N = 300$, $\varrho_1 = \varrho_2 = 0.5$.

For univariate PLS models, one approach to selecting the appropriate number of components K is to choose K as the number of components for which the minimum cross-validation mean square error of prediction (MSEP) is achieved (Denham 2000). To extend this to multivariate PLS models, we may obtain the number of components at which the minimum MSEP is attained for each $\mathbf{Y}$ variable, and choose K as the maximum of these numbers.

We fit PLS models to the simulated data using leave-one-out cross-validation and recorded the number of components chosen using this modified minimum rule for 1000 repetitions of each of the 3 scenarios described above. We restricted the number of components to be between 0 and 10 inclusive.

8.2.4 Real Data Analysis

The real dataset used is a subset of a larger dataset from a study of genetic and environmental risk factors for ADHD in children. The dataset is comprised of genotype and phenotype information on 323 children from different families. The genetic data consists of genotypes at 39 SNPs in two genes known to be associated with ADHD, NET/SLC6A2 and TPH2.

The phenotypes of interest are 16 scores from behavioral and cognitive testing. Prior to the PLS regression analysis, the phenotype variables were adjusted for the covariates age, sex, ADHD clinical subtype, and maternal smoking during pregnancy by fitting separate multiple linear regression models for each phenotype variable as the outcome and the four covariates as the predictors. The residuals from these models were used as the adjusted phenotypes for the PLS analysis.

A PLS regression model was fitted using 30 SNPs in the NET/SLC6A2 gene and 9 SNPs in the TPH2 gene as the $\mathbf{X}$ variables and the 7 behavioral traits and 9 cognitive traits as the $\mathbf{Y}$ variables. The model was fitted using leave-one-out cross-validation and 10 components. We applied the modified minimum rule to choose the appropriate number of components.

8.3 Results

The following presents the results obtained for the bias evaluation simulations, bias correction simulations, selection of the number of components, and the real data analysis.

8.3.1 Evaluating the Optimism Bias

Figure 8.2 presents the results of the simulations to evaluate the bias. In each plot, the bias is shown for $\varrho_1 = \text{cor}(\mathbf{t}_1, \mathbf{u}_1)$ in black and $\varrho_2 = \text{cor}(\mathbf{t}_2, \mathbf{u}_2)$ in gray for 3 different values of σ. Along the x axis are the four sub-scenarios based on how many SNPs in each gene and traits in each category are informative. The points indicating the amount of bias for each of the four sub-scenarios are joined with a line for clarity.

Figure 8.3 presents the bias results for the correlations ρ_1 (solid lines) and ρ_2 (dashed lines) when running simulations with an increased sample size or a larger true correlation. The black lines indicate the bias with $\rho_1 = \rho_2 = 0.2$ and a sample size of 300. The dark grey lines represent the situation where the sample size was increased to 1000 and the light grey lines represent the situation where the correlations were increased to 0.5. Again, the points for the four sub-scenarios are joined with lines for visual interpretability.

8.3.2 Correcting the Optimism Bias

Table 8.1 presents the results of incorporating the bootstrap-based and permutation-based bias correction methods into the simulations. The results are shown for the two cases considered: the first with $\varrho_1 = \varrho_2 = 0.2$ and the second with $\varrho_1 = \varrho_2 = 0$.

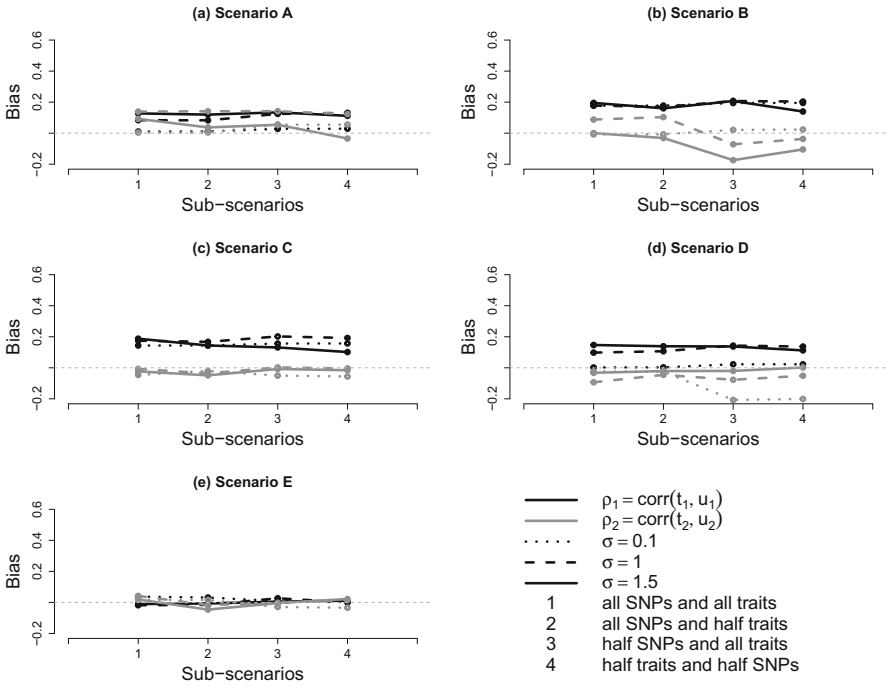

Fig. 8.2 Bias in the correlations. (**a**) Gene 1 is associated with behavioural traits and Gene 2 is associated with cognitive traits. (**b**) Genes 1 and 2 are associated with behavioral and cognitive traits. (**c**) Gene 1 is associated with behavioral and cognitive traits. (**d**) Gene 1 is associated with behavioral traits. (**e**) Neither gene is associated with behavioral and cognitive traits

8.3.3 Choosing the Number of Components

Figure 8.4 presents the results of the selection of the number of components for the three scenarios considered. The histograms show the number of iterations out of 1000 for which each number of components was chosen, for three values of σ.

8.3.4 Real Data Analysis

In Table 8.2, the number of components selected for each of the 16 $\mathbf{Y}$ variables is shown, chosen as the number of components at which the MSEP is at its minimum. Applying the modified minimum rule to the real data results in a choice of two components.

Assuming that two components is the appropriate choice, we obtain the correlations between the estimated scores $\mathbf{t}_1$ and $\mathbf{u}_1$, and $\mathbf{t}_2$ and $\mathbf{u}_2$, shown in Table 8.3.

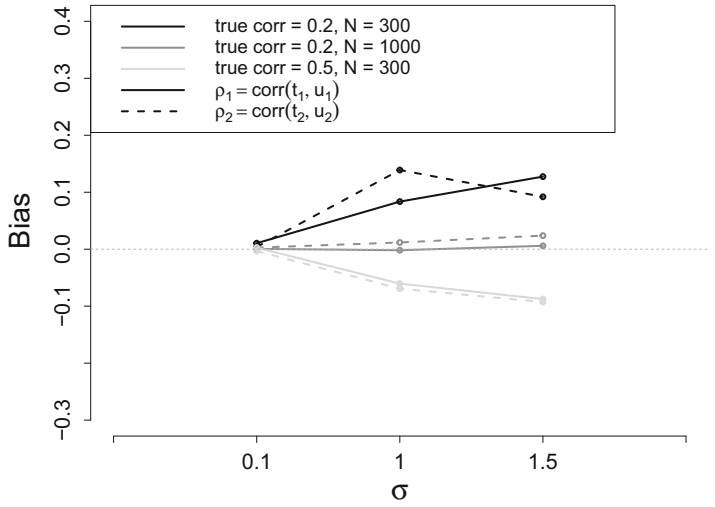

Fig. 8.3 Bias in the correlations when sample size or strength of the correlation increases

Table 8.1 Bias in correlations between **X** scores ($\mathbf{t}_1$ and $\mathbf{t}_2$) and **Y** scores ($\mathbf{u}_1$ and $\mathbf{u}_2$) in simulations

Simulated correlation	Estimated bias	Bootstrap-corrected bias	Permutation-corrected bias
$\varrho_1 = \text{cor}(\mathbf{t}_1, \mathbf{u}_1) = 0.2$	0.100	0.096	0.014
$\varrho_2 = \text{cor}(\mathbf{t}_2, \mathbf{u}_2) = 0.2$	0.005	0.100	0.013
$\varrho_1 = \text{cor}(\mathbf{t}_1, \mathbf{u}_1) = 0$	−0.122	0.454	0.363
$\varrho_2 = \text{cor}(\mathbf{t}_2, \mathbf{u}_2) = 0$	0.046	0.401	0.305

8.4 Discussion

It can be seen from Fig. 8.2 that bias is indeed present in the correlation estimates and is sometimes substantial. Consider, for example, scenario A, where we see a bias of approximately 0.2 when the true correlations are $\varrho_1 = \varrho_2 = 0.2$. This result indicates that the PLS correlation estimates are on average twice as large as the true correlation values, and consequently we would not be able to make any valid inference based on the correlation estimates we obtain. With this magnitude of bias, there is potential for false positives, as the **X** and **Y** variables will appear to be more highly correlated than they actually are. Only in scenario E, where neither gene is associated with the traits, does there appear to be minimal bias.

Considering Fig. 8.3, we see that with a larger sample size the bias decreases to almost 0. This result suggests that PLS regression estimates the correlation well when the sample size is large. With $N = 1000$, the amount of bias does not change much as σ changes. When the strength of the correlation increases to 0.5, the correlation is now slightly underestimated by PLS.

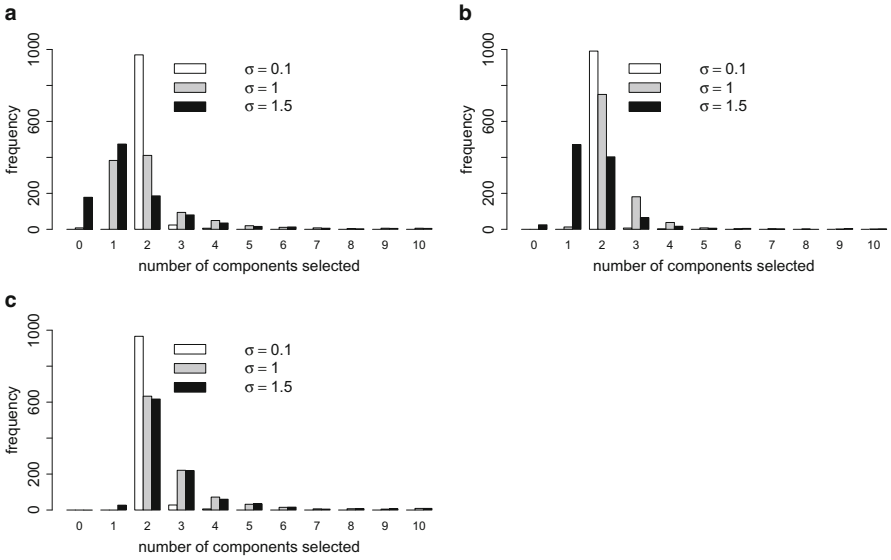

Fig. 8.4 Frequencies of components chosen by the minimum rule over 1,000 iterations. (**a**) $N = 300$, $\varrho_1 = \varrho_2 = 0.2$. (**b**) $N = 1,000$, $\varrho_1 = \varrho_2 = 0.2$. (**c**) $N = 300$, $\varrho_1 = \varrho_2 = 0.5$

Table 8.2 Number of components selected for each **Y** variable

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0

Table 8.3 Correlation estimates from the real data

Correlation of interest	Estimate
$\varrho_1 = \text{cor}(\mathbf{t}_1, \mathbf{u}_1)$ (NET/SLC6A2 and behavioral traits)	0.199
$\varrho_2 = \text{cor}(\mathbf{t}_2, \mathbf{u}_2)$ (TPH2 and cognitive traits)	0.214

The bias correction results presented in Table 8.1 indicate that the permutation-based method is more successful than the bootstrap-based approach. However, in some cases neither method works to reduce the bias and instead the bias is increased. The explanation for this inconsistency requires further investigation.

From Fig. 8.4, the modified minimum rule chooses the correct number of components in most cases when σ is small. With higher values of σ, the incorrect number of components is chosen much more frequently. By increasing ϱ, the correct number of components is selected more often for $\sigma = 1$ and $\sigma = 1.5$ than with lower ϱ. With an increase in N, the correct number of components is chosen often for all three values of σ, although 1 component is still frequently chosen for $\sigma = 1.5$, and models with more than 4 components are even more rare than in the previous two scenarios.

The results of the bias evaluation suggest that when using PLS for a real genetic association study, the correlation estimates may indicate that the associations between the genotypes and phenotypes are stronger than they actually are. Consequently, researchers should be cautious of the correlation estimates they obtain from an analysis using PLS and be wary of the potential for false positives. There is some indication that with a large enough sample size, the bias is minimal and the appropriate number of components can be chosen fairly reliably using the modified minimum rule shown here.

References

Abdi, H.: Partial least squares regression and projection on latent structure regression (PLS regression). Wiley Interdiscip. Rev.: Comput. Stat. **2**, 97–106 (2010)

Denham, M.C.: Choosing the number of factors in partial least squares regression: estimating and minimizing the mean square error of prediction. J. Chemom. **14**, 351–361 (2000)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, Boca Raton (1993)

Kim, C.-H., Hahn, M.K., Joung, Y., Anderson, S.L., ... Kim, K.-S.: A polymorphism in the norepinephrine transporter gene alters promoter activity and is associated with attention-deficit hyperactivity disorder. Proc. Natl. Acad. Sci. U.S.A. **103**, 19164–19169 (2006)

Mevik, B.H., Wehrens, R.: The `pls` package: principal component and partial least squares regression in `R`. J. Stat. Softw. **18**(2), 1–24 (2007)

Sheehan, K., Lowe, N., Kirley, A., Mullins, C., Fitzgerald, M., Gill, M., Hawi, Z.: Tryptophan hydroxylase 2 (TPH2) gene variants associated with ADHD. Mol. Psychiatry **10**, 944–949 (2005)

Chapter 9
Discriminant Analysis for Multiway Data

Gisela Lechuga, Laurent Le Brusquet, Vincent Perlbarg, Louis Puybasset, Damien Galanaud, and Arthur Tenenhaus

Abstract A multiway Fisher Discriminant Analysis (MFDA) formulation is presented in this paper. The core of MFDA relies on the structural constraint imposed to the discriminant vectors in order to account for the multiway structure of the data. This results in a more parsimonious model than that of Fisher Discriminant Analysis (FDA) performed on the unfolded data table. Moreover, computational and overfitting issues that occur with high dimensional data are better controlled. MFDA is applied to predict the long term recovery of patients after traumatic brain injury from multi-modal brain Magnetic Resonance Imaging. As compared to FDA, MFDA clearly tracks down the discrimination areas within the white matter region of the brain and provides a ranking of the contribution of the neuroimaging modalities. Based on cross validation, the accuracy of MFDA is equal to 77 % against 75 % for FDA.

G. Lechuga (✉) • L. Brusquet
Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506),
CentraleSupélec-CNRS-Université Paris-Sud, Paris, France
e-mail: gisela.lechuga@centralesupelec.fr; laurent.lebrusquet@centralesupelec.fr

V. Perlbarg
Bioinformatics/Biostatistics Platform IHU-A-ICM, Brain and Spine Institute,
47-83, Bd de l'Hôpital, Paris, France
e-mail: vperlbar@imed.jussieu.fr

L. Puybasset
AP-HP, Pitié-Salpêtrière Hospital, Department of Neuroradiology, Paris, France
e-mail: louis.puybasset@psl.aphp.fr

D. Galanaud
Department of Neuroradiology, AP-HP, Pitié-Salpêtrière Hospital,
Surgical Neuro-Intensive Care Unit, Paris, France
e-mail: galanaud@gmail.com

A. Tenenhaus
Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506),
CentraleSupélec-CNRS-Université Paris-Sud and Bioinformatics/Biostatistics Platform
IHU-A-ICM, Brain and Spine Institute, Paris, France
e-mail: arthur.tenenhaus@centralesupelec.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_9

Keywords Discriminant analysis • Multiway fisher discriminant analysis (MFDA) • Overfitting • Brain imaging

9.1 Introduction

In standard multivariate data analysis, an individuals × variables data table is usually considered. However, from a practical viewpoint this simple data structure appears to be somehow restricted. An example is found in multi-modal brain Magnetic Resonance Imaging (MRI) where K neuroimaging modalities (each characterized by J voxels), are collected on a set of I patients. In that context, an individuals × voxels × modalities data table can be considered and yields a three-way dataset (or tensor). A three-way dataset can be considered in terms of a stack of matrices as illustrated in Fig. 9.1. Most data analysis methods in their primary definition do not take into account this natural three-way structure. Indeed, such structure is lost by considering a $I \times JK$ unfolded version leading potentially (i) to a procedure that destroys the integrity of the structure of the data and (ii) to a very large parameter vector to estimate. These two aspects can yield a lack of relevant interpretations of the resulting model and additional structural constraints are required.

Many two-way data analysis methods have been extended to the multiway configuration. For instance PARAFAC proposed by Harshman (1970) is a generalization of Principal Component Analysis. PARAFAC relies on the maximization of a variance criterion but explicitly takes into account the multiway structure of the input data by imposing a special Kronecker structure on the weight vectors. A second approach is the Multi-linear Partial Least Squares Regression (N-PLS) proposed by Bro (1998) which is a generalization of the classic PLS regression method to multiway data. N-PLS relies on the maximization of a covariance criterion but has the same PARAFAC structural constraint on the weight vectors. N-PLS relies on SVD decomposition and is particularly well suited to the high dimensional setting. In this paper, a multiway formulation of Fisher Discriminant Analysis (MFDA) is presented. The structural constraint that is imposed to N-PLS and PARAFAC weight vectors constitutes the starting point of MFDA.

This paper is organized as follows: Fisher Discriminant Analysis (FDA) and its multiway counterpart (MFDA) are presented in Sect. 9.2. In Sect. 9.3, MFDA is

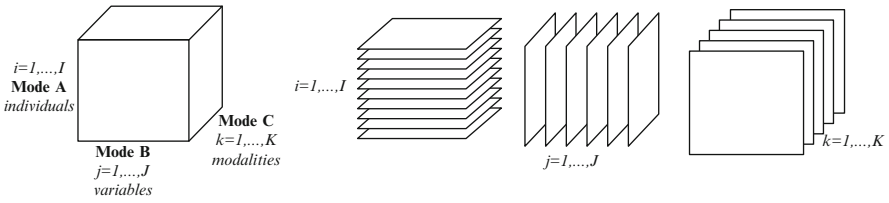

Fig. 9.1 Structure of the three-way dataset

illustrated on a multi-modal Magnetic Resonance Brain Imaging (MRI) dataset in order to predict the long-term recovery, of patients that suffered from traumatic brain injury. A comparison between MFDA and FDA is discussed in Sect. 9.4.

9.2 Multiway Fisher Discriminant Analysis

Let $\mathbf{X}$ be the individuals $\times$ variables $\times$ modalities tensor and $\mathbf{X}^{\mathrm{u}}$ the associated unfolded matrix where all the $I \times J$ two-way matrices are collected next to each other in an $I \times JK$ matrix. In addition, let $\mathbf{y}$ be the qualitative variable that encodes the class membership of each individual. Let $\mathbf{Y}$ be the matrix of dummy variables indicating the group memberships.

9.2.1 Regularized Fisher Discriminant Analysis

FDA consists in finding a projection vector $\mathbf{w}$ such that the between class variance is maximized relative to the within-class variance. Regularized FDA is defined by the optimization problem:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_T \mathbf{w} + \lambda \mathbf{w}^\top \mathbf{R} \mathbf{w}}, \tag{9.1}$$

where $\mathbf{S}_B = (\mathbf{X}^{\mathrm{u}})^\top \mathbf{Y}(\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X}^{\mathrm{u}} = (\mathbf{X}^{\mathrm{u}})^\top \mathbf{H}_B \mathbf{X}^{\mathrm{u}}$ is the between covariance matrix and $\mathbf{S}_T = (\mathbf{X}^{\mathrm{u}})^\top \mathbf{X}^{\mathrm{u}}$ is the total covariance matrix. A regularization term $\lambda \mathbf{w}^\top \mathbf{R} \mathbf{w}$ is added to improve the numerical stability when computing the inverse of $\mathbf{S}_T$ in high dimensional setting ($I \ll JK$). $\mathbf{R}$ is usually equal to the identity.

$\mathbf{w}^*$ is obtained as the first eigenvector of $(\mathbf{S}_T + \lambda \mathbf{R})^{-1} \mathbf{S}_B$.

Additional structural constraints should be added to the optimization problem (9.1) in order to account for the three-way structure of the data.

9.2.2 MFDA Criterion

Structural constraints are imposed in such way that the weight vector $\mathbf{w}$ will be decomposed in two vectors as $\mathbf{w} = \mathbf{w}_K \otimes \mathbf{w}_J$. $\mathbf{w}_K$ is a weight vector associated with the K modalities while $\mathbf{w}_J$ is the weight vector related to the J variables. This structural constraint results in a more parsimonious model ($J + K$ instead of $J \times K$ parameters to estimate), and allows to study separately the effects of the variables and the modalities. A possible reformulation of FDA that takes into account the three-way structure of the data is introduced through the optimization problem:

$$\operatorname*{argmax}_{\mathbf{w}} \frac{(\mathbf{w}_K \otimes \mathbf{w}_J)^\top \mathbf{S}_B (\mathbf{w}_K \otimes \mathbf{w}_J)}{(\mathbf{w}_K \otimes \mathbf{w}_J)^\top \mathbf{S}_T (\mathbf{w}_K \otimes \mathbf{w}_J) + \lambda (\mathbf{w}_K \otimes \mathbf{w}_J)^\top \mathbf{R} (\mathbf{w}_K \otimes \mathbf{w}_J)}. \tag{9.2}$$

where the matrix $\mathbf{R} = \mathbf{R}_K \otimes \mathbf{R}_J$ is introduced to avoid numerical issues as in (9.1). $\mathbf{R}_K$ is of dimension $K \times K$ and $\mathbf{R}_J$ is of dimension $J \times J$. The choices of $\mathbf{R}$ and λ are illustrated in Sect. 9.3.

9.2.3 MFDA Algorithm

The optimization problem (9.2) is solved by considering the following identities:

$$\mathbf{w}^\top \mathbf{S_B} \mathbf{w} = \mathbf{w}_J^\top \left(\sum_{k=1}^{K} (\mathbf{w}_K)_k \, \mathbf{X}_{..k} \right)^\top \mathbf{H}_B \left(\sum_{k=1}^{K} (\mathbf{w}_K)_k \, \mathbf{X}_{..k} \right) \mathbf{w}_J \tag{9.3}$$

$$= \mathbf{w}_K^\top \left(\sum_{j=1}^{J} (\mathbf{w}_J)_j \, \mathbf{X}_{.j.} \right)^\top \mathbf{H}_B \left(\sum_{j=1}^{J} (\mathbf{w}_J)_j \, \mathbf{X}_{.j.} \right) \mathbf{w}_K \tag{9.4}$$

$$\mathbf{w}^\top \mathbf{S_T} \mathbf{w} = \mathbf{w}_J^\top \left(\sum_{k=1}^{K} (\mathbf{w}_K)_k \, \mathbf{X}_{..k} \right)^\top \left(\sum_{k=1}^{K} (\mathbf{w}_K)_k \, \mathbf{X}_{..k} \right) \mathbf{w}_J \tag{9.5}$$

$$= \mathbf{w}_K^\top \left(\sum_{j=1}^{J} (\mathbf{w}_J)_j \, \mathbf{X}_{.j.} \right)^\top \left(\sum_{j=1}^{J} (\mathbf{w}_J)_j \, \mathbf{X}_{.j.} \right) \mathbf{w}_K \tag{9.6}$$

and

$$(\mathbf{w}_K \otimes \mathbf{w}_J)^\top \mathbf{R} (\mathbf{w}_K \otimes \mathbf{w}_J) = (\mathbf{w}_J^\top \mathbf{R}_J \mathbf{w}_J)(\mathbf{w}_K^\top \mathbf{R}_K \mathbf{w}_K) \tag{9.7}$$

Solving the optimization problem (9.2) with respect to $\mathbf{w}_J$ while maintaining $\mathbf{w}_K$ fixed, is achieved with a joint use of Eqs. (9.3) and (9.5). Similarly, solving the optimization problem (9.2) with respect to $\mathbf{w}_K$ while maintaining $\mathbf{w}_J$ fixed, is achieved with a joint use of Eqs. (9.4) and (9.6). The MFDA algorithm that solves optimization problem (9.2) is described in Algorithm 9.1. This algorithm starts by assigning random initial values for $\mathbf{w}_J$ or $\mathbf{w}_K$ and then iterates a sequence of FDA problems. More specifically, each update boils down to perform FDA on either $\mathbf{X}_J = \sum_{k=1}^{K} (\mathbf{w}_K)_k \, \mathbf{X}_{..k}$ or $\mathbf{X}_K = \sum_{J=1}^{J} (\mathbf{w}_J)_j \, \mathbf{X}_{.j.}$. From the expressions of $\mathbf{X}_J$ and $\mathbf{X}_K$, it becomes clear that $(\mathbf{w}_J)_j$ reflects the influence of the jth variable while $(\mathbf{w}_K)_k$ the influence of the kth modality. Notice that $\mathbf{X}_J$ (resp. $\mathbf{X}_K$) is a $I \times J$ (resp. $I \times K$) matrix as compared to the $I \times JK$ unfolded matrix $\mathbf{X}^u$.

Algorithm 9.1 yields $\mathbf{w}^1 = \mathbf{w}_K^1 \otimes \mathbf{w}_J^1$ corresponding to the first discriminant axis. Subsequent discriminant axes can be determined by imposing orthogonality constraints as detailed hereinafter.

Algorithm 9.1: Computation of the first multi-way FDA axis

Require: $\epsilon > 0$, $\mathbf{w}_K^{(0)}$
$q \leftarrow 0$
repeat
$\mathbf{X}_K = \sum_{k=1}^{K} \left(\mathbf{w}_K^{(q)}\right)_k \mathbf{X}_{..k}$, $\lambda_K = \left(\mathbf{w}_K^{(q)}\right)^\top \mathbf{R}_K \mathbf{w}_K^{(q)}$
$\mathbf{w}_J^{(q+1)} \leftarrow \underset{\mathbf{w}_J, \|\mathbf{w}_J\|=1}{\operatorname{argmax}} \frac{\mathbf{w}_J^\top \mathbf{X}_K^\top \mathbf{H}_B \mathbf{X}_K \mathbf{w}_J}{\mathbf{w}_J^\top \mathbf{X}_K^\top \mathbf{X}_K \mathbf{w}_J + \lambda_K \mathbf{w}_J^\top \mathbf{R}_J \mathbf{w}_J} \leftarrow \text{FDA}(\mathbf{y}, \mathbf{X}_K, \lambda_K)$
$\mathbf{X}_J = \sum_{j=1}^{J} \left(\mathbf{w}_J^{(q+1)}\right)_j \mathbf{X}_{.j.}$, $\lambda_J = \left(\mathbf{w}_J^{(q+1)}\right)^\top \mathbf{R}_J \mathbf{w}_J^{(q+1)}$
$\mathbf{w}_K^{(q+1)} \leftarrow \underset{\mathbf{w}_K, \|\mathbf{w}_K\|=1}{\operatorname{argmax}} \frac{\mathbf{w}_K^\top \mathbf{X}_J^\top \mathbf{H}_B \mathbf{X}_J \mathbf{w}_K}{\mathbf{w}_K^\top \mathbf{X}_J^\top \mathbf{X}_J \mathbf{w}_K + \lambda_J \mathbf{w}_K^\top \mathbf{R}_K \mathbf{w}_K} \leftarrow \text{FDA}(\mathbf{y}, \mathbf{X}_J, \lambda_J)$
$q \leftarrow q + 1$
until $\|\mathbf{w}_K^{(q-1)} - \mathbf{w}_K^{(q)}\| < \epsilon$
return $(\mathbf{w}_K^{(q)}, \mathbf{w}_J^{(q)})$

9.2.4 Additional Constraints

At the end of Algorithm 9.1, one discriminant vector $\mathbf{w}^1 = \mathbf{w}_K^1 \otimes \mathbf{w}_J^1$ is obtained. The following $C-1$ axes (where C is the number of classes): $\mathbf{w}_J^s$, $\mathbf{w}_K^s$, $s = 2, \ldots, C-1$, are obtained subject to orthogonality constraints expressed as follows:

$$(\mathbf{w}^s)^\top [\mathbf{w}^1 \ldots \mathbf{w}^{s-1}] = \mathbf{0} \iff (\mathbf{w}_K^s \otimes \mathbf{w}_J^s)^\top (\mathbf{w}_K^c \otimes \mathbf{w}_J^c) = 0 \quad \forall c \in [1, \ldots, s-1]$$
$$(\mathbf{w}_K^{s\top} \mathbf{w}_K^c)(\mathbf{w}_J^{s\top} \mathbf{w}_J^c) = 0 \quad \forall c \in [1, \ldots, s-1] \quad (9.8)$$

From Eq. (9.8), orthogonality can be obtained by either imposing $\mathbf{w}_K^{s\top} \mathbf{w}_K^c = 0$ or $\mathbf{w}_J^{s\top} \mathbf{w}_J^c = 0$. The construction of the next discriminant axes is derived below for the constraint $\mathbf{w}_J^{s\top} \mathbf{w}_J^c = 0$.

9.2.5 Second Discriminant Axis

Considering $\mathbf{H} = \text{span}\{\mathbf{w}_J^1\}$ and $\mathbf{P}_{\mathbf{H}^\perp}$ the projection matrix over $\mathbf{H}^\perp$. The orthogonality condition is equivalent to say that there exists a non unique $\mathbf{v} \in \mathbb{R}^J$ such that $\mathbf{w}_J^2 = \dfrac{\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}}{\|\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}\|}$. The orthogonality constraint on $\mathbf{w}_J^2$ yields the optimization problem:

$$\max_{\mathbf{w}_K, \mathbf{v}} \frac{(\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}))^\top \mathbf{S}_B (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}))}{(\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}))^\top \mathbf{S}_T (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v})) + \lambda (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}))^\top \mathbf{R} (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp} \mathbf{v}))}, \quad (9.9)$$

subject to $\|\mathbf{w}_K\| = 1$ and $\|\mathbf{v}\| = 1$ which is also a MFDA problem due to the following identities:

$$\begin{aligned}\mathbf{w}^\top \mathbf{S}_B \mathbf{w} &= (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp}\mathbf{v}))^\top \mathbf{X}^\top \mathbf{H}_B \mathbf{X} (\mathbf{w}_K \otimes (\mathbf{P}_{\mathbf{H}^\perp}\mathbf{v})) \\ &= \left(\sum_{k=1}^{K} \left(\mathbf{w}_K^2\right)_k (\mathbf{X}_{..k}\mathbf{P}_{\mathbf{H}^\perp})\mathbf{v}\right)^\top \mathbf{H}_B \left(\sum_{k=1}^{K} \left(\mathbf{w}_K^2\right)_k (\mathbf{X}_{..k}\mathbf{P}_{\mathbf{H}^\perp})\mathbf{v}\right) \\ &= \left(\sum_{j=1}^{K} \left(\mathbf{w}_J^2\right)_j (\mathbf{X}_{.j.}\mathbf{P}_{\mathbf{H}^\perp})\mathbf{v}\right)^\top \mathbf{H}_B \left(\sum_{j=1}^{J} \left(\mathbf{w}_J^2\right)_j (\mathbf{X}_{.j.}\mathbf{P}_{\mathbf{H}^\perp})\mathbf{v}\right)\end{aligned}$$

We emphasize that $\mathbf{P}_{\mathbf{H}^\perp}$ is of rank $J-1$ but does not pose any computational issues because $\mathbf{P}_{\mathbf{H}^\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{H}}$ with $\mathbf{P}_{\mathbf{H}} = \mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top = \mathbf{w}_J^1(\mathbf{w}_J^1)^\top$. It comes that the projection is now advantageously replaced by a deflation:

$$\mathbf{X}_{..k}\mathbf{P}_{\mathbf{H}^\perp} = \mathbf{X}_{.k} - (\mathbf{X}_{..k}\mathbf{w}_J^1)(\mathbf{w}_J^1)^\top.$$

9.2.6 Subsequent Discriminant Axes

The sth discriminant axis is obtained using the same deflation strategy considering the vector space

$$\mathbf{H} = \text{span}\left\{\mathbf{w}_J^1, \mathbf{w}_J^2, \ldots, \mathbf{w}_J^{s-1}\right\}.$$

Let $\mathbf{X}'$ be the three-way data obtained from the previous step. Since $\mathbf{X}'_{..k}$ has already been projected over

$$\text{span}\left\{\mathbf{w}_J^1, \mathbf{w}_J^2, \ldots, \mathbf{w}_J^{s-2}\right\}^\perp,$$

the vector $\mathbf{w}_J^{s-1}$ is thus obtained using Algorithm 9.1 on the deflated version of $\mathbf{X}'$ which is obtained from the following deflation:

$$\mathbf{X}'_{..k}\,(= \mathbf{X}_{..k}\mathbf{P}_{\mathbf{H}^\perp}) \leftarrow \mathbf{X}'_{..k} - (\mathbf{X}'_{..k}\mathbf{w}_J^{s-1})(\mathbf{w}_J^{s-1})^\top.$$

9.3 Application to Traumatic Brain Injury

Traumatic brain injury is one of the leading causes of death and disability in the industrialized world, generally requiring prolonged rehabilitation (Grubb et al. 1996).

In the scope of this paper MFDA is applied to a multi-modal brain MRI data set in order to predict, in the long term, the recovery of patients that suffered from traumatic brain injury. The I horizontal slices characterize the patients $i = 1, \ldots, I$,

the J lateral slices are related to the voxels $j = 1, \ldots, J$ and the K frontal slices correspond to the different modalities $k = 1, \ldots, K$. From this the data can be structured into the tensor $\mathbf{X} = \{\mathbf{X}_{ijk}\}_{1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K}$ of order 3. Due to the high dimensionality of the dataset, a kernel version of FDA is used (Mika et al. 1999). The optimal value for the regularization parameter λ is tailored through a leave-one-out cross-validation procedure. The $\mathbf{R}$ matrix is set to be the identity.

9.3.1 Data Description

The multi-modal MRI diffusion images were acquired on individuals divided into 3 classes: 39 controls, 65 coma patients with a positive outcome and 39 coma patients with a negative outcome ($I = 143$). Four diffusion images ($K = 4$), namely fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (L1) and radial diffusivity (Lt), images were acquired from the entire brain of both patients and controls. Each volumetric image has $91 \times 109 \times 91$ voxels which were then reshaped into a $1 \times 902,629$ vector ($J = 902,629$). The resulting tensor $\mathbf{X}$ considered as input for MFDA is of dimensions $143 \times 902,629 \times 4$, whereas the resulting unfolded tensor $\mathbf{X}^{\mathrm{u}}$ is of dimensions $143 \times 3,610,516$.

9.3.2 FDA Applied to the Entire Brain

A linear kernel version of FDA (Mika et al. 1999) applied to $\mathbf{X}^{u}$ results in 8 weight matrices (4 for each eigenvector). A leave-one-out cross-validation yields the optimal regularization parameter to be $\lambda = 400$ with an accuracy of 76 %. Moreover, the resulting FA weights matrix obtained by considering the segment of the eigenvectors corresponding to FA are shown in Fig. 9.2. These images are difficult to interpret since there is no focalized region used for the discrimination. We mention that other modalities (i.e., other segments) could be visualized but do not give additional discriminative information (results not shown).

9.3.3 MFDA Applied to the Entire Brain

MFDA applied to $\mathbf{X}$ results in 2 weight matrices associated with $\mathbf{w}_j^1$ and $\mathbf{w}_j^2$ which integrate all the modalities. This yields a single volumetric image that integrate the 4 modalities instead of one for each modality in FDA. After performing a leave-one-out cross-validation, the optimal regularization parameter for MFDA is found to be $\lambda = 10^4$ with an accuracy of 71 %. Table 9.1 shows the contribution of each modality for the construction of the single volumetric image. FA has the highest

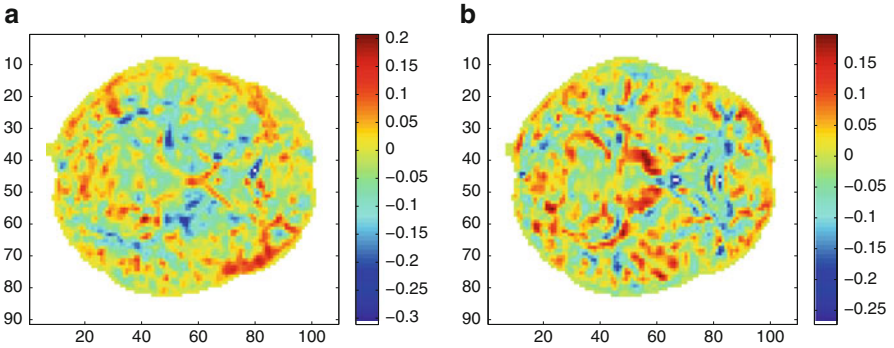

Fig. 9.2 Entire brain. FA segment of the FDA weights vectors $(\mathbf{w}_{FA}^1, \mathbf{w}_{FA}^2)$ for $\lambda = 400$. (**a**) FDA analysis FA, 1st eigenvector. (**b**) FDA analysis FA, 2nd eigenvector

Table 9.1 Entire brain. MFDA weights vectors $(\mathbf{w}_K^1, \mathbf{w}_K^2)$

Modality	$\mathbf{w}_K^1$	$\mathbf{w}_K^2$
FA	**0.9887**	−0.0066
MD	0.0036	0.5703
L1	0.0046	0.6094
Lt	0.0031	0.5508

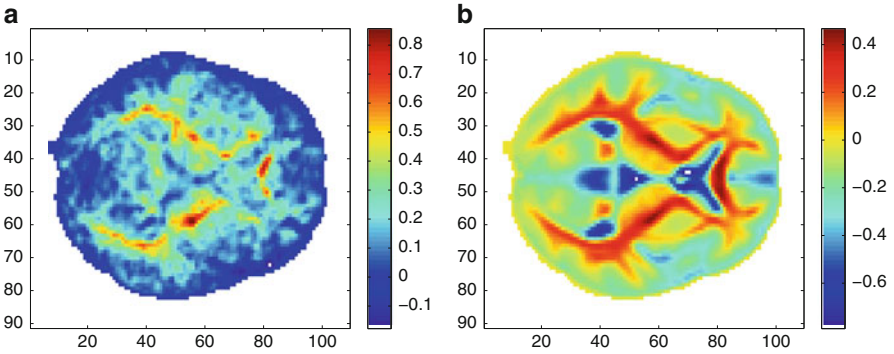

Fig. 9.3 Entire brain. MFDA weights vectors $(\mathbf{w}_J^1, \mathbf{w}_J^2)$ for $\lambda = 10,000$. (**a**) $\mathbf{w}_J^1$. (**b**) $\mathbf{w}_J^2$

weight in the discrimination for $\mathbf{w}_K^1$. For $\mathbf{w}_K^2$, all modalities but FA have been taken into account in the same proportion.

Figure 9.3 shows an example of MFDA weights $\mathbf{w}_J^1$ and $\mathbf{w}_J^1$ obtained on the entire brain (same plane as for FDA). Contrary to FDA, MFDA clearly locates the discriminative voxels in the white matter (in red). Since specific and smooth regions are selected, MFDA model is easier to interpret. In addition, $\mathbf{w}_J^2$ is reported in Table 9.1 and shows that all the modalities participate in the same proportion to the construction of the MFDA model. These results are consistent with the ones obtained by Sidaros et al. (2008) and Galanaud et al. (2012) regarding the importance of FA when assessing long-term recovery.

MFDA exhibits that the discriminating voxels are located, within the main white matter bundles, more specifically in the posterior limb of the internal capsule. Indeed, traumatic brain injury is characterized by the presence of diffuse axonal injury mainly located within deep and axial white matter bundle as found by Galanaud et al. (2012). For this reason, a second analysis based only on the white matter region is applied giving a $143 \times 20,764 \times 4$ tensor to analyze.

9.3.4 FDA and MFDA Applied to the White Matter

In Fig. 9.4, training and testing accuracies for FDA are reported for different values of λ. The optimal regularization parameter for FDA is equal to $\lambda = 400$ with an accuracy of 75 %. The associated confusion matrix is shown in Table 9.2. We note that the most frequent error is done between the positive and negative outcome, and that the distinction between patients and controls is very accurate. In Fig. 9.5, training and testing accuracies for MFDA are reported. The optimal regularization parameter is $\lambda = 100$ with an accuracy of 77 %. The associated confusion matrix is shown in Table 9.3.

The resulting weights obtained when analyzing the white matter are presented in Fig. 9.6, together with the corresponding w_K values in Table 9.4. These results are consistent with the ones obtained using the entire brain, where modality FA serves as the most discriminant modality.

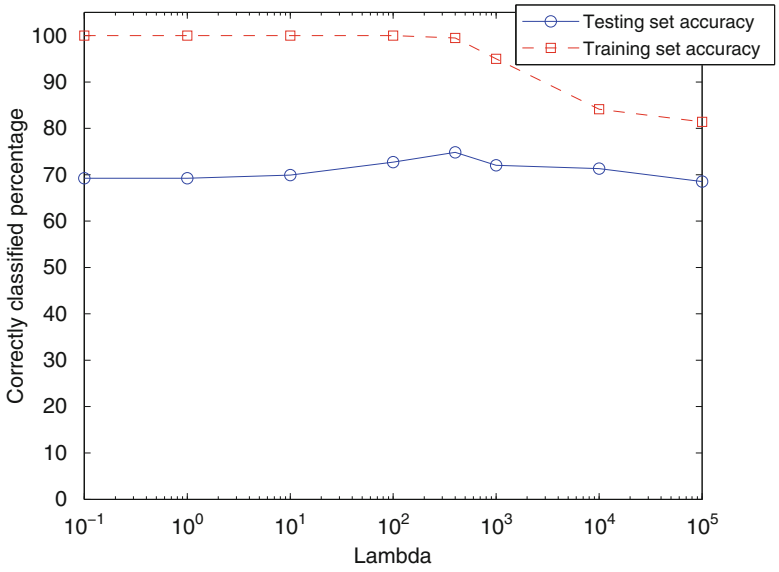

Fig. 9.4 FDA leave-one-out cross validation

Table 9.2 FDA confusion matrix with $\lambda = 400$

FDA	Predicted		
Observed	Controls	Positive outcomes	Negative outcomes
Controls	39	0	0
Positive outcomes	6	49	10
Negative outcomes	0	20	19

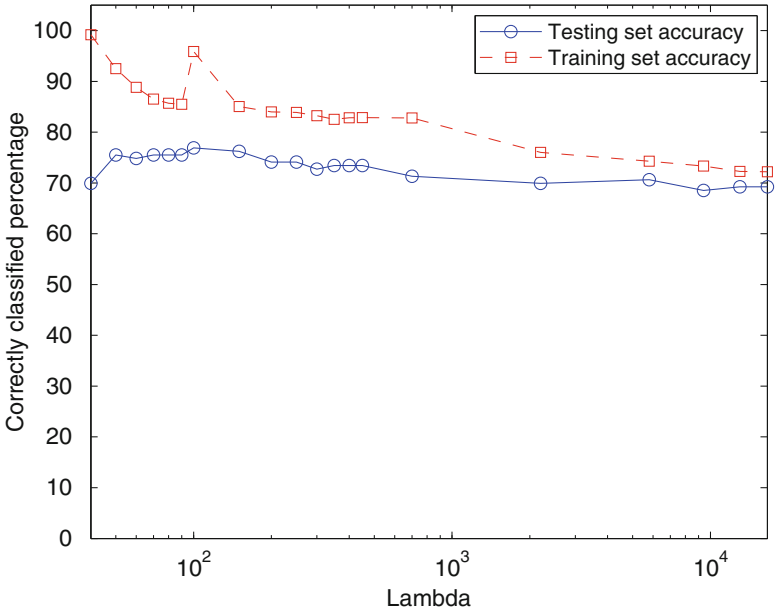

Fig. 9.5 MFDA leave-one-out cross validation

Table 9.3 MFDA confusion matrix with $\lambda = 100$

MFDA	Predicted		
Observed	Controls	Positive outcomes	Negative outcomes
Controls	37	2	0
Positive outcomes	0	49	16
Negative outcomes	0	15	24

9.4 Discussion

In this paper, we propose a multiway formulation of FDA that considers the intrinsic tensor structure of the data. MFDA was applied to multi-modal MRI diffusion images to predict the long term recovery of patients with traumatic brain injury, for which good accuracy rates were obtained, from 71 % for MFDA to 76 % for FDA,

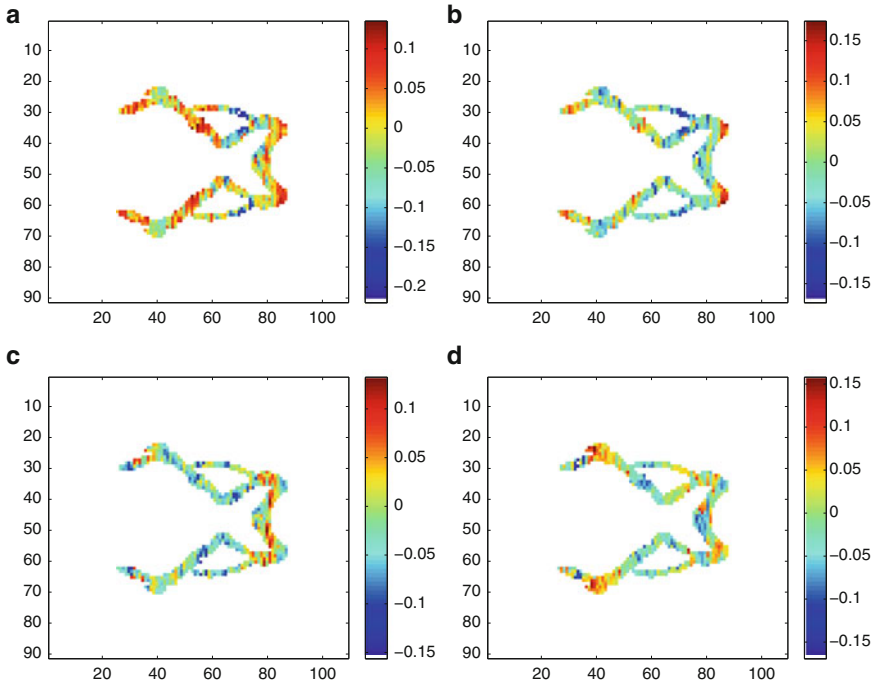

Fig. 9.6 White matter. MFDA weights vectors ($\mathbf{w}_J^1$, $\mathbf{w}_J^2$) for $\lambda = 100$. FA segment of the FDA weights vectors ($\mathbf{w}_{FA}^1$, $\mathbf{w}_{FA}^2$) for $\lambda = 400$. (**a**) FDA analysis FA, 1st eigenvector. (**b**) MFDA analysis, $\mathbf{w}_J^1$. (**c**) FDA analysis FA, 2nd eigenvector. (**d**) MFDA analysis, $\mathbf{w}_J^2$

Table 9.4 White matter. MFDA weights vectors ($\mathbf{w}_K^1$, $\mathbf{w}_K^2$)

Modality	$\mathbf{w}_K^1$	$\mathbf{w}_K^2$
FA	**0.8017**	−0.0681
MD	−0.2224	−0.5327
L1	0.2319	−0.8072
Lt	−0.5040	−0.2448

when using the entire brain. This loss in accuracy for MFDA is compensated by an improvement in the interpretability of the obtained classifier. This improvement is due to the introduced a priori structure that has been taken into account during the modelisation process. When analyzing the white matter we obtain a 75 % accuracy for FDA and 77 % for MFDA. MFDA separates the influence of spatial positions and the influence of the different modalities.

Another observation is that the FDA weights give higher importance to the borders of the brain, when the majority of the discriminant information should be found in the white matter since there is evidence that damage in this region is a distinctive feature of traumatic brain injury (Galanaud et al. 2012) as shown in the MFDA results. The MFDA weight matrices seem to supply more information on the location of the discrimination regions, as shown in Fig. 9.3. Moreover, FDA results

in 8 weight matrices ($J \times K$ classifier), complicating the interpretability, as opposed to only 2 weight matrices ($J + K$ classifier) obtained with MFDA which integrate all the modalities.

Future perspectives include an improvement of the accuracy of the classification of the positive and negative outcomes. In order to further improve the interpretability of the classifier a sparse MFDA algorithm is under development for reducing the number of active variables in the MFDA model.

Acknowledgements This study was funded by a grant from the French Ministry of Health (Projet Hospitalier de Recherche Clinique registration #P051061 [2005]) and from departmental funds from the Assistance Publique-Hôpitaux de Paris. The research leading to these results has also received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06 and LG acknowledges support from CONACYT.

References

Bro, R.: Multi-way analysis in the food industry: models, algorithms and applications. Ph.D. thesis, Royal Veterinary and Agricultural University (1998)

Galanaud, D., Perlbarg, V., Gupta, R., Stevens, R., Sanchez, P., Tollar, E., Champfleur, N., Dinkel, J., Faivre, S., Soto-Ares, G., Veber, B., Cottenceau, V., Masson, F., Tourdias, T., André, E., Audibert, G., Schmitt, E., Ibarrola, D., Dailler, F., Vanhaudenhuyse, A., Tshibanda, L., Payen, J., Bas, J.L., Krainik, A., Bruder, N., Girard, N., Laureys, S., Benali, H., Puybasset, L.: Assessment of white matter injury and outcome in severe brain trauma: a prospective multicenter cohort. Anesthesiology **117**, 1300–1310 (2012)

Grubb, A., Walsh, P., Lambe, N., Murrells, T., Robinson, S.: Survey of British clinicians' views on management of patients in persistent vegetative state. Lancet **348**, 35–40 (1996)

Harshman, R.A.: Foundations of the parafac procedure: models and conditions for an explanatory multi-mode factor analysis. UCLA Work. Pap. Phon. **16**, 1–84 (1970)

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Muller, K.B.: Fisher discriminant analysis with kernels. In: IEEE Conference on Neural Networks for Signal Processing IX, pp. 41–48. Institute of Electrical and Electronics Engineers, New York/Piscataway (1999)

Sidaros, A., Engberg, A., Sidaros, K., Liptrot, M., Herning, M., Petersen, P., Paulson, O., Jernigan, T., Rostrup, E.: Diffusion tensor imaging during recovery from severe traumatic brain injury and relation to clinical outcome: a longitudinal study. Brain **131**, 559–572 (2008)

Part III
New and Alternative Methods for Multitable and Path Analysis

Chapter 10
Structured Variable Selection for Regularized Generalized Canonical Correlation Analysis

Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe, Edouard Duchesnay, Vincent Frouin, and Arthur Tenenhaus

Abstract Regularized Generalized Canonical Correlation Analysis (RGCCA) extends regularized canonical correlation analysis to more than two sets of variables. Sparse GCCA (SGCCA) was recently proposed to address the issue of variable selection. However, the variable selection scheme offered by SGCCA is limited to the covariance ($\tau = 1$) link between blocks. In this paper we go beyond the covariance link by proposing an extension of SGCCA for the full RGCCA model ($\tau \in [0, 1]$). In addition, we also propose an extension of SGCCA that exploits pre-given structural relationships between variables within blocks. Specifically, we propose an algorithm that allows structured and sparsity-inducing penalties to be included in the RGCCA optimization problem.

Keywords RGCCA • Variable selection • Structured penalty • Sparse penalty

T. Löfstedt (✉)
Computational Life Science Cluster (CLiC), Department of Chemistry,
Umeå University, Umeå, Sweden
e-mail: lofstedt.tommy@gmail.com

F. Hadj-Selem • E. Duchesnay • V. Frouin
NeuroSpin, CEA Saclay, Gif-sur-Yvette, France
e-mail: fouad.hadjselem@vedecom.fr; edouard.duchesnay@cea.fr; vincent.frouin@cea.fr

V. Guillemot
Bioinformatics/Biostatistics Core Facility, IHU-A-ICM, Brain and Spine Institute, Paris, France
e-mail: v.guillemot-ihu@icm-institute.org

C. Philippe
Gustave Roussy, Villejuif, France
e-mail: cathy.philippe@gustaveroussy.fr

A. Tenenhaus
Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506),
CentraleSupélec-CNRS-Université Paris-Sud and Bioinformatics/Biostatistics Platform
IHU-A-ICM, Brain and Spine Institute, Paris, France
e-mail: arthur.tenenhaus@centralesupelec.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_10

10.1 Introduction

Regularized Generalized Canonical Correlation Analysis (RGCCA) (Tenenhaus and Tenenhaus 2011) is a generalization of regularized canonical correlation analysis (Vinod 1976) to more than two sets of variables. RGCCA relies on a sound theoretical foundation with a well-defined optimization criterion (as is customary in multiblock data analysis) and at the same time allows the incorporation of prior knowledge or hypotheses about the relationships between the blocks (e.g., as in PLS path modeling).

Sparse GCCA (SGCCA) (Tenenhaus et al. 2014) was recently proposed to address the issue of variable selection. The RGCCA criterion was modified to include L_1 penalties (with L_1 being the degree 1 norm of a vector defined such that the L_1-norm of vector $\mathbf{x}$ is $||\mathbf{x}||_1 = \sum |x_i|$) on the outer weights vectors in order to promote sparsity. For technical reasons, concerning the RGCCA algorithm, the variable selection offered by SGCCA is limited to the covariance link between blocks (i.e., with all $\tau_k = 1$).

In this paper we go beyond the covariance link and allow any $\tau \in [0, 1]$. More specifically, we present an extension of SGCCA that allows variable selection to be performed for the full RGCCA model. In addition, we also propose an extension of SGCCA that exploits a pre-given structural relationships between variables within blocks. This is achieved by introducing structured and sparsity-inducing penalties in the model. Such penalties have recently become popular in machine learning and related fields (Hadj-Selem et al. 2016) and encourage the resulting models to have a particular structure.

Structured penalties have previously been considered in a two-block setting with canonical correlation analysis (Chen and Liu 2011). However, to combine such structured penalties with RGCCA poses new optimization challenges that we propose to tackle. Specifically, we propose a general multiblock algorithm that allows structured and sparsity-inducing penalties to be included in the RGCCA model.

10.2 Method

We consider K data matrices, $\mathbf{X}_1, \ldots, \mathbf{X}_K$. Each $n \times p_k$ data matrix $\mathbf{X}_k$ is called a block and represents a set of p_k centered variables observed on n observations. The number and the nature of the variables usually differ from one block to another but the observations are the same across all the blocks.

Let $\mathbf{C} = \left[c_{kj}\right]$ be an adjacency matrix, where $c_{kj} = 1$ if blocks $\mathbf{X}_k$ and $\mathbf{X}_j$ are connected, and $c_{kj} = 0$ otherwise. RGCCA investigates the relationships between these blocks, while taking into account the structural connection between blocks defined by the adjacency matrix. For that purpose, RGCCA is defined by the following optimization problem

$$\underset{\mathbf{w}_k \in \mathbb{R}^{p_k},\, k=1,\ldots,K}{\text{minimize}} \varphi(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{k=1}^{K}\sum_{j=1}^{K} c_{kj} g\left(\text{cov}\left(\mathbf{X}_k\mathbf{w}_k,\, \mathbf{X}_j\mathbf{w}_j\right)\right) \tag{10.1}$$

$$\text{subject to} \quad \tau_k \|\mathbf{w}_k\|_2^2 + (1-\tau_k)\text{var}(\mathbf{X}_k\mathbf{w}_k) = 1, \quad k = 1, \ldots, K, \tag{10.2}$$

where $\mathbf{w}_k$ are the weight vectors (also called "outer" weight vectors), $\mathbf{X}_k\mathbf{w}_k$ are the block components, $||.||_2$ is the L_2 (a.k.a., "Euclidean") norm (i.e., $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\top\mathbf{x}}$), and g is called the inner-weighting scheme and is usually the identity (Horst's scheme), the absolute value (Centroid scheme), or the square function (Factorial scheme).

The regularization parameters $\tau_k \in [0, 1]$ provide a way to maximize either the correlation ($\tau_k = 0$) or the covariance ($\tau_k = 1$). A trade-off between covariance and correlation is achieved for all other values of $\tau_k \in (0, 1)$.

In the framework of SGCCA, all τ_k, for $k = 1, \ldots, K$, are assumed to be equal to 1. This means that variable selection is only possible for the covariance link in Eq. 10.1. Also, L_1 constraints are added on the weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ and yield the SGCCA optimization problem which is defined as:

$$\begin{aligned} &\underset{\mathbf{w}_k \in \mathbb{R}^{p_k},\, k=1,\ldots,K}{\text{minimize}} \varphi(\mathbf{w}_1, \ldots, \mathbf{w}_K) \\ &\text{subject to} \quad \|\mathbf{w}_k\|_1 \le s_k \;\text{ and }\; \|\mathbf{w}_k\|_2^2 = 1, \quad k = 1, \ldots, K, \end{aligned} \tag{10.3}$$

where $s_k > 0$ is the radius of the L_1 ball and determines the amount of sparsity for $\mathbf{w}_k$; the smaller s_k is, the larger the degree of sparsity for $\mathbf{w}_k$.

However, the L_1 constraint is blind to any structure between the variables in $\mathbf{X}_k$, and is thus not able to account for cases such as groups or similarities between variables in the RGCCA model. We therefore add structured penalties to the objective function, that account for such structured prior knowledge or assumptions. The general optimization problem considered in this paper is

$$\underset{\mathbf{w}_k \in \mathbb{R}^{p_k},\, k=1,\ldots,K}{\text{minimize}} \varphi(\mathbf{w}_1, \ldots, \mathbf{w}_K) + \sum_{k=1}^{K} \omega_k \Omega_k(\mathbf{w}_k) \tag{10.4}$$

$$\text{subject to} \quad \|\mathbf{w}_k\|_1 \le s_k, \quad k = 1, \ldots, K;$$

$$\tau_k \|\mathbf{w}_k\|_2^2 + (1-\tau_k)\text{var}(\mathbf{X}_k\mathbf{w}_k) \le 1, \quad k = 1, \ldots, K \tag{10.5}$$

where Ω_k are the structured penalties with regularization parameters ω_k. For technical reasons, we must constrain the inner-weighting function to the identity $g(x) = x$ (i.e., to Horst's scheme). Note that the equality in Eq. 10.2 has been changed to an inequality in Eq. 10.5 because the algorithm presented below requires the constraints to be convex. This is not a relaxation, however, because the Karush-Kuhn-Tucker conditions require all constraints to be active at the solution, and it is always possible to find constraint parameters s_k such that both constraints are active for each block (Witten et al. 2009).

When the structured penalties, Ω_k, are convex, Eq. 10.4 is a multiconvex function with convex constraints. However, the structured penalties are usually non-smooth and non-separable (i.e., they cannot be written as a separable sum). Therefore we cannot minimize the penalties together with the smooth loss function by using a smooth minimization algorithm, and would thus have to resort to non-smooth minimization algorithms such as, for example, proximal methods. However, there is no explicit solution for computing the proximal operator of the structured penalties without the separability condition and it is therefore difficult to find a minimum in the general case. Solutions exist for some particular structured penalties, but they are tailored towards a particular formulation, and can not be used for the general problem that was defined in Eq. 10.4. We therefore adapt a very efficient smoothing technique proposed in Nesterov (2004) to resolve both the non-smoothness and non-separability issues for a very wide and general class of structured penalties. This smoothing technique is presented in the next section.

10.2.1 Nesterov Smoothing

The structured penalties, Ω_k, considered in this paper are convex but possibly non-differentiable. The functions Ω_k must fit Nesterov's framework, as described in Hadj-Selem et al. (2016), and are therefore required to have the form

$$\Omega_k(\mathbf{w}_k) = \max_{\boldsymbol{\alpha} \in \mathscr{K}_k} \langle \boldsymbol{\alpha} \mid \mathbf{A}_k \mathbf{w}_k \rangle,$$

where $\mathscr{K}_k$ is a compact convex set and $\mathbf{A}_k$ a linear operator between two finite-dimensional vector spaces. The Nesterov smoothing of Ω_k is then defined as

$$\widehat{\Omega}_k(\mu_k, \mathbf{w}_k) = \langle \boldsymbol{\alpha}_k^* \mid \mathbf{A}_k \mathbf{w}_k \rangle - \frac{\mu_k}{2} \|\boldsymbol{\alpha}_k^*\|_2^2,$$

for all $\mathbf{w}_k \in \mathbb{R}^{p_k}$, with μ_k a positive real smoothing parameter and where $\boldsymbol{\alpha}_k^* = \arg\max_{\boldsymbol{\alpha} \in \mathscr{K}_k} \left\{ \langle \boldsymbol{\alpha} \mid \mathbf{A}_k \mathbf{w}_k \rangle - \frac{\mu_k}{2} \|\boldsymbol{\alpha}\|_2^2 \right\}$.

Using Nesterov's smoothing on the functions Ω_k yields

$$\lim_{\mu_k \to 0} \widehat{\Omega}_k(\mu_k, \mathbf{w}_k) = \Omega_k(\mathbf{w}_k).$$

An immediate consequence is that, since the functions $\widehat{\Omega}_k$ are convex and differentiable, they may (for a sufficiently small value of μ_k) be used instead of Ω_k. The gradients of the Nesterov smoothed functions $\widehat{\Omega}_k(\mu_k, \mathbf{w}_k)$ are $\nabla_{\mathbf{w}_k} \widehat{\Omega}_k(\mu_k, \mathbf{w}_k) = \mathbf{A}_k^\top \alpha_k^*$. It can be shown that these gradients are Lipschitz continuous with Lipschitz constant $L\big(\nabla_{\mathbf{w}_k} \widehat{\Omega}_k(\mu_k, \mathbf{w}_k)\big) = \|\mathbf{A}_k\|_2^2 / \mu_k$, where $\|\mathbf{A}_k\|_2$ is the spectral norm of $\mathbf{A}_k$.

10.2.2 Reformulating the Objective

The Nesterov's smoothing technique allows us to have a smooth objective function with convex constraints. We rephrase the constraints as a single indicative constraint over a convex set. The optimization problem thus becomes

$$\begin{aligned} &\underset{\mathbf{w}_k \in \mathbb{R}^{p_k},\, k=1,\ldots,K}{\text{minimize}} \; \hat{f}(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \varphi(\mathbf{w}_1, \ldots, \mathbf{w}_K) + \sum_{k=1}^{K} \omega_k \widehat{\Omega}_k(\mu_k, \mathbf{w}_k) \\ &\text{subject to} \quad \mathbf{w}_k \in \mathscr{W}_k = \{\mathbf{x} \mid \mathbf{x} \in \mathscr{P}_k \cap \mathscr{S}_k\}, \quad k = 1, \ldots, K, \end{aligned} \tag{10.6}$$

where $\mathscr{P}_k = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^{p_k} \wedge \|\mathbf{x}\|_1 \leq s_k\}$ and $\mathscr{S}_k = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^{p_k} \wedge \tau_k \|\mathbf{x}\|_2^2 + (1 - \tau_k)\mathrm{Var}(\mathbf{X}_k\mathbf{x}) \leq 1\}$. Note that in general $\mathscr{W}_k \neq \{\}$ because we have, al least, $\mathbf{0} \in \mathscr{W}_k$.

The gradient of the objective function in Eq. 10.6 with respect to $\mathbf{w}_k$ is

$$\nabla_{\mathbf{w}_k} \hat{f}(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{j=1}^{K} c_{kj} \underbrace{g'\big(\mathrm{cov}(\mathbf{X}_k\mathbf{w}_k,\, \mathbf{X}_j\mathbf{w}_j)\big)}_{=1, \text{ because } g(x)=x} \frac{1}{n-1} \mathbf{X}_k^\top \mathbf{X}_j \mathbf{w}_j + \sum_{k=1}^{K} \omega_k \mathbf{A}_k^\top \alpha_k^*, \tag{10.7}$$

where $\mathrm{cov}(\mathbf{X}_k\mathbf{w}_k,\, \mathbf{X}_j\mathbf{w}_j) = \frac{1}{n-1} \mathbf{w}_k^\top \mathbf{X}_k^\top \mathbf{X}_j \mathbf{w}_j$ (i.e., the unbiased sample covariance).

10.3 Algorithm

The optimization problem in Eq. 10.6 is characterized by a multiconvex objective function with an indicative constraint over a convex set. This suggests an optimization procedure that minimizes the objective function one parameter vector at a time and treats the other parameter vectors as constants during this minimization. If each update improves the function value, gradually the function will be (locally) optimized over the entire set of parameter vectors. This principle is called block relaxation (De Leeuw 1994). The corresponding Algorithm 10.2 is related to the algorithm presented in Witten et al. (2009). However, several details need to be introduced before we discuss the proposed algorithm.

10.3.1 Projection Operators

At each iteration of the algorithm, and for each block, orthogonal projections onto the convex sets $\mathscr{W}_k$ are required. The projection onto the set $\mathscr{W}_k$ is the unique point that minimizes the problem

$$\text{proj}_{\mathscr{W}_k}(\mathbf{x}) = \underset{\mathbf{y} \in \mathscr{W}_k = \mathscr{P}_k \cap \mathscr{S}_k}{\arg\min} \|\mathbf{y} - \mathbf{x}\|_2. \tag{10.8}$$

The projection onto the intersection of these two sets can be computed using Dykstra's projection algorithm (Combettes and Pesquet 2011), stated in Algorithm 10.1. The sequence $(\mathbf{x}^{(s)})_{s \in \mathbb{N}}$ generated by Algorithm 10.1 converges to the unique solution of Eq. 10.8. Three key points need to be detailed in order to make Algorithm 10.1 understandable: (i) the projection onto $\mathscr{P}$ (Line 3), (ii) the projection onto $\mathscr{S}$ (Line 5), and (iii) the stopping criterion (Line 7). These 3 points are discussed below.

Algorithm 10.1: Dykstra's projection algorithm

Require: $\mathbf{x}^{(0)}, \mathscr{P}, \mathscr{S}, \varepsilon > 0$
Ensure: $\mathbf{x}^{(s)} \in \mathscr{P} \cap \mathscr{S}$
1: $\mathbf{p}^{(0)} \leftarrow \mathbf{0}$,;
$\mathbf{q}^{(0)} \leftarrow \mathbf{0}$
2: **for** $s = 0, 1, 2, \ldots$ **do**
3: $\mathbf{y}^{(s)} = \text{proj}_{\mathscr{P}}(\mathbf{x}^{(s)} + \mathbf{p}^{(s)})$
4: $\mathbf{p}^{(s+1)} = \mathbf{x}^{(s)} + \mathbf{p}^{(s)} - \mathbf{y}^{(s)}$
5: $\mathbf{x}^{(s+1)} = \text{proj}_{\mathscr{S}}(\mathbf{y}^{(s)} + \mathbf{q}^{(s)})$
6: $\mathbf{q}^{(s+1)} = \mathbf{y}^{(s)} + \mathbf{q}^{(s)} - \mathbf{x}^{(s+1)}$
7: **if** $\max\left(\|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{P}}(\mathbf{x}^{(s+1)})\|_2,\ \|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{S}}(\mathbf{x}^{(s+1)})\|_2\right) \leq \varepsilon$ **then break**
8: **end for**

Projection onto $\mathscr{S}$. The $\mathscr{S}$ *constraint* is a quadratic constraint

$$\tau_k \|\mathbf{w}_k\|_2^2 + (1 - \tau_k)\text{Var}(\mathbf{X}_k \mathbf{w}_k) = \tau_k \mathbf{w}_k^\top \mathbf{w}_k + \frac{1 - \tau_k}{n - 1} \mathbf{w}_k^\top \mathbf{X}_k^\top \mathbf{X}_k \mathbf{w}_k = \mathbf{w}_k^\top \mathbf{M}_k \mathbf{w}_k,$$

where $\mathbf{M}_k = \tau_k \mathbf{I}_{p_k} + \frac{1-\tau_k}{n-1} \mathbf{X}_k^\top \mathbf{X}_k$ is a positive-semidefinite matrix. The projection operator onto $\mathscr{S}_k$ is defined as

$$\text{proj}_{\mathscr{S}_k}(\mathbf{x}) = \underset{\mathbf{y} \in \mathbb{R}^{p_k}}{\arg\min} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda_k^* \mathbf{y}^\top \mathbf{M}_k \mathbf{y} = \left(\mathbf{I}_{p_k} + 2\lambda_k^* \mathbf{M}_k\right)^{-1} \mathbf{x}, \tag{10.9}$$

with λ_k^* the smallest λ_k such that $\mathbf{y}^\top \mathbf{M}_k \mathbf{y} \leq 1$. It is not feasible to numerically find λ_k^* directly from Eq. 10.9, especially when the number of variables is large. We have therefore devised an efficient algorithm that rephrases the problem and then uses the Newton-Raphson method to compute λ_k^* from a simple univariate auxiliary function that only depends on the eigenvalues of $\mathbf{M}_k$.

Projection onto $\mathscr{P}$. The projection onto an L_1 ball of radius s_k is achieved by using the *soft thresholding operator* (Parikh and Boyd 2013) defined as

$$\left(\text{proj}_{\mathscr{P}_k}(\mathbf{x})\right)_i = \begin{cases} x_i - \lambda_k^*, & \text{if } x_i > \lambda_k^*, \\ 0, & \text{if } |x_i| \leq \lambda_k^*, \\ x_i + \lambda_k^*, & \text{if } x_i < -\lambda_k^*. \end{cases} \tag{10.10}$$

where λ_k^* satisfies $\sum_{i=1}^{p_k} \max(0, |x_i| - \lambda_k^*) = s_k$. For details concerning the determination of λ^*, see for example van den Berg et al. (2008).

Stopping Criterion. Since the projection on Line 6 of Algorithm 10.2 is approximated by Algorithm 10.1, we are actually performing an inexact projected descent (Schmidt et al. 2011).

At step s of Algorithm 10.2, after projection onto $\mathscr{W}$ with Algorithm 10.1, the following inequality must be respected to ensure convergence to the minimum of the objective function (Schmidt et al. 2011):

Algorithm 10.2: Algorithm for structured variable selection in RGCCA

Require: $\hat{f}$, $\nabla\hat{f}$, $\mathbf{w}_k = \mathbf{w}_k^{(0)} \in \mathscr{W}_k$, $\varepsilon > 0$
Ensure: $\mathbf{w}_k^{(s)} \in \mathscr{W}_k$ such that $\varepsilon \in \partial\hat{f}(\mathbf{w}_1^{(s)}, \ldots, \mathbf{w}_K^{(s)})$
1: **repeat**
2: **for** $k = 1$ **to** K **do**
3: $\mathbf{w}_k^{(1)} = \mathbf{w}_k^{(0)} = \mathbf{w}_k$
4: **for** $s = 1, 2, \ldots$ **do**
5: $\mathbf{y} = \mathbf{w}_k^{(s)} + \frac{k-2}{k+1}\left(\mathbf{w}_k^{(s)} - \mathbf{w}_k^{(s-1)}\right)$
6: $\mathbf{w}_k^{(s+1)} = \text{proj}_{\mathscr{W}_k}\left(\mathbf{y} - t_k \nabla_{\mathbf{w}_k^{(s)}}\hat{f}\left(\mathbf{w}_1^{(s)}, \ldots, \mathbf{y}, \ldots, \mathbf{w}_K^{(s)}\right)\right)$
7: **if** $\|\mathbf{w}_k^{(s+1)} - \mathbf{y}\|_2 \leq t_k\varepsilon$ **then break**
8: **end for**
9: $\mathbf{w}_k = \mathbf{w}_k^{(s+1)}$
10: **end for**
11: **until** $\|\mathbf{w}_k - \text{proj}_{\mathscr{W}_k}\left(\mathbf{w}_k - t_k\nabla_{\mathbf{w}_k}\hat{f}(\mathbf{w}_1, \ldots, \mathbf{w}_K)\right)\|_2 < t_k\varepsilon$, **for all** $k = 1, \ldots, K$

$$\left\|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{W}}\left(\mathbf{x}^{(s+1)}\right)\right\|_2 < \varepsilon^{(s)},$$

where $\varepsilon^{(s)}$ must decrease like $\mathscr{O}\left(1/i^{4+\delta}\right)$, for some $\delta > 0$ and i the iteration count for the inner-most loop (which is actually an application of the fast iterative-shrinkage-thresholding algorithm (FISTA) with optimal convergence speed for first-order methods).

Since we cannot compute $\left\|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{W}}\left(\mathbf{x}^{(s+1)}\right)\right\|_2$ directly (this is essentially the problem we are trying to solve) and since $\mathscr{W}$ is the intersection of $\mathscr{S}$ and $\mathscr{P}$, we approximate it by

$$\max\left(\|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{S}}(\mathbf{x}^{(s+1)})\|_2, \quad \|\mathbf{x}^{(s+1)} - \text{proj}_{\mathscr{P}}(\mathbf{x}^{(s+1)})\|_2\right).$$

10.3.2 Algorithm for Structured Variable Selection in RGCCA

We are now ready to describe the full RGCCA accelerated projected gradient method. This algorithm is presented in Algorithm 10.2.

The step sizes, t_k, appearing in the FISTA loop, are computed (Hadj-Selem et al. 2016) such that $t_k = \left(\sum_f L(\nabla_{\mathbf{w}_k} f)\right)^{-1}$, where the sum is over the Lipschitz constants of the gradients in the loss function in Eq. 10.6 (i.e., the Lipschitz constants of the gradients in Eq. 10.7). If some gradient is not Lipschitz continuous, or if the sum of Lipschitz constants is zero, the step size can also be found efficiently using backtracking line search.

Finally, the main stopping criterion on Line 11 is actually performing a step of the iterative soft-thresholding algorithm (ISTA). Thus, the stopping criterion stems from the subgradient definition and optimality condition of proximal operators (Qin et al. 2013).

10.4 Example

The proposed method is illustrated on a 3-block data-set. The objective is to predict the location of pediatric brain tumors from a $53 \times 15{,}702$ gene expression (GE) data matrix $\mathbf{X}_1$ and a $53 \times 41{,}996$ genome-wide array of comparative genomic hybridation (CGH) data matrix $\mathbf{X}_2$ (Philippe et al. 2012). A 53×3 dummy matrix $\mathbf{X}_3$ encodes the locations of the tumors. The tumors are divided into three locations: Diffuse Intrinsic Pontine Gliomas (DIPG, i.e., these tumor are localized in brain stem), central nuclei (Midline) and supratentorial (Hemisphere).

The design was chosen to be oriented towards the prediction of the location: $\mathbf{X}_1$ and $\mathbf{X}_2$ are connected to $\mathbf{X}_3$ ($c_{13} = c_{23} = 1$), but there is no connection between $\mathbf{X}_1$ and $\mathbf{X}_2$ and therefore ($c_{12} = 0$). Tenenhaus et al. (2014) have shown on similar data that this design yields the best prediction performances among all possible designs. The regularization constants, τ_k, were obtained from the Schäfer and Strimmer formula (Schäfer and Strimmer 2005), and were $\tau_1 = 1$, $\tau_2 = 0.3$. The dummy matrix, $\mathbf{X}_3$, had but one constraint, namely the $\mathscr{S}$ *constraint* with $\tau_3 = 1$.

An L_1 and a group $L_{1,2}$ (Qin et al. 2013) constraint were associated with GE. An L_1 and a total variation constraint (Michel et al. 2011) were associated with $\mathbf{X}_2$ to smooth the often noisy CGH data. The associated parameters were found by grid search in a seven-fold cross-validation scheme for maximizing $\mathrm{R}^2_{\mathbf{X}_3}$, defined as

$$\mathrm{R}^2_{\mathbf{X}_3} = \mathrm{R}^2_{\mathbf{X}_1 \to \mathbf{X}_3} \cdot \mathrm{R}^2_{\mathbf{X}_2 \to \mathbf{X}_3} = \left(1 - \frac{\|\hat{\mathbf{X}}_{1\to 3} - \mathbf{X}_3\|_F^2}{\|\mathbf{X}_3\|_F^2}\right)\left(1 - \frac{\|\hat{\mathbf{X}}_{2\to 3} - \mathbf{X}_3\|_F^2}{\|\mathbf{X}_3\|_F^2}\right) \tag{10.11}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\hat{\mathbf{X}}_{k\to 3}$ is the prediction of $\mathbf{X}_3$ from the model associated with $\mathbf{X}_k$. The predictions were computed in the standard

way using the inner relation as $\hat{\mathbf{X}}_{k\to 3} = \mathbf{T}_k(\mathbf{T}_k^\top \mathbf{T}_k)^{-1}\mathbf{T}_k^\top \mathbf{T}_3 \mathbf{W}_3^\top$, where $\mathbf{T}_k = [\mathbf{X}_k\mathbf{w}_{k,1}|\cdots|\mathbf{X}_k\mathbf{w}_{k,A}]$ is the matrix containing the A components associated with $\mathbf{X}_k$ and $\mathbf{W}_3 = [\mathbf{w}_{3,1}|\cdots|\mathbf{w}_{3,A}]$ is a matrix containing the A weight vectors associated with $\mathbf{X}_3$. Therefore, $\mathrm{R}^2_{\mathbf{X}_3}$ is the combined prediction rate from the models of $\mathbf{X}_1$ and $\mathbf{X}_2$, forcing both blocks to predict $\mathbf{X}_3$ well.

The regularization constant for group $L_{1,2}$ was $\omega_1 = 0.8$ and the L_1 norm constraint had a radius of $s_1 = 18$. This selected roughly 2 % of the variables in $\mathbf{X}_1$ in the first component, and roughly 4 % in the second component. The regularization constant for total variation was $\omega_2 = 0.005$, and the L_1 norm constraint had a radius of $s_2 = 10$. This selected roughly 11 % of the variables in $\mathbf{X}_2$ in the first component, and roughly 13 % in the second component.

Therefore, the optimization problem for this example was

$$\begin{aligned}
\min_{\mathbf{w}_i\in\mathbb{R}^{p_i}} \hat{f}(\mathbf{w}_1,\ldots,\mathbf{w}_K) = {} & -\mathrm{cov}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_3\mathbf{w}_3) - \mathrm{cov}(\mathbf{X}_2\mathbf{w}_2, \mathbf{X}_3\mathbf{w}_3) \\
& + 0.8\cdot\widehat{\Omega}_{GL}(\mu_1,\mathbf{w}_1) + 0.005\cdot\widehat{\Omega}_{TV}(\mu_2,\mathbf{w}_2), \\
\text{subject to } \|\mathbf{w}_1\|_1 \le 18, \quad & \|\mathbf{w}_1\|_2^2 \le 1 \\
\|\mathbf{w}_2\|_1 \le 10, \quad & 0.3\cdot\|\mathbf{w}_2\|_2^2 + 0.7\cdot\mathrm{var}(\mathbf{X}_2\mathbf{w}_2) \le 1 \\
& \|\mathbf{w}_3\|_2^2 \le 1,
\end{aligned}$$

in which $\mu_1 = \mu_2 = 5\cdot 10^{-4}$. Two components were extracted using the deflation rule

$$\mathbf{X}_k \leftarrow \mathbf{X}_k - \frac{\mathbf{X}_k\mathbf{w}_k\mathbf{w}_k^\top}{\mathbf{w}_k^\top\mathbf{w}_k}.$$

10.5 Results

The final model had an $\mathrm{R}^2_{\mathbf{X}_3} = \mathrm{R}^2_{\mathbf{X}_1\to\mathbf{X}_3}\cdot\mathrm{R}^2_{\mathbf{X}_2\to\mathbf{X}_3} = 0.51\times 0.47 = 0.24$ (means of seven-fold cross-validation). This is similar to the prediction rates reported in Tenenhaus et al. (2014) but in this case with an additional structure imposed to the weight vectors.

The locations were predicted using three different approaches: From $\mathbf{X}_1$ only, from $\mathbf{X}_2$ only and from both $\mathbf{X}_1$ and $\mathbf{X}_2$. The block $\mathbf{X}_1$ was able to predict $42/53 \approx 79$ % of the locations correctly; $\mathbf{X}_2$ was able to predict $38/53 \approx 72$ % of the locations correctly; and $49/53 \approx 92$ % of the locations were correctly identified when predicting from both $\mathbf{X}_1$ and $\mathbf{X}_2$ simultaneously (these are cross-validated values).

To evaluate the stability of the signatures, we decided to use Fleiss' κ indicator (Fleiss 1971). The number of times a variable is selected or not selected is counted across the 100 bootstrap samples. These frequencies are summarized by

the Fleiss' κ score and measures the agreement among the bootstrap samples. The higher the value of κ, the more stable the method is with respect to sampling. Fleiss' κ was 0.63 for the first component of $\mathbf{X}_1$, and 0.54 for the second component; 0.47 for the first component of $\mathbf{X}_2$ and 0.23 for the second component.

The group $L_{1,2}$ penalty selected 98.7 out of the 199 identified groups in the first component and 146.5 in the second (bootstrap averages). Groups were considered "strong" if they had a high ratio between the number of selected gene expressions within the group over the total number of selected gene expressions. Among the top ranking groups were: "Axon guidance" (hsa04360)—a function which implies a relation to DIPG, because of the abundance of axons in the brain stem—and Alzheimer's disease (hsa05010)—a result which implies a relation to a supratentorial tumor (in the hemispheres) because Alzehiemer disease affects the cortex and the hippocampus.

Among the groups that were excluded from the model were the Citrate cycle (TCA cycle) (hsa00020). Citrate seems to be abundant in DIPG (unpublished results), but its occurrence in other locations is unknown and so it could be similarly found in the other locations or cancer types.

10.6 Discussion and Conclusions

We propose a promising approach for taking into account prior information within RGCCA. The proposed optimization problem subsumes many well-known multi-block-based methods as special cases. Examples include PLS-R, PCA, RGCCA, SGCCA, but with the addition of sparse and structured penalties. This generalized RGCCA method was applied to chromosomal imbalances and gene expression data to predict the location of brain tumors. We used a group $L_{1,2}$ penalty for GE data and a total variation penalty for the CGH data. Both data sets were also subject to an L_1 and a quadratic constraint. The obtained results illustrate the benefit of adding sparse and structured constraints.

Acknowledgements This work was supported by grants from the French National Research Agency: ANR IA BRAINOMICS (ANR-10-BINF-04), and a European Commission grant: MESCOG (FP6 ERA-NET NEURON 01 EW1207).

References

Chen, X., Liu, H.: An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. Stat. Biosci. **4**, 3–26 (2011)

Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York (2011)

De Leeuw, J.: Block relaxation algorithms in statistics. In: Bock, H.-H., Lenski, W., Richter, M.M. (eds.) Information Systems and Data Analysis, pp. 308–325. Springer, Berlin (1994)

Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)

Hadj-Selem, F., Löfstedt, T., Duchesnay, E., Frouin, V., Guillemot, V.: Iterative Smoothing Algorithm for Regression with Structured Sparsity (2016, Submitted paper)

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for *f*MRI-based prediction of behavior. IEEE Trans. Med. Imaging **30**, 1328–1340 (2011)

Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**, 127–152 (2004)

Parikh, N., Boyd, S.: Proximal Algorithms. Now Publishers Inc., New York (2013)

Philippe, C., Puget, S., Bax, D.A., Job, B., Varlet, P., Junier, M.P., Andreiuolo, F., Carvalho, D., Reis, R., Guerrini-Rousseau, L., Roujeau, T., Dessen, P., Richon, C., Lazar, V., Le Teuff, G., Sainte-Rose, C., Geoerger, B., Vassal, G., Jones, C., Grill, J.: Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. PloS one **7**, 1–14 (2012)

Qin, Z., Scheinberg, K., Goldfarb, D.: Efficient block-coordinate descent algorithms for the Group Lasso. Math. Program. Comput. **5**, 143–169 (2013)

Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**, 1–30 (2005)

Schmidt, M., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization (2011). ArXiv:1109.2415

Tenenhaus, A., Tenenhaus, M.: Regularized generalized canonical correlation analysis. Psychometrika **76**, 257–284 (2011)

Tenenhaus, A., Philippe, C., Guillemot, V., Lê Cao, K.A., Grill, J., Frouin, V.: Variable selection for generalized canonical correlation analysis. Biostatistics **15**, 569–583 (2014)

van den Berg, E., Schmidt, M., Friedlander, M.P., Murphy, K.: Group sparsity via linear-time projection. Technical report TR-2008-09, Department of Computer Science, University of British Columbia, Vancouver (2008)

Vinod, H.: Canonical ridge and econometrics of joint production. J. Econom. **4**, 147–166 (1976)

Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**, 515–534 (2009)

Chapter 11
Supervised Component Generalized Linear Regression with Multiple Explanatory Blocks: THEME-SCGLR

Xavier Bry, Catherine Trottier, Fréderic Mortier, Guillaume Cornu, and Thomas Verron

Abstract We address component-based regularization of a multivariate Generalized Linear Model (GLM). A set of random responses Y is assumed to depend, through a GLM, on a set X of explanatory variables, as well as on a set T of additional covariates. X is partitioned into R conceptually homogeneous blocks $X_1, \ldots, X_R$, viewed as explanatory *themes*. Variables in each X_r are assumed many and redundant. Thus, generalized linear regression demands regularization with respect to each X_r. By contrast, variables in T are assumed selected so as to demand no regularization. Regularization is performed searching each X_r for an appropriate number of orthogonal components that both contribute to model Y and capture relevant structural information in X_r. We propose a very general criterion to measure structural relevance (SR) of a component in a block, and show how to take SR into account within a Fisher-scoring-type algorithm in order to estimate the model. We show how to deal with mixed-type explanatory variables. The method, named THEME-SCGLR, is tested on simulated data, and then applied to rainforest data in order to model the abundance of tree-species.

X. Bry (✉)
Institut Montpelliérain Alexander Grothendieck, UM2, Place Eugène Bataillon CC 051 - 34095 Montpellier, France
e-mail: xavier.bry@univ-montp2.fr

C. Trottier
Université Montpellier 3, route de Mende – 34095 Montpellier, France
e-mail: catherine.trottier@univ-montp3.fr

F. Mortier • G. Cornu
Cirad – UR Biens et Services des Ecosystèmes Forestiers tropicaux – Campus International de Baillarguet -TA C-105/D – 34398 Montpellier, France
e-mail: frederic.mortier@cirad.fr; guillaum.cornu@cirad.fr

T. Verron
ITG-SEITA – Centre de recherche SCR – 4 rue André Dessaux – 45404 Fleury-les-Aubrais, France
e-mail: Thomas.VERRON@fr.imptob.com

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_11

Keywords Component-based regularization • Generalized linear model (GLM) • Regularization

11.1 Data, Model and Problem

A set of q random responses $Y = \{y^1, \ldots, y^q\}$ is assumed to depend on p numeric regressors, partitioned into R blocks $X_1, \ldots, X_R$, with: $\forall r$, $X_r = \{x_r^1, \ldots, x_r^{p_r}\}$, plus one block T of additional covariates. X and T may include the indicator variables of nominal explanatory variables. Every X_r contains an unknown number of structurally relevant dimensions important to predict Y. Variables in T are assumed to have been selected so as to preclude redundancy, while variables in X_r's have not: T gathers all explanatory variables to be kept as such in the model, whereas dimension reduction and regularization are needed in the X_r's. Each X_r is thus to be searched for an appropriate number of orthogonal components that both capture relevant structural information in X_r and contribute to model Y.

Let $X := [X_1, \ldots, X_R]$. Each y^k is modeled through a GLM taking $X \cup T$ as regressor set. The y's are assumed independent conditional on $X \cup T$, but their linear predictors are constrained to lean on a small set of unknown common directions, which implies that all y's be included in a single model. The conceptual model stating that Y depends on $X \cup T$, and that structurally relevant dimensions should be identified in the X_r's, will be referred to as *Thematic Model* and denoted by symbolic equation: $Y = \langle X_1, \ldots, X_R; T \rangle$ (cf Fig. 11.2 for an example).

In the particular context of $R = 1$ with T empty, Bry et al. (2012, 2013) introduced a technique named Supervised Component Generalized Linear Regression (SCGLR), extending the work by Marx (1996). The basic principle of SCGLR is to replace the weighted least squares step of the Fisher Scoring Algorithm (FSA) with an extended partial least squares step. That way, component-based regularization was introduced into generalized linear regression. The interest of operating at FSA level is that, since the FSA mimics MLE, estimation weights keep consistent with the component-model being estimated. In this work, we propose to extend SCGLR by:

1. Introducing additional covariates.
2. Extending the notion of *structural relevance* of a component, so as to track various kinds of structures.
3. Extending SCGLR to the multiple-explanatory-block situation.

Notations:

Π_E := orthogonal projector on space E, with respect to some metric to be specified.
$\langle X \rangle$:= space spanned by the column-vectors of X.

11.2 Adapting the FSA to Estimate a Multivariate GLM with Partially Common Predictor

Consider that $y^1, \ldots, y^q$ depend on linear predictors, the X-parts of which are collinear:

$$\forall k = 1, \ldots, q: \ \eta_k \ = \ X\gamma_k u + T\delta_k$$

Denote component $f = Xu$. Mark that f is common to all the y's and does not depend on k. For identification, we impose $u'Au = 1$, where A may be any symmetric positive definite (p.d.) matrix. In view of the conditional independence assumption and independence of units, the likelihood is:

$$l(y|\eta) = \prod_{i=1}^{n} \prod_{k=1}^{q} l_k(y_{ki}|\eta_{ki})$$

The classical FSA in GLM's (see Nelder and Wedderburn 1972) can be viewed as an iterated weighted least squares on a linearized model, which reads here, on iteration $[t]$:

$$\forall k = 1, \ldots, q: \ z_k^{[t]} = X\gamma_k u + T\delta_k + \zeta_k^{[t]} \tag{11.1}$$

where $\zeta_k^{[t]}$ is an error term and the working variables are obtained as: $z_k^{[t]} = X\gamma_k^{[t]} u^{[t]} + T\delta_k^{[t]} + \left(\frac{\partial \eta_k}{\partial \mu_k}\right)^{[t]} (y - \mu_k^{[t]})$. Denoting g the link function, we have: $\frac{\partial \eta_k}{\partial \mu_k} = diag\left(g'(\mu_{k,i})\right)_{i=1,n}$ and $W_k = diag\left(g'(\mu_{k,i})^2 v(\mu_{k,i})\right)_{i=1,n}$, where μ and v are the expectation and variance of the corresponding GLM.

In this model, it is assumed that: $\forall k$: $E(\zeta_k) = 0$; $V(\zeta_k^{[t]}) = W_k^{[t]-1}$. In our context, model (11.1) is not linear, owing to the product $\gamma_k u$. So, it must be dealt with through an Alternated Least Squares step, estimating in turn the following two linear models:

$$z_k^{[t]} = [X\hat{u}]\,\gamma_k + T\delta_k + \zeta_k^{[t]}$$
$$z_k^{[t]} = [X\hat{\gamma}_k]\,u + T\delta_k + \zeta_k^{[t]}$$

Let $\Pi^k_{\langle f,T\rangle}$ be the projector onto $\langle f, T\rangle$ with respect to W_k. The estimation of model (11.1) may be viewed as the solution of the following program:

$$Q: \min_{f \in \langle X\rangle} \sum_k \|z_k - \Pi^k_{\langle f,T\rangle} z_k\|^2_{W_k}$$
$$\Leftrightarrow Q': \max_{u'Au=1} \psi(u)\,,$$

$$\text{where} \qquad \psi(u) = \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k\,,\ \langle Xu, T\rangle) \tag{11.2}$$

In order to later deal with multiple X_r's, we have yet to replace Q' by another equivalent program:

$$Q'' : \max_{\forall j,\, u_j' A_j u_j = 1} \psi(u_1, \ldots, u_R)$$

where $A_1, \ldots, A_R$ are any given p.d. matrices, and

$$\psi(u_1, \ldots, u_R) = \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k\ ,\ \langle X_1 u_1, \ldots, X_R u_R, T\rangle) \tag{11.3}$$

$\psi(u_1, \ldots, u_R)$ is a goodness-of-fit measure, now to be combined with some structural relevance measure to get regularization.

11.3 Structural Relevance (SR)

Consider a given weight matrix W, e.g. $W = n^{-1} I_n$, reflecting the a priori importance of units. Let X be an $n \times p$ variable block endowed with a $p \times p$ metric matrix M, the purpose of which is to "weight" variables appropriately (informally, PCA of (X, M, W) must be relevant, see Sect. 11.4.3.2 for details). Component $f = Xu$ is constrained by: $\|u\|^2_{M^{-1}} = 1$ (M^{-1} will thus be our choice of the aforementioned matrix A). We may consider various measures of SR, according to the type of structure we want f to align with.

Definition 11.1. Given a set of J symmetric positive semi-definite (p.s.d.) matrices $N = \{N_j\,;\, j = 1, \ldots, J\}$, a weight system $\Omega = \{\omega_j\,;\, j = 1, \ldots, J\}$, and a scalar $l \geq 1$, we define the associated SR measure as:

$$\phi(u) := \left(\sum_{j=1}^{J} \omega_j (u' N_j u)^l \right)^{\frac{1}{l}}$$

Various particular measures can be recovered from this general formula.

Example 11.1. Component Variance:

$$\phi(u) = V(Xu) = \|Xu\|^2_W = u'(X'WX)u\ ,$$

implying $J = 1$, $\omega_1 = 1$ and $N_1 = X'WX$. This quantity is obviously maximized by the first eigenvector in the PCA of (X, M, W).

Example 11.2. Variable Powered Inertia (VPI): We impose $\|f\|^2_W = 1$ through $M = (X'WX)^{-1}$. For a block X consisting of p standardized numeric variables x^j:

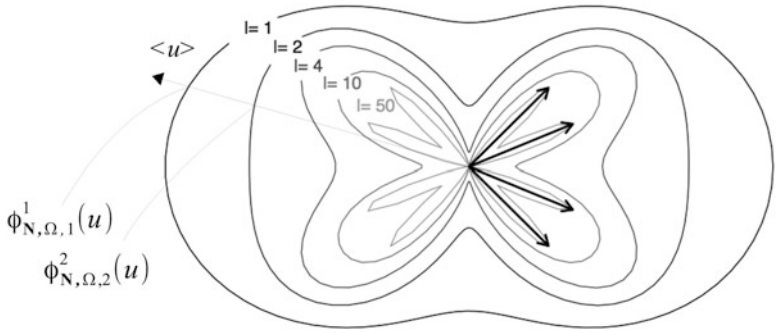

Fig. 11.1 Polar representation of the Variable Powered Inertia according to the value of l

$$\phi(u) = \left(\sum_{j=1}^{p} \omega_j \rho^{2l}(Xu, x^j)\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} \omega_j (u'X'Wx^j x^{j'}WXu)^l\right)^{\frac{1}{l}},$$

implying $J = p$ and $\forall j,\ N_j = X'Wx^j x^{j'}WX$.

For a block X consisting of p categorical variables X^j, each of which is coded through the set of its centered indicator variables (less one to avoid singularity of $X^{j'}WX^j$), we take:

$$\phi(u) = \left(\sum_{j=1}^{p} \omega_j \cos^{2l}(Xu\,,\,\langle X^j\rangle)\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} \omega_j \langle Xu|\Pi_{X^j}Xu\rangle_W^l\right)^{\frac{1}{l}},$$

where: $\Pi_{X^j} = X^j(X^{j'}WX^j)^{-1}X^{j'}W$. Here, we have $J = p$ and $\forall j,\ N_j = X'W\Pi_{X^j}X$.

VPI is the measure we stick to, from here on. The role of l is easy to understand in the case of numerical variables. For $l = 1$, we get the part of X's variance captured by component f, which is also maximized by the first eigenvector in the PCA of (X, M, W). More generally, tuning parameter l allows to draw components towards more (greater l) or less (smaller l) local variable bundles. Figure 11.1 graphs $\phi^l(u)$ in polar coordinates ($z(\theta) = \phi^l(e^{i\theta})e^{i\theta}$; $\theta \in [0, 2\pi]$) for various values of l in the elementary case of 4 coplanar variables x with $\forall j, \omega_j = 1$. Note that $\phi^l(u)$ was graphed instead of $\phi(u)$ so that curves would be easier to distinguish. One can see how the value of l tunes the locality of bundles considered.

11.4 THEME-SCGLR

We shall first consider the simpler case of a single explanatory block ($R = 1$), and then turn to the general case.

11.4.1 Dealing with a Single Explanatory Block

In this sub-section, we consider the thematic model $Y = \langle X; T\rangle$.

11.4.1.1 The Criterion and Program

In order to regularize the regression corresponding to program Q' at step k of the FSA, we consider program:

$$R: \max_{u'M^{-1}u=1} S(u) \qquad \text{with} \quad S(u) = \psi(u)^{1-s}\phi^s(u) \tag{11.4}$$

where $\psi(u)$ is given by (11.2) and s is a parameter tuning the relative importance of the SR with respect to the goodness of fit. Taking $s = 0$ equates the criterion with the goodness of fit, while at the other end, taking $s = 1$ equates it with the mere SR. The product form of the criterion is a straightforward way to make the solution insensitive to "size effects" of $\phi(u)$ and $\psi(u)$.

11.4.1.2 Analytical Expression of $S(u)$

$$\langle Xu, T\rangle = \langle \tilde{X}u, T\rangle$$

with $\tilde{X} := \Pi_{T^\perp} X$

$$\langle \tilde{X}\rangle \perp \langle T\rangle \;\Rightarrow\; \Pi_{\langle \tilde{X}u,T\rangle} = \Pi_{\langle \tilde{X}u\rangle} + \Pi_{\langle T\rangle}$$

$$\Rightarrow \quad \cos^2_{W_k}(z_k\,,\,\langle \tilde{X}u, T\rangle) = \frac{1}{\|z_k\|^2_{W_k}}\left(\langle z_k|\Pi_{\langle \tilde{X}u\rangle} z_k\rangle_{W_k} + \langle z_k|\Pi_{\langle T\rangle} z_k\rangle_{W_k}\right)$$

Now:

$$\langle z_k|\Pi_{\langle \tilde{X}u\rangle} z_k\rangle_{W_k} = z_k' W_k \Pi_{\langle \tilde{X}u\rangle} z_k = \frac{u'\tilde{X}' W_k z_k z_k' W_k \tilde{X}u}{u'\tilde{X}' W_k \tilde{X}u}$$

Let: $A_k := \frac{\tilde{X}' W_k z_k z_k' W_k \tilde{X}}{\|z_k\|^2_{W_k}}$; $B_k := \tilde{X}' W_k \tilde{X}$; $c_k := \frac{\langle z_k|\Pi_{\langle T\rangle} z_k\rangle_{W_k}}{\|z_k\|^2_{W_k}}$. We have:

$$\psi(u) = \sum_k \left(\frac{u'A_k u}{u'B_k u} + c_k\right) \tag{11.5}$$

From (11.1) and (11.5), we get the general matrix form of $S(u)$.

11.4.1.3 Rank 1 Component

THEME-SCGLR's rank 1 component is obtained by solving program (11.4) instead of performing the current step of the modified FSA used to estimate the multivariate GLM of Sect. 11.2. We give an algorithm to maximize, at least locally, any criterion on the unit-sphere: the Projected Iterated Normed Gradient (PING) algorithm (cf. appendix). For component 1, PING is used with $D = 0$.

11.4.1.4 Rank $h > 1$ Component

The role of each extra-component must be clear. We adopt the local nesting principle (LocNes) presented in Bry et al. (2012). Let $F^h := \{f^1, \ldots, f^h\}$ be the set of the first r components. According to LocNes, extra component f^{h+1} must best complement the existing ones plus T, i.e. $T^h := F^h \cup T$. So f^{h+1} must be calculated using T^h as a block of extra-covariates. Moreover, we must impose that f^{h+1} be orthogonal to F^h, i.e.:

$$F^{h\prime} W f^{h+1} = 0 \tag{11.6}$$

To ensure (11.6), we add it to program (11.4). To calculate component $f^{h+1} = Xu$, we would now solve:

$$R : \max_{\substack{u'M^{-1}u=1 \\ D^{h\prime}u=0}} S(u)$$

where $D^h := X'WF^h$. Again, the PING algorithm allows to solve this program.

11.4.2 Dealing with $R > 1$ Explanatory Blocks

Consider now the complete thematic equation: $Y = \langle X_1, \ldots, X_R\,;\, T\rangle$

11.4.2.1 Rank 1 Components

Estimating the multivariate GLM of Sect. 11.2 led to currently solving program Q''. Introducing SR in it, we will now solve:

$$R'' : \max_{\forall r,\, u_r' M_r^{-1} u_r = 1} \psi(u_1, \ldots, u_R)^{1-s} \prod_{r=1}^{R} \phi^s(u_r) \tag{11.7}$$

where $\psi(u_1, \ldots, u_R)$ is given by (11.3). Equation (11.7) can be solved by iteratively maximizing in turn the criterion on every u_r. Now, we have:

$$\forall r: \ \cos^2_{W_k}(z_k \, , \, \langle X_1 u_1, \ldots, X_R u_R, T\rangle) = \cos^2_{W_k}(z_k \, , \, \langle X_r u_r, \tilde{T}_r\rangle)$$

where $\tilde{T}_r = T \cup \{f_s; s \neq r\}$. So, (11.7) can be solved by iteratively solving:

$$R''_r : \max_{u'_r M_r^{-1} u_r = 1} \psi(u_r)^{(1-s)} \phi^s(u_r)$$

using $\tilde{T}_r$ as additional covariates. Section 11.4.1 already showed how to solve this program.

11.4.2.2 Rank $h > 1$ Components

Suppose we want H_r components in X_r. $\forall r \in \{1, \ldots, R\}, \forall l < H_r$, let $F_r^l := \{f_r^h ; h = 1, \ldots, l\}$. LocNes states that f_r^{h+1} must best complement the "existing" components (by "existing", we mean components with rank $< h+1$ ones in X_r plus *all* components of all other blocks) plus T, i.e.: $T_r^h := F_r^h \cup_{s \neq r} F_s^{H_s} \cup T$. So, the current value of f_r^{h+1} is calculated solving:

$$R_r^{h\prime\prime} : \max_{\substack{u'_r M_r^{-1} u_r = 1 \\ D_r^{h\prime} u_r = 0}} \psi(u_r)^{(1-s)} \phi^s(u_r)$$

where $D_r^h := X'_r W F_r^h$ and taking T_r^h as additional covariates.

Informally, the algorithm consists in currently calculating all H_r components in X_r as done in Sect. 11.4.1, taking $T \cup_{s \neq r} F_s^{H_s}$ as extra-covariates—and then loop on r until overall convergence of the component-system is reached.

11.4.3 Further Issues

11.4.3.1 Models with Offset

In count data, units may not have the same "size". As a consequence, the corresponding variables may not have the same offset. Models with offset call for elementary developments, which are not included here.

11.4.3.2 Dealing with Mixed-Type Covariates

In practice, covariates are most often a mixture of numeric and categorical variables. This situation is dealt with by adapting matrix M. Consider a particular block $X = \left[x^1, \ldots, x^K, X^1, \ldots, X^L\right]$ (the block-index is omitted here), where: $x^1, \ldots, x^K$ are column-vectors coding the numeric regressors, and $X^1, \ldots, X^L$ are blocks of centered indicator variables, each block coding a categorical regressor (X^l has $q_l - 1$ columns if the corresponding variable has q_l levels, the removed level being taken as "reference level"). In order to get a relevant PCA of (X, M, W), we must consider the metric block-diagonal matrix:

$$M := diag\left\{(x^{1\prime}Wx^1)^{-1}, \ldots, (x^{K\prime}Wx^K)^{-1}, (X^{1\prime}WX^1)^{-1}, \ldots, (X^{L\prime}WX^L)^{-1}\right\}$$

The regressor matrix is then transformed as follows: $\tilde{X} = XM^{\frac{1}{2}}$ and $\tilde{X}$ is used in THEME-SCGLR in place of X.

11.4.3.3 Coefficients of Original Variables in Linear Predictors

Let $\tilde{X} := [\tilde{X}_1, \ldots, \tilde{X}_R]$ and M be the block-diagonal matrix having $(M_r)_{r=1,\ldots,R}$ as diagonal blocks. Once the components f_r^h have been calculated, a generalized linear regression of each y^k is performed on $[F, T]$, where $F := \{F_r^{H_r}\}_{1 \le r \le R}$, yielding linear predictor: $\eta_k = \theta_k + T\delta_k + F\gamma_k = \theta_k + T\delta_k + \tilde{X}U\gamma_k = \theta_k + T\delta_k + X\beta_k$, where $\beta_k = M^{\frac{1}{2}}U\gamma_k$.

11.5 Model Assessment

11.5.1 Principle

Assessment of a model **M** is based on its predictive capacity on a test-sample in a cross-validation routine. The latter uses an error indicator e suitable to each response-type. It is measured on and averaged over test-samples, yielding an average cross-validation error rate $CVER(\mathbf{M})$ allowing to compare models.

11.5.2 Error Indicators

To every type of y may correspond one or more error indicators. For instance, for a binary output $y \sim B(p(x,t))$, AUC denoting the corresponding area under ROC curve, we would advise to take:

$$e = 2(1 - AUC)$$

Whereas for a quantitative variable, we had rather consider indicators based on the mean quadratic error, such as:

$$e = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{E}(y_i|x_i, t_i))^2}{\hat{V}(y_i|x_i, t_i)}$$

But these error indicators are not necessarily comparable across the y's, and must yet be pooled into an overall indicator. We propose to use *geometric* averaging, since it allows *relative* compensations of indicators.

11.5.3 Backward Component Selection

Let $\mathbf{M}(h_1, \ldots, h_R)$ denote the model of Y based on h_1 (resp. $\ldots h_R$) components in X_1 (resp. $\ldots X_R$). Starting with "large enough" numbers of components in every block allows to better focus on components *proper* effects, minimizing the risk of confusion between effects. But, to ensure having "large enough" numbers, one should start with "too large" ones, hence an over-fitting model. So, some high-rank components should be removed. This is enabled by LocNes, in that it makes every component complement all lower-rank ones in its block, and all components in other blocks. Thus, every new component should improve the overall quality of prediction of the y's, unless it contributes to over-fitting. Consider the loss in $CVER(\mathbf{M}(h_1, \ldots, h_R))$ related to the highest rank component in X_r: $f_r^{h_r}$. It is measured through:

$$\Gamma(r, h_r) = CVER(\mathbf{M}(h_1, \ldots, h_r - 1, \ldots, h_R)) - CVER(\mathbf{M}(h_1, \ldots, h_r, \ldots, h_R))$$

11.5.3.1 Backward Selection Algorithm

Starting with too large component numbers $\{H_r\}_r$, we consider in turn the removal of every higher rank component in every block. We remove the one with the higher $\Gamma(r, h_r)$. This is iterated until $\Gamma(r, h_r)$ becomes negative.

11.5.4 Model Selection

Models not only differ by the number of components in each block, but also by the choice of SR. Let us first split the observation sample S into two subsamples S_1 and S_2. S_1 has to be large relative to S_2, because S_1 is used to determine (calibrate, test

and validate) the best model for each choice of *SR* and select the one leading to the smallest error, when S_2 is only used to validate this best *SR*.

Consider a set $SSR = \{s_1, \ldots, s_L\}$ of SR measures. Given S_1, one gets for each $s \in SSR$, through backward selection, a sequence of nested models, the CVER of which are calculated. The model $\mathbf{M}^*(s)$ exhibiting the lowest value is selected. Then, $\mathbf{M}^*(s)$ is used to predict the y's on validation sample V and its average error rate (*AER*) is calculated on V. $\mathbf{M}^*(s)$ is validated when this *AER* is close enough to its *CVER*. *CVER*'s of all $\mathbf{M}^*(s)$ are then compared and the value s^* leading to the best performance is selected. Finally, $\mathbf{M}^*(s^*)$ is validated on S_2.

11.6 Applications to Data

We shall first sum up the results of tests performed on data simulated so as to emphasize the role of parameters. Then, we shall describe an application to rainforest-data.

11.6.1 Tests on Simulated Data

We considered $n = 100$ units, and thematic model given by:

$$Y = \langle X_1, X_2, X_3; T \rangle \tag{11.8}$$

Each X_r contained 3 variable-bundles of tunable width: B_r^1, B_r^2, B_r^3, respectively structured about 3 latent variables a_r^1, a_r^2, a_r^3, having tunable angles. Moreover, each X_r contained a large number of noise-variables. Only a_r^1, a_r^2 played any role in the model of Y, so that B_r^3 be a nuisance-bundle, with as many variables in it as to "make" the block's first PC by itself. The role of a_r^1 was made more important than that of a_r^2, so that every f_r^1 should align to a_r^1. Every X_r was made to contain 100 variables. T was made of a random categorical variable having 3 levels. Y contained 50 conditionally independent indicator variables, known to be the worst type in GLM-estimation.

The simulation study led to no convergence problem except when the $\langle a_r^1, a_r^2 \rangle$'s were much too close *between* blocks, which is only fair, since the influences of blocks can then theoretically not be separated. It demonstrated that the estimation results are not very sensitive to s, except in the vicinity of values 0 and 1. It also showed that l is of paramount importance to identify the truly explanatory bundles: $l = 1$ tends to make f_r^1 very close to PC1 (so, a_r^3) in X_r, whereas taking $l \geq 2$ allows f_r^1 to focus on a_r^1.

11.6.2 Application to Rainforest Data

We considered $n = 1000$ $8 \times 8\,\text{km}^2$ plots sampled in the Congo Basin rainforests, and divided it 5 times into 800 plots for calibration and 200 for prediction and cross-validation. Responses Y were counts of $q = 27$ common tree species. Each count was assumed to be Poisson-distributed conditional on 41 covariates, the plot's surface standing as offset. Covariates were partitioned into 3 sets: one containing all geographic variables (topography and climate), one containing satellite measures of photosynthetic activity over a year, and finally, an extra-covariate: the geologic type of the plot (cf. Fig. 11.2).

With $l = 1$ and $s = 1/2$ (even balance between GoF and SR), 2 components were found necessary in both X_1 and X_2 to model Y. While components in X_1 are easy to interpret in terms of rain-patterns, components in X_2 are not (cf. Fig. 11.3).

It appears on Fig. 11.3 that, in X_2, components may have been "trapped" by PC's, so, we raised l to 4. The new components are shown on Fig. 11.4. It appears that one

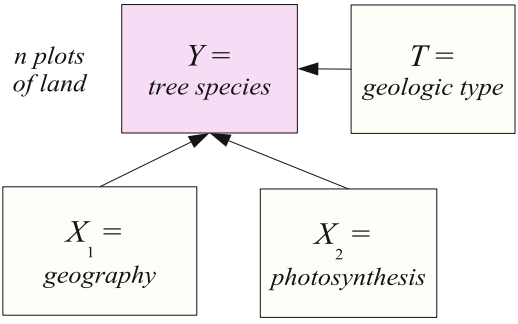

Fig. 11.2 Thematic model of tree species in the Congo basin

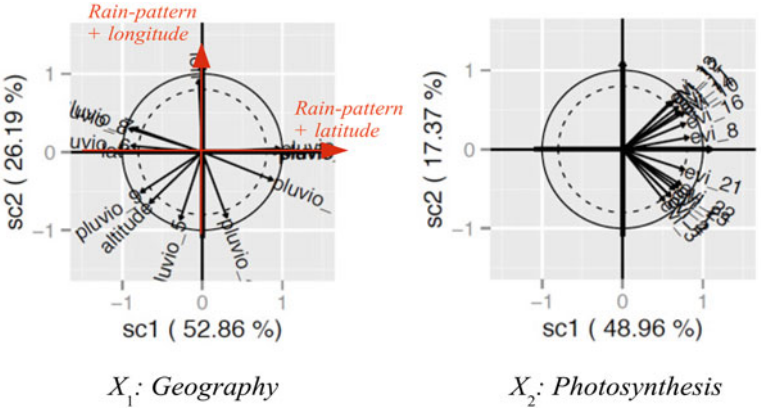

Fig. 11.3 Correlation scatterplots of the blocks' first 2 components for $l = 1$

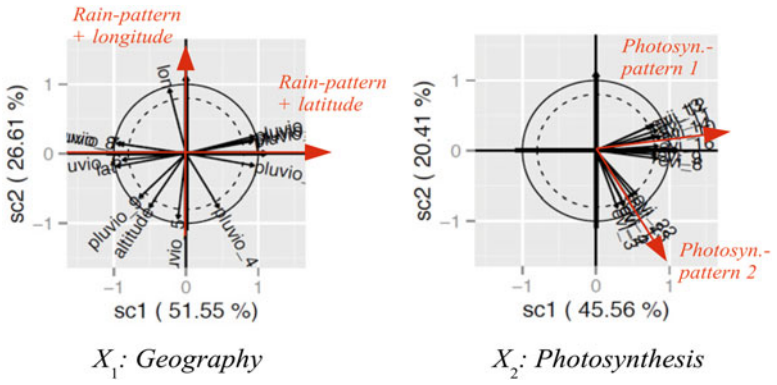

Fig. 11.4 Correlation scatterplots of the blocks' first 2 components for $l = 4$

photosynthetic pattern is more important than the other (even if ultimately, they are both important), and the corresponding bundle attracts f_2^1, letting the other bundle attract f_2^2. The model obtained with $l = 4$ also having a lower *CVER*, it was retained as the final model.

11.7 Conclusion

THEME-SCGLR is a powerful trade-off between Multivariate GLM estimation (which cannot afford many and redundant covariates) and PCA-like methods (which take no explanatory model into account). Given a thematic model of the phenomenon under attention, it provides robust predictive models based on interpretable components. It also allows, through the exploration facilities it offers, to gradually refine the design of the thematic model.

Acknowledgements This research was supported by the CoForChange project (http://www.coforchange.eu/) funded by the ERA-Net BiodivERsA with the national funders ANR (France) and NERC (UK), part of the 2008 BiodivERsA call for research proposals involving 16 European, African and international partners including a number of timber companies (see the list on the website, http://www.coforchange.eu/partners), and by the CoForTips project funded by the ERA-Net BiodivERsA with the national funders FWF (Austria), BelSPO (Belgium) and ANR (France), part of the 2011–2012 BiodivERsA call for research proposals (http://www.biodiversa.org/519).

Appendix: The Projected Iterated Normed Gradient (PING) Algorithm

Consider program:

$$\max_{\substack{u'M^{-1}u=1 \\ D'u=0}} h(u)$$

Putting $v = M^{-1/2}u$, $g(x) = h(M^{1/2}x)$ and $C = M^{-1/2}D$, this is strictly equivalent to:

$$R_C : \max_{\substack{v'v=1 \\ C'v=0}} g(v)$$

Applying the first order conditions to the Lagrangian, we get that the solution satisfies the stationary equation:

$$v = \frac{\Pi_{C^\perp}\Gamma(v)}{\|\Pi_{C^\perp}\Gamma(v)\|}$$

where $\Pi_{C^\perp} := I - C(C'C)^{-1}C'$. This gives the basic iteration of the ING algorithm:

$$\begin{aligned} m_{[t+1]} &= \frac{\Pi_{C^\perp}\Gamma(v_{[t]})}{\|\Pi_{C^\perp}\Gamma(v_{[t]})\|} \\ v_{[t+1]} &= m_{[t+1]} \end{aligned} \tag{11.9}$$

It can be shown that this iteration follows a direction of ascent. Now, picking a point on a direction of ascent does not guarantee that g actually increases, since one may "go too far". But staying "close enough" to the current starting point on the arc $(v_{[t]}, m_{[t+1]})$ guarantees that g increases. We may thus replace (11.9) with $v_{[t+1]} = \arg\max_{v\in(v_{[t]},m_{[t+1]})} g(v)$, the search for this maximum being obtained through a unidimensional maximization procedure.

References

Bry, X., Trottier, C., Verron, T., Mortier, F.: Supervised component generalized linear regression using a PLS-extension of the fisher scoring algorithm. In: COMPSTAT 2012 Proceedings, Limassol (2012)

Bry, X., Trottier, C., Verron, T., Mortier, F.: Supervised component generalized linear regression using a PLS-extension of the fisher scoring algorithm. J. Multivar. Anal. **119**, 47–60 (2013)

Marx, B.D.: Iteratively reweighted partial least squares estimation for generalized regression. Technometrics **38**(4), 374–381 (1996)

Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Stat. Soc.: Ser. A **135**, 370–384 (1972)

Chapter 12
Partial Possibilistic Regression Path Modeling

Rosaria Romano and Francesco Palumbo

Abstract This paper introduces structural equation modeling for imprecise data, which enables evaluations with different types of uncertainty. Coming under the framework of *variance-based* analysis, the proposed method called *Partial Possibilistic Regression Path Modeling* (PPRPM) combines the principles of PLS path modeling to model the network of relations among the latent concepts, and the principles of possibilistic regression to model the vagueness of the human perception. Possibilistic regression defines the relation between variables through possibilistic linear functions and considers the error due to the *vagueness* of human perception as reflected in the model via interval-valued parameters. PPRPM transforms the modeling process into minimizing components of uncertainty, namely randomness and vagueness. A case study on the motivational and emotional aspects of teaching is used to illustrate the method.

Keywords Structural equation modeling (SEM) • Possibilistic regression (PR) • Partial possibilistic regression path modeling (PPRPM)

12.1 Introduction

Structural equation models (SEMs) include various statistical methodologies that aim to estimate a network of causal relationships among latent variables (LVs) defined by blocks of manifest variables (MVs) (Bollen 1989). The SEM research paradigm is grounded on psychometric (covariance-based, CBSEM) and chemometric research tradition (variance-based, VBSEM). With increasing popularity in several areas, under the *variance-based* framework estimation methods (Wold 1975), Partial Least Squares Path Modeling (PLSPM) represents a statistical

R. Romano (✉)
University of Calabria, Cosenza, Italy
e-mail: rosaria.romano@unical.it

F. Palumbo
University of Naples Federico II, Naples, Italy
e-mail: fpalumbo@unina.it

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_12

approach to SEM (Tenenhaus et al. 2005). PLSPM formulates the causality dependencies between LVs in terms of linear conditional expectation and estimates the LVs through a system of interdependent equations based on simple/multiple regressions.

As with classical least squares regression, in PLSPM the process of data analysis is represented by the simple equation: $data = model + error$ (Judd and McClelland 2009). In a very general definition, *error*, which is also called *uncertainty*, refers to the information that is not explained by the model itself. In statistical thinking the error component corresponds to *randomness*, and is related to the natural variability of the analyzed phenomena. However, there are other sources of uncertainty besides randomness (Coppi 2008): for example, human judgments are subjective measurements that are generally affected by *vagueness* (Zadeh 1973).

Vagueness characterizes phenomena that are vague in their own nature, which means they have no objective measurement scales. Indeed, concepts such as satisfaction, trust, happiness, stress, etc. well define the underlying phenomenon; yet they cannot be quantified. Research methodology generally defines these concepts in terms of LVs measured through MVs that are equally measured on subjective perception scales. Different approaches have been proposed to cope with *vagueness* in regression analysis. For the sake of simplicity they can be grouped into two broad categories: Fuzzy Least Square Regression (FLSR) and Possibilistic Regression (PR). Two papers can be considered seminal for each of them, while many others have proposed further developments. Diamond's papers (1988; 1990) introduced the FLSR approach (see also Coppi et al. 2006), which is closer to the traditional statistical approach. In fact, following the Least Squares line of thought, the aim is to minimize the distance between the observed and the estimated fuzzy data. This approach has been extended to interval data analysis (Blanco-Fernndez et al. 2011; Billard and Diday 2000; Marino and Palumbo 2002) and to symbolic data analysis (see Lima Neto and de Carvalho 2010). The paper by Tanaka et al. (1982) and that by Tanaka (1987) introduced the PR approach. For an exhaustive overview of possibilistic data analysis, we refer the reader to the book by Tanaka and Guo (1999). In PR the error term is embedded in the interval parameters that model the *vagueness* in the relation among the variables and the solutions are defined through numerical optimization.

In a previous work, following the PLSPM approach, Romano and Palumbo (2013) proposed a new method termed Partial Possibilistic Regression Path Modeling (PPRPM). PPRPM aims to explain at best the residual variance in any regression inside the model, but it is based on the use of PR to model relations among the LVs. This paper shows how PPRPM can properly gear *randomness* as well *vagueness* in path models.

PPRPM is a method to analyze phenomena whose description requires the analysis of a complex structure of relations among the variables inside the system, and where there is an additional source of complexity arising from the involvement of influential human beings. This is achieved by combining the principles of PLSPM (Tenenhaus et al. 2005) and PR (Tanaka and Guo 1999). Such a combination was proposed by Palumbo and Romano (2008) and Palumbo et al. (2008). The novelty

of PPRPM consists in the use of quantile regression (Koenker and Basset 1978; Davino et al. 2013) to model the relations between each LV and its respective block of indicators. This choice of combining PR and quantile regression allows us to have a robust measure of the latent variables (measurement model), on the one hand, and to take into account the imprecision inherent in systems where human estimation is influential and the observations cannot be described accurately, on the other.

Under the Judd and McClelland paradigm 2009, VBSEM considers the error component as the sum of the error due to MVs, termed *measurement model error* and the error due to LVs, termed *structural model error*. VBSEM reaches the solution by alternately minimizing the two components, whereas CBSEM focuses on the whole covariance structure considering the error as a whole.

Like PLSPM, the PPRPM approach independently considers the measurement model error and structural model error. However, it assumes that the *randomness* can be referred to the measurement model and the *vagueness* to the structural model and it uses different methods to minimize the error. In this partial approach it is assumed that the randomness component can only be referred to the measurement model, whereas the uncertainty component is part of the model itself.

In PPRPM the process of data analysis is represented by the equation: *data* = *possibilisticmodel* + *randomness*. Unlike the classical statistical paradigm, where only *randomness* is considered an additional element to the deterministic relation among the variables, PPRPM also considers *vagueness* as being reflected in the structural interval-valued model parameters.

In the following, we will first introduce the PR and SEM, and then we will present the basic PPRPM algorithm. A case study on a meta-cognitive questionnaire for teachers will be illustrated. The paper will end with the main conclusions.

12.2 Possibilistic Regression

The purpose of PR is to explain a dependent variable as an interval output in terms of the variation of explanatory variables. Specifically, PR defines the relation between one dependent variable $\mathbf{y}$ and a set of M predictors $\mathbf{x}_1, \mathbf{x}_m, \ldots, \mathbf{x}_M$, observed on N statistical units, through a linear function holding interval valued coefficients

$$\mathbf{y} = \tilde{\omega}_1 \mathbf{x}_1 + \ldots + \tilde{\omega}_m \mathbf{x}_m + \ldots + \tilde{\omega}_M \mathbf{x}_M, \tag{12.1}$$

where $\tilde{\omega}_m$ denotes the generic interval-valued coefficient. Interval-valued coefficients are defined in terms of midpoint and spread: $\tilde{\omega}_m = \{c_m; a_m\}$ and will be referred to as interval coefficients in the rest of the paper. There are no restrictive assumptions on the model. Unlike statistical regression, the deviations between data and linear models are assumed to depend on the vagueness of the parameters and not on measurement errors (Kim et al. 1996). This means that in PR there is no external error component but the spread of the coefficients embeds all uncertainty, such that PR minimizes the total spread of the interval coefficients

$$\min_{a_m} \sum_{m=1}^{M} \left(\sum_{n=1}^{N} a_m |x_{nm}| \right), \ \forall\, m = 1, \ldots, M, \tag{12.2}$$

under the following linear constraints

$$\begin{aligned} &\sum_{m=1}^{M} c_m x_{nm} + (1-\alpha) \sum_{m=1}^{M} a_m |x_{nm}| \geq y_n \ , \\ &\sum_{m=1}^{M} c_m x_{nm} - (1-\alpha) \sum_{m=1}^{M} a_m |x_{nm}| \leq y_n \ , \qquad \forall n = 1, \ldots, N, \end{aligned} \tag{12.3}$$

satisfying the following conditions: $(i)\ a_m \geq 0$; $(ii)\ c_m \in R$; $(iii)\ x_{n1} = 1$.

Constraints in (12.3) guarantee the inclusion of the whole given data set in the estimated boundaries. The degree of possibility α is a subjective measure that depends on the context: increasing the α coefficient expands the estimated intervals (see Tanaka and Guo 1999 on the choice of α).

Wang and Tsaur (2000) provided a suitable interpretation of the regression interval. The basic idea was to find a representative value of the interval among the infinite values enclosed within the interval boundaries. Let $\underline{y}_n$ and $\overline{y}_n$ be the lower and upper bound of the estimated value $\tilde{y}_n^*$. The authors proved that in models with symmetric coefficients the mean value of $\tilde{y}_n^*$ is given by

$$\check{y}_n = \frac{\underline{y}_n + \overline{y}_n}{2},$$

and that it is equal to the value occurring with the higher possibility level ($\alpha = 1$) denoted by $\tilde{y}_n^1$. In other words, $\tilde{y}_n^1$ is the best representative value of the possibilistic interval and, more generally, the regression line $\tilde{Y}^1$ has the best ability to interpret the given data. Starting from these results the following quantities were defined.

- *Total Sum of Squares* (SST)
  a measure of the total variation of y_n in $\tilde{y}_n^*$

$$SST = \sum_{n=1}^{N} \left(y_n - \underline{y}_n \right)^2 + \sum_{n=1}^{N} \left(\overline{y}_n - y_n \right)^2 \tag{12.4}$$

- *Regression Sum of Squares* (SSR)
  a measure of the variation of $\tilde{y}_n^1$ in $\tilde{y}_n^*$

$$SSR = \sum_{n=1}^{N} \left(\tilde{y}_n^1 - \underline{y}_n \right)^2 + \sum_{n=1}^{N} \left(\overline{y}_n - \tilde{y}_n^1 \right)^2 \tag{12.5}$$

- *Error Sum of Squares* (SSE)
  an estimate of the difference when $\tilde{y}_n^1$ is used to estimate y_n

$$SSE = 2 \sum_{n=1}^{N} \left(\tilde{y}_n^1 - y_n \right)^2 \tag{12.6}$$

Thus, using (12.4) and (12.5), an index of confidence is built, which is similar to the traditional R^2 in statistics. The index is defined as: IC=SSR/SST, with

$0 \leq IC \leq 1$, and gives a measure of the variation of Y between $\underline{Y}$ and $\overline{Y}$. The higher the IC, the better the $\tilde{Y}^1$ used to represent the given data. A high value of IC means that a well estimated PR is modeled and can support a better prediction.

12.3 Modeling Uncertainty in Structural Equation Modeling

SEMs allow simultaneous use of both latent and observed variables within one framework. The basic structural equation model can be described as

$$\mathbf{y} = \Lambda_y \eta + \epsilon, \tag{12.7a}$$

$$\mathbf{x} = \Lambda_x \xi + \delta, \tag{12.7b}$$

$$\eta = B\eta + \Gamma\xi + \zeta, \tag{12.7c}$$

where $\mathbf{y}$ is a $(p \times 1)$-dimensional vector containing p endogenous observed variables, $\mathbf{x}$ is $(q \times 1)$-dimensional vector with q exogenous observed variables, η is an $(r \times 1)$-dimensional vector containing r endogenous latent variables, ξ is an $(s \times 1)$-dimensional vector containing s exogenous latent variables; ϵ and δ are error vectors, respectively, in $(p \times 1)$ dimensions and $(q \times 1)$ dimensions, and ζ is a residual vector of $(r \times 1)$ dimensions; Λ_x and Λ_y are respectively loading matrices in $(p \times r)$ and $(q \times s)$ dimensions, and B and Γ are respectively coefficient matrices of $(r \times r)$ and $(r \times s)$ dimensions. Both, Eqs. (12.7a) and (12.7b) form the measurement equation (also referred to as outer relations or measurement model), and Eq. (12.7c) is called as the structure equation (also referred to as inner relation or structural model). Focusing on the error terms, ζ represents the error in the inner relations, i.e. disturbance in the prediction of the endogenous latent variables from their respective explanatory latent variables, whereas ϵ and δ represent imprecision in the measurement process. Let us denote Φ as the covariance matrix $(s \times s)$ of ξ, Ψ as the covariance matrix $(r \times r)$ of ζ, and Θ_ϵ $(p \times p)$ and Θ_δ $(q \times q)$ are respectively covariance matrices of ϵ and δ. Let θ be the unknown parameter vector including $\Lambda_x, \Lambda_y, B, \Gamma, \Phi, \Psi, \Theta_\epsilon, \Theta_\delta$, which is estimated in the modeling process. If the assumed model (see Eqs. in 12.7) is true, in the sense of explaining the covariation of all the indicators, its inherent population covariance matrix $\Sigma(\theta)$ shall be equal to the population covariance matrix of manifest variables denoted by Σ, [i.e., $\Sigma(\theta) = \Sigma$]. Because Σ is unknown, it is usually replaced by the empirical covariance matrix C. As a consequence, the modeling process of SEM is converted into the estimation of unknown parameter θ in $\Sigma(\theta)$. Typically using a Maximum Likelihood (*ML*) function, the covariance-based procedure provides optimal estimations of the model parameters under the assumptions that indicators

follow a multivariate normal distribution and that observations are independent of one another. The parameters are estimated by minimizing the following *discrepancy function*

$$F_{ML} = \log|\Sigma(\theta)| + \text{trace}\,(C\Sigma^{-1}(\theta)) - \log|C| - (p+q). \tag{12.8}$$

Since in *covariance-based* approaches there is a unique minimization function related to the ability of the model to reproduce the sample covariance matrix, it is possible to have a global measure of fit that is defined as

$$\min(C - \Sigma(\theta)). \tag{12.9}$$

CBSEMs consider the three residual terms ϵ, δ, and ζ in a unique minimization problem such that all the parameters are estimated simultaneously.

In VBSEMs the three residual terms ϵ, δ, and ζ play a crucial role in the modeling process. In practice, PLSPM aims to minimize the sum of residual variances of all the dependent variables in the model, both latent and observed ones, rather than explain the covariance structure of all the indicators. Hence, PLSPM is more strongly oriented to prediction than to parameter estimation. The logic behind the PLSPM is to partially estimate parameters by minimizing in each step of the procedure a residual variance with respect to a subset of the parameters being estimated given proxies or fixed estimates for other parameters (Chin 1998). For this reason PLSPM uses a three-stage estimation algorithm: first it performs an iterative scheme of simple/multiple regressions until the solution converges to a set of weights that are used for estimating the latent variables scores, and then uses these scores for obtaining loadings and path coefficients, using OLS regressions. PLSPM lacks a global optimization criterion but separately minimizes the following residual variances

$$\min\left(\text{trace}\,(\Theta_\epsilon); \text{trace}\,(\Theta_\delta); \text{trace}\,(\Psi)\right). \tag{12.10}$$

PPRPM differs from both CBSEM and VBSEM in that elements in coefficient matrices, (i.e., B and Γ in Eq. 12.7c) are interval-valued, yet vector residual ζ is no longer covered in the model. PPRPM treats differently the vagueness in the prediction of the LVs (error term in the structural model) and the imprecision in the measurement of MVs (error term in the measurement model). The first type of error is assumed to depend on the indefiniteness/vagueness of the parameters which govern the system structure, not on its measurement errors. PPRPMs give rise to possibilistic regressions that account for the imprecise nature or vagueness in our understanding of phenomena, which is manifested by yielding interval path coefficients of the structural model. The second type of error is still considered as a measurement error, but the estimation process minimizes the sum of the absolute values and not the squares, considered in the PLSPM approach. The minimization problem in (12.10) therefore does not include the structural residual variance trace (Ψ), which is part of the modeling process of the structural model through the PR, and becomes

$$\mathbf{u}^{\mathsf{T}}\epsilon + \mathbf{u}^{\mathsf{T}}\delta, \tag{12.11}$$

where $\mathbf{u}$ is a $p \times 1$ unitary vector.

12.4 Partial Possibilistic Regression Path Modeling

In PPRPM, an iterative procedure permits the LV scores and the outer weights to be estimated, while path coefficients are obtained from PR between the estimated LVs. Since in PLSPM notation there is no difference between endogenous and exogenous LVs or between their respective MVs, in the following any block of MVs is referred to as $\mathbf{X}_h$ and each LV as ξ_h.

The algorithm computes the latent variables' scores alternating the *outer* and *inner* estimation till convergence (Jöreskog 1970). The procedure starts on centered (or standardized) MVs by choosing arbitrary weights w_{ph}. In the external estimation, the h-th latent variable is estimated as a linear combination of its own MVs

$$\mathbf{v}_h \propto \sum_{p=1}^{P_h} w_{ph}\mathbf{x}_{ph} = \mathbf{X}_h\mathbf{w}_h, \tag{12.12}$$

where $\mathbf{v}_h$ is the standardized outer estimation of the latent variable $\boldsymbol{\xi}_h$ and the symbol $\propto$ means that the left-hand side of the equation corresponds to the standardized right-hand side. In the internal estimation, the latent variable is estimated by considering its links with the other adjacent h' latent variables

$$\boldsymbol{\vartheta}_h \propto \sum_{h'} e_{hh'}\mathbf{v}_{h'}, \tag{12.13}$$

where $\boldsymbol{\vartheta}_h$ is the standardized inner estimation of the latent variable $\boldsymbol{\xi}_h$ and the inner weights, according to the so-called *centroid scheme* (Tenenhaus et al. 2005), are equal to the sign of the correlation between $\mathbf{v}_h$ and $\mathbf{v}_{h'}$ (with $h, h' = 1, \ldots, H$). These first two steps allow us to update the outer weights w_{ph}. In PPRPM the weight w_{ph} is the regression coefficient in the quantile regression of the p-th manifest variable of the $h-$th block $\mathbf{x}_{ph}$ on the inner estimate of the h-th latent variable $\boldsymbol{\vartheta}_h$

$$\mathbf{x}_{ph} = w_{ph}\boldsymbol{\vartheta}_h + \boldsymbol{\epsilon}_{ph}. \tag{12.14}$$

The quantile regression is an extension of the classical estimation of the conditional mean to the estimation of a set of conditional quantiles (Koenker and Basset 1978; Davino et al. 2013)

$$Q_\tau(\mathbf{x}_{ph}|\boldsymbol{\vartheta}_h) = \boldsymbol{\vartheta}_h w_{ph}(\tau) + \boldsymbol{\epsilon}_{ph}, \tag{12.15}$$

where $0 < \tau < 1$ and $Q_\tau(.|.)$ denotes the conditional quantile function for the τ-th quantile. In particular, PPRPM considers only the case in which $\tau = 0.5$, i.e. the median is the single chosen quantile.

The algorithm iterates till convergence. After convergence, structural (or path) coefficients are estimated through PR among the estimated LVs

$$\boldsymbol{\xi}_j = \tilde{\beta}_{0j} + \sum_{h:\boldsymbol{\xi}_h \to \boldsymbol{\xi}_j} \tilde{\beta}_{hj} \boldsymbol{\xi}_h, \tag{12.16}$$

where $\boldsymbol{\xi}_j$ ($j = 1, \ldots, J$ and $J < H$) is the generic endogenous (dependent) latent variable and $\tilde{\beta}_{hj}$ is the generic *interval path coefficient* in terms of midpoint and range $\tilde{\beta}_{hj} = \{c_{hj}; a_{hj}\}$, or equivalently $[\underline{\beta}_{hj}, \overline{\beta}_{hj}] = [c_{hj} \pm a_{hj}]$, interrelating the h-th exogenous (independent) variable to the j-th endogenous one (with $h \neq j$). The higher the midpoint coefficient the higher the contribution to the prediction of the endogenous LV. At the same time, the higher the spread coefficient the higher the vagueness in the relation among the considered LVs.

An important aspect to note is that in PPRPM the model can be validated using the same criteria defined in the PLSPM framework. In particular, this applies to the assessment of the measurement model, which can be validated by means of the *communality index* (Tenenhaus et al. 2005). However, this reasoning cannot be extended to the validation of the structural model, and even less to the global model. In PPRPM each individual structural equation is modeled by PR which includes the error term in the parameters; thus no residuals are provided. The quality of the model is here measured by the IC index presented in Sect. (12.2).

12.5 An Empirical Evidence: The MESI Questionnaire

The case study presents research carried out in the administrative area of Naples, which set itself the objective of investigating some dimensions that affect the quality of teaching in high schools (Palumbo et al. 2014). In particular, we examined the motivational and emotional aspects of teachers depending on the type of high school, their working position, gender and the socio-cultural context in which the teacher operates. The tool used to conduct this study was the questionnaire known as MESI (Motivation, Emotions, Strategies, Teaching) (Moè et al. 2010), which consists of six scales that investigate job satisfaction, practices, teaching strategies, emotions, self-efficacy, and incrementality. The idea is that effective teachers are those with a high sense of self-efficacy, satisfied with their work and able to sustain themselves through the activation of positive emotions in the workplace and in their personal life. The questionnaire was administered to 216 teachers working in high schools of the province of Naples. Fifteen high schools joined the research, divided into three different categories: Liceo (5), Technical Institute (6) and Professional Institute (4). In the following, the focus will be only on some of the scales composing the questionnaire: job satisfaction, emotions, and self-efficacy.

The first scale (satisfaction) is used to assess how job satisfaction is perceived from the point of view of the teachers. It consists of five items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The second scale (emotions) comprises two subscales each of which measures what emotions teachers experience when they teach (teach-emotions) and what emotions they experience in the role of teacher (role-emotions). The scale comprises a total of 30 items, each of which is constituted by a specific positive or negative emotion, and for each the teacher's frequency in experiencing the emotion is recorded on a 5-point scale (1 = hardly ever, 5 = almost always). In this study, we will focus only on the positive emotions measured by 13 items, the same for both subscales. Finally, the third scale (self-efficacy) explores the perception of self-efficacy of teaching by presenting a number of situations. Originally, it consisted of 24 items to which the teacher had to respond with a 9-point scale (1 = not at all, 9 = very much), how she/he felt able to deal with certain situations. However, a reduced subset of items is used in this study (9 items). According to theoretical assumptions, we propose an empirical framework (see Fig. 12.1) for analyzing the relationships among the subscales composing the MESI. PPRPM was adopted to check the research framework. An exploratory analysis of the observed indicators shows how the distribution of the subjective measurements is typically highly skewed (see Fig. 12.2). Thus, the choice of adopting the quantile regression in the measurement model seems appropriate for such type of data. Indicator reliability is assessed by looking at the standardized loadings in Table 12.1, where it is shown that all indicators are highly correlated with the respective constructs. To assess construct reliability, we calculate Dillon-Goldstein's ρ (DG.rho) and the communality indexes. As we show in Table 12.1, both indexes for all constructs are above the cut-off value of 0.7 and 0.5, respectively. This means constructs are homogeneous and capture on average 64 %, 59 %, 47 % and 49 % of the variance of their indicators in relation to the amount of variance due to measurement error. Consistent with the communality, the satisfaction and self-efficacy scales present the highest loadings.

The results of the structural model are shown in Fig. 12.3, where interval path coefficients are reported in terms of midpoints and spreads. As can be seen, there

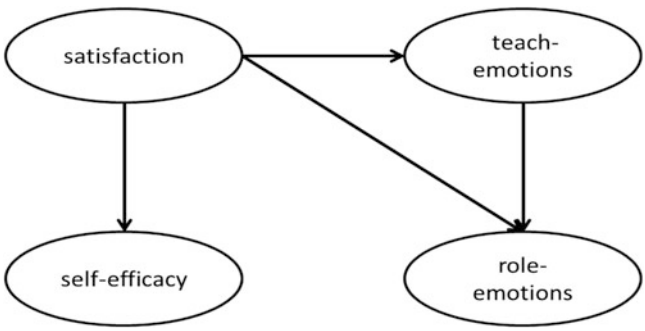

Fig. 12.1 Structural model of the MESI questionnaire

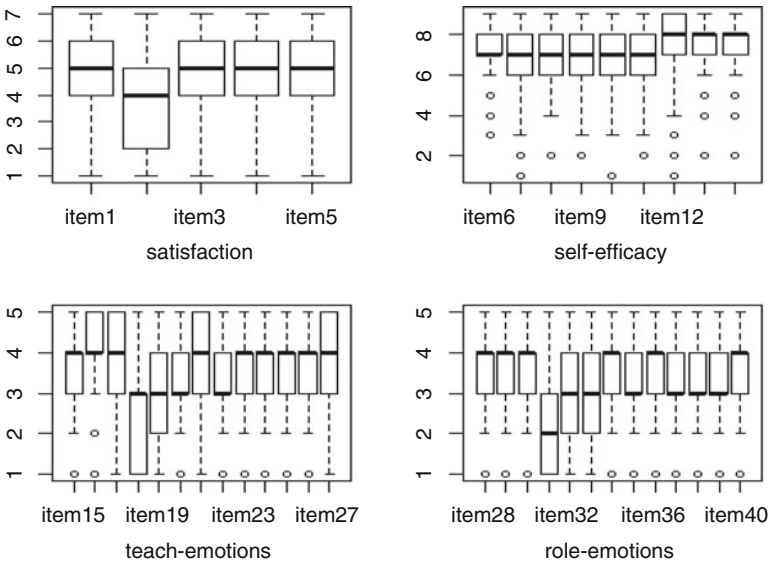

Fig. 12.2 Boxplots of the observed indicators

is no relation between satisfaction and self-accuracy, since the path coefficient is equal to 0.Teach-emotions is positively related to satisfaction with a path coefficient equal to 0.69, which means that when a teacher is satisfied he/she feels more frequently positive emotions while teaching. Both satisfaction and teach-emotions are good predictors of role-emotions, with path coefficients equal to 0.39 and 0.22, respectively. In other words, when a teacher is satisfied he/she feels positive emotions more frequently also in his/her role as a teacher. In addition, the increase in positive emotions while teaching also increases positive emotions in the role of teacher. It is worth noting that some relations indicate a certain imprecision. This holds for the relationship between satisfaction and teach-emotions, whose path coefficient has a spread equal to 0.23, and the relationship between the latter and the role-emotion, whose path coefficient has a spread of 0.16.

In Table (12.2) the results of the PPRPM are compared with those of the classical PLSPM. In particular, the table shows the values of the path coefficients and of the goodness of fit indexes. As can be seen, PPRPM results are consistent with the results obtained on the classical single-valued parameter model. The weak relationship between satisfaction and self-efficacy highlighted by a path coefficient close to zero in the PPRPM approach, is underlined by the low value of the R^2 index in PLSPM. The coefficient between satisfaction and teach-emotions is very similar in the two approaches, but PPRPM also provides information on the vagueness of the relation. In other words, the spread of the coefficient shows that the variation in the opinions of the respondents with respect to these two scales is not sufficient to arrive at a precise measurement of the dependent relationship between the two scales. Finally, both approaches show that role-emotions depend on the satisfaction

Table 12.1 Indicator and construct reliability

LV	MV	Standardized loadings	DG.rho	Communality
Satisfaction	Item1	0.817	0.899	0.641
	Item2	0.750		
	Item3	0.867		
	Item4	0.836		
	Item5	0.726		
Self-efficacy	Item6	0.675	0.934	0.586
	Item7	0.841		
	Item8	0.647		
	Item9	0.816		
	Item10	0.713		
	Item11	0.758		
	Item12	0.829		
	Item13	0.854		
	Item14	0.726		
Teach-emotions	Item15	0.724	0.917	0.469
	Item16	0.756		
	Item17	0.535		
	Item18	0.419		
	Item19	0.626		
	Item20	0.572		
	Item21	0.700		
	Item22	0.709		
	Item23	0.741		
	Item24	0.778		
	Item25	0.784		
	Item26	0.756		
	Item27	0.701		
Role-emotions	Item28	0.737	0.926	0.493
	Item29	0.735		
	Item30	0.502		
	Item31	0.420		
	Item32	0.697		
	Item33	0.609		
	Item34	0.769		
	Item35	0.728		
	Item36	0.798		
	Item37	0.763		
	Item38	0.805		
	Item39	0.829		
	Item40	0.692		

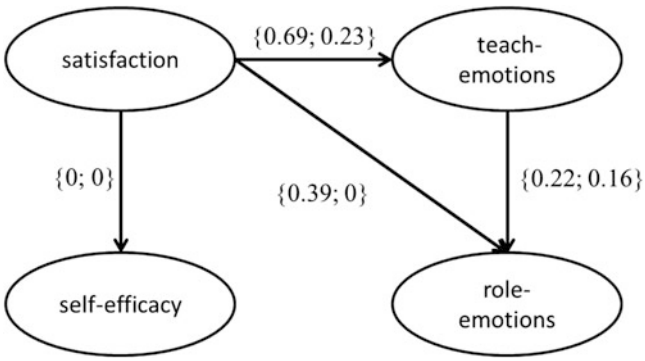

Fig. 12.3 Structural model results of the MESI questionnaire

and teach-emotions, but the PPRPM approach highlights the fact that there is a greater margin of vagueness in the second relation (higher spread).

Table 12.2 PLSPM and PPRPM structural model results

Relations	PLSPM path	R^2	PPRPM path	IC
Satisfaction > self-efficacy	0.22	0.05	{0.00; 0.00}	0.77
Satisfaction > teach-emotions	0.60	0.36	{0.69; 0.23}	0.88
Satisfaction > role-emotions	0.29	0.58	{0.39; 0.00}	0.80
Teach-emotions > role emotions	0.55		{0.22; 0.16}	

12.6 Conclusion and Perspectives

The present work presented the use of PPRPM for handling different types of uncertainty in the SEM context. After discussing the methodological aspects, the work focused on a case study and interpretation of the findings. It was shown that the use of PPRPM highlights the component of uncertainty inherent in subjective evaluations, besides the classical randomness. On-going research concerns the possibility of considering all structural equations simultaneously, such that the interval path coefficients would be estimated by optimizing a single objective function based on the spreads of all the coefficients inside the structural model.

References

Billard, L., Diday, E.: Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.) Data Analysis, Classification and Related Methods, Proceedings of 7th Conference IFCS, Namur, pp. 369–374 (2000)

Blanco-Fernndez, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput. Stat. Data Anal. **55**, 2568–2578 (2011)

Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)

Chin, W.W.: The partial least squares approach for structural equation modeling. In: Macoulides, G.A. (ed.) Modern Methods for Business Research, pp. 295–336. Lawrence Erlbaum Associates, Mahwah (1998)

Coppi, R.: Management of uncertainty in statistical reasoning: the case of regression analysis. Int. J. Approx. Reason. **47**, 284–305 (2008)

Coppi, R., D'Urso, P., Giordani, P., Santoro, A.: Least squares estimation of a linear regression model with LR fuzzy. Comput. Stat. Data Anal. **51**, 267–286 (2006)

Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Applications. Wiley, Chichester (2013)

Diamond, P.: Fuzzy least squares. Inf. Sci. **46**, 141–157 (1988)

Diamond, P.: Least squares fitting of compact set-valued data. J. Math. Anal. Appl. **147**, 531–544 (1990)

Jöreskog, K.G.: A general method for analysis of covariance structures. Biometrika **57**, 239–251 (1970)

Judd, C.M., McClelland, G.H.: Data Analysis: A Model Comparison Approach. Routledge, New York (2009)

Kim, K.J., Moskowitz, H., Koksalan, D.: Fuzzy versus statistical linear regression. Eur. J. Oper. Res. **92**, 417–434 (1996)

Koenker, R., Basset, G.W.: Regression quantiles. Econometrica **46**, 33–50 (1978)

Lima Neto, E.A., de Carvalho, F.A.T.: Constrained linear regression models for symbolic interval-valued variables. Comput. Stat. Data Anal. **54**, 333–347 (2010)

Marino, M., Palumbo, F.: Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. Ital. J. Appl. Stat. **14**, 277–291 (2002)

Moè, A., Pazzaglia, F., Friso, G.: MESI, Motivazioni, Emozioni, Strategie e Insegnamento. Questionari metacognitivi per insegnanti. Erickson, Trento (2010)

Palumbo, F., Romano, R.: Possibilistic PLS path modeling: a new approach to the multigroup comparison. In: Brito, P. (ed.) Compstat 2008, pp. 303–314. Physica-Verlag, Heidelberg (2008)

Palumbo, F., Romano, R., Esposito Vinzi, V.: Fuzzy PLS path modeling: a new tool for handling sensory data. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, pp. 689–696. Springer, Berlin/Heidelberg (2008)

Palumbo, F., Strollo, M.R., Melchiorre, F.: Stress and burnout in the school teachers: a study on the motivations to teach in the Neapolitan district. (in Italian). In: Strollo, M.R. (ed.) La motivazione nel contesto scolastico. pp. 3–47. Franco Angeli, Milan (2014)

Romano, R., Palumbo, F.: Partial possibilistic regression path modeling for subjective measurement. QdS – J Methodol. Appl. Stat. **15**, 177–190 (2013)

Tanaka, H.: Fuzzy data analysis by possibilistic linear models. Fuzzy Sets Syst. **24**, 363–375 (1987)

Tanaka, H., Guo, P.: Possibilistic Data Analysis for Operations Research. Physica-Verlag, Wurzburg (1999)

Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Trans. Syst. Man Cyber. **12**, 903–907 (1982)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)

Wang, H.F., Tsaur, R.C.: Insight of a fuzzy regression model. Fuzzy Sets Syst. **112**, 355–369 (2000)

Wold, H.: Modelling in complex situations with soft information. In: Third World Congress of Econometric Society, Toronto (1975)
Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. **1**, 28–44 (1973)

Chapter 13
Assessment and Validation in Quantile Composite-Based Path Modeling

Cristina Davino, Vincenzo Esposito Vinzi, and Pasquale Dolce

Abstract The paper aims to introduce assessment and validation measures in Quantile Composite-based Path modeling. A quantile approach in the Partial Least Squares path modeling framework overcomes the classical exploration of average effects and highlights how and if the relationships among observed and unobserved variables change according to the explored quantile of interest. A final evaluation of the quality of the obtained results both from a descriptive (assessment) and inferential (validation) point of view is needed. The functioning of the proposed method is shown through a real data application in the area of the American Customer Satisfaction Index.

Keywords Quantile composite-based path modeling • PLS-PM • Quantile regression

13.1 Introduction

Quantile Composite-based Path modeling (QC-PM) has been recently introduced (Davino and Esposito Vinzi 2014; Davino 2014; Davino and Esposito Vinzi 2016) as a complementary approach to the classical methods used to analyzed a network of relationships between unobserved and observed variables. In this framework, Partial Least Squares path modeling (PLS-PM) (Wold 1985; Tenenhaus 1998; Tenenhaus et al. 2005; Esposito Vinzi et al. 2010) is a consolidated method. Basically, PLS-

C. Davino (✉)
University of Macerata, Macerata, Italy
e-mail: cristina.davino@unimc.it

V. Esposito Vinzi
ESSEC Business School, Avenue Bernard Hirsch, B.P. 50105 95021
Cergy Pontoise Cedex, France
e-mail: vinzi@essec.edu

P. Dolce
University of Naples "Federico II", Naples, Italy
e-mail: pasquale.dolce@unina.it

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_13

PM algorithm consists of an iterative procedure in which simple and multiple ordinary least squares (OLS) regressions are applied. In several applications it can be advisable to broaden the analysis beyond the estimation of average effects in the network of relationships among variables. A QC-PM aims to highlight how and if the relationships among observed and unobserved variables as well as among the unobserved variables change according to the explored quantile of interest, thus providing an exploration of the whole dependence structure. To this purpose Quantile regression (QR) and Quantile Correlation (QC) are introduced in all the estimation phases of a PLS-PM algorithm.

In this paper we go through the assessment and the validation of the QC-PM. The goodness of fit measures typically used in PLS-PM are extended to QC-PM and a non parametric approach is used to validate the significance of the estimates. QC-PM is applied to real data in the area of the American Customer Satisfaction Index (American Customer Satisfaction Index 2000; Anderson and Fornell 2000).

The paper is organized as follows. Sections 13.2 presents the basic notations and the methodological framework. Section 13.3 is devoted to the description of the dataset used in the real data application. In Sects. 13.4, 13.5, and 13.6, QC-PM is introduced and the measures for assessing and evaluating the estimation results as well as the real data application results are presented.

13.2 Basic Notations and Methodological Framework

The methodological framework of the paper is represented by PLS-PM (Wold 1985; Tenenhaus 1998; Esposito Vinzi et al. 2010) and QR (Koenker 2005; Davino et al. 2013). The former is a consolidated method used to analyze a network of relationships between concepts that cannot be directly measured, while the latter is proposed as alternative methodology in the estimation procedure of PLS-PM. QC (Li et al. 2014) is also exploited and it will be described in Sect. 13.4.

PLS-PM aims at studying the relationships among Q blocks $\mathbf{X}_1, \ldots, \mathbf{X}_q, \ldots, \mathbf{X}_Q$ of manifest variables (MVs), each one summarized by an unobservable variable ξ_q, $(q = 1 \ldots Q)$, that is usually called latent variable (LV).

The general model consists of a measurement (or outer) model that specifies the relationships between MVs and LVs and a structural (or inner) model that specifies the relationships of LVs among each other. In the measurement model each MV $\mathbf{x}_{pq}$ $(p = 1, \ldots, P_q; q = 1, \ldots, Q)$ of the q-th block (P_q is the number of MVs in the q-th block) is assumed to be generated as a linear function of its LV ξ_q and its measurement error variable ϵ_{pq} (Lohmöller 1989),

$$\mathbf{x}_{pq} = \lambda_{pq0} + \lambda_{pq}\xi_q + \epsilon_{pq} \tag{13.1}$$

where λ_{pq0} is a location parameter and λ_{pq} is the loading coefficient.

In the structural model, LVs that depend on other LVs are called endogenous LVs. LVs that appear as predictors in every structural equation are called exogenous LVs. A generic endogenous LV, ξ_m $(m = 1 \ldots M)$, is linked to corresponding latent predictors by the following multiple regression model:

$$\xi_m = \beta_{m0} + \sum_{q \to m} \beta_{mq} \xi_q + \zeta_m \tag{13.2}$$

where β_{mq} is the so-called path coefficient capturing the effects of the predictor ξ_q on the dependent LV ξ_m, and ζ_m is the inner residual variable.

As a vehicle for the estimation of the parameters of the model, the scores of the qth LV are estimated as a linear combination of the corresponding MVs through the so-called weight relationship:

$$\hat{\xi}_q = \sum_{p=1}^{p_q} w_{pq} \mathbf{x}_{pq} \tag{13.3}$$

where w_{pq} are the outer weights and measure the contribution of each MV to the corresponding LV.

The outer weights w_{pq} are estimated by an iterative procedure alternating outer and inner estimation steps. In the outer estimation step each outer LV approximation is obtained as a standardized weighted aggregate ($\mathbf{v}_q$) of its own manifest variables, i.e. $\mathbf{v}_q \propto \sum_p w_{pq} \mathbf{x}_{pq} = \mathbf{X}_q \mathbf{w}_q$ (outer estimation). Generally, two different schemes are utilized for the computation of the outer weights (Esposito Vinzi and Russolillo 2012). In the *mode A* scheme (also called outward directed or reflective scheme) each MV is regressed on the corresponding so-called inner approximation, $\mathbf{z}_q$. In the *mode B* scheme (also called inward directed or formative scheme) the weights are computed as the regression coefficients in the multiple regression of $\mathbf{z}_q$ on its own MVs $\mathbf{x}_{pq}$ $(p = 1, \ldots, P_q)$. Then, the weights are normalized such as $\text{var}(\mathbf{X}_q \mathbf{w}_q) = 1$.

In the inner estimation step, each inner LV approximation is obtained as a weighted linear combination of the outer approximation of the connected LVs. Two LVs are connected if there exists a link between the two blocks: an arrow goes from one variable to the other in the Path diagram, independently of the direction (Esposito Vinzi and Russolillo 2012).

One of the schemes for the estimation of the inner weights is named *path weighting scheme* and it exploits the direction of the links between LVs. Such a scheme differently computes the weights according to the role played by a given LV with respect to the other LVs it is connected to. The LVs connected to a generic endogenous LV ξ_m are divided into two groups: the predecessors of ξ_m ($\xi_{q \to m}$), which are LVs explaining ξ_m, and the successors, which are LVs explained by ξ_m ($\xi_{q \to}$ m). The weights among the mth LV and its successor LVs are determined by their correlations while for its predecessor LVs the weights are the coefficients of a multiple regression, $\xi_m = \Xi_{\to m} \beta$, where $\Xi_{\to m}$ is the matrix of the all ξ_m's predecessor LVs. Possible alternatives are the *centroid* and the *factorial scheme*.

These schemes are based respectively on the sign and the value of the correlations between LVs. Therefore, they disregard the direction of the links between LVs.

An extension of OLS to the estimation of a set of conditional quantile functions is represented by QR. For a given quantile θ, a QR model can be formulated as follows:

$$Q_\theta\left(\hat{\mathbf{y}}|\mathbf{X}\right) = \mathbf{X}\hat{\beta}\left(\theta\right) \tag{13.4}$$

where $\mathbf{y}$ is the response variable observed on n individuals, $\mathbf{X} = [\mathbf{1}, \mathbf{X}_p]$ is a matrix with p regressors and a vector of ones for the intercept estimation, $0 < \theta < 1$ and $Q_\theta(.|.)$ denotes the conditional quantile function for the θth quantile.

Although different functional forms can be used, the paper will refer to linear regression models. In a QR, no probabilistic assumptions are required for the error. The parameter estimates in QR linear models have the same interpretation as those of any other linear model. As a consequence, the estimated values of the response variable conditioned to given values of the regressors, reconstruct the conditioned quantile of the dependent variable.

13.3 Dataset Description

The proposed methodology is applied to a real dataset concerning the ACSI (American Customer Satisfaction Index 2000; Anderson and Fornell 2000).[1] This index was established in 1994 and it is the only national cross-industry measure of customer satisfaction in the United States. The index measures the satisfaction of U.S. household consumers with the quality of products and services offered by both foreign and domestic firms with significant share in U.S. markets. Our application refers to the food processing sector including 1617 observations. The customer satisfaction is driven by three factors (customer expectations, perceived value and perceived quality) and has loyalty as outcome. The complaints LV has been excluded because the number of complaints was very small (1 %). The relationships among the five LVs are represented in the path diagram in Fig. 13.1. Each LV is measured through a set of MVs measured on a scale from 1 to 10 (see Table 13.1).

A preliminary analysis of the MV right tails is advisable before estimating a QC-PM because data deriving from customer satisfaction surveys are often characterised by a very high concentration of the responses on the upper values or even the maximum of the used scales. The deriving effect is an absence of variability in a given part of the distribution which is not interesting to explore. This information is not evident exploring the MV means while it is highlighted by the quantile values (Table 13.1).

[1] http://www.theacsi.org/

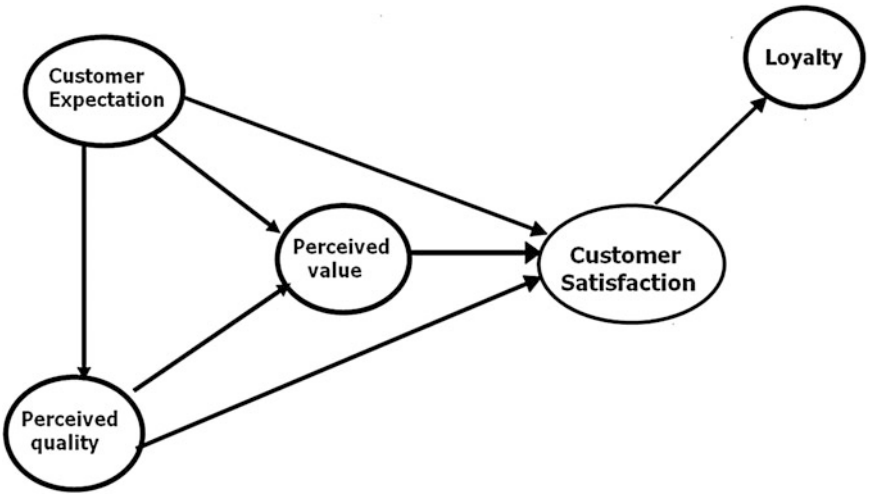

Fig. 13.1 Structural model describing driving factors and outcomes of customer satisfaction

In Fig. 13.2, the distribution of the maximum quantile for each MV is shown. It represents a threshold value because it limits the part of the variable distribution with variability from that with constant values. In the ACSI dataset all the MVs show a considerable percentage of customers expressing an evaluation equal to 10. We notice, for example, that it is not interesting to explore the variable WRONGQ from the 0.41 quantile forward because all the quantile values will be equal to 10.

Even if the maximum quantile value is different for each MV, QC-PM cannot be performed beyond the minimum threshold quantile which corresponds to 0.41, as for each quantile of interest QC-PM applies regression models for all the equations of the model. The requirement to confine the analysis at a lower quantile cannot be considered a limit of the proposed method, because QC-PM aims at the exploration of the different parts of the dependent variables distribution when they are characterised by different and not constant effects of the regressors. Moreover Table 13.1 shows that this choice is not detrimental to the treatment and interpretation of other MVs, because the maximum quantile is able to catch the most satisfied customers (values equal or greater than 7) and all the values of the quantiles greater than the maximum quantile are quite similar. It is worth noticing that this question only arises in case of discrete MVs.

13.4 QPLS-PM: Methodology and Results

In QC-PM (Davino and Esposito Vinzi 2014, 2016) all the estimation steps are carried out using a quantile approach. In particular, a QC-PM introduces, for each quantile θ of interest, either a QR or QC in both the inner estimation and the outer

Table 13.1 LVs and MVs of the ACSI dataset and means and main quantile values

LV	MV	Label	Mean	θ =0.1	θ =0.25	θ =0.41	θ =0.5	θ =0.75	θ =0.9
Customer Expectations	Expectations about overall quality	OVERALLX	8	6	8	8	9	10	10
	Expectations about customization	CUSTOMX	9	7	8	9	9	10	10
	Expectation about reliability	WRONGX	8	3	7	9	9	10	10
Perceived Quality	Meeting personal requirements	CUSTOMQ	9	7	8	9	9	10	10
	Things went wrong	WRONGQ	9	6	9	10	10	10	10
Perceived Value	Price given Quality	PQ	8	5	7	7	8	9	10
	Quality given Price	QP	8	6	7	8	8	9	10
Customer Satisfaction	Customer Satisfaction	SATIS	9	7	8	9	9	10	10
	Overall Quality	OVERALLQ	9	7	8	9	9	10	10
	Confirmation to Expectations	CONFIRM	8	5	6	8	8	9	10
	Close to ideal product/service	IDEAL	8	5	7	8	8	9	10
Customer Loyalty	Repurchase Intention	REPUR	8	6	8	9	9	10	10

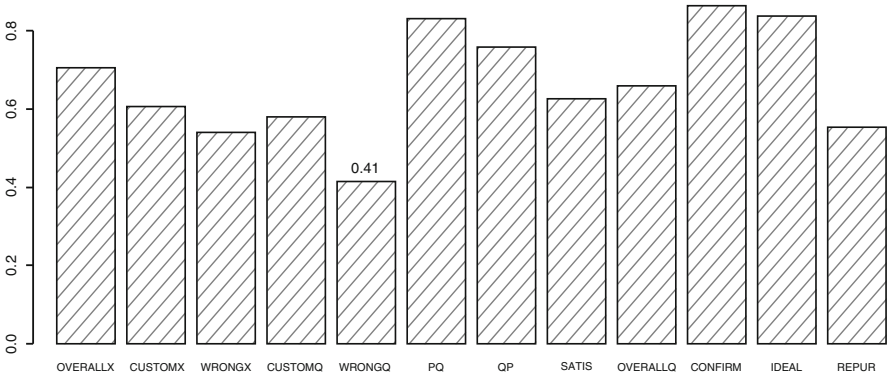

Fig. 13.2 Maximum quantile for each MV

estimation as well in the estimation of the path coefficients and loadings. Hence, for each quantile of interest θ we have estimates for all the parameters of the model.

According to the choice adopted in the various estimation phases, different versions of the QC-PM are available.

In the outer estimation, simple (Mode A) or multiple (Mode B) QR allows to compute the LV scores for each quantile of interest.

The inner estimation exploits the outer LV scores defined as the linear combination of the outer weights and the MVs belonging to each block. The way the inner weights are computed depends on the adopted weighting scheme. If the path weighting scheme is chosen, the inner weights linking and endogenous mth LV to its predecessors are computed through a QR:

$$Q_\theta\left(\hat{\boldsymbol{\xi}}_m | \Xi_{\to m}\right) = \Xi_{\to m}\hat{\beta}\left(\theta\right) \tag{13.5}$$

where $\Xi_{\to m}$ is the matrix of the ξ_m's predecessor LVs. Instead, the weights among the mth LV and its successor LVs are determined using the QC proposed by Li et al. (2014). Since in the quantile framework even the correlation is a not symmetric measure, the use of QC distinguishes between predecessors and successors. Let ξ_m and $\xi_{q\to m}$ be respectively a LV and one of its predecessor LVs, the former plays the role of the dependent variable and the latter is the regressor. The QC proposed by Li et al. (2014) and adapted in the QC-PM framework, is defined as:

$$\text{qcor}_\theta = \frac{\text{qcov}_\theta\left\{\xi_m, \xi_{q\to m}\right\}}{\sqrt{\left(\theta - \theta^2\right) var\left(\xi_{q\to m}\right)}} \tag{13.6}$$

where

$$\text{qcov}_\theta\left\{\xi_m, \xi_{q\to m}\right\} = \text{cov}\left\{I\left(\xi_{q\to m} - Q_\theta(\xi_{q\to m}) > 0\right), \xi_m\right\}, \tag{13.7}$$

$Q_\theta(\cdot)$ is the θth unconditional quantile, and $I(\cdot)$ is the indicator function.

QC is also proposed as an alternative to the Pearson correlation coefficient if either the centroid or the factorial scheme is adopted.

A new mode (named Mode Q) is introduced in the outer estimation. In Mode Q weights are obtained by computing QC between LVs and their own MVs (Davino and Esposito Vinzi 2014). Since QC is an asymmetric correlation coefficient, Mode Q allows us to handle both outwards-directed and inwards-directed measurement models. Once convergence is reached and LV scores are computed, the *path coefficients* related to endogenous LVs are estimated by means of QR.

The ACSI application is carried out using standardised MVs, the factorial scheme in the inner estimation and the outwards-directed relationship in the outer estimation. In both the estimation phases QR is used. Table 13.2 shows the outer weights estimated using Mode Q in an outwards-directed measurement model. Significant weights at $\alpha = 0.10$ are in bold (details about the validation of the coefficients are postponed to Sect. 13.6).

Differences in the weights sizes can be appreciated using a graphical representation. Figure 13.3 (left-hand side) depicts, for the *customer satisfaction* LV, the PLS-PM and QC-PM normalised outer weights with respect to the average values of the corresponding MVs. Labels 10, 25 and 41 refer to QC-PM weights for quantiles equal to 0.10, 0.25 and 0.41, respectively. PLS-PM weights are pointed out with the MV names. QC-PM and PLS-PM weights related to the same MV are vertically aligned with respect to the MV average. According to the PLS-PM results, it is not possible to identify how to improve satisfaction because IDEAL and CONFIRM show the lowest average values but also the lowest weights. QC-PM complements such a result suggesting that an improvement of the judgment on IDEAL and CONFIRM has a higher impact on the most satisfied customers. Moreover, as regards to CONFIRM, the impact is irrelevant on the most unsatisfied customers.

Table 13.3 shows the path coefficients obtained using the factorial scheme in PLS-PM and in QC-PM for a selected grid of quantile of interest ($\theta = [0.1, 0.25, 0.41]$). Significant coefficients at $\alpha = 0.1$ are in bold (details about the validation of the coefficients are postponed to Sect. 13.6).

A graphical representation of the path coefficients is more effective in highlighting differences among PLS-PM and QC-PM results and among QC-PM path coefficients at different quantiles. Figure 13.3 (right-hand side) shows the path coefficients of the *customer satisfaction* LV, the horizontal axis refers to the estimated quantile, the vertical axis to the corresponding coefficient and each segment represents the QC-PM coefficients of each LV impacting on the *customer satisfaction* LV. Full circles refer to the PLS-PM path coefficients while stars represent significant QC-PM path coefficients for each quantile of interest. For the sake of interpretation, PLS-PM results are vertically alligned to the last considered quantile (0.41). It is worth noting that path coefficients vary in the extreme parts of the distribution, meaning that the impact of a given LV changes for either very low

Table 13.2 Outer weights and correlations LV-MVs

		Outer weights				Correlations LV-MVs			
LV	MV	PLS-PM	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.41$	PLS-PM	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.41$
Expectation	OVERALLX	**0.477**	**0.438**	**0.484**	**0.535**	0.825	0.652	0.770	0.689
	CUSTOMX	**0.575**	**0.672**	**0.533**	**0.472**	0.894	0.908	0.804	0.589
	WRONGX	**0.232**	0.054	**0.303**	**0.325**	0.401	0.083	0.491	0.406
Quality	CUSTOMQ	**0.810**	**0.818**	**0.767**	**0.713**	0.945	0.849	0.796	0.595
	WRONGQ	**0.354**	**0.343**	**0.413**	0.481	0.661	0.688	0.647	0.593
Value	PQ	**0.463**	**0.454**	**0.481**	**0.494**	0.883	0.763	0.832	0.736
	QP	**0.630**	**0.638**	**0.613**	**0.600**	0.938	0.864	0.793	0.734
Satisfaction	SATIS	**0.374**	**0.388**	**0.375**	**0.326**	0.877	0.884	0.803	0.543
	OVERALLQ	**0.375**	**0.378**	**0.366**	**0.311**	0.846	0.816	0.775	0.509
	CONFIRM	**0.248**	**0.221**	**0.225**	**0.309**	0.697	0.640	0.568	0.629
	IDEAL	**0.254**	**0.259**	**0.286**	**0.323**	0.714	0.712	0.715	0.636

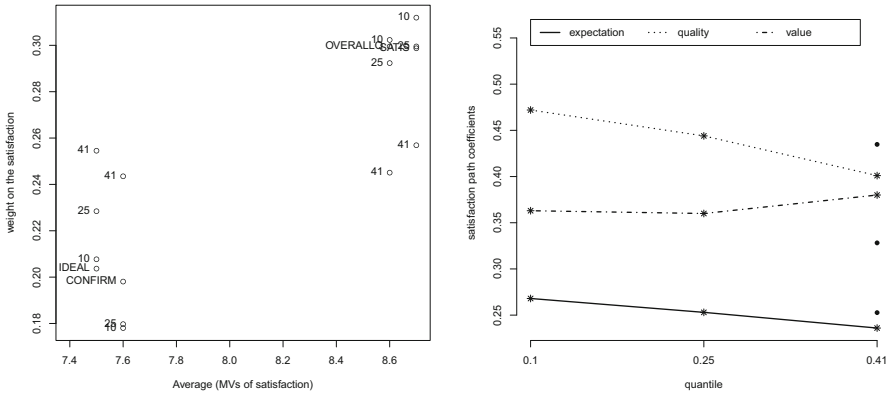

Fig. 13.3 Normalised outer weights with respect to the MV averages of the *customer satisfaction* LV (*left-hand side*) and QC-PM path coefficients for a set of selected quantiles (*right-hand side*)

Table 13.3 Path coefficients from a classical PSLPM and from a QC-PM for a selected set of quantiles ($\theta = [0.1, 0.25, 0.41]$)

LV	MV	PLS-PM	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.41$
Quality	Intercept	0.000	**−0.930**	**−0.300**	0.004
	Expectation	**0.585**	**0.757**	**0.822**	**0.708**
Value	Intercept	0.000	**−1.106**	**−0.502**	**−0.110**
	Expectation	**0.174**	**0.152**	**0.162**	**0.216**
	Quality	**0.401**	**0.462**	**0.433**	**0.410**
Satisfaction	Intercept	0.000	**−0.674**	**−0.317**	**−0.083**
	Expectation	**0.253**	**0.268**	**0.253**	**0.239**
	Quality	**0.435**	**0.472**	**0.444**	**0.398**
	Value	**0.328**	**0.363**	**0.360**	**0.378**
Loyalty	Intercept	0.000	**−0.877**	**−0.301**	**−0.045**
	Satisfaction	**0.604**	**0.868**	**0.828**	**0.687**

and very high satisfied customers. For example, considering the *expectation* LV, its effect decreases moving from the first 10 % of the distribution to the last considered quantile.

13.5 Model Assessment

The assessment of QC-PM is performed exploiting the main indexes proposed in the PLS-PM framework (Esposito Vinzi et al. 2010; Gotz et al. 2010; Henseler and Sarstedt 2013): communality and average communality, multiple linear determination coefficient (R^2), redundancy index, average redundancy index and global criterion of goodness of fit (GoF) for the structural model. It is worth noticing

that QC-PM is estimated for each quantile θ of interest thus it provides a set of assessment measures for each estimated model.

At first, we consider the correlations among MVs and LVs. The results are expected to show higher correlations between a LV with its own block of MVs than with other LVs representing different blocks of MV (cross-correlations). The aim is to measure if the concept underlying each LV differs from the other theoretical concepts. In Table 13.2 (last four columns) PLS-PM and QC-PM correlations between MVs and LVs are shown. QC-PM correlations are computed as QCs where each MV plays the role of dependent variable in the block it belongs to. The results are satisfactory for all the LVs (for the sake of brevity cross-correlations are not shown but they are in all cases lower than the correlations). It is worth noting the change of the correlation values across the quantiles. For example, the correlation of CUSTOMX to the Expectation LV is higher in the lower part of the distribution ($\theta = 0.1$) and even greater than the PLS-PM loading.

In the PLS-PM framework, the communality index measures the amount of the variability of a MV explained by its LV, and it is obtained as the square of the correlation between each MV and its LV. Therefore, for a generic $\mathbf{x}_{pq}$ MV belonging to the q_{th} block, the communality is equivalent to the R^2 of the simple regression $\mathbf{x}_{pq} = \alpha_0 + \alpha_1 \xi_q$. In a quantile framework, an index analogous to the R^2 of the classical regression analysis is the *pseudo*R^2 index (Koenker and Machado 1999). For each considered quantile θ, it compares a residual absolute sum of weighted differences using the selected model (*RASW*) (corresponding to the residual sum of squares in classical regression) with a total absolute sum of weighted differences (*TASW*) (corresponding to the total sum of squares of the dependent variable in classical regression) using a model with only the intercept (Davino et al 2013).

Let us consider the simplest regression model with one explanatory variable:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}. \tag{13.8}$$

For each considered quantile θ, *RASW* is the corresponding minimizer:

$$\begin{aligned} RASW(\theta) = & \sum_{y_i \geq \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} \theta \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right| \\ & + \sum_{y_i < \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} (1-\theta) \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right| \end{aligned} \tag{13.9}$$

where ρ_θ is the so-called check function which weights positive and negative residuals asymmetrically, respectively with weights equal to $(1-\theta)$ and θ.

The *TASW* is:

$$TASW(\theta) = \sum_{y_i \geq \theta} \theta \left| y_i - \hat{\theta} \right| + \sum_{y_i < \theta} (1-\theta) \left| y_i - \hat{\theta} \right|. \tag{13.10}$$

and the obtained $pseudoR^2$ can be computed as follows:

$$pseudoR^2(\theta)(\mathbf{y}, \mathbf{x}) = 1 - \frac{RASW(\theta)}{TASW(\theta)}. \tag{13.11}$$

As $RASW(\theta)$ is always less than $TASW(\theta)$, the $pseudoR^2(\theta)$ ranges between 0 and 1. For each considered quantile, the corresponding $pseudoR^2$ indicates whether the presence of the covariates influences the considered conditioned quantile of the response variable. It is worth noticing that $pseudoR^2$ is not a symmetric measure so that it assumes a different value inverting the role of the variables.

In QC-PM, for a generic $\mathbf{x}_{pq}$ MV of the q_{th} block and a quantile θ of interest, the communality expresses the quality of each simple regression $\mathbf{x}_{pq} = \alpha_0 + \alpha_1 \xi_q$, at the specific quantile, in terms of weighted residuals and can be defined as:

$$Com_{pq}(\theta) = pseudoR^2(\theta)(\mathbf{x}_{pq}, \xi_q) \tag{13.12}$$

The model assessment can also be carried out for the generic q_{th} block with p_q MVs (Com_q) or for the whole measurement part of the model ($\overline{Com}$) through averages respectively of the communalities related to the block and to all the MVs (weighted by the number of MVs in each block):

$$Com_q(\theta) = \frac{1}{p_q} \sum_{p=1}^{p_q} pseudoR^2(\theta)(\mathbf{x}_{pq}, \xi_q), \quad \overline{Com}(\theta) = \frac{1}{\sum_q p_q} \sum_q p_q Com_q(\theta) \tag{13.13}$$

Table 13.4 shows the indexes for the measurement model assessment provided by PLS-PM and QC-PM. The highest values across the quantiles with respect to a given MV are in italics, while QC-PM communality values higher than PLS-PM corresponding ones are in bold, even though we recommend not to compare communalities from QC-PM to those from PLS-PM as they are based on different residuals. Unsatisfactory communalities are to be taken into account in the interpretation of the results (like a warning on the use of the results related to that quantile) but they do not lead to the elimination of a MV unless all the communalities (from all the estimated QC-PMs and from PLS-PM) related to that MV are unsatisfactory.

With respect to the structural model, the $pseudoR^2$ index is proposed just like the essential criterion in PLS-PM, the coefficient of determination of the endogenous LVs (Chin 1998). A $pseudoR^2$ index is computed for each structural equation and each of them measures the amount of variability of a given endogenous LV explained by its predecessor LVs. The average of all the $pseudoR^2$ indexes ($\overline{pseudoR^2(\theta)}$) provides a synthesis of the evaluations regarding the structual part of the model.

Another important index is the *redundancy* because it is able to take into account also the contribution of the MVs related to the q_{th} endogenous LV thus linking the prediction performance of the measurement model to the structural one (Amato et al. 2004). In the QC-PM framework the redundancy of a generic q_{th} endogenous LV is

Table 13.4 Measurement model assessment indexes

LV	MV	Communality			
		PLS-PM	θ=0.1	θ=0.25	θ=0.41
Expectation	OVERALLX	0.680	0.503	0.494	*0.520*
	CUSTOMX	0.799	*0.759*	0.639	0.563
	WRONGX	0.161	0.016	***0.232***	0.209
	$Com_{Expectation}$	0.546	0.426	*0.455*	0.431
Quality	CUSTOMQ	0.892	*0.851*	0.768	0.670
	WRONGQ	0.438	***0.550***	0.464	0.450
	$Com_{Quality}$	0.665	***0.701***	0.616	0.560
Value	PQ	0.779	0.516	*0.587*	0.616
	QP	0.881	0.749	*0.774*	0.731
	Com_{Value}	0.830	0.632	*0.681*	0.674
Satisfaction	SATIS	0.768	*0.617*	0.590	0.515
	OVERALLQ	0.716	*0.537*	0.533	0.462
	CONFIRM	0.486	0.235	0.328	*0.385*
	IDEAL	0.510	0.356	*0.381*	0.402
	$Com_{Satisfaction}$	0.620	0.436	*0.458*	0.441
$\overline{Com}$		0.646	0.517	*0.526*	0.502

proposed as:

$$Red_q(\theta) = Com_q(\theta) \times pseudoR^2(\theta)(\hat{\xi}_q; \hat{\Xi}_{\rightarrow q}) \tag{13.14}$$

where $\hat{\Xi}_{\rightarrow q}$ is the matrix of the predictor LVs for the qth LV.

An overall assessment of the quality of the structural part is provided by the average redundancy ($\overline{Red}(\theta)$) obtained as a mean of the redundancies associated to the set of endogenous LVs.

With respect to the goodness-of-fit of the model, it is worth noticing that PLS-PM is not based on the optimization of a global function. Tenenhaus et al. (2004) have solved the lack of a global goodness-of-fit measure by proposing an index, the GoF, able to take both the measurement and the structural part of the model into account.

In QC-PM the absolute GoF is obtained as geometric mean of the average communality and the average $pseudoR^2$:

$$GoF(\theta) = \sqrt{\overline{Com}(\theta) \times \overline{pseudoR^2}(\theta)} \tag{13.15}$$

The first and the second term in Eq. 13.15 measure the predictive performance respectively of the measurement and the structural model (Amato et al. 2004; Esposito Vinzi et al. 2008). Further research will be devoted to the extension of the relative GoF to the QC-PM.

Table 13.5 shows the indexes for the structural model assessment provided by PLS-PM and QC-PM (in italics the highest values across the quantiles with respect to a given MV). Notwithstanding the interesting variability of the indexes across the quantiles, the overall assessment of the structural part shows rather low values of the R^2, $pseudoR^2$ and consequently redundancy values. This is probably due to the presence of endogenous LVs explained by few (or even one) LVs (Chin 1998). Moreover, in case of the QC-PM, it is well known that the typical determination index is not a satisfactory assessment index (Koenker and Machado 1999).

Further developments will regard the exploration of different goodness of fit measure in the quantile framework and the adjustment to the QPLS-PM of further assessment indexes proposed in PLS-PM framework (Henseler et al. 2009) (e.g. the average variance extracted (Fornell and Larcker 1981), the Stone-Geisser's Q^2 using blindfolding procedures (Stone 1974), the relative GoF Amato et al. 2004).

13.6 Model Validation

The evaluation of the statistical significance of the coefficients related to the different quantiles can be carried out exploiting the asymptotically normal distribution of the QR estimators as well as the bootstrap approach classically used in PLS-PM and QR.

QR estimators are asymptotically normal distributed with different forms of the covariance matrix depending on the model assumptions (independent and identically distributed errors or non-identically distributed errors) (Koenker and Basset 1978, 1982a,b). Resampling methods (Efron and Tibshirani 1993) can represent a valid alternative to the asymptotic inference (among many see Kocherginsky et al. 2005) because they allow the estimation of parameter standard errors without requiring any assumption in relation to the error distribution. Several bootstrap procedures have been proposed in the QR framework. The simplest and widespread is the *xy*-pair method or design matrix bootstrap (Parzen et al. 1994). The model parameters are estimated through the average of the bootstrap values. The standard error of the vector of parameter bootstrap estimates represents an estimate of the QR standard error useful in confidence intervals and hypothesis tests.

A bootstrap approach is also applied to obtain a variability measure of the QR estimates obtained choosing Mode Q in the measurement model and/or factorial or centroid scheme in the structural model.

In future work, a jackknife approach could be explored especially in case of small samples to estimate the standard errors of the parameter estimators and statistical tests could be introduced in a QPR-PM to test if coefficients at different quantiles can be considered statistically different (Gould 1997).

Table 13.5 Structural model assessment indexes

		Redundancy				R^2	$pseudoR^2$		
LV	MV	PLS-PM	θ=0.1	θ=0.25	θ=0.41	PLS-PM	θ=0.1	θ=0.25	θ=0.41
Quality	CUSTOMQ	0.299	0.204	*0.229*	0.184				
	WRONGQ	0.146	0.132	*0.138*	0.124				
	$Red_{Quality}$	0.223	0.102	*0.136*	0.118	0.335	0.240	*0.298*	0.275
Value	PQ	0.194	0.093	*0.106*	0.095				
	QP	0.220	*0.134*	0.140	0.112				
	Red_{Value}	0.207	*0.126*	0.112	0.086	0.250	0.180	*0.181*	0.153
Satisfaction	SATIS	0.506	*0.310*	0.293	0.221				
	OVERALLQ	0.472	*0.270*	0.264	0.198				
	CONFIRM	0.320	0.118	0.163	*0.165*				
	IDEAL	0.336	0.179	*0.189*	0.172				
	$Red_{Satisfaction}$	0.409	0.317	*0.337*	0.289	0.659	*0.502*	0.496	0.429
Loyalty	REPUR	0.364	0.275	*0.297*	0.282				
	$Red_{Loyalty}$	0.364	0.120	*0.136*	0.124	0.364	0.276	*0.297*	0.282
Mean		0.301	0.166	*0.180*	0.154	0.402	0.299	*0.318*	0.285

References

Amato, S., Esposito Vinzi, V., Tenenhaus, M.: A global goodness-of-fit index for PLS structural equation modeling. Oral Communication to PLS Club, HEC School of Management, France (2004)

American Customer Satisfaction Index: LLC. Food processing sector (2000)

Anderson, E.W., Fornell, C.: Foundations of the American customer satisfaction index. J. Total Qual. Manag. **11**(7), 869–882 (2000)

Chin, W.W.: The partial least squares approach to structural equation modeling. In: Marcoulides, G.A. (ed.) Modern Methods for Business Research, pp. 295–358. Lawrence Erlbaum Associates, Mahwah (1998)

Davino, C.: Combining PLS path modeling and quantile regression for the evaluation of customer satisfaction. Ital. J. Appl. Stat. **26**, 93–116 (2014) (published in 2016)

Davino, C., Esposito Vinzi, V.: Quantile PLS path modeling. In: Book of Abstract of the 8th International Conference on Partial Least Squares and Related Methods, Paris (2014)

Davino, C., Esposito Vinzi, V.: Quantile Composite-based Path Modelling, Advances in Data Analysis and Classification. Theory, Methods, and Applications in Data Science (2016) DOI 10.1007/s11634-015-0231-9

Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Applications. Wiley, Chichester (2013)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman Hall, New York (1993)

Esposito Vinzi, V., Russolillo, G.: Partial least squares algorithms and methods. WIREs Comput. Stat. **5**, 1–19 (2012)

Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M.: REBUS-PLS: a response-based procedure for detecting unit segments in PLS path modelling. Appl. Stoch. Models Bus. Ind. **24**(5), 439–458 (2008)

Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H.: Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer, Berlin/New York (2010)

Fornell, C., Larcker, D.F.: Structural equation models with unobservable variables and measurement error: algebra and statistics. J. Mark. Res. **18**(3), 328–388 (1981)

Götz, O., Liehr-Gobbers, K., Krafft, M.: Evaluation of structural equation models using the partial least squares (PLS) approach. In: Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (eds.) Handbook of Partial Least Squares: Concepts, Methods, and Applications. Springer, Berlin (2009)

Gould, W.: sg70: interquantile and simultaneous-quantile regression. Stata Tech. Bull. **38**, 14–22 (1997)

Henseler, J., Sarstedt, M.: Goodness-of-fit indices for partial least squares path modeling. Comput. Stat. **28**, 565–580 (2013)

Henseler, J., Ringle, C.M., Sinkovics, R.R.: The use of partial least squares path modeling in international marketing. Adv. Int. Mark. **20**, 277–319 (2009)

Kocherginsky, M., He, H., Mu, Y.: Practical confidence intervals for regression quantiles. J. Comput. Graph. Stat. **14**(1), 41–55 (2005)

Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge/New York (2005)

Koenker, R., Basset, G.W.: Regression quantiles. Econometrica **46**, 33–50 (1978)

Koenker, R., Basset, G.W.: Robust tests for heteroscedasticity based on regression quantiles. Econometrica **50**, 43–61 (1982a)

Koenker, R., Basset, G.W.: Tests for linear hypotheses and L1 estimation. Econometrica **46**, 33–50 (1982b)

Koenker, R., Machado, J.: Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. **94**, 1296–1310 (1999)

Li, G., Li, Y., Tsai, C.L.: Quantile correlations and quantile autoregressive modeling. J. Am. Stat. Assoc. **110**(509), 233–245 (2015)

Lohmöller, J.B.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)

Parzen, M.I., Wei, L., Ying, Z: A resampling method based on pivotal estimating functions. Biometrika **18**, 341–350 (1994)

Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. **36**, 111–147 (1974)

Tenenhaus, M.: La Régression PLS: Théorie et Pratique. Technip, Paris (1998)

Tenenhaus, M., Amato, S., Esposito Vinzi, V.: A global goodness-of-fit index for PLS structural equation modelling. In: Proceedings of the XLII SIS scientific meeting, pp. 739–742 (2004)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**(1), 159–205 (2005)

Wold, H.: Partial least squares. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences. John Wiley, New York (1985)

Part IV
Advances in Partial Least Square Regression

Chapter 14
PLS-Frailty Model for Cancer Survival Analysis Based on Gene Expression Profiles

Yi Zhou, Yanan Zhu, and Siu-wai Leung

Abstract Partial least squares (PLS) and gene expression profiling are often used in survival analysis for cancer prognosis; but these approaches show only limited improvement over conventional survival analysis. In this context, PLS has mainly been used in dimension reduction to alleviate the overfitting and collinearity issues arising from the large number of genomic variables. To further improve the cancer survival analysis, we developed a new PLS-frailty model that considers frailty as a random effect when modeling the risk of death. We used PLS regression to generate K PLS components from genomic variables and added the frailty of censoring as a random effect variable. The statistically significant PLS components were used in the frailty model for survival analysis. The genomic components representing the frailty followed a Gaussian distribution. Ten-fold cross-validation was used to evaluate the risk discrimination (between high risk and low risk) and survival prediction based on two breast cancer datasets. The PLS-frailty model performed better than the traditional PLS-Cox model in discriminating between the high and low risk clinical groups. The PLS-frailty model also outperformed the conventional Cox model in discriminating between high and low risk breast cancer patients according to their gene expression profiles.

Y. Zhou (✉)
State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

Department of Clinical Epidemiology and Biostatistics, Graduate School of Medicine, Osaka University, Osaka, Japan
e-mail: yi.zhou@stat.med.osaka-u.ac.jp

Y. Zhu
State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China
e-mail: mb35838@umac.mo

S.-w. Leung
State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China

School of Informatics, University of Edinburgh, Edinburgh, UK
e-mail: siu@inf.ed.ac.uk

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_14

Keywords PLS regression • Microarray • Genomic • Cancer • PLS frailty

14.1 Introduction

With the advent of DNA microarray technology in biomedical research, gene expression profiles are increasingly being used to predict cancer survival (Pawitan et al. 2004; Van De Vijver et al. 2002). In survival analysis, Cox regression (Cox 1972) is the tool of choice for analyzing the prognostic outcomes of patients in the presence of censoring. Cox regression estimates the regression parameters by maximizing Cox's partial likelihood but it only works when the number of patients is smaller than the number of covariates (Lee et al. 2013). However, in DNA microarray experiments, the number of genomic variables is typically much larger than the number of patients and this configuration causes collinearity and model overfitting. One way to palliate these problems is to implement effective dimension reduction methods on gene expression data prior to the analysis.

Partial least square (PLS) method is often applied to reduce the dimensionality of gene expression data and it has the additional benefit of modeling the relations of numerous genomic variables and observations. PLS assumes that the observed data is generated by a small number of latent variables (Rosipal and Krämer 2006), these latent variables were used, for example, by Nguyen and Rocke (2002) in a Cox regression model for diagnostic prediction, but these authors did not consider the censoring information when generating PLS components. Park et al. (2002) used PLS in generalized linear models to link the genomic variables and reformulated binary survival response from Poisson regression, but the number of dimensions increased with the number of iterations. Li and Gui (2004) proposed an algorithm to construct the PLS components by estimating the coefficients of the Cox regression model. Bastien (2004) used Cox regression to obtain PLS components, but the computational scheme of the coefficients of the genomic variable was complex and computationally intensive. In related work Lee et al. (2013) proposed a sparse PLS procedure to select genomic variables for dimension reduction; and Lambert-Lacroix et al. (2011) combined PLS and ridge penalized Cox regression for genomic variables selection.

By contrast with these previous attempts, our goal was to build a simple model for survival prediction from gene expression data. We favored a PLS regression model for this goal because it is simple, can efficiently analyze gene expression data with numerous (colinear) variables, and generates only a few genomic latent components based on the structure of predictors and response variables (Wold et al. 2001). Considering the possible random effects from gene expression profiles, shared frailty model is used by adding frailties into Cox regression. The frailties are unobserved covariates representing excess risk of early death (Therneau et al. 2003). In our PLS regression implementation, we used survival rates as the dependent variable and added the frailty of censoring as covariates. In our PLS-frailty model, we used genomic PLS components to represent frailty, and rejected the assumption of Cox regression that the survival times are independent of each other (Ripatti and Palmgren 2000).

14.2 Methods

14.2.1 Partial Least Squares Frailty Model for Survival Data

Our PLS-frailty model included (1) PLS regression with frailty of censoring to obtain PLS components, (2) univariate Cox regression to select PLS components with statistical significance, and (3) multivariate Cox regression with frailty for survival prediction.

14.2.1.1 Partial Least Square Regression with Frailty

Let $X_1, X_2, \ldots, X_P$ represent P genomic variables, $S(S_i, i = 1, 2, \ldots, N)$ is the survival rate vector of N patients from follow-up times. PLS regression generates K PLS components, denoted as $T_1, T_2, \ldots, T_K$ ($K \ll P$). The PLS components are regarded as predictors of S with weights $C_1, C_2, \cdots, C_K$ (see, Eq. (14.1), below) and also linear combinations of $X_1, X_2, \ldots, X_P$ with weights $W_1, W_2, \ldots, W_K$. A frailty factor of censoring with weight W_{j+1} is added as a random effect variable into the PLS regression (Eq. (14.2)) as:

$$S = \sum_{j=1}^{K} C_j T_j + F \tag{14.1}$$

$$T_j = \sum_{m=1}^{P} W_j X_m + W_{j+1}\text{frailty(censor)}, \tag{14.2}$$

where F is the residual of S. The tuning parameter K is set to ten. The model can be tested as cross-validation. PLS regression computations were performed by the `plsr` function in the `R` package `pls` (http://cran.r-project.org/web/packages/pls/index.html), and the frailty term was added by the frailty function in the `R` package "survival" (http://cran.r-project.org/web/packages/survival/index.html).

14.2.1.2 Univariate Cox Regression

Let the variable t $(t_i, i = 1, 2, \ldots, n)$ denote the follow-up time vector of patients and let the variable c $(c_i, i = 1, 2, \ldots, n)$ represent the censoring time vector. The right censored survival time is denoted (y, δ) with $y_i = \min(t_i, c_i)$ and $\delta_i (i = 1, 2, \ldots, N)$ the indicator of event: $\delta_i = 1$ indicates death and $\delta_i = 0$ indicates censoring. The individual PLS component, significant in univariate Cox regression

(Eq. (14.3)), is selected and denoted $T_1, T_2, \ldots, T_q$ from $T_1, T_2, \ldots, T_K (q \leq K)$. The significant level α was set to the value of $p < .05$. The coefficient of each $T_j (j = 1, 2, \cdots, K)$ is computed by maximizing the Cox's partial log-likelihood (PL) such as

$$h(y) = h_0(y) \exp(T_j \beta_j) \tag{14.3}$$

and specified with h being the proportional hazard function, h_0 the baseline hazard function, and β_j the coefficient of T_j.

14.2.1.3 Multivariate Cox Regression and Shared Frailty Model

Cox regression is a popular method in survival prediction (Gui and Li 2005). According to PLS-Cox model (Nguyen and Rocke 2002), the PLS components $T_1, T_2, \cdots, T_q$ are used as covariates in the multivariate Cox regression model (Eq. (14.4))

$$h(y) = h_0(y) \exp(\sum_{j=1}^{q} T_j \beta) \tag{14.4}$$

and specified with β being the coefficient vector. The shared frailty model (Therneau et al. 2003) (see also Eq. (14.5)) is the extension of Cox regression. It treats the frailty term as an additional covariate and produces estimates of the model parameters faster. We use the last selected PLS component T_q to present the frailty as a Gaussian distribution because Gaussian distribution is essential for models to converge (Therneau et al. 2003). The estimates of coefficients are obtained from penalized partial log-likelihood (PPL) as:

$$\lambda(y) = \lambda_0(y) \text{ frailty}\left(T_q\right) \exp\left(\sum_{j=1}^{q} T_j \alpha\right). \tag{14.5}$$

To make a distinction from the classic multivariate Cox regression, we specify λ the proportional hazard function and λ_0 the baseline hazard function, and α the coefficient vector. The coefficients used in the Cox regression model and the shared frailty model can be easily computed using the `coxph` function and frailty function in the `R` package `survival`.

14.2.2 Datasets of Breast Cancer

14.2.2.1 NKI Breast Cancer Dataset

The breast cancer dataset collected at Netherlands Cancer Institute (NKI) contains 24,481 genomic variables, and follow-up times of 295 patients with or without censoring. In previous work, van't Veer et al. (2002) found that a model with 70 genes outperformed all clinical variables in predicting the likelihood of distant metastases that occurred within five years. Without a predetermined gene expression profiles, we selected 2,448 genomic variables at random. To explore whether the genomic variables were associated with the follow-up times of the patients, we used the global test of Goeman et al. (2005) as first screening for the results with p value smaller than .05.

14.2.2.2 Stockholm Breast Cancer Dataset

Another breast cancer gene expression dataset was collected at the Karolinska Hospital in Stockholm. This data-base contains 44,928 genomic variables collected on 159 patients' follow-up times with or without censoring. Pawitan et al. (2004) used 64 genes to develop a prognostic model for breast cancer. We randomly chose 4,493 genomic variables for further analysis. The global test was used for pre-screening to explore if there was some association between gene expression profiles and survival times of patients.

14.2.3 Performance Evaluation

14.2.3.1 Performance in Discrimination

We applied the same dataset samples to both the PLS-frailty and PLS-Cox models and compared the diagnostic outcomes. Because the baseline hazard function was not estimated by PL or PPL, we decided to predict the linear prediction part ($\sum_{j=1}^{q} T_j\beta$; or $\sum_{j=1}^{q} T_j\alpha$) of the survival model, prognostic index (PI). The gene expression dataset was split into a training dataset and a testing dataset. We built a diagnostic model from the training dataset that we then used to predict PI in the testing dataset. The patients in the testing dataset were clustered by the model into two groups: the low risk group (PI $\leq$ 0) and the high risk group (PI$>$ 0), and the Kaplan-Meier survival curves of each group was then estimated. The log-rank test was adopted to test the difference between survival rates. In order to assess the quality of the predictive models, the dataset was randomly divided into ten blocks for a ten-fold cross-validation scheme. The results of PLS-frailty and PLS-Cox model in every testing dataset were compared.

14.2.3.2 Performance in Prediction

The linear association between the patient survival rates and PI was used to assess the prediction performance. We performed ten-fold cross-validation with the whole dataset to compute, in both models, the prediction errors (PEs) in terms of root mean squared prediction error (RMSPE). Each of the ten blocks was left out once to fit the generalized linear model between survival rates and PIs, and the survival rates of patients were computed in the left-out block to obtain the RMSPE for each observation. Ten-fold cross-validation was performed by the `cvTool` function in the `R` package `cvTools`. We got the RMSPEs (van Houwelingen and Putter 2011) of the PLS-frailty and PLS-Cox models (denoted as ε) and RMSPE of the null Cox regression (denoted as ε_0), which had no genomic covariates, to calculate the relative prediction error reductions (RPERs). These RPERs (computed as $\frac{\varepsilon_0 - \varepsilon}{\varepsilon_0}$) were used to compare the relative changes of PEs from the null Cox regression. Lower PEs and higher RPERs indicated better prediction performance.

14.3 Results and Discussion

A total of 2,448 genomic variables were selected from 24,481 gene expression profiles in the NKI dataset. The global test showed that the genomic variables had high significant relations with the patient survival rates (statistics = 0.935, $p < .001$). In the discrimination results from the testing data, the Kaplan-Meier survival curves (Fig. 14.1) obtained in the ten-fold cross-validation showed significant discrimination between the low risk and high risk groups in the PLS-frailty model. The log-rank test results and p values are shown in Table 14.1. In predicting the performance assessment from the overall dataset, we considered age as a covariate along the genomic covariates in the survival models. After the univariate Cox regression, the PLS-frailty model had four significant PLS components and the PLS-Cox model had six. The ten-fold cross-validation showed that PLS-frailty model did not outperform the PLS-Cox model in predicting the survival rates with higher PEs and lower RPERs (Table 14.2).

A total of 4,493 genomic variables were selected as covariates from 44,928 gene expression profiles in the Stockholm breast cancer dataset. The global test showed significant association between the genomic variables and the patient follow-up times (statistics = 0.86, $p = .004$). The Kaplan-Meier survival curves of the testing data (Fig. 14.2) with the ten-fold cross-validation scheme showed that the PLS-frailty model distinguished the low and high risk groups best. The log-rank test statistics and p values are shown in Table 14.1. The genomic PLS components from the overall dataset were used for prediction assessment. The univariate Cox regression selected three significant PLS components in the PLS-frailty model and four in the PLS-Cox model. Ten-fold cross-validation showed that the PLS-frailty model outperformed the PLS-Cox model in terms of lower PEs and higher RPERs (Table 14.3).

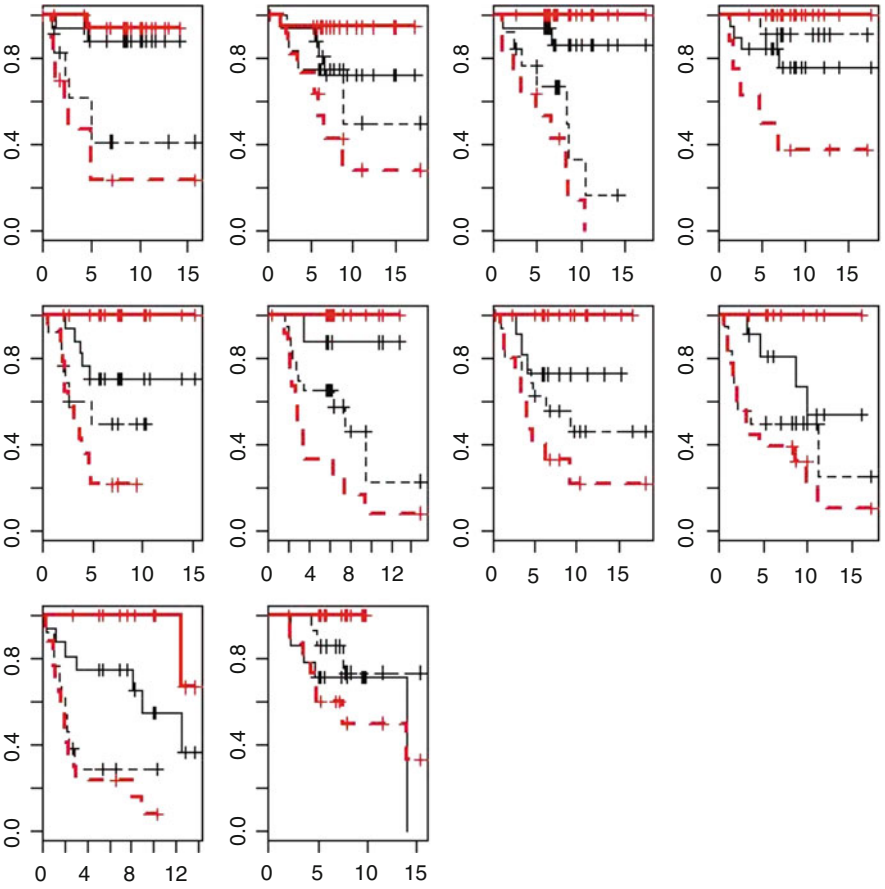

Fig. 14.1 Kaplan-Meier survival curves between high risk group and low risk group in NKI dataset in ten-fold cross-validation

14.4 Conclusion

Our proposed PLS-frailty model effectively represented the frailty as a random effect to improve the survival prediction of breast cancer and significantly reduced the number of dimensions of the genomic variables of gene expression profiles. As a result, it outperformed PLS-Cox model in differentiating low risk from high risk patients. Thus, the proposed PLS-frailty model would able to discriminate between low risk and high risk patients to identify stratified patient populations and to inform clinical decisions on precise and personalized treatments.

Acknowledgements This study was supported by the research grants (MYRG 2014-00117-ICMS-QRCM and MYRG190-Y3-L3-ICMS11-LSW) received from the University of Macau.

Table 14.1 Ten-fold cross-validation of log-rank test results between high risk and low risk groups in NKI dataset and Stockholm dataset

	NKI dataset				Stockholm dataset			
	PLS-Cox		PLS-frailty		PLS-Cox		PLS-frailty	
Run	Log-rank	P value	Log-rank	P value	Log-rank	P value	Log-rank	P value
1	6.398	0.011*	16.773	< 0.001*	0.017	0.896	3.849	0.050
2	0.437	0.509	10.491	0.001*	3.798	0.051	6.717	0.010*
3	6.951	0.008*	25.526	< 0.001*	3.527	0.060	11.633	0.001*
4	0.846	0.358	17.052	< 0.001*	4.602	0.032*	4.602	0.032*
5	1.954	0.162	19.416	< 0.001*	1.097	0.295	4.213	0.040*
6	3.723	0.054	21.958	< 0.001*	1.195	0.274	1.067	0.302
7	1.059	0.303	14.044	< 0.001*	1.790	0.181	5.013	0.025*
8	2.130	0.144	13.545	< 0.001*	0.343	0.558	3.849	0.050
9	4.950	0.026*	19.378	< 0.001*	0.328	0.567	1.988	0.159
10	1.112	0.292	7.327	0.007*	0.475	0.491	9.338	0.002*

* $p < .05$.

Table 14.2 Ten-fold cross-validation of prediction errors (PEs) and relative prediction error reductions (RPERs) in NKI breast cancer dataset

	PE			RPER	
Run	Null Cox	PLS-Cox	PLS-frailty	PLS-Cox	PLS-frailty
1	0.8087	0.0653	0.0703	0.9193	0.9130
2	0.8074	0.0656	0.0703	0.9188	0.9129
3	0.8132	0.0656	0.0707	0.9194	0.9131
4	0.8095	0.0657	0.0704	0.9189	0.9131
5	0.8064	0.0655	0.0702	0.9188	0.9129
6	0.8111	0.0656	0.0702	0.9192	0.9135
7	0.8123	0.0657	0.0704	0.9191	0.9133
8	0.8103	0.0656	0.0706	0.9190	0.9129
9	0.8119	0.0655	0.0701	0.9193	0.9136
10	0.8110	0.0654	0.0702	0.9193	0.9134

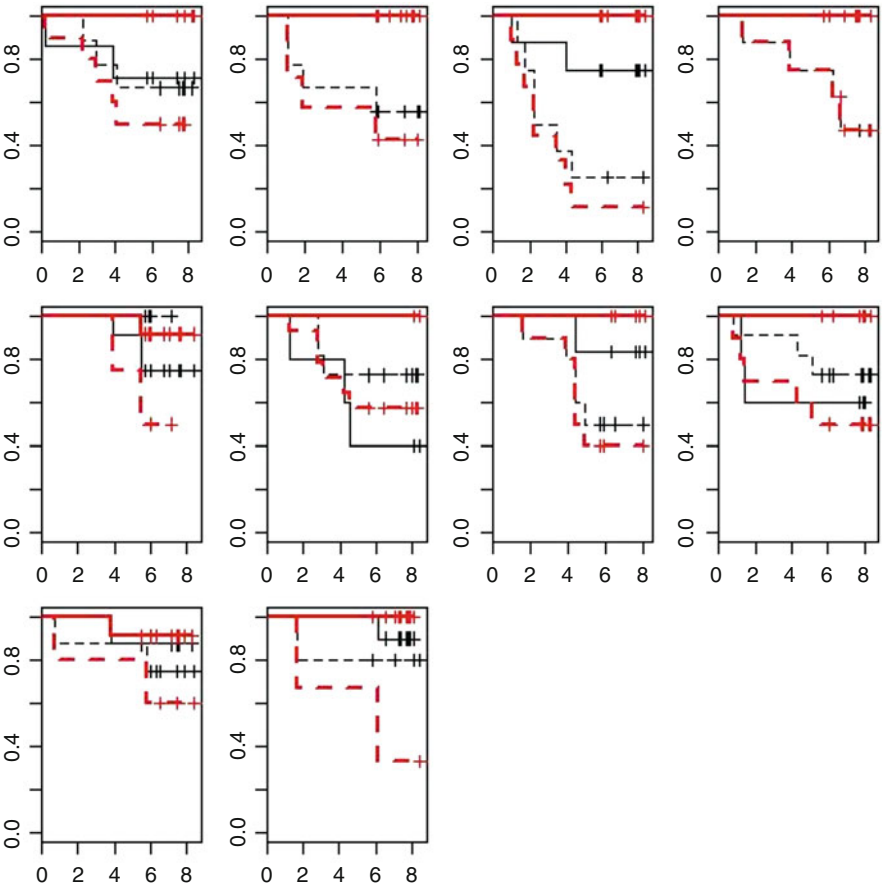

Fig. 14.2 Kaplan-Meier survival curves between the high and low risk groups in the Stockholm dataset obtained from ten-fold cross-validation

Table 14.3 Ten-fold cross-validation of prediction errors (PEs) and relative prediction error reductions (RPERs) in Stockholm breast cancer dataset

Run	PE			RPER	
	Null Cox	PLS-Cox	PLS-frailty	PLS-Cox	PLS-frailty
1	0.7558	0.0285	0.0205	0.9623	0.9729
2	0.7558	0.0290	0.0207	0.9616	0.9726
3	0.7558	0.0284	0.0203	0.9624	0.9731
4	0.7558	0.0286	0.0205	0.9622	0.9729
5	0.7558	0.0286	0.0205	0.9621	0.9729
6	0.7558	0.0284	0.0202	0.9624	0.9733
7	0.7558	0.0285	0.0203	0.9623	0.9732
8	0.7558	0.0288	0.0207	0.9620	0.9727
9	0.7558	0.0285	0.0205	0.9623	0.9729
10	0.7558	0.0284	0.0204	0.9624	0.9730

References

Bastien, P.: PLS-Cox model: application to gene expression. In: Antoch, J. (ed.) Proceedings in Computational Statistics, COMPSTAT 2004, pp. 655–662. Springer, Berlin (2004)

Cox, D.R.: Regression models and life tables. J. R. Stat. Soc. Ser. B (Methodol.) **34**, 187–220 (1972)

Goeman, J.J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J.K., Van Houwelingen, H.C.: Testing association of a pathway with survival using gene expression data. Bioinformatics **21**, 1950–1957 (2005)

Gui, J., Li, H.: Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics **21**, 3001–3008 (2005)

Lambert-Lacroix, S., Letué, F., et al.: Partial least squares and Cox model with application to gene expression, working paper (2011)

Lee, D., Lee, Y., Pawitan, Y., Lee, W.: Sparse partial least-squares regression for high-throughput survival data analysis. Stat. Med. **32**, 5340–5352 (2013)

Li, H., Gui, J.: Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics **20**(Suppl 1), i208–i215 (2004)

Nguyen, D.V., Rocke, D.M.: Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics **18**, 1625–1632 (2002)

Park, P.J., Tian, L., Kohane, I.S.: Linking gene expression data with patient survival times using partial least squares. Bioinformatics **18**(suppl 1), S120–S127 (2002)

Pawitan, Y., Bjöhle, J., Wedren, S., Humphreys, K., Skoog, L., Huang, F., Amler, L., Shaw, P., Hall, P., Bergh, J.: Gene expression profiling for prognosis using Cox regression. Stat. Med. **23**, 1767–1780 (2004)

Ripatti, S., Palmgren, J.: Estimation of multivariate frailty models using penalized partial likelihood. Biometrics **56**, 1016–1022 (2000)

Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) Subspace, Latent Structure, and Feature Selection, pp. 34–51. Springer, New York (2006)

Therneau, T.M., Grambsch, P.M., Pankratz, V.S.: Penalized survival models and frailty. J. Comput. Graph. Stat. **12**, 156–175 (2003)

Van De Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al.: A gene-expression signature as a predictor of survival in breast cancer. New Engl. J. Med. **347**, 1999–2009 (2002)

van Houwelingen, H., Putter, H.: Dynamic Prediction in Clinical Survival Analysis. CRC Press, Boca Raton (2011)

Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**, 109–130 (2001)

Chapter 15
Functional Linear Regression Analysis Based on Partial Least Squares and Its Application

Huiwen Wang and Lele Huang

Abstract Functional linear model with functional predictors and scalar response is a simple and popular model in the field of functional data analysis. The slope function is usually expanded on some basis functions, such as spline and functional principal component (FPC) basis, and then the model can be converted into a multivariate linear model. The FPC basis can keep most variance information of the functional data, but the correlation with response is not considered. Motivated by this, we use partial least square basis to expand the slope function. Meanwhile, considering the functional predictors are not all significant and variable selection procedure is implemented. In this process, group variable selection is introduced to identify the significant predictors. Then the proposed method is used to analyse the relationship between number of monthly emergency patients and some environmental factors in functional form, and some meaningful results are obtained.

Keywords Partial least squares regressions (PLSR) • Functional data analysis • Basis function

15.1 Introduction

With the rapid development of modern technology in the era of big data, especially the computer-related technique, data can be recorded densely over time (can be seen in the continuum), such as the prices of financial products (stock, futures and so on). Such type of data is termed functional or curve data. There are broad application prospects for functional data in a wide range of areas, which promotes the development of functional data analysis (FDA). There are vast literature on FDA, such as Ramsay and Silverman (1997, 2002) and Ferraty and Vieu (2006). In this paper, we discuss the estimation for functional linear model with functional predictors and scalar response (see Ramsay and Dalzell 1991).

H. Wang (✉) • L. Huang
School of Economics and Management, Beihang University, Beijing, China
e-mail: wanghw@vip.sina.com; nanhuabiren@163.com

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_15

Let Y be a real-valued random variable, $\{X_j(t) : t \in F\}$ be a zero mean , second-order stochastic process defined on $(\Omega, \mathscr{B}, P)$ and $EX_j^2(t) < \infty$ for all $t \in F, 1 \leq j \leq p$. The sample paths of $\{X_j(t) : t \in F\}$ are in $L^2(F)$, the set of all square integrable functions on F. To express different kinds of functional data, such as curves, images and arrays, F can be subsets of R, R^p or other spaces. It's assumed that the scalar response Y is linearly related to the functional predictor $X_j(t)$ through the relationship

$$Y = \alpha + \sum_{j=1}^{p} \int_F \beta_j(t) X_j(t) dt + \epsilon \tag{15.1}$$

where the intercept α and ϵ are scalars, ϵ is a random error variable.

In the functional linear regression model (15.1), the infinite diemnsional functional coefficients $\beta_j(t)$s of functional predictors are to be estimated, which is much different from traditional linear models where the unknown parameters are finite dimensional. But there are also some connections. In fact, in semiparametric and nonparametric regression models, functional forms appear and many estimators for them are studied. With functional predictors in consideration, we have to convert them based on basis expansion, then the next procedure is similar to multivariate problems. At last we reconstruct functional coefficients according to the basis function. During this procedure, it's crucial to choose basis function sequences and truncating parameter, see Cai and Hall (2006), Hall and Horowitz (2007), and Crambes et al. (2009) for details.

We can express $\beta_j(t), X(t)$ in terms of orthonormal basis chosen independently of the data. For example, Crambes et al. (2009) studied smoothing spline estimator for functional linear models. But a drawback is that the given basis function can not keep the information as much as possible. Then functional principal components (FPC) basis has attracted much attention of many statisticians and many theoretical results and practical applications are reported. These researches include contributions to FPC analysis (Silverman 1996; Boente and Fraiman 2000; Kneip and Utikal 2001; Hall and Hosseini-Nasab 2005; Hall et al. 2006; Jiang and Wang 2010; Berrendero et al. 2011). Yuan and Cai (2010) and Cai and Yuan (2012) obtained the optimal convergence rate of the estimator for slope function in the framework of reproducing kernel Hilbert space. While other methods could be used, the FPC technique is currently the most popular.

However, Delaigle and Hall (2012) pointed out that there is no reason why the first FPCs capture the most important information about the regression function and they proposed functional partial least square basis (FPLS) function. FPLS basis function makes it reasonable to truncate the infinite dimensional functional slope functions $\beta_j(t)$ into finite dimensional space spanned by the first basis functions while keeping information as much as possible.

It's well known that including unnecessary predictors can degrade the efficiency of the resulting estimation and yield less accurate predictions in the regression setting. After projecting the functional data and functional coefficients $X_j(t), \beta_j(t)$

into finite dimensional spaces, variable selection is necessary. Many methods on variable selection in linear model have been proposed based on the idea of penalization, see Tibshirani (1996), Fan and Li (2001), and Zou and Hastie (2005). In the framework of FDA, if there are multiple functional predictors, functional variable selection is needed. But the predictors are in functional form and the penalty term can not be utilized directly. One feasible solution is to project these functional predictors into lower dimensional spaces and then introduce group variable selection methods see Lian (2013). Tutz and Gertheiss (2010) proposed blockwise procedure to select sub-intervals in the domain of functional predictors and estimate the effect on the response simultaneously. These functional variable selection methods do not use data-driven basis functions, or in obtaining the data-driven basis they do not consider the response. In this paper, we select functional variables based on FPLS basis functions by the help of group variable selection techniques.

The remainder of the paper is organized as follows. Section 15.2 introduces variable selection and estimation method based on FPLS basis function. The application of the proposed method and some results are given in Sect. 15.3. Section 15.4 is a simple summary.

15.2 FPLS Basis Function and Variable Selection

For selection and estimation, we first expand the predictors and their slope functions on the basis functions, which are obtained by PLS method, and then grouped variable selection method is utilized to select and estimate simultaneously.

15.2.1 PLS Basis Function

Preda and Saporta (2005) proved the existence of FPLS components as well as some convergence properties towards the classical linear regression. Delaigle and Hall (2012) gave the asymptotical properties of FPLS basis. In the functional setting, if there is a single functional predictor, then the standard PLS basis is defined iteratively by choosing ψ_p in a sequential manner, to maximize the covariance functional

$$\gamma_k(\psi_k) = cov\left(Y - g_{k-1}(X), \int X\psi_k\right) \tag{15.2}$$

subject to

$$||\psi_k|| = 1, \quad \int\int \psi_j(s)K(s,t)\psi_k(t)dsdt = 0, \quad 1 \le j \le k-1,$$

where $g_{k-1}(x) = E(Y) + \sum_{j=1}^{k-1} \int_F x\psi_j$. According to Delaigle and Hall (2012), if $\int EX^2 < \infty$, then the function ψ_k that maximizes γ_k in (15.2), expressed by $\psi_1, \cdots, \psi_{k-1}$, is determined by

$$\psi_k = c_0[K(b - \sum_{j=1}^{k-1} \psi_j \int b\psi_j) + \sum_{j=1}^{k-1} c_j\psi_j], \tag{15.3}$$

where, for $1 \le j \le k-1$, the constants c_j are derived by solving the linear system of $k-1$ equations

$$\int\int \psi_j K \psi_p = 0, \quad j = 1, \cdots, k-1,$$

and where c_0 is determined by $||\psi_k|| = 1$, K is the covariance function defined by $K(s,t) = cov(X(s), X(t)) = E[X(s)X(t)]$.

For each functional predictor, we can obtain the associated FPLS basis function $\{\psi_{jk}(t), k = 1, \cdots, m_j, 1 \le j \le p\}$. The functional predictor and corresponding functional coefficients $X_j(t)$, $\beta_j(t)$ can be expanded as

$$\begin{aligned} X_j(t) &\approx \sum_{k=1}^{m_j} \xi_{jk}\psi_{jk}(t) \\ \beta_j(t) &\approx \sum_{k=1}^{m_j} b_{jk}\psi_{jk}(t) \end{aligned} \tag{15.4}$$

Assume that samples $(Y_i, X_{ij}(t)), j = 1, \cdots, p, i = 1, \cdots, n$ are observed, then the model (15.1) can be rewritten as

$$\begin{aligned} Y &\approx \alpha + \sum_{j=1}^{p} \sum_{k=1}^{m_j} \xi_{jk} b_{jk} \\ Y_i &\approx \alpha + \sum_{j=1}^{p} \sum_{k=1}^{m_j} \hat{\xi}_{ijk} \hat{b}_{jk} \end{aligned} \tag{15.5}$$

where $\xi_{jk} = \int x_j\psi_{jk}, \hat{\xi}_{ijk} = \int x_{ij}\hat{\psi}_{jk}$, and truncating parameter m_j means that m_j FPLS scores are utilized.

Remark 15.1. Jacques and Preda (2014) proposed to approximate the density of functional random variables and they considered the dependency of components of functional data. In our problem, we concentrate more on multiple functional data, not the components of the same functional data. Then we neglect the correlation between components. In the further study, we should take this aspect into consideration.

15.2.2 Group Variable Selection and Estimation

In many regression problems, it is more meaningful to identify significant factors, rather than individual predictors, where each factor is represented by a group of predictor variables, and the details can be seen in Yuan and Lin (2006). In our problem, the $\{\xi_{jk}\}_{k=1}^{m_j}$ in (15.5) enter the model or should be removed wholly, considering they denote the functional predictor $X_j(t)$.

We concern the following criterion:

$$Q(\hat{\beta}) = \sum_{j=1}^{n}(Y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_j}\hat{\xi}_{ijk}\hat{b}_{jk})^2 + n\sum_{j=1}^{p}P_{\lambda_j}(||\hat{b}_j||_\infty), \tag{15.6}$$

where $\hat{b}_j = [\hat{b}_{j1}, \cdots, \hat{b}_{jm_j}]$, $||\cdot||_\infty$ is L_∞ norm and $||\hat{b}_j||_\infty = \max\{|\hat{b}_{j1}|, \cdots, |\hat{b}_{jm_j}|\}$. λ_j is the tuning parameter and in practice we allow different regularization parameters for different coefficients. P_λ is a penalty term and in this paper, $P_\lambda(t) = \lambda|t|$, which means it is a Lasso-type penalty. Note that each $\hat{b}_j$ is a coefficient vector and this criteria means group variable selection. The penalty term P_λ can be in other forms, such as SCAD (Fan and Li 2001) or elastic net (Zou and Hastie 2005).

15.2.3 Tuning Parameter Selection

In every regularized model fitting procedure, it is well known that tuning parameters play an important role in the performance of the proposed model. Once the solution path of the proposed group Lasso has been constructed, we should choose the final estimator according to some criteria measuring the accuracy of prediction. Specifically, the tuning parameter λ_j and truncating parameter m_j should be given.

To choose the optimal tuning parameters m_j, AIC criteria is used and define AIC(k) by

$$AIC_j(k) = \log[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)/n] + 2k/n, \tag{15.7}$$

where $\hat{Y}_i$ is the fitted valued based on just one functional predictor $X_j(t)$. In fact, some $AIC_j(k)$ may be near to zero, considering its correlation with the response. Other criteria can also be used, such as AICc, BIC and so on.

As to the selection of λ_j, there are several types of tuning methods, such as the Schwarz information criterion (SIC) (Schwarz et al. 1978), the generalized approximate cross-validation criterion (Yuan 2006) and k-fold cross-validation. Let $\lambda_j = \lambda||\tilde{b}_j||_\infty^{-1}$ (where $\tilde{b}_j$ is the unpenalized least square estimator) and then we need select λ. Similar to Bang and Jhun (2012), we use the following SIC-type objective function

$$\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}\sum_{k=1}^{m_j}\hat{\xi}_{ijk}\hat{b}_{jk}\right)^2\right)+\frac{\log n}{2n}|E_\lambda|, \tag{15.8}$$

where $E_\lambda=\{i: Y_i=\sum_{j=1}^{p}\sum_{k=1}^{m_j}\hat{\xi}_{ijk}\hat{b}_{jk}\}$ and $|E_\lambda|$ means the card of E_λ.

15.3 Numerical Results

15.3.1 Simulation

In this section, we investigate the finite sample performances of the proposed estimators with Monte Carlo simulation studies. For comparison, we also list the results based on FPC.

Our design is similar to Matsui and Konishi (2011). The predictors $X_{\alpha mi}$ corresponding to mth predictor $\mathbf{X}_m$ are generated according to the following rule:

$$X_{\alpha mi}=\mu_{\alpha m}(t_{mi})+\epsilon_{\alpha mi},\quad \epsilon_{\alpha mi}\sim N(0,0.025r^2_{\chi\alpha m}),\alpha=1,\cdots,n,$$

where t_{mi} is the observation time points, $r_{\chi\alpha m}=\max_i(\mu_{\alpha m}(t_{mi}))-\min_i(\mu_{\alpha m}(t_{mi}))$ and $\mu_{\alpha m}(t)$ is assumed as follows:

$$\begin{aligned}
&X_1:\mu_{\alpha1}(t)=\cos(2\pi(t-a_1))+a_2t, t\in[0,1],\\
&\quad a_1\sim N(-5,3^2), a_2\sim N(7,1),\\
&X_2:\mu_{\alpha2}(t)=b_1\sin(2t)+b_2, t\in(0,\pi/3),\\
&\quad b_1\sim U(3,7), b_2\sim N(0,1),\\
&X_3:\mu_{\alpha3}(t)=c_1t^3+c_2t^2+c_3t+c_4, t\in[-1,1],\\
&\quad c_1\sim N(-3,1.2^2), c_2\sim N(2,0.5^2), c_3\sim N(-2,1), c_4\sim N(2,1.5^2).
\end{aligned}$$

The scalar response Y is generated by $Y_\alpha=\int_0^1\mu_{\alpha1}(t)\beta_1(t)dt+\sigma\epsilon_\alpha$, where $\beta_1(t)=\sin(2\pi t)$, $\sigma=c(\max(\int_0^1\mu_{\alpha1}(t)\beta_1(t)dt)-\min(\int_0^1\mu_{\alpha1}(t)\beta_1(t)dt))$, and $\sigma=0.1,0.3$. The distribution of ϵ_α is normal.

We considered the following error criteria: $RMSE=\sqrt{\frac{1}{n_1}\sum_{k=1}^{n_1}[\hat{\beta}(t_k)-\beta(t_k)]^2}$, where $\{t_k\}$ are equally spaced grid points; average number of correctly identified nonvanishing coefficients (TP), average number of incorrectly identified nonvanishing coefficients (FP). The *ORMSE* is oracle root of mean square error where the true nonzero coefficients are assumed to be known and no shrinkage is applied.

By Table 15.1, we can know that both of these two methods can select the nonzero variables, but the FPC method is opt to select more variables. The *RMSE* based on PLS are smaller relatively, while the FPC method's is a little larger.

Table 15.1 Variable selection results for simulation

n	σ		pls.RMSE	pca.RMSE	pls.TP	pls.FP	pca.TP	pca.FP
50	0.1	Mean	0.0903	0.1156	1.0000	0.2600	1.0000	1.1400
		Sd	0.0388	0.0449	0.0000	0.4845	0.0000	0.6360
	0.3	Me	0.2410	0.3068	1.0000	0.3900	0.9800	0.7100
		Sd	0.1364	0.1450	0.0000	0.5667	0.1407	0.6711
100	0.1	Mean	0.0687	0.0846	1.0000	0.2000	1.0000	0.9300
		Sd	0.0300	0.0353	0.0000	0.4264	0.0000	0.6705
	0.3	Mean	0.1997	0.2269	1.0000	0.5600	1.0000	0.7600
		Sd	0.0811	0.0936	0.0000	0.6407	0.0000	0.6980
200	0.1	Mean	0.0524	0.0706	1.0000	0.2300	1.0000	0.9100
		Sd	0.0214	0.0226	0.0000	0.4462	0.0000	0.5877
	0.3	Mean	0.1498	0.1723	1.0000	0.6500	1.0000	0.7000
		Sd	0.0672	0.0701	0.0000	0.5417	0.0000	0.6113

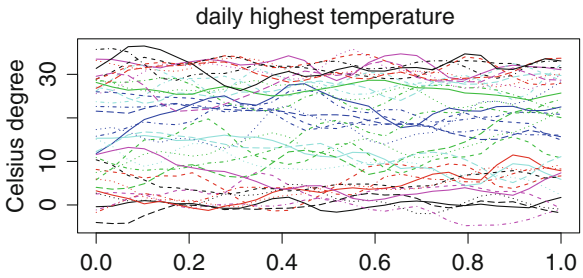

Fig. 15.1 The daily highest temperature curves. Each curve denotes the daily highest temperature of one month

15.3.2 Environmental Data

It is a meaningful issue to study the relationship between people's health and the environmental factors. Here we make use of patients numbers in emergency department of hospitals to measure people's health and some meteorological data is utilized to denote the environmental factors. Specifically, the data includes the total patient numbers monthly of emergency department in 21 main hospitals of a big city in China (which is a scalar response) and some environmental records, such as the daily highest temperature, daily maximum air pressure, daily highest wind speed, daily minimum relative humidity and hourly PM 2.5 (air pollutant) level (which are all can be seen as functional data). The period lasts from January, 2011 to March, 2014. Figures 15.1, 15.2 and 15.3, show the smoothed daily highest temperature, daily maximum air pressure and hourly PM 2.5 level curves, respectively. We regularize the time domain to $[0, 1]$, which can avoid the problems caused by different days in different months.

It is clear that the multivariate methods are not feasible here, considering the sample size 39 is far less than predictors number (thousands). Besides, the day numbers of these months are not all the same and the predictors are not recorded

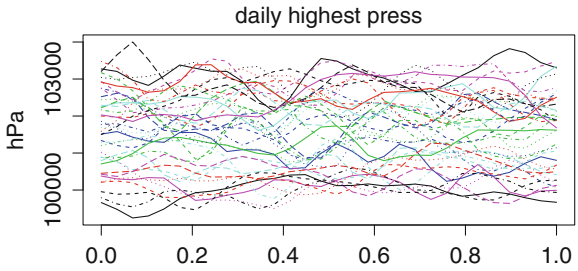

Fig. 15.2 The daily highest press curves. Each curve denotes the daily highest press of one month

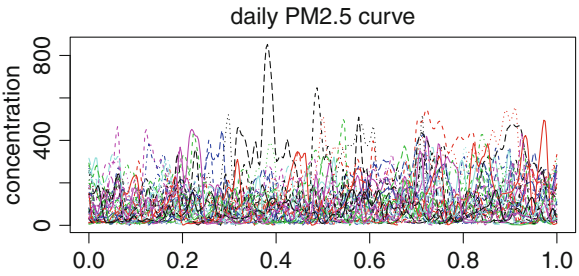

Fig. 15.3 The daily highest press curves. Each curve denotes the daily highest press of one month

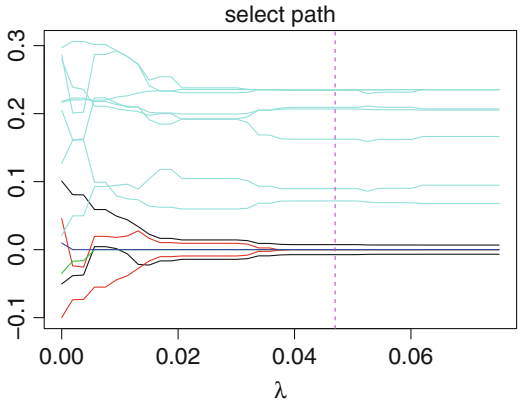

Fig. 15.4 The selection path of variable selection based on FPLS. Each color denoted one gruop

in the same frequency (daily or hourly). Meanwhile, there are many records about the environment and selecting out significant ones to patient numbers is interesting. The monthly total patient numbers are counting random variables and actually are discrete, but the numbers are large. Thus here the transformation using logarithmic scale is implemented and then they are treated as continuous random variables. The Poisson regression model may be an alternative model.

The number of these five FPLS components according to *AIC* above are 2, 2, 1, 1 and 8, respectively. Now in our model, we have 14 variables in 5 groups. By the SIC criteria above, we select $\lambda = 0.047$ and just the functional coefficients of PM 2.5 are not zero, which can be seen in Fig. 15.4. To demonstrate the performance, we plot the fitted values of different estimators with different predictors in Fig. 15.5.

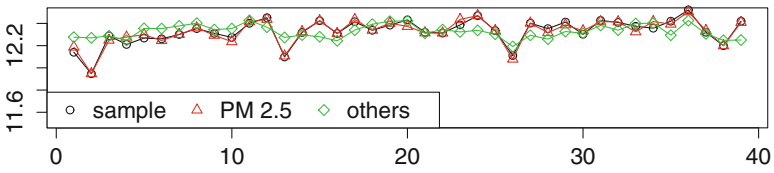

Fig. 15.5 The prediction with the selected PM 2.5 curve and with other curves

We can see that with just PM 2.5 as a functional predictor, the fitted values are good and there is little explanation ability with other 4 predictors.

Statistically, we say the emergency patients number is correlated with the functional PM 2.5 for the whole population from the view of large samples, but it is not casual. We cannot say that some factors promotes the increase of emergency patients. Besides, the patients number is not directly correlated with daily highest temperature and other weather indices above. We just verify that the air pollution is harmful to people's health from the view of large samples in statistics. Furthermore, we have to admit that just using these data above and just data of one city can not obtain general conclusions. It is just for illustrating the effectiveness of our proposed method.

15.4 Remarks and Conclusion

We propose a group regularization method for shrinkage estimation of multiple functional linear regression models based on PLS basis function. To demonstrate the effectiveness of the method, we apply it in analysing the relationship between emergency patients number and weather record curves, providing an interesting alternative perspective to the previously used kernel regression on this data.

We would like to end this paper by discussing some possible topics for future study. In fact, in the PM 2.5 curves there are many outliers and how to deal with functional data outliers by robust statistics will be an interesting issue. Meanwhile, robust variable selection procedures will be appreciated in selecting significant functional predictors. Besides, in a regression model with both scalar and functional variables, how to estimate the coefficient and select the significant variables simultaneously will be very meaningful.

Acknowledgements Wang and Huang's research was supported by the National Natural Science Foundation of China (No:71031001, 71420107025) and the Innovation Foundation of BUAA for Ph.D. Graduates (YWF-14-YJSY-027). The content is solely the responsibility of the authors and does not necessarily represent the official views of organizations.

References

Bang, S., Jhun, M.: Simultaneous estimation and factor selection in quantile regression via adaptive sup-norm regularization. Comput. Stat. Data Anal. **56**, 813–826 (2012)

Berrendero, J.R., Justel, A., Svarc, M.: Principal components for multivariate functional data. Comput. Stat. Data Anal. **55**, 2619–2634 (2011)

Boente, G., Fraiman, R.: Kernel-based functional principal components. Stat. Probab. Lett. **48**, 335–345 (2000)

Cai, T.T., Hall, P.: Prediction in functional linear regression. Ann. Stat. **34**, 2159–2179 (2006)

Cai, T.T., Yuan, M.: Minimax and adaptive prediction for functional linear regression. J. Am. Stat. Assoc. **107**, 1201–1216 (2012)

Crambes, C., Kneip, A., Sarda, P.: Smoothing splines estimators for functional linear regression. Ann. Stat. **37**, 35–72 (2009)

Delaigle, A., Hall, P.: Methodology and theory for partial least squares applied to functional data. Ann. Stat. **40**, 322–352 (2012)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer, New York (2006)

Hall, P., Horowitz, J.L.: Methodology and convergence rates for functional linear regression. Ann. Stat. **35**, 70–91 (2007)

Hall, P., Hosseini-Nasab, M.: On properties of functional principal components analysis. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **68**, 109–126 (2005)

Hall, P., Muller, H.G., Wang, J.L.: Properties of principal component methods for functional and longitudinal data analysis. Ann. Stat. **34**, 1493–1517 (2006)

Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Comput. Stat. Data Anal. **71**, 92–106 (2014)

Jiang, C.R., Wang, J.L.: Covariate adjusted functional principal components analysis for longitudinal data. Ann. Stat. **38**, 1194–1226 (2010)

Kneip, A., Utikal, K.J.: Inference for density families using functional principal component analysis. J. Am. Stat. Assoc. **96**, 519–542 (2001)

Lian, H.: Shrinkage estimation and selection for multiple functional regression. Stat. Sin. **23**, 51–74 (2013)

Matsui, H., Konishi, S.: Variable selection for functional regression models via the regularization. Comput. Stat. Data Anal. **55**, 3304–3310 (2011)

Preda, C., Saporta, G.: PLS regression on a stochastic process. Comput. Stat. Data Anal. **48**, 149–158 (2005)

Ramsay, J.O., Dalzell, C.: Some tools for functional data analysis. J. R. Stat. Soc. Ser. B (Methodol.) **53**, 539–572 (1991)

Ramsay, J.O., Silverman, B.W.: Functional data analysis. Springer, New York (1997)

Ramsay, J.O., Silverman, B.W.: Applied functional data analysis: methods and case studies, vol. 77. Springer, New York (2002)

Schwarz, G., et al.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Silverman, B.W.: Smoothed functional principal components analysis by choice of norm. Ann. Stat. **24**, 1–24 (1996)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**, 267–288 (1996)

Tutz, G., Gertheiss, J.: Feature extraction in signal regression: a boosting technique for functional data regression. J. Comput. Graph. Stat. **19**, 154–174 (2010)

Yuan, M.: Gacv for quantile smoothing splines. Comput. Stat. Data Anal. **50**(3), 813–829 (2006)

Yuan, M., Cai, T.T.: A reproducing kernel Hilbert space approach to functional linear regression. Ann. Stat. **38**, 3412–3444 (2010)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **68**, 49–67 (2006)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **67**, 301–320 (2005)

Chapter 16
Multiblock and Multigroup PLS: Application to Study Cannabis Consumption in Thirteen European Countries

Aida Eslami, El Mostafa Qannari, Stéphane Legleye, and Stéphanie Bougeard

Abstract We address the problem of investigating the relationships between $(K+1)$ blocks of variables (i.e., K blocks of independent variables and one block of dependent variables), where the observations are a priori divided into several known groups. We propose a simple procedure called multiblock and multigroup PLS regression—which is a straightforward extension of multiblock PLS regression—that takes into account the group structure of the observations. This method of analysis is illustrated with a large, questionnaire based, survey exploring, in 2011, the cannabis consumption of teenagers of thirteen European countries (the European School Survey Project on Alcohol and other Drugs).

Keywords Multigroup and multiblock PLS regression • Multiblock analysis • Multigroup analysis

A. Eslami (✉) • E.M. Qannari
LUNAM University, ONIRIS, USC Sensometrics and Chemometrics Laboratory, Rue de la Géraudière, Nantes F-44322, France
e-mail: aida.eslami@yahoo.fr; elmostafa.qannari@oniris-nantes.fr

S. Legleye
National institute of demographic studies (Ined), Paris, France

Inserm, U669 Paris, France

University Paris-Sud and University Paris Descartes, UMR-S0669 Paris, France
e-mail: stephane.legleye@ined.fr

S. Bougeard
French Agency for Food, Environmental and Occupational Health Safety (ANSES), Unit of Epidemiology, Technopole Saint Brieuc Armor, F-22440 Ploufragan, France
e-mail: stephanie.bougeard@anses.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_16

16.1 Introduction

Often, in methods of multivariate data analysis such as Principal Component Analysis (PCA), Partial Least Square (PLS) Regression and multiblock data analysis, the observation are a priori divided into groups. In the literature, this group structure is known by different names such as multilevel, hierarchical, nested data (Hox 2010), or even clustered data (Liang and Zeger 1986). Following Krzanowski (1984) and Kiers and Ten Berge (1994), we call such data sets "multigroup" data.

In the case of two-block and multigroup data (denoted by $\mathbf{X}$ and $\mathbf{Y}$), the classical approaches to investigate the relationship between $\mathbf{X}$ and $\mathbf{Y}$ either (1) ignore the group structure and perform a PLS regression of $\mathbf{Y}$ upon $\mathbf{X}$ or (2) analyze each group separately with PLS regression. Clearly, these strategies are unsatisfactory because with the first strategy, the total variance recovered by the latent variables mixes both the between and the within-group variances and with the second strategy, it is difficult to have an integrated vision of the data since the analysis does not provided a common structure among the separate analyses. To palliate these problems in the study of two-block and multigroup data, Eslami et al. (2013b) recently proposed a simple strategy—called multigroup PLS (mgPLS)—that extends multigroup PCA Eslami et al. (2013a) and predicts $\mathbf{Y}$ from $\mathbf{X}$ taking into account the groups difference and conjointly seeks common parameters (e.g., common loading weights) across the groups. As shown by Eslami et al. (2014a) mgPLS outperformed classical PLS regression when used with multigroup data.

In addition to standard structure of the two-block and multigroup data, there is often some external information associated with the independent variables ($\mathbf{X}$) such that $\mathbf{X}$ can be split into several known blocks of variables. These data-sets—called *multiblock* and *multigroup* data—are quite commonplace, and can be illustrated by an example from veterinary epidemiology where several measures are obtained on animals that are divided into known groups corresponding to various farms, where the independent blocks are related to different potential risk factors (breeding factors, environment, feeding, farm management, ...) and the dependent data ($\mathbf{Y}$) relate to the outbreak of a disease. With this example, the aim would be to investigate the relationships between the disease and the potential risk factors (organized into meaningful blocks) beyond the diversity among the farms.

In this article we propose an original multivariate method to deal with multiblock and multigroup data that are organized in several independent blocks of variables and one block of dependent variable in presence of several groups of individuals known a priori. The proposed method—called multiblock and multigroup PLS (mbmgPLS)—extends mgPLS to the case of multiblock and multigroup data. This is achieved by maximizing a criterion which explicitly shows that we seek, step by step, common vectors of loading weights across the groups for each block of variables, block and global components in the dependent variables that are optimally linked to components from the independent variables. The solution of the maximization criterion of mbmgPLS is given by means of an iterative algorithm.

This paper is organized as follows. In Sect. 16.2.1 the notations and the aim of the study are presented. The proposed multiblock and multigroup PLS method is described in Sect. 16.2.3. Some alternative methods to study multiblock and multigroup data are discussed in Sect. 16.3. Finally, in Sect. 16.4, mbmgPLS is illustrated on the basis of a questionnaire aiming at studying the problem of cannabis consumption among teenagers in thirteen countries. The survey was conducted in 2011 within the European School Survey Project on Alcohol and other Drugs (www.espad.org). The discussion of the merits and limitations of the proposed approach and the perspectives are outlined in Sect. 16.5.

16.2 Method

16.2.1 Notations and Aims

The datasets $\mathbf{X}$ and $\mathbf{Y}$ collect (respectively) P and Q quantitative variables measured on the same set of N observations. In the case of multiblock setting, we have K blocks of datasets denoted $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$ corresponding to $(P^{(1)}, \ldots, P^{(K)})$ variables such that the total number of variables is equal to $\sum_k P^{(K)} = P$. We define the overall concatenated dataset as

$$\mathbf{X} = \left[\mathbf{X}^{(1)} | \ldots | \mathbf{X}^{(K)}\right].$$

In addition, the set of N observations is partitioned in M subsets of N_m observations ($N = \sum_m N_m$) such that the K blocks of the independent variables for the m-th group of observations are denoted: $\mathbf{X}_m^{(1)}, \ldots, \mathbf{X}_m^{(K)}$, whereas the dependent variables of the m-th group are stored in $\mathbf{Y}_m$. The structure of the data is illustrated in Fig. 16.1. We also assume that all datasets $\mathbf{X}_m^{(k)}$ ($k = 1, \ldots, K$) and $\mathbf{Y}_m$ are centered.

Multiblock and multigroup PLS analysis seeks common loading weights denoted $\mathbf{a}^{(k)}$ (for $k = 1, \ldots, K$) and $\mathbf{b}$ to investigate the relationships between the $(P^{(1)}, \ldots, P^{(K)})$ independent variables with the Q dependent variables. We denote by $\mathbf{a}_m^{(k)}$ the vector of dimensions $(P^{(k)} \times 1)$ of the loading weights specific to each block $k = (1, \ldots, K)$ and each group $m = (1, \ldots, M)$. For a given block $\mathbf{X}^{(k)}$ (resp. $\mathbf{Y}$), a common vector of loading weights $\mathbf{a}^{(k)}$ (resp. $\mathbf{b}$) is sought. Thereafter, the associated group and block scores $\mathbf{t}_m^{(k)} = \mathbf{X}_m^{(k)}\mathbf{a}^{(k)}$ are computed in connection with the common loading weights $\mathbf{a}^{(k)}$. For the independent blocks, the global group component $\mathbf{t}_m$ associated with the N_m individuals in group $m = (1, \ldots, M)$ is defined as a weighted sum of the block and group scores: $\mathbf{t}_m = \sum_k \omega^{(k)}\mathbf{t}_m^{(k)}$. This latter constraint allows us to link the blocks to each other. We denote by $\mathbf{t}$ the global component of the N individuals formed by the vertical concatenation of the group scores $\mathbf{t}_m$. This concatenation is possible because all the group components $\mathbf{t}_m$ share the same loading weights: $\mathbf{t} = \mathbf{X}\mathbf{a}$. This global component $\mathbf{t}$ can be used to depict the N individuals. For the dependent data, group component $\mathbf{u}_m$ is calculated

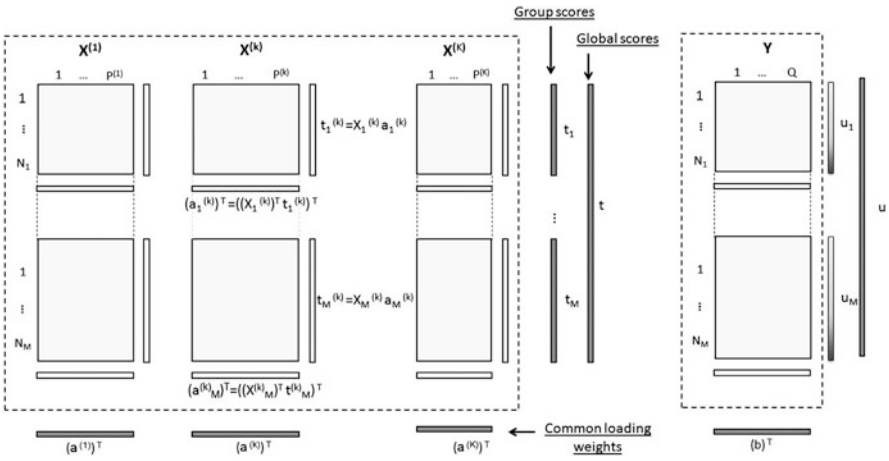

Fig. 16.1 Graphical display of multiblock and multigroup data

as $\mathbf{u}_m = \mathbf{Y}_m\mathbf{b}$. The global component $\mathbf{u}$ of the N individuals is formed by the vertical concatenation of the group scores $\mathbf{u}_m$: $\mathbf{u} = \mathbf{Yb}$. All the loading weights and components are associated with a given dimension $h = (1, \ldots, H)$ where H is the maximum dimension of the analysis. All these elements are illustrated in Fig. 16.1.

16.2.2 Preprocessing

Depending on the nature of the variables, different pre-treatments can be applied. For instance, if the variables are not in the same scale unit, they can be standardized to unit variance or norm before the analysis. Alternatively, in order to give the same importance to all the groups, the variables can be standardized within each group. However, such a choice requires that the sample size in each group is large enough to ensure correct estimations of the group variances. In addition, in order to put all the blocks on the same footing (i.e., same total variance), the data blocks are normalized by dividing all the elements of block $\mathbf{X}^{(k)}$ by its norm defined as $\sqrt{\text{trace}\left(\mathbf{X}^{(k)T}\mathbf{X}^{(k)}\right)}$.

16.2.3 Multiblock and Multigroup PLS

We consider the multigroup context, where a dependent dataset, $\mathbf{Y}$, is predicted by several explanatory ones $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)})$, both these datasets being a priori partitioned into the same M groups of individuals. We propose an original method called multiblock and multigroup PLS (mbmgPLS). Multigroup and muliblock

PLS seeks vectors of loading weights $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(K)})$ and $\mathbf{b}$ which are respectively common to all the groups within each independent block and dependent block. The mbmgPLS method is based on a criterion that aims at maximizing the link between the dependent group component $\mathbf{u}_m = \mathbf{Y}_m\mathbf{b}$ and the independent group component $\mathbf{t}_m$ common to all the blocks. Formally, we seek to maximize the following criterion:

$$\sum_{m=1}^{M} N_m \text{cov}(\mathbf{t}_m, \mathbf{Y}_m\mathbf{b}) \tag{16.1}$$

under the constraints

$$\mathbf{t}_m = \sum_{k=1}^{K} \omega^{(k)} \mathbf{X}_m^{(k)} \mathbf{a}^{(k)}, \quad \sum_{k=1}^{K} (\omega^{(k)})^2 = 1, \quad \|\mathbf{a}^{(k)}\| = \|\mathbf{b}\| = 1$$

Clearly, criterion (16.1) consists in determining for each group $m = (1, \ldots, M)$, a global group score $\mathbf{t}_m$ which is a linear combination of block and group components. This global component is sought in such a way that it is tightly linked to a latent component from $\mathbf{Y}$. Moreover, within each $\mathbf{X}$ block and $\mathbf{Y}$ block, the loading weights are considered to be identical across the various groups. The criterion (16.1) can be written in terms of the common loading weights as follows.

$$\begin{aligned} \sum_{m=1}^{M} N_m \text{cov}(\mathbf{t}_m, \mathbf{Y}_m\mathbf{b}) &= \sum_{m=1}^{M} \sum_{k=1}^{K} \omega^{(k)} \text{cov}(\mathbf{X}_m^{(k)}\mathbf{a}^{(k)}, \mathbf{Y}_m\mathbf{b}) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \omega^{(k)} (\mathbf{a}^{(k)})^T (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m\mathbf{b} \\ &\text{with } \sum_{k=1}^{K} (\omega^{(k)})^2 = 1, \text{ and } \|\mathbf{a}^{(k)}\| = \|\mathbf{b}\| = 1 \end{aligned} \tag{16.2}$$

In the case of one block of independent data ($K = 1$) with a group structure among the individuals, mbmgPLS leads to mgPLS (Eslami et al. 2013b). If there is no group structure among the individuals ($M = 1$), mbmgPLS leads to multiblock PLS (Wold 1984, 1966; Wangen and Kowalski 1988). If there is no group structure among the individuals ($M = 1$) and if there is only one block ($K = 1$) of variables mbmgPLS leads to standard PLS regression.

We propose an iterative algorithm to solve the maximization problem of mbmgPLS. The proposed algorithm for each dimension is presented in the following (see Algorithm 16.1).

The convergence of the algorithm is guaranteed by the fact that at each stage the parameters are computed so as to maximize the criterion at hand, the other parameters being held constant. Thus, at each stage, the criterion increases and

Algorithm 16.1: Algorithm for mbmgPLS for each dimension

[STEP-0] Initialization: Choose starting vectors of common loading weights $\mathbf{a}^{(k)}$ for $k = (1, \ldots, K)$ and $\mathbf{Y}$ common loading weight vector $\mathbf{b}$;

[STEP-1] Compute the weights $\omega^{(k)}$ for $k = (1, \ldots, K)$:

$$\omega^{(k)} = \frac{(\mathbf{a}^{(k)})^T \sum_m (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b}}{\sqrt{\sum_k ((\mathbf{a}^{(k)})^T \sum_m (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b})^2}} = \frac{(\mathbf{a}^{(k)})^T (\mathbf{X}^{(k)})^T \mathbf{Y}\mathbf{b}}{\sqrt{\sum_k ((\mathbf{a}^{(k)})^T (\mathbf{X}^{(k)})^T \mathbf{Y}\mathbf{b})^2}}$$

[STEP-2] Compute the block and group scores $\mathbf{t}_m^{(k)} = \mathbf{X}_m^{(k)} \mathbf{a}^{(k)}$ for $m = (1, \ldots, M)$ and $k = (1, \ldots, K)$;
[STEP-3] Compute the global group scores $\mathbf{t}_m = \sum_k \omega^{(k)} \mathbf{X}_m^{(k)} \mathbf{a}^{(k)}$ for $m = (1, \ldots, M)$;
[STEP-4] Compute the dependent vector of common loading weights:

$$\mathbf{b} = \frac{\sum_m \mathbf{Y}_m^T \mathbf{t}_m}{\| \sum_m \mathbf{Y}_m^T \mathbf{t}_m \|}$$

[STEP-5] Compute the dependent vector of group loading weights: $\mathbf{b}_m = \frac{\mathbf{Y}_m^T \mathbf{t}_m}{\| Y_m^T \mathbf{t}_m \|}$ for $m = (1, \ldots, M)$;
[STEP-6] Compute the dependent component $\mathbf{u} = \mathbf{Y}\mathbf{b}$;
[STEP-7] Compute the common independent loading weights $\mathbf{a}^{(k)}$, for $k = (1, \ldots, K)$:

$$\mathbf{a}^{(k)} = \frac{\sum_m \omega^{(k)} (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b}}{\| \sum_m \omega^{(k)} (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b} \|}$$

[STEP-8] Compute the block and group loading weights

$$\mathbf{a}_m^{(k)} = \frac{(\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b}}{\| (\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b} \|}$$

for $m = (1, \ldots, M)$ and $k = (1, \ldots, K)$;
[STEP-9] Repeat the process starting from Step-1 until convergence (i.e., insignificant variation in criterion $\sum_{m=1}^{M} N_m \text{cov}(\mathbf{t}_m, \mathbf{Y}_m \mathbf{b})$ between two successive iterations).

because it is bounded, the process converges. It is noteworthy that in the course of the algorithm we compute block and group loadings given by:

$$\mathbf{a}_m^{(k)} = \frac{(\mathbf{X}_m^{(k)})^T \mathbf{Y}_m \mathbf{b}}{\| (X_m^{(k)})^T \mathbf{Y}_m \mathbf{b} \|}$$

for block k and group m and

$$\mathbf{b}_m = \frac{\mathbf{Y}_m^T \mathbf{t}_m}{\| \mathbf{Y}_m^T \mathbf{t}_m \|}$$

for group m. These vectors can be useful to highlight differences among the groups as discussed in a subsequent section.

A second order component and loading weights vectors can be determined by means of a deflation strategy. The idea which is nowadays popular within the framework of PLS regression is to remove from each block $\mathbf{X}^{(k)}$ $(k = 1, \ldots, K)$ and $\mathbf{Y}$ the information already accounted for by the first latent variable $\mathbf{t}$. More precisely, this consists in replacing the dataset $\mathbf{X}^{(k)}$ (resp. $\mathbf{Y}$) by its residuals in the orthogonal projection onto the subspace spanned by the global component $\mathbf{t}$ (i.e., $(\mathbf{I}-(\mathbf{t}\mathbf{t}^T/\mathbf{t}^T\mathbf{t}))\mathbf{X}^{(k)}$) and $(\mathbf{I}-(\mathbf{t}\mathbf{t}^T/\mathbf{t}^T\mathbf{t}))\mathbf{Y}$; $\mathbf{I}$ being the identity matrix. By applying the mbmgPLS algorithm to these new datasets, we obtain scores and loading weights for the second dimension. Subsequent components and loading weights can be found by reiterating this process.

16.2.4 Similarity Among Group and Common Loading Weights

In order to measure the similarity between the group loading vectors and the common loading vector, we set up, for each group, a sequence of indices indexed by the number of components retained in the model (Eslami et al. 2014b). Each index ranges between 0 and 1 and reflects the extent to which the group vectors of loadings and the associated common block vectors of loadings are similar. Specifically, let $\mathbf{A}_m^k = [\mathbf{a}_m^{(k)1}, \ldots, \mathbf{a}_m^{(k)H}]$ and $\mathbf{A}^k = [\mathbf{a}^{(k)1}, \ldots, \mathbf{a}^{(k)H}]$ be (respectively) the group and common loadings matrices. We will investigate whether these matrices lead to similar vectors of loadings up to a given dimension h for $h = (1, \ldots, H)$. For this purpose we consider the sequence of similarity indices given by Eq. (16.3).

$$S^h = \frac{1}{h}\sum_{r=1}^{h}\left|\left(\mathbf{a}_m^{(k)r}\right)^T\mathbf{a}^{(k)r}\right| = \frac{1}{h}\sum_{r=1}^{h}\left|\cos\left(\mathbf{a}_m^{(k)r}, \mathbf{a}^{(k)r}\right)\right| \quad \text{for } h = (1, \ldots, H) \quad (16.3)$$

16.3 Alternative Methods

The first alternative method stems from the idea that multiblock and multigroup datasets can be considered as data with external information on the individuals, and on the variables. The external information on the individuals is reflected by the indicator matrix $\mathbf{G}$ (of size $N \times M$) which indicates the membership of the individuals to the various groups. Likewise, the external information on the variables is reflected by the indicator matrices $\mathbf{L}_\mathbf{X}$ (of size $K \times P$) and $\mathbf{L}_\mathbf{Y}$ (of size $K' \times Q$) which reflects the block structure. Thereupon, several methods have been proposed, related to the framework of Canonical Correlation Analysis (Takane and Hwang 2002; Takane et al. 2006), Redundancy Analysis (Takane and Jung 2006) and Co-Inertia Analysis (Amenta 2008). These methods mainly consist of two steps: the first (external) step aims at partitioning the datasets into several additive (orthogonal) parts according to the external information, and the second (internal) one aims at applying a two-block method to each pair of the decomposed matrices.

Other alternative methods are dedicated to the prediction purpose with a stepwise determination of common parameters. An extension of Ordinary Least Squares and Partial Least Squares to the multiblock framework is proposed by Jorgensen et al. (2007). This method aims at seeking common regression coefficients (to all the groups) by means of an iterative algorithm where $\mathbf{Y}$ is explained both by the design matrix $\mathbf{G}$ (group membership) and a summary of the explanatory datasets $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)})$ oriented towards the explanation of the within-group variation of $\mathbf{Y}$. However, because this method performs poorly in presence of correlation between $\mathbf{G}$ and $\mathbf{X}$ and because the convergence of the iterative algorithm is not guaranteed, sequential and orthogonalized PLS have been proposed by Måge et al. (2008). The first step of this latter strategy of analysis consists in finding a common subspace by means of Generalized Canonical Correlation Analysis. The second step aims at orthogonalizing each explanatory block with respect to this common space in order to identify specific information to each block. Although the method is mainly devoted to seeking common regression coefficients, the components and loading weights computed in the course of the algorithm can be used to depict the relationship among the individuals and the variables. In the Structural Equation Modeling framework, the method proposed by Chin et al. (2003) to incorporate interaction effects in PLS Path Modeling is of particular interest for multigroup and multiblock setting. It aims at modeling the relationships between a dependent dataset $\mathbf{Y}$ and one or several explanatory datasets by taking account of the group structure summed up with the design matrix $\mathbf{G}$ (group memberships). However, this method may be time consuming particularly in presence of a large number of groups, variables and blocks since the number of interaction terms highly increases.

Finally, Martin et al. (2002) proposed multigroup PLS as an extension of the ideas of Flury (1984) based on the algorithm presented by Lindgren et al. (1993). This method aims at finding common loadings to all the groups from a pooled variance-covariance matrix. This method is interesting because the common loading weights are proposed to depict the explanatory variables in a common space. However, the method does not exhibit common dependent loading weights associated with the dependent block, and is not based on a clear optimization criterion.

16.4 Illustration

16.4.1 Data and Aims

The proposed method of analysis is illustrated on multiblock and multigroup data from the analysis of the 2011 European School Survey Project on Alcohol and Drugs (ESPAD) which aims at collecting data on alcohol and drugs following the same protocol in various countries (*www.espad.org*). One aspect in the questionnaire was the Cannabis Abuse Screening Test (CAST) (Legleye et al. 2011) chosen by

thirteen countries. The database consists of 5,204 teenagers who reported having smoked cannabis in the previous 12 months. The multigroup structure of the individuals derives from the fact that the data are sampled from 13 countries namely; Belgium ($N_1 = 331$), Cyprus ($N_2 = 177$), Czech Republic ($N_3 = 1{,}013$), France ($N_4 = 723$), Germany ($N_5 = 365$), Italy ($N_6 = 617$), Kosovo ($N_7 = 55$), Latvia ($N_8 = 292$), Liechtenstein ($N_9 = 52$), Poland ($N_{10} = 1,113$), Romania ($N_{11} = 93$), Slovak Republic ($N_{12} = 246$) and Ukraine ($N_{13} = 127$). These data are multiblock because they comprise five blocks: one block of dependent variables and ($K = 4$) blocks of independent variables. The block of dependent variables relates to the drug consumption, namely $\mathbf{Y}$ ($Q = 6$ variables; Table 16.1) (Legleye et al. 2007). Four blocks of independent variables ($K = 4$) related to the use and the context of cannabis, namely $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ respectively with ($P^{(1)} = 5$), ($P^{(2)} = 4$), ($P^{(3)} = 5$) and ($P^{(4)} = 2$) variables (see Table 16.2). All these variables are considered as quantitative variables and when the case applies, the categorical variables were replaced by the indicator (or, 0/1) variables of their categories. The main aim is to describe the link between the cannabis consumption variables (CAST) and the independent variables in ($K = 4$) blocks which describe the drug use and consumption context, beyond the diversity among the countries.

For data prepossessing, all the variables were scaled. Moreover, in order to set the ($K = 4$) independent blocks on the same footing, blocks were divided by the square root of their total variance. The proposed method and the associated interpretations tools are performed using code programs developed in the free software `R` (Development Core Team 2012).

16.4.2 Overall Interpretation of mbmgPLS Outputs

We applies mbmgPLS to investigate cannabis consumption ($\mathbf{Y}$) by the use and context variables ($\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(4)}$), taking into account the diversity between countries. Two dimensions were retained as they explained together 63 % of the total variance. The importance of the blocks in the determination of the successive global components can be reflected by the squared values of block weights $(\omega^{(k)})^2$, for ($k = 1, \ldots, 4$), where $\sum_{k=1}^{4} (\omega^{(k)})^2 = 1$. These values are shown in Table 16.3.

Table 16.1 Cannabis consumption: dependent variables $\mathbf{Y}$

Abbreviation	Dependent variables
CAST1	Have you smoked cannabis before midday?
CAST2	Have you smoked cannabis when you were alone?
CAST3	Did You have memory problems when you smoked cannabis?
CAST4	Do you have friends or relatives who told you to reduce or stop cannabis consumption?
CAST5	Have you tried to reduce or stop the cannabis use without succeeding?
CAST6	Have you had problems because of your cannabis use (argument, fight, accident, poor results at school, etc.)?

Table 16.2 Cannabis consumption: independent variables ($\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}, \mathbf{X}^{(4)}$)

Abbreviation	Independent variables
$\mathbf{X}^{(1)}$ alcohol and tobacco consumption:	
C09	Frequency of cigarette smoking
C12b, C12c	Frequency of alcohol consumption in the last 12 months and in the past 30 days
C19a,C19b	Frequency of drunkenness in life and in the last 12 months
$\mathbf{X}^{(2)}$ frequency and history of cannabis use:	
C25b, C25c	Frequency of cannabis use in the last 12 months and in the past 30 days
C26	Age at first cannabis use
C27ab	Frequency of refusals of cannabis smoking in life
$\mathbf{X}^{(3)}$ drug use by peers and availability of cannabis:	
C34a, C34b, C34c, C34d	Proportion of friends who smoke cigarettes, drink alcoholic beverages, get drunk, smoke cannabis
C24	Perceived availability of cannabis: How difficult do you think it would be for you to get marijuana or hashish (cannabis) if you wanted?
$\mathbf{X}^{(4)}$ Perceived risk of cannabis consumption:	
C36f, C36h	How much do you think people risk harming themselves (physically or in other ways), if they try marijuana or hashish (cannabis) once or twice, smoke marijuana or hashish (cannabis) regularly

Table 16.3 Percentage of squared block weights $(\omega^{(k)})^2$ for the first five dimensions

	Dim1	Dim2	Dim3	Dim4	Dim5
$\mathbf{X}^{(1)}$	16.95	7.70	29.98	18.16	9.68
$\mathbf{X}^{(2)}$	**55.90**	17.62	23.15	55.83	57.31
$\mathbf{X}^{(3)}$	17.63	13.35	18.93	15.07	28.06
$\mathbf{X}^{(4)}$	9.52	**61.33**	27.94	10.94	4.96

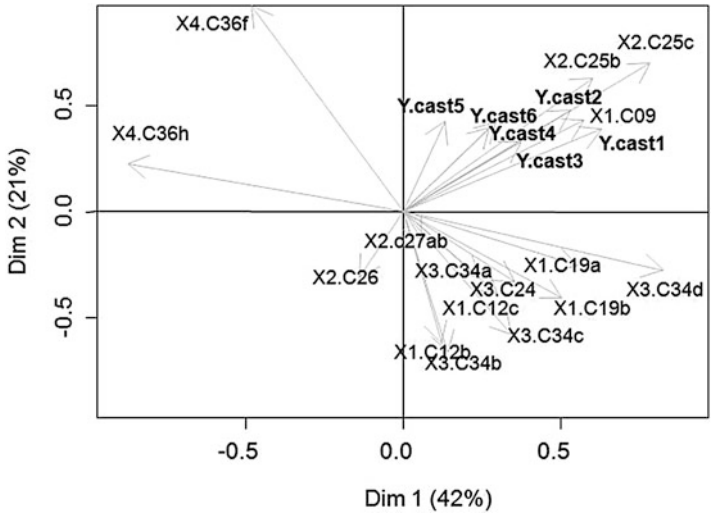

Fig. 16.2 Common loadings plot for the first two global components

For the first dimension the global component is mainly explained by the variables from the block $\mathbf{X}^{(2)}$, whereas the second dimension is mainly related to the variables from the block $\mathbf{X}^{(4)}$.

16.4.2.1 Interpretation at the Variable Level

The loadings plots associated with the ($K = 4$) independent blocks and $\mathbf{Y}$ block are depicted in Fig. 16.2. The main finding is that five CAST variables, especially non-recreational use CAST1 and CAST2 (smoking before midday and smoking alone), are positively correlated with C25c and C25b (cannabis consumption in the previous year or month) in block $\mathbf{X}^{(2)}$ and C34d (number of friends who take cannabis) in block $\mathbf{X}^{(3)}$ and, to a lesser extent, with C09 (cigarette smoking) and C19a and C19b (the frequency of drunkenness during life and year) and negatively linked with C26 (age when the teenagers start taking cannabis) in block $\mathbf{X}^{(2)}$.

The similarity between the group loading weights (associated with each country) and common loading weights (across the countries) for each block for the first two components is shown in Fig. 16.3. The results show that in general there is a high similarity between the group loading weights (countries) and the common loading

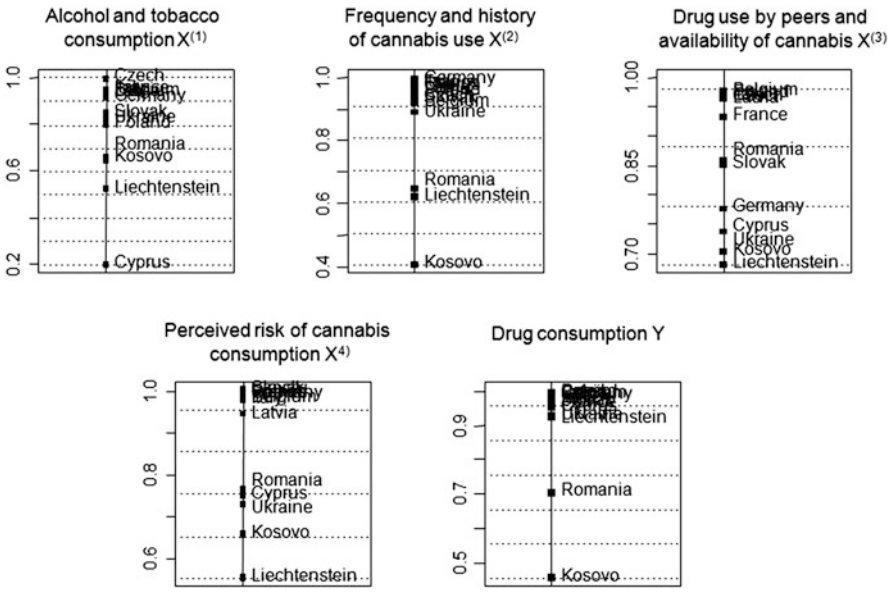

Fig. 16.3 Similarity index between group and block loading weights $\mathbf{a}_m^{(k)}$ (resp. $\mathbf{b}_m$) and common block loading weights $\mathbf{a}^{(k)}$ (resp. $\mathbf{b}$) for the first two (cumulative) global components of mbmgPLS ($h = 1$ and 2)

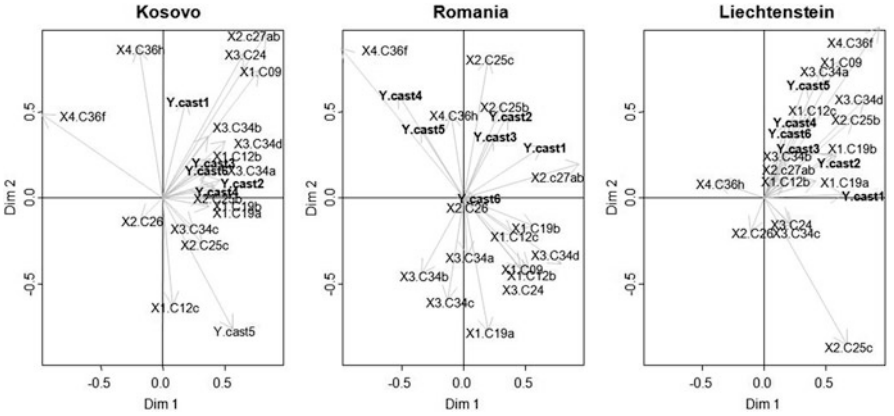

Fig. 16.4 Graphical displays of group loading weights from Kosovo, Romania and Liechtenstein for the first two components

weights expect for the countries 'Lichtenstein,' 'Kosovo,' and 'Romania.' Moreover, in blocks $\mathbf{X}^{(1)}$, $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$, 'Cyprus' is far from the common structure. It follows that for each block, the common loading weights are good representations of their associated group and block loadings. The particularity of 'Kosovo,' 'Romania,' and 'Lichtenstein' is illustrated in Fig. 16.4.

16.5 Conclusion and Perspectives

In this paper we present a new method called multiblock and multigroup PLS (mbmgPLS)—an extension of multigroup PLS—specifically tailored for the analysis of multiblock and multigroup data. In order to investigate the relationships between the variables, this method finds common loadings for all the groups of individuals for each block of variables and provides parsimonious models which ensure a better stability than separate group analyses.

For future research, it would be interesting to develop and assess the prediction ability of the proposed method and to compare its performance to other alternative methods. The proposed multigroup method can also be extended to the case of several levels of hierarchy within individuals (e.g., animals are nested within farms and farms are nested within regions). Furthermore, it would be interesting to extend mbmgPLS to the framework of several independent blocks $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)})$ to explain several dependent datasets $(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(L)})$.

References

Amenta, P.: Generalized constrained co-inertia analysis. Adv. Data Anal. Class. **2**, 81–105 (2008)

Chin, W.W., Marcolin, B.L., Newsted, P.R.: A partial least squares latent variable modeling approach for measuring interaction effects: results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. Inf. Syst. Res. **14**, 189–217 (2003)

Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: General overview of methods of analysis of multi-group datasets. Revue des Nouvelles Technologies de l'Information **25**, 108–123 (2013a)

Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: Two-block multi-group data analysis: application to epidemiology. In: Abdi, H., Chin, W.W., Esposito Vinzi, V., Russolillo, G., Trinchera, L. (eds.) New Perspectives in Partial Least Squares and Related Methods, pp. 243–255. Springer, New York (2013b)

Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: Algorithms for multi-group PLS. J. Chemom. **28**, 192–201 (2014a)

Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: Multivariate analysis of multiblock and multigroup data. J. Chemom. Intel. Lab. Syst. **133**, 63–69 (2014b)

Flury, B.N.: Common principal components in K groups. J. Am. Stat. Assoc. **79**, 892–898 (1984)

Hox, J.: Multilevel Analysis: Techniques and Applications, 2nd edn. Taylor & Francis, New York (2010)

Jorgensen, K., Mevik, B.H., Naes, T.: Combining designed experiments with several blocks of spectroscopic data. Chemom. Intel. Lab. Syst. **88**, 143–212 (2007)

Kiers, H.A.L., Ten Berge, J.M.F.: Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. Br. J. Math. Stat. Psychol. **47**, 109–126 (1994)

Krzanowski, W.J.: Principal component analysis in the presence of group structure. Appl. Stat. **33**, 164–168 (1984)

Legleye, S., Karila, L., Beckand, F., Reynaud, M.: Validation of the cast, a general population cannabis abuse screening test. J. Subst. Use **12**, 233–242 (2007)

Legleye, S., Piontek, D., Kraus, L.: Psychometric properties of the cannabis abuse screening test (cast) in a french sample of adolescents. Drug Alcohol Depend **113**, 229–235 (2011)

Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
Lindgren, F., Geladi, P., Wold, S.: The kernel algorithm for PLS. J. Chemom. **7**, 45–59 (1993)
Mage, I., Mevik, B.H., Naes, T.: Regression models with process variables and parallel blocks of raw material measurements. J. Chemom. **22**, 443–456 (2008)
Martin, E., Morris, J., Lane, S.: Monitoring process manufacturing performance. IEEE Control Syst. **22**, 26–39 (2002)
R Development Core Team: A Language and Environment of Statistical Computing. R Foundation for Statistical Computing, Vienna (2012). http://cran.r-project.org/
Takane, Y., Hwang, H.: Generalized constrained canonical correlation analysis. Multivar. Behav. Res. **37**, 163–195 (2002)
Takane, Y., Jung, S.: Generalized constrained redundancy analysis. Behaviormetrika **33**, 179–192 (2006)
Takane, Y., Yanai, H., Hwang, H.: An improved method for generalized constrained canonical correlation analysis. Comput. Stat. Data Anal. **50**, 221–241 (2006)
Wangen, L.E., Kowalski, B.R.: A multiblock partial least squares algorithm for investigating complex chemical systems. J. Chemom. **3**, 3–20 (1988)
Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (ed.) Multivariate Analysis, pp. 391–420. Academic Press, New York (1966)
Wold, S.: Three PLS algorithms according to SW. In: Symposium MULDAST (Multi-Variate Analysis in Science and Technology), Umea University, Umea, pp. 26–30 (1984)

Chapter 17
A Unified Framework to Study the Properties of the PLS Vector of Regression Coefficients

Mélanie Blazère, Fabrice Gamboa, and Jean-Michel Loubes

Abstract In this paper we propose a new approach to study the properties of the Partial Least Squares (PLS) vector of regression coefficients. This approach relies on the link between PLS and discrete orthogonal polynomials. In fact many important PLS objects can be expressed in terms of some specific discrete orthogonal polynomials, called the residual polynomials. Based on the explicit analytical expression we have stated for these polynomials in terms of signal and noise, we provide a new framework for the study of PLS. We show that this approach allows to simplify and retrieve independent proofs of many classical results (proved earlier by different authors using various approaches and tools). This general and unifying approach also sheds light on PLS and helps to gain insight on its properties.

Keywords Partial least squares regression (PLSR) • Orthogonal polynomial • Krylov subspaces

17.1 Introduction

The PLS regression method, introduced and developped by S. Wold and his coauthors in the early 1980s, is an alternative to Ordinary Least Squares (OLS) when the explanatory variables are highly collinear or when they outnumber the observations. This method has been successfully applied in a wide variety of fields and has gained an increasing attention especially in chemical engineering and genetics. The idea behind PLS is to first reduce the data to a well adapted low dimensional space to then perform prediction. Originally, it is a sequential procedure that leads to orthogonal latent components maximizing both the variance of the predictors and the covariance with the response variable. Early references on PLS are Helland (1988), Martens and Naes (1992) and Frank and Friedman (1993). For more details on PLS we also refer to Helland (2001) and Rosipal and Krämer (2006).

M. Blazère (✉) • F. Gamboa • J.-M. Loubes
Institut de mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse, France
e-mail: melanie.blazere@math.univ-toulouse.fr; fabrice.gamboa@math.univ-toulouse.fr; jean-michel.loubes@math.univ-toulouse.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_17

PLS has been mainly investigated but its properties are still little known. This is mainly due to the fact that the PLS estimate depends in a non linear way of the response. In Blazère et al. (2014), we have suggested a new way of thinking PLS based on its connections with orthogonal polynomials. In this paper, we consider again these connections to provide a general and unified framework for the study of the PLS properties.

First, in Sect. 17.2, we set the framework and the notations. Then, we recall in Sect. 17.3 the link between PLS and Krylov subspaces and also its connections with orthogonal polynomials. In Sect. 17.4, we provide an explicit formula for the PLS vector of regression coefficients which only depends on the noise on the observations and on the spectrum of the design matrix. We show that this new expression helps to gain more insight into PLS and sheds lights on this method. We also give a new expression for the filter factors. Finally, in Sect. 17.5, we show how it is obvious to then recover most of the main properties of the PLS filter factors (Lingjaerde and Christophersen 2000; Butler and Denham 2000) and also the fact that PLS globally shrinks the Least Squares estimator (De Jong 1995; Goutis 1996).

17.2 Framework

17.2.1 The Regression Model

We consider the classical linear regression model

$$Y = X\beta + \varepsilon \tag{17.1}$$

where X is the (n, p) matrix of design, $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is the vector of the observed outcome and $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the unknown parameter vector. The vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ contains the errors. To simplify we assume that X and Y are centered in such a way that there is no intercept. We allow p to be larger than n and we denote by r the rank of X. Of course $r \leq \min(n, p)$. When r is not equal to p (i.e. X is not full column rank), β is not uniquely determined by the linear predictor $X\beta$. However, because PLS is a predictive tool and not an estimation one, we are not really concerned by β in itself but by $X\beta$ which remains estimable. Therefore, even when $p > n$ the PLS procedure is still valid and provides an estimate of the response.

17.2.2 Singular Value Decomposition of the Design Matrix

An important and useful tool to study the properties of the PLS estimate is the Singular Value Decomposition (SVD). The SVD of X is given by $X = UDV^T$ where

- U is a (n, n) matrix whose columns $u_1, \ldots, u_n$ form an orthonormal basis of $\mathbb{R}^n$.
- V is a (p, p) matrix and its columns $v_1, \ldots, v_p$ form an orthonormal basis of $\mathbb{R}^p$.
- $D \in \mathbb{M}_{n,p}$ is a matrix which contains $(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ on the diagonal and zero anywhere else (i.e. $d_{ii} = \sqrt{\lambda_i}$ for $i = 1, \ldots, r$ and $d_{ij} = 0$ otherwise).

$\lambda_1, \ldots, \lambda_r$ represent the non-zero positive eigenvalues of the predictor sample covariance matrix $X^T X$. Without loss of generality we assume that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$.

Two important quantities in this paper are $p_i = (X\beta)^T u_i,\ i = 1, \ldots, n$ and $\hat{p}_i = Y^T u_i,\ i = 1, \ldots, n$. We also denote by $\tilde{\varepsilon}_i := \varepsilon^T u_i,\ i = 1, \ldots, n$ and $\tilde{\beta}_i := \beta^T v_i,\ i = 1, \ldots, p$ the projections of ε and β respectively onto the left and right eigenvectors of X.

17.3 Connections Between PLS and Discrete Orthogonal Polynomials

17.3.1 PLS and Krylov Subspaces

We recall that the minimum length least squares estimator is defined by

$$\hat{\beta}_{LS} := (X^T X)^- X^T Y = \sum_{i=1}^{r} \frac{\hat{p}_i}{\sqrt{\lambda_i}} v_i,$$

where $\hat{p}_i = Y^T u_i$. When some λ_i are small the LS estimator has a high variance. A solution can be to use Principal Components Regression. However, this method fails when the principal components corresponding to small eigenvalues have high correlations with Y. In this case an alternative can be to use the PLS method. As mentioned before, this procedure takes into account the value of the response to build a low dimensional space by maximizing both the variance of the predictors and the covariance with the response variable. Then, the data are projected into this lower space to sequentially build latent components. For the algorithmic construction we refer to Wold et al. (1984) and to Frank and Friedman (1993). In our work we do not consider the sequential construction of the PLS components. We rather use that PLS is the minimization of least squares over some Krylov subspaces.

Proposition 17.1 (Helland 1988).

$$\hat{\beta}_k^{PLS} = \underset{b \in \mathscr{K}^k(X^T X, X^T Y)}{argmin} \|Y - Xb\|^2 \tag{17.2}$$

where $\mathscr{K}^k(X^T X, X^T Y) = \left\{X^T Y, (X^T X)X^T Y, \ldots, (X^T X)^{k-1} X^T Y\right\}$, $k = 1, \ldots, r$.

The space $\mathscr{K}^k(X^T X, X^T Y)$ is called the kth Krylov subspace with respect to $X^T Y$ and $X^T X$ (Saad 1992).

17.3.2 The Discrete Orthogonal Polynomials Approach

In this section, we recall the link between PLS and orthogonal polynomials we have stated in a previous paper (see Blazère et al. 2014).

17.3.2.1 Link Between PLS and Discrete Orthogonal Polynomials

We denote by $\mathscr{P}_k$ the set of all polynomials of degree at most k and by $\mathscr{P}_{k,1}$ the subset of $\mathscr{P}_k$ constituted by polynomials with constant term equals to one. To simplify the notations we just denote by $\hat{\beta}_k$ the PLS estimate at step k.

From Proposition 17.1, it is easy to see that the PLS estimate has a polynomial representation in terms of X^TX. It is a straightforward consequence of the fact that $\hat{\beta}_k \in \mathscr{K}^k$.

Proposition 17.2. *For $1 \leq k \leq r$, we have*

$$\hat{\beta}_k = \hat{P}_k(X^TX)X^TY \quad \text{and} \quad \|Y - X\hat{\beta}_k\|^2 = \|\hat{Q}_k(XX^T)Y\|^2 \tag{17.3}$$

where $\hat{Q}_k(t) = 1 - t\hat{P}_k(t) \in \mathscr{P}_{k,1}$ satisfies $\hat{Q}_k \in \underset{Q\in\mathscr{P}_{k,1}}{argmin}\|Q(XX^T)Y\|^2$.

We call the sequence $(\hat{Q}_k)_{1\leq\leq r}$ the residual polynomials. We can show that these polynomials form a sequence of discrete orthogonal polynomials.

Proposition 17.3. *$\hat{Q}_0 := 1, \hat{Q}_1, \ldots, \hat{Q}_r$ is a sequence of discrete orthogonal polynomials with respect to the measure $d\hat{\mu}(\lambda) = \sum_{j=1}^{r} \lambda_j (u_j^T Y)^2 \delta_{\lambda_j}$.*

Notice that the weights are positive and the magnitude of the point masses depends both on the variation in X and on the correlation between X and Y along each eigenvector direction.

17.3.2.2 Expression of the Residual Polynomials

An explicit formula for the residual polynomials $\left(\hat{Q}_k\right)_{1\leq k\leq r}$ easier to interpret and well tailored to the study of the PLS properties can be stated.

Theorem 17.4. *Let $1 \leq k \leq r$ and*

$$I_k^+ = \{(j_1, \ldots, j_k) : r \geq j_1 > \ldots > j_k \geq 1\}.$$

We have

$$\hat{Q}_k(x) = \sum_{(j_1,\ldots j_k)\in I_k^+} \left[\frac{\hat{p}_{j_1}^2 \ldots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \ldots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2}{\sum_{(j_1,\ldots j_k)\in I_k^+} \hat{p}_{j_1}^2 \ldots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \ldots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2} \right] \prod_{l=1}^{k} (1 - \frac{x}{\lambda_{j_l}}). \tag{17.4}$$

where $V(\lambda_{j_1}, \ldots, \lambda_{j_k})$ *denotes the Vandermonde determinant of* $(\lambda_{j_1}, \ldots, \lambda_{j_k})$ *and we recall that* $\hat{p}_i := p_i + \tilde{\varepsilon}_i$ *with* $p_i := (X\beta)^T u_i = \sqrt{\lambda_i}\tilde{\beta}_i$ *and* $\tilde{\varepsilon}_i := \varepsilon^T u_i$.

This formula clearly shows how the disturbance on the observations and the distribrution of the spectrum impact on the residuals. This expression of the residual polynomials contains all the information necessary to study the PLS properties.

17.4 A New Expression for the PLS Estimate

17.4.1 An Explicit and Developed Formula

Using Proposition 17.2 and expanding $X^T X$ and XX^T in terms of the right and left eigenvectors of X, we can write the PLS estimate just in terms of the eigenelements of X and the residual polynomials:

$$\hat{\beta}_k = \sum_{i=1}^{r} \left(1 - \hat{Q}_k(\lambda_i)\right) \frac{\hat{p}_i}{\sqrt{\lambda_i}} v_i. \tag{17.5}$$

Of course we can state a similar expression for the linear predictor and the empirical risk.

From this decomposition of $\hat{\beta}_k$, we recover that the PLS estimate is a shrinkage estimator. In addition, we have an alternative representation of the filter factors in terms of the residual polynomials: $f_i^{(k)} := 1 - \hat{Q}_k(\lambda_i)$. Then, using Theorem 17.4, we can expand the filter factors and provide a new expression as follow:

$$f_i^{(k)} := \sum_{(j_1,\ldots,j_k)\in I_k^+} \hat{w}_{(j_1,\ldots,j_k)} \left[1 - \prod_{l=1}^{k} (1 - \frac{\lambda_i}{\lambda_{j_l}})\right], \tag{17.6}$$

where $\hat{w}_{j_1,\ldots,j_k} := \dfrac{\hat{p}_{j_1}^2 \ldots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \ldots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2}{\sum_{(j_1,\ldots,j_k)\in I_k^+} \hat{p}_{j_1}^2 \ldots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \ldots \lambda_{j_k}^2 V(\lambda_{j_1}, \ldots, \lambda_{j_k})^2}$. We can interpret the weights $(\hat{w}_{(j_1,\ldots,j_k)})_{I_k^+}$ as probabilities on polynomial $\mathscr{P}_{k,1}$ supported by polynomials having their roots in the spectrum of the design matrix. This is an alternative representation to the one of Lingjaerde and Christophersen (2000) who consider the following implicit expression for the filter factors to study the shrinkage properties of PLS (see Theorem 1 in Lingjaerde and Christophersen 2000):

$$f_i^{(k)} = \frac{(\theta_1^{(k)} - \lambda_i) \ldots (\theta_k^{(k)} - \lambda_i)}{\theta_1^{(k)} \ldots \theta_k^{(k)}} \tag{17.7}$$

where $(\theta_i^{(k)})_{1\leq i\leq k}$ are the eigenvalues of $W_k({W_k}^T X^T X W_k){W_k}^T$ and W_k contains a basis of $\mathscr{K}^k$. The interest of Formula (17.6) compared to (17.7) is that it clearly and explicitly shows how the filter factors depend on the noise and the spectrum of X. We can notice that they are completely determined by these last quantities.

17.4.2 Properties of the Filter Factors

From Eq. (17.6), we easily see that the PLS filter factors are polynomials of degree k that strongly depend on the response in a non linear and complicated way (product of the projections of the response onto the right eigenvectors and normalization factor). Furthermore, because the PLS filter factors depend on the noise ε, usual results for linear spectral methods such as PCA or Ridge, cannot be applied in this case. Contrary to those of PCA or Ridge regression, the PLS filter factors are not easy to interpret. This is closely linked to the intrinsic idea of the method that takes into account at the same time the variance of the explanatory variables and their covariance with the response. However, we have a control of the distance of the filter factors to one.

Proposition 17.5. *For all $k \leq r$, we have*

$$\left|1-f_i^{(k)}(\lambda_i)\right| \leq \left(\frac{\lambda_1-\lambda_r}{\lambda_r}\right)^n \left(1+\hat{p}_i^2\lambda_i^2 \frac{\sum_{I_{k-1,i}^+}\prod_{l=1}^k(\hat{p}_{j_l}^2\lambda_{j_l}^2)V(\lambda_{j_1},\ldots,\lambda_{j_{k-1}})^2}{\sum_{I_{k,i}^+}\prod_{l=1}^k(\hat{p}_{j_l}^2\lambda_{j_l}^2)V(\lambda_{j_1},\ldots,\lambda_{j_k})^2}\right)^{-1},$$

where $I_{k,i}^+ := \left\{(j_1,\ldots,j_k) \in I_k^+ \mid j_l \neq i, l=1,\ldots,k\right\}$.

So the highest are the λ_i and $\hat{p}_i$ the closest to one is $f_i^{(k)}$ and the largest is the amount of expansion in this eigenvector direction. Actually, the PLS filter factors are not only related to the singular values but also to the magnitude of the covariance between the principal components and the response: what seems to be important it is not the order of decrease fo λ_i but the order of decrease of $\lambda_i\hat{p}_i^2$.

We can also notice that a rough bound is $\mid f_i^{(k)} \mid \leq \max_{I_k^+}\left|1-\prod_{l=1}^k\left(1-\frac{\lambda_i}{\lambda_{j_l}}\right)\right|$ for all $k \leq r$ and all $1 \leq i \leq r$. In addition, if we have $\lambda_1(1-\varepsilon) < \lambda_i < \lambda_n(1+\epsilon)$ then a straightforward calculation leads to

$$\mid 1-f_i^{(k)} \mid < \epsilon^k.$$

17.5 Shrinkage Properties of PLS: New Proofs of Known Results

In this section, we explain how we can easily recover (once Theorem 17.4 is stated) most of the main known results on the PLS filter factors.

17.5.1 Some Peculiar Properties of the PLS Filter Factors

In this section we investigate the shrinkage properties of the PLS estimate.

1. From Formula (17.6), we easily see that there is no order on the filter factors and no link between them at each step. Furthermore, they are not always in $[0, 1]$, contrary to those of PCR or Ridge regression which always shrink in all the eigenvectors directions. In particular the PLS filter factors can be greater than one and even negative. This is one of their very particular features. PLS shrinks in some direction but can also expand in others in such a way that $f_i^{(k)}$ represents the magnitude of shrinkage or expansion of the PLS estimate in the ith eigenvectors direction. Frank and Friedman (1993) were the first to notice this peculiar property of PLS. This result was proved by Butler and Denham (2000) and independently the same year by Lingjaerde and Christophersen (2000) using Ritz eigenvalues.

   The shrinkage properties of the PLS estimate were mainly investigated by Lingjaerde and Christophersen (2000). From Formula (17.6), we easily recover the main properties they have stated for the filter factors (but without using the Ritz eigenvalues). It is for instance the case for the behaviour of the filter factors associated to the largest and smallest eigenvalue. Indeed, if $k \leq r$ and $i = r$ then $0 < \prod_{l=1}^{k}(1 - \frac{\lambda_r}{\lambda_{j_l}}) < 1$. Because $\sum_{(j_1,\ldots,j_k)\in I_k^+} \hat{w}_{(j_1,\ldots,j_k)} = 1$, we can conclude directly that $0 < f_r^{(k)} < 1$.

   On the other hand, if $k \leq r$ and $i = 1$ then

$$\begin{cases} \prod_{l=1}^{k}(1 - \frac{\lambda_1}{\lambda_{j_l}}) < 0 \text{ if } k \text{ is odd} \\ \prod_{l=1}^{k}(1 - \frac{\lambda_1}{\lambda_{j_l}}) > 0 \text{ if } k \text{ is even} \end{cases} \text{ so that } \begin{cases} f_1^{(k)} > 1 \text{ if } k \text{ is odd} \\ f_1^{(k)} < 1 \text{ if } k \text{ is even} \end{cases}.$$

   This is exactly Theorem 3 of Lingjaerde and Christophersen (2000).

   Hence, the filter factor associated to the largest eigenvalues oscillates around one depending on the parity of the index of the factors. For the other filter factors we can have either $f_i^{(k)} \leq 1$ (PLS shrinks) or $f_i^{(k)} \geq 1$ (PLS expands) depending on the distribution of the spectrum.
2. Notice that for orthogonal polynomials of a finite supported measure there exists a point of the support of the discrete measure between any two of their zeros (Baik et al. 2007). Moreover, the roots of these polynomials belong to the interval whose bounds are the extreme values of the support of the discrete measure. Therefore, from Proposition 17.3, we deduce that all the k zeros of $\hat{Q}_k$ lie in $[\lambda_r, \lambda_1]$ and no more than one zeros lies in $[\lambda_i, \lambda_{i-1}]$, where $i = 1, \ldots, r+1$ and by convention $\lambda_{r+1} := 0$ and $\lambda_0 := +\infty$. We immediately deduce that the

eigenvalues $[\lambda_r, \lambda_1]$ can be partitioned into $k+1$ consecutive disjoint non empty intervals denoted by $(I_l)_{1\le l\le k+1}$ that first shrink and then alternately expand or shrink the OLS. In other words

$$\begin{cases} f_i^{(k)} \le 1 \text{ if } \lambda_i \in I_l, & l \text{ odd} \\ f_i^{(k)} \ge 1 \text{ if } \lambda_i \in I_l, & l \text{ even} \end{cases}.$$

This is Theorem 1 of Butler and Denham (2000). Notice that this result has been also proved independently by Lingjaerde and Christophersen (2000) using the Ritz eigenvalues theory (see Theorem 4).

3. Furthermore, we also recover Theorem 2 of Butler and Denham (2000):

Theorem 17.6. *For $i = 1, \ldots, n$*

$$f_i^{(r-1)} = 1 - C\left(\hat{p}_i\lambda_i \prod_{j=1, j\neq i}^{r} (\lambda_j - \lambda_i)\right)^{-1},$$

where C does not depend on i.

In addition we have the exact expression for the constant which is equal to

$$C = \left[\left(\prod_{j=1}^{r}\lambda_j\right)\sum_{l=1}^{r}\left(p_l^2\lambda_l^2\prod_{\substack{j=1\\ j\neq l}}^{r}(\lambda_l-\lambda_j)^2\right)^{-1}\right]^{-1} \tag{17.8}$$

Proof. Based on Formula (17.6), we have

$$f_i^{(r-1)} = 1 - \frac{\prod_{j=1,j\neq i}^{r}\left(\hat{p}_j^2\lambda_j(\lambda_j-\lambda_i)^{-1}\right)V(\lambda_1,\ldots,\lambda_r)^2}{\sum_{l=1}^{r}\left[\prod_{\substack{j=1\\ j\neq l}}^{r}\left(\hat{p}_j^2\lambda_j^2\right)V(\lambda_1,\ldots,\lambda_{l-1},\lambda_{l+1},\ldots,\lambda_r)^2\right]}$$

$$= 1 - \left(\hat{p}_i^2\lambda_i\prod_{\substack{j=1\\ j\neq i}}^{r}(\lambda_j-\lambda_i)\right)^{-1}\frac{1}{\left(\prod_{j=1}^{r}\lambda_j\right)\sum_{l=1}^{r}\left(p_l^2\lambda_l^2\prod_{\substack{j=1\\ j\neq l}}^{r}(\lambda_l-\lambda_j)^2\right)^{-1}}. \tag{17.9}$$

So the highest is $\hat{p}_i^2\lambda_i\prod_{j=1,j\neq i}^{r}(\lambda_j-\lambda_i)$ the closest to one is $f_i^{(r-1)}$. Using similar arguments, we can also provide an independant proof of Theorem 3 of Butler and Denham (2000).

In conclusion, we have showed that, based on our new expression of the PLS filter factors, we easily recover some of their main properties. Our approach provides a unified background to all these results.

Lingjaerde and Christophersen (2000) mentioned that, using their approach based on the Ritz eigenvalues, it appears difficult to establish the fact that PLS shrinks in a global sense. Butler and Denham (2000) also considered the shrinkage properties of the PLS estimate along the eigenvector directions but again they did not prove that the PLS estimate globally shrinks the LS one. With our approach we are able to prove this fact too. This is the aim of the next section.

17.5.2 Global Shrinkage Property of PLS

As seen in the previous section, PLS expands the LS estimator in some eigen-directions leading to an increase of the LS's projected length in these directions. However, PLS globally shrinks the LS in the sense that its Euclidean norm is lower than the LS one.

Proposition 17.7. *For all $k \leq r$, we have*

$$\| \hat{\beta}_k \|^2 \leq \| \hat{\beta}_{LS} \|^2 .$$

This global shrinkage feature of PLS was first proved algebraically by De Jong (1995) and a year later Goutis (1996) proposed a new independant proof based on the PLS iterative construction algorithm by taking a geometric point of view. In addition De Jong (1995) proved the more stronger following result:

Lemma 17.8. $\| \hat{\beta}_{k-1} \|^2 \leq \| \hat{\beta}_k \|^2$ *for all $k \leq r$.*

An alternative proof of Lemma (17.8) is given below using the residual polynomials. Even if this proof follows the guidelines of an independent proof given by Phatak and de Hoog (2002), we detail it to emphasize some of the powerful properties of the residual polynomials.

Proof. The vectors $X^T\hat{Q}_0(XX^T)Y,\ldots,X^T\hat{Q}_{k-1}(XX^T)Y$ belongs to $\mathscr{K}^k(X^TX, X^TY)$ and are orthogonals (because $(\hat{Q}_k)_{0\leq k\leq r}$ is a sequence of orthogonal polynomials with respect to the discrete measure $\hat{\mu}$). Therefore, they formed an orthogonal basis for $\mathscr{K}^k(X^TX, X^TY)$. As $\hat{\beta}_k \in \mathscr{K}^k(X^TX, X^TY)$, we have

$$\| \hat{\beta}_k \|^2 := \sum_{j=0}^{k-1} \frac{\left(\hat{\beta}_k^T X^T \hat{Q}_j(XX^T)Y\right)^2}{\| X^T\hat{Q}_j(XX^T)Y \|^2}.$$

Further, because $X\hat{\beta}_k = \sum_{i=1}^{r}(1-\hat{Q}_k(\lambda_i))\hat{p}_i u_i$, we may write

$$\hat{\beta}_k^T X^T \hat{Q}_j(XX^T)Y = \sum_{i=1}^{r}(1-\hat{Q}_k(\lambda_i))\hat{Q}_j(\lambda_i)\hat{p}_i^2 = \sum_{i=1}^{r}\hat{Q}_j(\lambda_i)\hat{p}_i^2 - \sum_{i=1}^{r}\hat{Q}_k(\lambda_i)\hat{p}_i^2$$

using that

$$\sum_{i=1}^{r}\hat{Q}_j(\lambda_i)\hat{Q}_k(\lambda_i)\hat{p}_i^2 = \sum_{i=1}^{r}\hat{Q}_j(\lambda_i)\hat{p}_i^2,\ j \le k. \tag{17.10}$$

This is a very important property of the residual polynomials. This interesting feature is due to the fact that $\hat{Q}_k(XX^T)X^TY = \left[I - \hat{\Pi}_k\right]Y$ where $\hat{\Pi}_k$ is the orthogonal projector onto the space spanned by $\mathscr{K}^k(XX^T, XX^TY)$. Then, based on $\hat{\Pi}_k\hat{\Pi}_j = \hat{\Pi}_j$, we get (17.10). Thus, we have

$$\hat{\beta}_k^T X^T \hat{Q}_j(XX^T)Y = \| Y - X\hat{\beta}_j \|^2 - \| Y - X\hat{\beta}_k \|^2 = \| X\hat{\beta}_k \|^2 - \| X\hat{\beta}_j \|^2 .$$

Furthermore, for $1 \le l < k \le r$, we have $\| X\hat{\beta}_l \|^2 < \| X\hat{\beta}_k \|^2$ (because $X\hat{\beta}_l$ and $X\hat{\beta}_l$ are the orthogonal projection of Y onto two Krylov subspaces, the first one included in the other). Therefore, we deduce

$$\| \hat{\beta}_k \|^2 \le \sum_{j=0}^{k-1} \frac{\left(\| X\hat{\beta}_{k+1} \|^2 - \| X\hat{\beta}_j \|^2\right)^2}{\| X^T\hat{Q}_j(XX^T)Y \|^2} := \| \hat{\beta}_{k+1} \|^2 .$$

Finally, Proposition 17.7 follows from the fact that $\| \hat{\beta}_r \|^2 = \| \hat{\beta}_{LS} \|^2$.

17.6 Conclusion

We have proposed a general and unifying approach to study the properties of the Partial Least Squares (PLS) vector of regression coefficients. This approach relies on the link between PLS and discrete orthogonal polynomials. The explicit analytic expression of the residual polynomials sheds new light on PLS and helps to gain insight on its properties. Furthermore, we have shown that this new approach provides a better understanding for several distinct classical results.

References

Baik, J., Kriecherbauer, T., McLaughlin, K.D.-R., Miller, P.D.: Discrete Orthogonal Polynomials.(AM-164): Asymptotics and Applications (AM-164). Princeton University Press, Princeton (2007)

Blazère, M., Gamboa, F., Loubes, J.-M.: PLS: a new statistical insight through the prism of orthogonal polynomials (2014). arXiv preprint arXiv:1405.5900

Butler, N.A., Denham, M.C.: The peculiar shrinkage properties of partial least squares regression. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **62**(3), 585–593 (2000)

De Jong, S.: PLS shrinks. J. Chemom. **9**(4), 323–326 (1995)

Frank, I.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**(2), 109–135 (1993)

Goutis, C.: Partial least squares algorithm yields shrinkage estimators. Ann. Stat. **24**(2), 816–824 (1996)

Helland, I.S.: On the structure of partial least squares regression. Commun. Stat.-Simul. Comput. **17**, 581–607 (1988)

Helland, I.S.: Some theoretical aspects of partial least squares regression. Chemom. Intell. Lab. Syst. **58**(2), 97–107 (2001)

Lingjaerde, O.C., Christophersen, N.: Shrinkage structure of partial least squares. Scand. J. Stat. **27**(3), 459–473 (2000)

Martens, H., Naes, T.: Multivariate Calibration. Wiley, New York (1992)

Phatak, A., de Hoog, F.: Exploiting the connection between PLS, lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. J. Chemom. **16**(7), 361–367 (2002)

Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Subspace, Latent Structure and Feature Selection, pp. 34–51. Springer, Berlin/New York (2006)

Saad, Y.: Numerical Methods for Large Eigenvalue Problems, vol. 158. SIAM, Manchester, UK (1992)

Wold, S., Ruhe, A., Wold, H., Dunn, III, W.: The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. **5**(3), 735–743 (1984)

Chapter 18
A New Bootstrap-Based Stopping Criterion in PLS Components Construction

Jérémy Magnanensi, Myriam Maumy-Bertrand, Nicolas Meyer, and Frédéric Bertrand

Abstract We develop a new universal stopping criterion in components construction, in the sense that it is suitable both for Partial Least Squares Regressions (PLSR) and its extension to Generalized Linear Regressions (PLSGLR). This criterion is based on a bootstrap method and has to be computed algorithmically. It allows to test each successive components on a significant level α. In order to assess its performances and robustness with respect to different noise levels, we perform intensive datasets simulations, with a preset and known number of components to extract, both in the case $N > P$ (N being the number of subjects and P the number of original predictors), and for datasets with $N < P$. We then use t-tests to compare the predictive performance of our approach to some others classical criteria. Our conclusion is that our criterion presents better performances, both in PLSR and PLS-Logistic Regressions (PLS-LR) frameworks.

Keywords Partial least squares regressions (PLSR) • Bootstrap • Cross-validation • Inference

J. Magnanensi (✉)
Institut de Recherche Mathématique Avancée, UMR 7501, LabEx IRMIA, Université de Strasbourg et CNRS, 7, Rue René Descartes 67084 Strasbourg Cedex, France

Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430, Université de Strasbourg, 4, Rue Kirschleger 67085 Strasbourg Cedex, France
e-mail: magnanensi@math.unistra.fr

M. Maumy-Bertrand • F. Bertrand
Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, 7, Rue Rene Descartes 67084 Strasbourg Cedex, France
e-mail: mmaumy@math.unistra.fr; fbertrand@math.unistra.fr

N. Meyer
Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430, Université de Strasbourg, 4, Rue Kirschleger 67085 Strasbourg Cedex, France
e-mail: nmeyer@unistra.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_18

18.1 Introduction

Performing usual linear regressions between an univariate response $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^{N\times 1}$ and highly correlated predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P) \in \mathbb{R}^{N\times P}$, with N the number of subjects and P the number of predictors, or on datasets including more predictors than subjects, is not suitable or even possible. However, with the huge technological and computer science advances, providing consistent analysis of such datasets has become a major challenge, especially in domains such as medicine, biology or chemistry. To deal with them, statistical methods have been developed, especially the PLS Regression (PLSR) which was introduced by Wold et al. (1983, 1984) and described precisely by Höskuldsson (1988) and Wold et al. (2001).

PLSR consists in building $K \leqslant \text{rk}(\mathbf{X})$ orthogonal "latent" variables $\mathbf{T}_K = (\mathbf{t}_1, \ldots, \mathbf{t}_K)$, also called components, in such a way that $\mathbf{T}_K$ describes optimally the common information space between $\mathbf{X}$ and $\mathbf{y}$. Thus, these components are build up as linear combinations of the predictors, in order to maximize the covariances $\text{cov}(\mathbf{y}, \mathbf{t}_h)$ so that:

$$\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^* = \sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j, \quad 1 \leqslant h \leqslant K \tag{18.1}$$

where $\mathbf{w}_h^* = (w_{1h}{}^*, \ldots, w_{Ph}{}^*)^T$ is the vector of predictors weights in the hth component (Wold et al. 2001) and $(.)^T$ represents the transpose.

Let K be the number of components. The final regression model is:

$$\mathbf{y} = \sum_{h=1}^{K} c_h \mathbf{t}_h + \epsilon = \sum_{h=1}^{K} c_h \left(\sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j \right) + \epsilon, \tag{18.2}$$

with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ the N by 1 error vector, verifying $\mathbb{E}(\epsilon \,|\mathbf{T}_K) = 0_N$, $\mathbb{V}\text{ar}(\epsilon \,|\mathbf{T}_K) = \sigma_\epsilon^2 \times Id_N$ and $(c_1, \ldots, c_K)$ the coefficients of regression of $\mathbf{y}$ on $\mathbf{T}_K$.

An extension to Generalized Linear Regression models, noted PLSGLR, has been developed by Bastien et al. (2005), with the aim of taking into account the specific distribution of $\mathbf{y}$. In this context, the regression model is the following one:

$$g(\theta) = \sum_{h=1}^{K} c_h \left(\sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j \right), \tag{18.3}$$

with θ the conditional expected value of $\mathbf{y}$ for a continuous distribution or the probability vector of a discrete law with a finite support. The link function g depends on the distribution of $\mathbf{y}$.

The determination of the optimal number of components K, which is equal to the exact dimension of the link between $\mathbf{X}$ and $\mathbf{y}$, is crucial to obtain correct estimations of the original predictors coefficients. Indeed, concluding $K_1 < K$ leads to a loss of information so that links between some predictors and $\mathbf{y}$ will not be correctly modelled. Concluding $K_2 > K$ involves that useless information in $\mathbf{X}$ will be used to model knowledge in $\mathbf{y}$, which leads to overfitting.

18.2 Criteria Compared Through Simulations

18.2.1 Existing Criteria Used for Comparison

- In PLSR:
  1. $\mathbf{Q^2}$. This criterion is obtained by Cross-Validation (CV) with q, the number of parts the dataset is divided, chosen equal to five (5-CV), according to results obtained by Kohavi (1995) and Hastie et al. (2009). For a new component $\mathbf{t}_h$, Tenenhaus (1998) considers that it improves significantly the prediction if:

$$\sqrt{PRESS_h} \leqslant 0.95\sqrt{RSS_{h-1}} \iff Q_h^2 \geqslant 0.0975.$$

  2. **BICdof**. Krämer and Sugiyama (2011) define a *dof* correction in the PLSR framework (without missing data) and apply it to the BIC criterion. We used the *R* package *plsdof*, based on Krämer and Sugiyama (2011) work, to obtain values of this corrected BIC and selected the model which realizes the first local minimum of this BICdof criterion.
- In PLSGLR:
  1. $\mathbf{CV - MClassed}$. This criterion could only be used for PLS-Logistic Regressions (PLS-LR). Through a 5-CV, it determines for each model the number of predicted missclassified values. The selected model is the one linked to the minimal value of this criterion.
  2. **p_val**. Bastien et al. (2005) define a new component $\mathbf{t}_h$ as non-significant if there is not any significant predictors within it. An asymptotic Wald test is used to conclude to the significance of the different predictors.

18.2.2 Bootstrap Based Criterion

All the criteria described just above have major flaws including arbitrary bounds dependency, results based on asymptotic laws or derived from q-CV which naturally depends on the value of q and on the way the group will be randomly drawn. For this purpose, we adapted non-parametric bootstrap techniques in order to test directly, with some confidence level $(1 - \alpha)$, the significance of the different coefficients c_h by extracting confidence intervals (CI) for each of them.

The significance of a new component $\mathbf{t}_H$ can not be tested by simulating the usual conditional distribution given $\mathbf{X}$ of its regression coefficient linked to $\mathbf{y}$ since it would be a positive one. Since $\mathbf{t}_H$ maximizes $\mathrm{Cov}\,(\mathbf{y}, \mathbf{t}_H\,|\mathbf{T}_{H-1}\,)$, we approached the conditional distribution given $\mathbf{T}_{H-1}$ to test each new component. We define the significance of a new component as resulting from its significance for both $\mathbf{y}$ and $\mathbf{X}$, so that the extracted number of components K is defined as the last one which is significant for both of them.

Bootstrapping pairs was introduced by Freedman (1981). This technique relies on the assumption that the originals pairs $(y_i, \mathbf{t}_{i\bullet})$, where $\mathbf{t}_{i\bullet}$ represents the ith row of $\mathbf{T}_H$, are randomly sampled from some unknown $(H+1)$-dimensional distribution. This technique was developed to treat the so called correlation models, in which predictors are considered as random and ϵ may be related to them.

In order to adapt it to PLSR and PLSGLR frameworks, we designed the following double bootstrapping pairs algorithmic implementation, with $R = 500$, which will be graphically reported as **BootYT**. To avoid confusions between the number of predictors and the coefficients of the regressions of $\mathbf{X}$ on $\mathbf{T}_H$, we set M as the total number of predictors.

- Bootstrapping $(\mathbf{X}, \mathbf{T}_H)$: let $H = 1$ and $l = 1, \ldots, M$.
  1. Compute the H first components $(\mathbf{t}_1, \ldots, \mathbf{t}_H)$.
  2. Bootstrap pair $(\mathbf{X}, \mathbf{T}_H)$, returning R bootstrap samples $(\mathbf{X}, \mathbf{T}_H)^{b_r}$, $1 \leqslant r \leqslant R$.
  3. For each $(\mathbf{X}, \mathbf{T}_H)^{b_r}$, do M least squares regressions $\mathbf{x}_l^{b_r} = \overset{[h=1]}{H} \sum \left(\hat{p_{lh}}^{b_r}.\mathbf{t}_h^{b_r}\right) + \hat{\delta}_{lH}^{b_r}$.
  4. $\forall p_{lH}$, construct a $(100 \times (1-\alpha))\%$ bilateral BC_a CI, noted $\mathrm{CI}_l = \left[\mathrm{CI}_{l,1}^H, \mathrm{CI}_{l,2}^H\right]$.
  5. **If** $\exists l \in \{1, \ldots, M\}$, $0 \notin \mathrm{CI}_l$, **then** $H = H + 1$ and return to Step 1. **Else**, $K_{max} = H - 1$.
- Bootstrapping $(\mathbf{y}, \mathbf{T}_H)$: let $H = 1$. Note that for PLSGLR, a generalized regression is performed at Step 3.
  1. Compute the H first components $(\mathbf{t}_1, \ldots, \mathbf{t}_H)$.
  2. Bootstrap pair $(\mathbf{y}, \mathbf{T}_H)$, returning R bootstrap samples $(\mathbf{y}, \mathbf{T}_H)^{b_r}$, $1 \leqslant r \leqslant R$.
  3. For each pair $(\mathbf{y}, \mathbf{T}_H)^{b_r}$, do the LS regression $\mathbf{y}^{b_r} = \overset{[h=1]}{H} \sum \left(\hat{c}_h^{b_r}.\mathbf{t}_h^{b_r}\right) + \hat{\epsilon}_H^{b_r}$.
  4. Since $c_H > 0$, construct a $(100 \times (1-\alpha))\%$ unilateral BC_a CI $= \left[\mathrm{CI}_1^H, +\infty\right[$ for c_H.
  5. **While** $\mathrm{CI}_1^H > 0$ and $H \leqslant K_{max}$, **do** $H = H + 1$, and return to Step 1. **Else**, the final extracted number of components is $K = H - 1$.

18.2.3 Simulation Plan

To compare these different criteria, datasets simulations have been performed by adapting the *simul_data_UniYX* function, available in the *R* package *plsRglm* (Bertrand et al. 2014).

Simulations were performed to obtain a three dimensions common space between $\mathbf{X}$ and $\mathbf{y}$, leading to an optimal number of components equal to three. They were performed under two different cases, both in PLSR and PLSGLR framework. The first one is the $N > P$ situation with $N = 200$ and $P \in \Omega_{200} = \{7, \ldots, 50\}$. The second one is the $N < P$ situation where $N = 20$ and $P \in \Omega_{20} = \{25, \ldots, 50\}$. For each fixed couple (σ_4, σ_5), which respectively represents the standard deviation owned by the useless fourth component present in $\mathbf{X}$ and the random noise standard deviation in $\mathbf{y}$, we simulated 100 datasets with P_l predictors, $l = 1, \ldots, 100$, obtained by sampling with replacement in Ω_N.

18.3 PLSR Results

18.3.1 PLSR: Case N > P

Results are stored in three tables (one per criterion) of dimension 2255×100. The first 1230 rows correspond to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.21, \ldots, 5.81\}$ and $\sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}$. The 1025 remaining rows correspond to results for $\sigma_4 \in \{6.01, 7.01, \ldots, 30.01\}$. Columns correspond to the 100 datasets simulated per couple.

We extract each row means and report them in Fig. 18.1 as a function of σ_4 and σ_5. Each row variances were also extracted and reported in Fig. 18.2.

Considering the extracted number of components as a discriminant factor, we conclude that the Q^2 criterion is the less efficient criterion by being the most sensitive one to the increasing value of σ_5 so that it globally underestimates the number of components. Comparing BICdof and BootYT, or advertising one of them

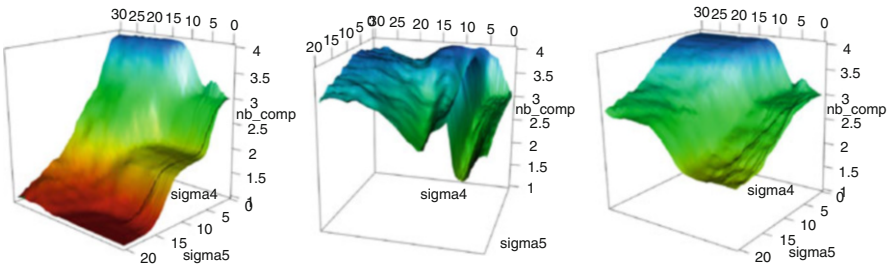

Fig. 18.1 *Left*: Q^2 row means. *Center*: BICdof row means. *Right*: BootYT row means

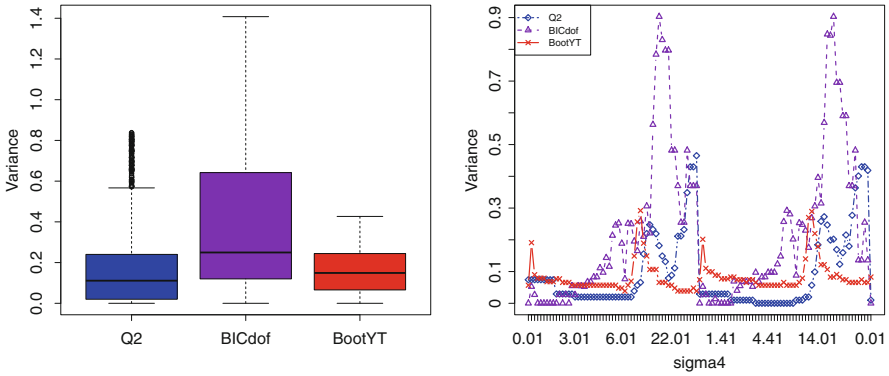

Fig. 18.2 *Left*: Boxplots of each row variances. *Right*: Evolution of variances for $\sigma_5 = \{5.01, 5.51\}$

is quite difficult in this large N case. BICdof has a low computational runtime and is the less sensitive one to the increasing value of σ_5. However, referring to Fig. 18.2, the variability of results linked to the BICdof is globally higher than the one linked to our new bootstrap based criterion, especially on datasets with large values of σ_4. BootYT is more robust than the BICdof to the increasing noise level in $\mathbf{X}$ and also directly applicable to the PLSGLR case. However, its computational runtime is clearly higher since, for each dataset, it requires $(K \times ((P_l + 1) \times R))$ least squares regressions.

18.3.2 PLSR: Case N < P

This small training sample size allows us to consider high-dimensional settings and is very interesting since usually least squares regression could not be performed.

Results are stored in three tables of dimension 287×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 1.01, \ldots, 6.01\}$ and $\sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}$. Row means are represented as a function of σ_4 and σ_5 in Fig. 18.3 and graphical representations of row variances were performed in Fig. 18.4.

In this particular case, based on Fig. 18.4, the BootYT criterion returns results with low variability for fixed couple (σ_4, σ_5) contrary to the BICdof criterion, which moreover is the most sensitive one to the increasing noise level in $\mathbf{y}$. Q^2 has a comparable attractive feature of stability but is less robust to noise level in $\mathbf{y}$ than our new bootstrap based criterion. So, by considering the number of extracted components as a discriminant factor, we conclude that the BootYT criterion is the best one to deal with these $N < P$ datasets.

However, we wanted to assess the predictive performances of each of these three criteria. Thus, for each of the 287,000 simulated datasets, we simulated 80 more

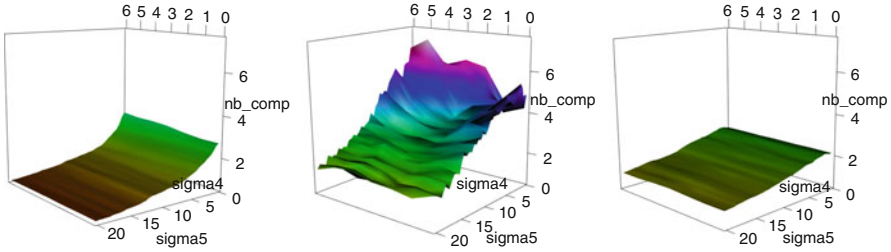

Fig. 18.3 *Left*: Q^2 row means. *Center*: BICdof row means. *Right*:BootYT row means

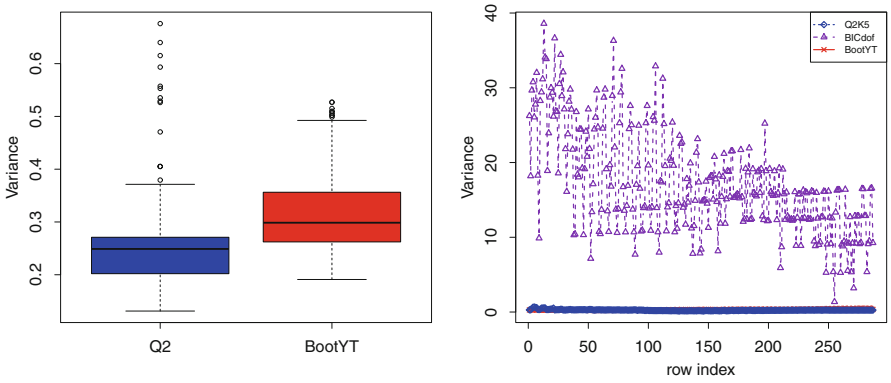

Fig. 18.4 *Left*: Boxplot of row variances. *Right*: Evolution of row variances

observations as test points and computed testing Normalized Mean Square Error (NMSE). The normalisation was done by dividing the testing MSE of the obtained model with the MSE linked to the trivial one (constant model equal to the mean of the training data). Furthermore, as mentioned by Krämer and Sugiyama (2011) (p. 702), "the large test sample size ensures a reliable estimation of the test error."

In order to compare the predictive performances of the three criteria depending on noise levels, we treat these predictive results for each couple of values (σ_4, σ_5) by testing the equalities of NMSE means with asymptotic *t*-tests with Welch-Satterthwaite *dof* approximation (Welch 1947). All these tests were performed at the alpha level equal to 0.05. Results of these *t*-tests are graphically reported in Fig. 18.5.

Concerning BootYT vs Q^2, the Q^2 has a better predictive ability for some very low values of σ_5. This result is not surprising since, in this case, the Q^2 criterion returns numbers of components closer to three than BootYT does (Fig. 18.3). However, tests results between the BICdof and the Q^2 criterion are not concluding to a significant better predictive performance of the Q^2 criterion for small values of σ_5 despite the BICdof globally overestimates the number of components in this case (Fig. 18.3). In fact, due to the small values of σ_5, the 80 supplementaries responses we simulated almost follow the same model than the first 20 ones. Thus, predictive

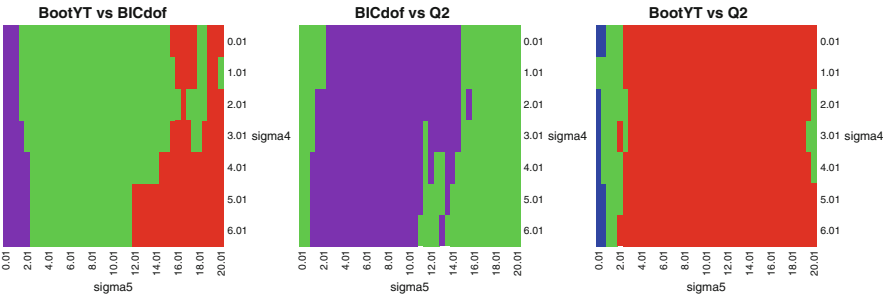

Fig. 18.5 *t*-tests results: BootYT better (*red*), BICdof better (*purple*), Q2K5 better (*blue*), no significant difference (*green*)

NMSE react in the same way than the training ones i.e. the higher the extracted number of components is, the lower the predictive NMSE are. This fact lead us to only focus on the extracted number of components when $\sigma_5 \simeq 0$, leading the Q^2 criterion to be the best one.

Finally, in all others cases, the BootYT criterion returns models with, at least, comparable or better predictive abilities than the two others.

18.3.3 PLSR: Conclusion

In the $N > P$ case, the BootYT criterion offers a better robustness to noise in $\mathbf{y}$ than the Q^2. It is also more robust to the increasing noise level in $\mathbf{X}$ than the BICdof, which moreover has some variance issues for high values of σ_4. We also conclude the BootYT criterion as a good compromise between the two others criteria, owning their advantages without their drawbacks. Concerning the $N < P$ case, our bootstrap-based criterion is globally the best one since it is less sensitive than the others to the increasing noise level in $\mathbf{y}$ and is linked to low variance results, leading to global better predictive performances.

18.4 PLS-LR Results

In this framework, due to the specific distribution of $\mathbf{y}$ and link-function $g = inv.logit$, the increase of σ_5 does not lead to a linear increase of noise level in $\mathbf{y}$. The bijectivity of g insures the presence of three common components between $\mathbf{X}$ and $\mathbf{y}$.

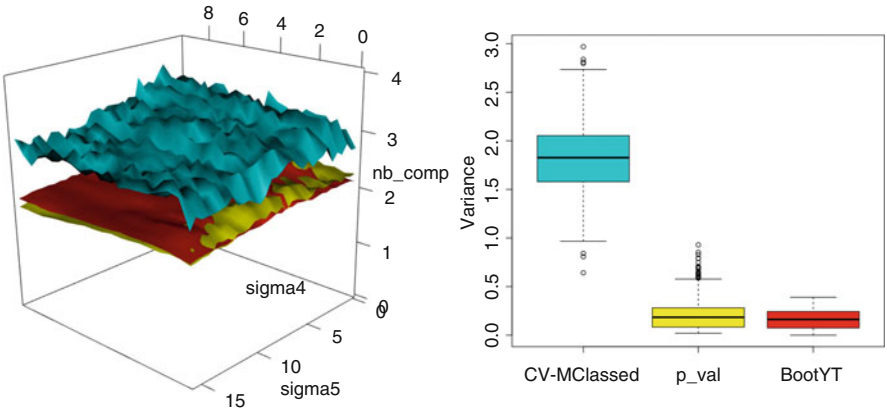

Fig. 18.6 *Left*: Row means surfaces, from top to bottom: CV-MClassed, BootYT, p_val. *Right*: Boxplots of row variances

18.4.1 PLS-LR: Case N > P

Results are stored in three tables of dimension 640×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$ and $\sigma_5 \in \{0.01, 0.51, 1.01, \ldots, 15.51\}$. We graphically report row means as a function of σ_4 and σ_5 as well as boxplots of row variances in Fig. 18.6.

Based on these graphics, the CV-MClassed performs well in estimating the optimal number of components in average. However, this good property has to be nuanced by the high variances linked to its results and which lead this criterion to be used with caution. The BootYT and p_val criteria return similar results in this asymptotic case. Both of them slightly underestimate the optimal number of components but with the advantage of low variances of their results.

18.4.2 PLS-LR: Case N < P

Results are stored in three tables of dimension 400×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$ and $\sigma_5 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$. We set the maximal value of σ_5 to 9.51, and not to 15.51 as for the $N > P$ case, in order to save computational runtime since an increasing value of σ_5 does not really affect the choice of the number of extracted components.

We graphically report row means as a function of σ_4 and σ_5 as well as boxplots of row variances in Fig. 18.7.

The CV-MClassed criterion conserves the same property of well estimating in average and issue of variability as in the $N > P$ framework. Concerning the two

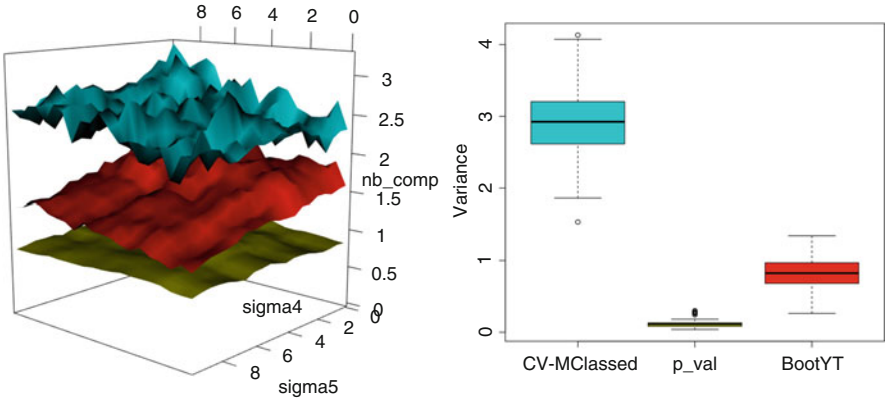

Fig. 18.7 *Left*: Row means surfaces, from top to bottom: CV-MClassed, BootYT, p_val. *Right*: Boxplots of row variances

others criteria, we observed a higher underestimating issue linked to the p_val criterion than for the BootYT one. Furthermore, they both had low variability in results they return.

In order to test their predictive performances, we simulated 80 more observations for each simulated datasets (40 000), and computed the predictive NMSE linked to each models established by the three criteria. Furthermore, since the binary response obtained by the model is equal to 1 if the estimated response is over 0.5, 0 if not, returning higher NMSE does not necessarily lead to higher number of missclassified values. Thus, we also computed the number of predictive missclassified values (*M_classed*) for each of these three criteria. Then, *t*-tests were computed for each fixed values of (σ_4, σ_5). Results of these tests are graphically reported in Fig. 18.8.

The bootstrap-based criterion is never less efficient than the other criteria. If there is globally no significant differences between bootstrapping pairs or the p_val criterion concerning the predictive NMSE, BootYT is better than this criterion concerning the predictive missclassified values. Then, there is few cases where bootstrapping pairs is significantly better than the CV-MClassed criterion concerning the predictive number of missclassified values. But, concerning the predictive NMSE, the BootYT criterion is better than this last one by returning significant smallest NMSE values, especially for high σ_5 values.

The boostrap-based criterion is also the best one by having, at least, similar predictive performances compared to the two others.

18.4.3 PLS-LR: Conclusion

Through these simulations, we can reasonably assume that the bootstrap-based criterion is globally more efficient than the other ones. In the $N > P$ case, it offers a similar stability compared to the p_val criterion. However, it globally

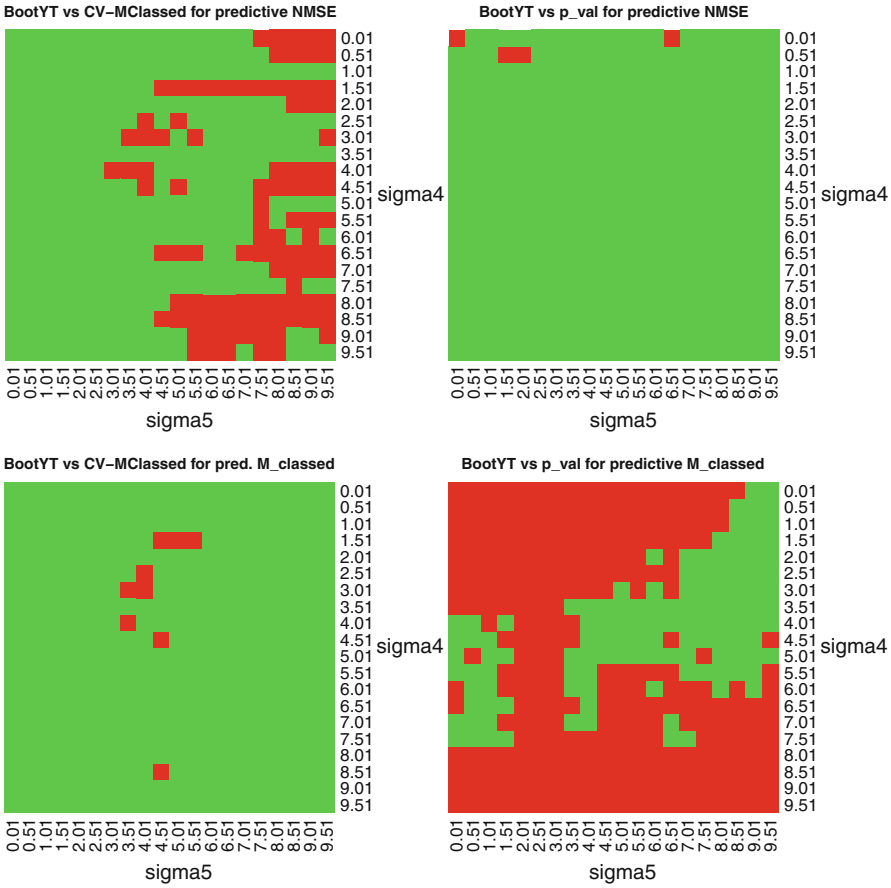

Fig. 18.8 *t*-tests results: BootYT better (*red*), no significant difference (*green*)

underestimates the optimal number of components when the CV-MClassed criterion retains it on average but with high variability. Concerning the $N < P$ case, the BootYT criterion has better predictive performances than the two others studied criteria in terms of predictive NMSE and predictive missclassified values. It also keeps a quite low variability, which is really important for a future routine implementation.

18.5 Discussion

Our new bootstrap based criterion requires huge computational runtime, so that an optimization of the algorithm seems necessary. Furthermore, the development of corrected *dof* in PLSGLR framework would also permit to develop a corrected

BIC formulation in this framework. This corrected BIC could provide an interesting alternative to the bootstrap-based criterion since it could save an important computational runtime conditionally to the fact that it would have at least similar properties to those we conclude in Sect. 18.3.

However, this new criterion represents a reliable, consistent and universal stopping criterion in order to select the optimal number of *PLS* components. It also allows users to test the significance of a new component with a preset risk level α.

In the $N > P$ PLSR framework, our simulations confirm the BICdof as being an appropriate and well designed criterion. However, our new bootstrap-based criterion is an appropriate alternative in the $N < P$ case, since the BICdof criterion suffers from overestimating issues for models with low random noise levels in $\mathbf{y}$ and returns results linked to high variances. Furthermore, both BICdof and Q^2 criteria are more sensitive than the bootstrap-based criterion to the increasing noise level in $\mathbf{y}$.

Concerning the PLSGLR framework, our simulations results lead to advertise this new bootstrap-based criterion. Indeed, in this PLS-LR case, we show that depending on the statistic we used (testing NMSE or predictive number of missclassified values) to test its predictive ability, the bootstrap-based is never significantly worse than both the CV-MClassed and p_val criteria.

References

Bastien, P., Vinzi, V.E., Tenenhaus, M.: PLS generalised linear regression. Comput. Stat. Data Anal. **48**, 17–46 (2005)

Bertrand, F., Magnanensi, J., Maumy-Bertrand, M., Meyer, N.: Partial least squares regression for generalized linear models. http://www-irma.u-strasbg.fr/~fbertrand/ (2014). Book of abstracts, User2014!, Los Angeles, p. 150

Freedman, D.A.: Bootstrapping regression models. Ann. Stat. **9**, 1218–1228 (1981)

Hastie, T., Tibshirani, R., Friedman, J.J.H.: The Elements of Statistical Learning, vol. 1, 2nd edn. Springer, New York (2009)

Höskuldsson, A.: PLS regression methods. J. Chemom. **2**, 211–228 (1988)

Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, vol. 2, pp. 1137–1143. Morgan Kaufmann Publishers Inc (1995)

Krämer, N., Sugiyama, M.: The Degrees of Freedom of Partial Least Squares Regression. J. Am. Stat. Assoc. **106**, 697–705 (2011)

Tenenhaus, M.: La Régression PLS, Théorie et pratique. Editions Technip, Paris (1998)

Welch, B.L.: The generalization of student's problem when several different population variances are involved. Biometrika **34**, 28–35 (1947)

Wold, S., Martens, H., Wold, H.: The multivariate calibration problem in chemistry solved by the PLS method. In: Matrix Pencils, pp. 286–293. Springer, Berlin/New York (1983)

Wold, S., Ruhe, A., Wold, H., Dunn, III, W.J.: The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci Stat. Comput. **5**, 735–743 (1984)

Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**(2), 109–130 (2001)

Part V
PLS Path Modeling: Breakthroughs and Applications

Chapter 19
Extension to the PATHMOX Approach to Detect Which Constructs Differentiate Segments and to Test Factor Invariance: Application to Mental Health Data

Tomas Aluja-Banet, Giuseppe Lamberti, and Antonio Ciampi

Abstract In this paper we propose an extension to the PATHMOX segmentation algorithm to detect which endogenous latent variables and predictors are responsible for heterogeneity. We also address the problem of factor invariance in the terminal nodes of PATHMOX. We demonstrate the utility of such methodology on real mental health data by investigating the relationship between *dementia*, *depression* and *delirium*.

Keywords PATHMOX • Latent variables • Segmentation

19.1 The PATHMOX Algorithm as Solution to the Heterogeneity Problem

When collecting data for a specific study, the focus is the variables, which correspond to the scientific questions raised by that study. However, in addition to the main variables, it is usual to collect some background information, in the form, for example, of socio-demographic variables such as sex, social status, or age. In our context, these variables will be referred to as segmentation variables, since they may be useful in identifying potential sources of heterogeneity. Resolving the heterogeneity may mean to perform distinct analyses based on the main variables for distinct segments of the data, defined in terms of the segmentation variables. Often heterogeneity may be controlled by defining a priori segments according

T. Aluja-Banet (✉) • G. Lamberti
Universitat Politecnica de Catalunya, Barcelona Tech, Barcelona, Spain
e-mail: tomas.aluja@upc.edu; giuseppelamb@hotmail.com

A. Ciampi
Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, Montreal, QC, Canada
e-mail: antonio.ciampi@mcgill.ca

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_19

to domain knowledge. However, it is not always possible to proceed this way, as domain knowledge may not be sufficient to suggest any a priori segmentation. On the other hand, many segmentation variables may be available, which could be used to identify and resolve heterogeneity by some appropriate algorithm. One algorithm with this aim was developed in 2009 by Gastón Sánchez, known as PATHMOX (Sanchez 2009). This technique, based on recursive partitioning, produces a segmentation tree with a distinct path models in each node. At each node PATHMOX searches among all splits based on the segmentation variables and chooses the one resulting in the maximal difference between the PLS-PM models in the children nodes. Our measure of goodness-of-split is an adaptation of Fisher's F for testing the equality of regression models (Lebart et al. 1979; Chow 1960), which permits comparing structural models.We will call it F-global test. The algorithm can be summarized as follows:

The algorithm repeats these three steps iteratively until of the following stop conditions is verified: (1) the number of individuals in the group falls below a fixed level; (2) the test's p-values are not significant at a pre-specified threshold; (3) a pre-specified maximum tree depth is attained. The conditions 1 and 3 are left to the researcher, whereas condition 2 is related to the statistical test used as split criterion in the algorithm. The output of PATHMOX is a binary tree, where each split is defined such as the PLS-PM inner models of the children node are most significant different. Notice that the outer model doesn't intervene in the split criterion, so the outer model is recalculated for every node. Also, PATHMOX is either adopted to formative and reflective construct since it only compares the path coefficients of the inner model.

19.1.1 Split Criterion: Testing the Equality of Two PLS Path Models

To better understand the split criterion used in PATHMOX, let us consider a structural model (Fig. 19.1) with two endogenous variables, η_1 and η_2, and two exogenous variables ξ_1, ξ_2. Its generalization into more complex models is straightforward, with the inconvenient of complicating the notation

The structural equations for both endogenous constructs are:

$$\eta_1 = \beta_1\xi_1 + \beta_2\xi_2 + \zeta_1 \tag{19.1}$$

$$\eta_2 = \beta_3\xi_3 + \beta_4\xi_4 + \beta_5\eta_1 + \zeta_2 \tag{19.2}$$

The disturbance terms ζ_1 and ζ_2 are assumed to be normally distributed with zero mean and costant variance, that is, $E(\zeta_1) = E(\zeta_2) = 0$ and $Var(\zeta_1) = Var(\zeta_2) = \sigma^2$. It is also assumed that $Cov(\zeta_1, \zeta_2) = 0$.

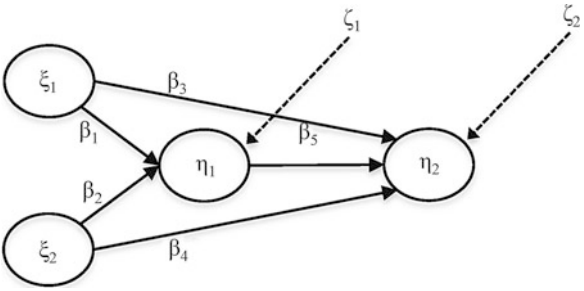

Fig. 19.1 Path diagram of a PLS model with two endogenous variables

We can define the following matrices:

$$X_1 = [\xi_1, \xi_2] \quad \text{a column matrix with the explicative latent variables of } \eta_1 \tag{19.3}$$

$$B_1 = [\beta_1, \beta_2] \quad \text{a vector of path coefficients for the regression of } \eta_1 \tag{19.4}$$

$$X_2 = [\xi_1, \xi_2, \eta_1] \quad \text{a column matrix with the explicative latent variables of } \eta_2 \tag{19.5}$$

$$B_2 = [\beta_3, \beta_4, \beta_5] \text{a vector of path coefficients for the regression of } \eta_2 \tag{19.6}$$

Then, supposing that a node splits into two children nodes A and B performing a split in n_A observations belonging to the A segment and n_B observations belonging to the B segment ($n = n_A + n_B$), we can test the null hypothesis H_0 of equality of path coefficients B_1 and B_2 for each structural equation in both segments A and B against the alternative hypothesis H_1 of having different path coefficients in each segment as:

$$H_0: \underset{[2n,1]}{\begin{bmatrix} \eta_1^A \\ \eta_2^A \\ . \\ \eta_1^B \\ \eta_2^B \end{bmatrix}} + \underset{[2n,p_1+p_2]}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_2^A \\ \cdots & \\ X_1^B & 0 \\ 0 & X_2^B \end{bmatrix}} \underset{[p_1+p_2,1]}{\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix} \zeta_1^A \\ \zeta_2^A \\ . \\ \zeta_1^B \\ \zeta_2^B \end{bmatrix}} \tag{19.7}$$

$$H_1: \underset{[2n,1]}{\begin{bmatrix} \eta_1^A \\ \eta_2^A \\ . \\ \eta_1^B \\ \eta_2^B \end{bmatrix}} + \underset{[2n,2p_1+2p_2]}{\begin{bmatrix} X_1^A & 0 & 0 & 0 \\ 0 & X_2^A & 0 & 0 \\ & \cdots\cdots & & \\ 0 & 0 & X_1^B & 0 \\ 0 & 0 & 0 & X_2^B \end{bmatrix}} \underset{[2p_1+2p_2,1]}{\begin{bmatrix} B_1^A \\ B_1^B \\ . \\ B_2^A \\ B_2^B \end{bmatrix}} + \underset{[2n,1]}{\begin{bmatrix} \zeta_1^A \\ \zeta_2^A \\ . \\ \zeta_1^B \\ \zeta_2^B \end{bmatrix}} \tag{19.8}$$

Then, assuming that the random perturbations associated to the latent variables are uncorrelated with equal variance, we can apply the Lemmas 1 and 2 of Lebart et al. (1979, pp. 201:214). Hence, the F statistic measuring the discrepancy between the two models is:

$$F_{Global} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ (p_1 + p_2)}{SS_{H_1} \Big/ [2(n_1 + n_2) - 2(p_1 + p_2)]} \tag{19.9}$$

where SS_{H_0} and SS_{H_1} stands for the corresponding sum of squares of residuals in both models, follows, under the null hypothesis, an F distribution with $p_1 + p_2$ and $2(n_1 + n_2) - 2(p_1 + p_2)$ degrees of freedom.

19.2 Extended PATHMOX

As we can see, the PATHMOX approach allows us to detect the existence of different models for different subsets of a data-set without defining a priori segments: segments are revealed as branches of the segmentation tree. However there are two possible improvements in Algorithm 19.1:

1. The F-global test used as split criterion, is a global criterion that provides only a global comparison of two PLS-PMs: it tests whether or not all the path coefficients for two structural models are equal, but, when it detects a difference, it does not provide which path coefficients are responsible for it.
2. The Pathmox approach does not assume factor invariance (i.e., it does not impose equality of the measurement model across nodes). Therefore, at different nodes the measurement models may also differ substantially. As a result the meaning of the latent variables may vary from node to node. Although such an occurrence is not necessarily undesirable, it is important to detect it to provide correct interpretations.

Algorithm 19.1: PATHMOX

Step 1. Start with the global PLS path model at the root node
Step 2. Establish a set of admissible partitions for each segmentation variable in each node of the tree
Step 3. Detect the best partition:
- **3.1.** Compare all binary partitions in all segmentation variables
- **3.2.** Apply the F-global test, calculating for each comparison a p-value
- **3.3.** Sort the p-values in a descending order
- **3.4.** Chose as the best partition the one associated to the lowest p-value

To address the first issue we asses the significantly distinct regression equations forming the PLS-PM model; to this purpose we have introduced the F-Block test (Aluja et al. 2013a):

$$F_{Block} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ p_1}{SS_{H_1} \Big/ 2(n - p_1 - p_2)} \tag{19.10}$$

On the other hand, when the previous test gives a significant results,we detect which are the path coefficients responsible for the split by the F-coefficient test (Aluja et al. 2013a):

$$F_{Coeffcient} = \frac{(SS_{H_0} - SS_{H_1}) \Big/ 1}{SS_{H_1} \Big/ 2(n \sum_{j=1}^{P} p_j)} \tag{19.11}$$

These tests are an adaptation of the F-global test, which makes it possible to investigate the causes of the difference between PLS-PMs in greater depth;

Regarding the second issue, to overcome the problem of factor invariance, we have suggested the invariance measurement test:

$$SS_{HO} - SS_{H1} \sim \chi^2_{(s-1)\sum_{k=1}^{s} p_k} \tag{19.12}$$

where s is the number of the terminal nodes and p_k is the number of manifest variables in the block k. This test enables the analyst to verify the equality of the coefficients of the measurement models in the terminal nodes (Aluja et al. 2013b) i.e. if we can suppose the same measurement model for all the terminal nodes of the tree, or if the latent variables are distinct depending of the node, so no comparison can be done between the individuals of the identified subgroups. In our extension of PATHMOX we have introduced these statistics in the tree construction to provide aid to the interpretation of the PATHMOX's results.

19.3 The Mental Health Data-Set

Using the new PATHMOX, we analyzed data on a cohort of 138 elderly patients from seven Quebec long-term care facilities, observed between July 2005 and January 2007. The data were assembled by a team of St. Marys Hospital Research Centre and were previously analyzed to answer a number of specific research questions (Voyer et al. 2011). In this analysis data was collected at the first assessment of the cohort, excluding participants with moderate-severe cognitive impairment

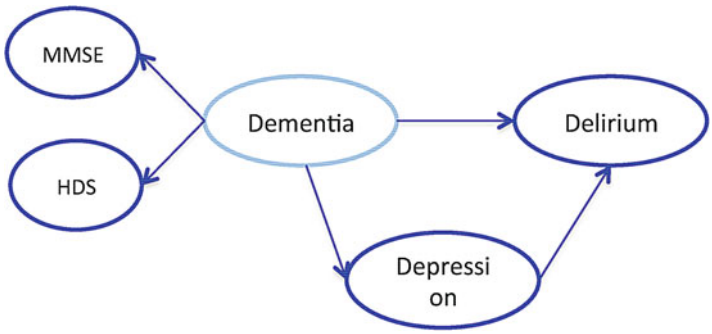

Fig. 19.2 Path diagram of *3D* latent variables relationship

and those with missing data, giving a sample size of 138. Our aim is to demonstrate our approach by investigating the relationship between three constructs representing three mental disorders that are common in elderly populations: *dementia*, *delirium* and *depression*. A total of 27 variables were available, divided in two groups; one group is formed by 24 manifest variables, which are the elements that define our PLS-PM; and the other is formed by 3 segmentation variables: patients gender, duration of hospitalizations, patients's age. We defined a measurement model for dementia based on the items of two well-known instruments: the Hierarchical Dementia Scale (HDS) (Dastoor and Cole 1988) and the Mini Mental State Examination MMSE (Folstein et al. 1975). Similarly, the items of the Cornell scale to assess depression (Alexopoulos et al. 1988) and of the Delirium Index (Pompei et al. 1995) as measure of delirium severity, were used to define the measurement models for depression and delirium respectively.

19.3.1 Manifest and Latent Variables Relationship

In our model, the *dementia* construct is treated as a second order latent variable estimated among *HDS* and *MMSE* (treated as first order latent variables). We estimated it following the repeated indicator approach (Lohmoller 1989; Wold 1982). As we show in Fig. (19.2), we have considered *dementia* as antecedent of *depression* and *delirium*, and *depression*, as antecedent of *delirium*. All latent variables are formative as the corresponding indicators describe different facets of the diseases; between *HDS* and *MMSE* and *dementia* we have considered a reflective relation since these two indices reflect the presence or not of the disease in the patients.

A description of the manifest variables used to estimate the latent variables is shown in the following table (Table 19.1):

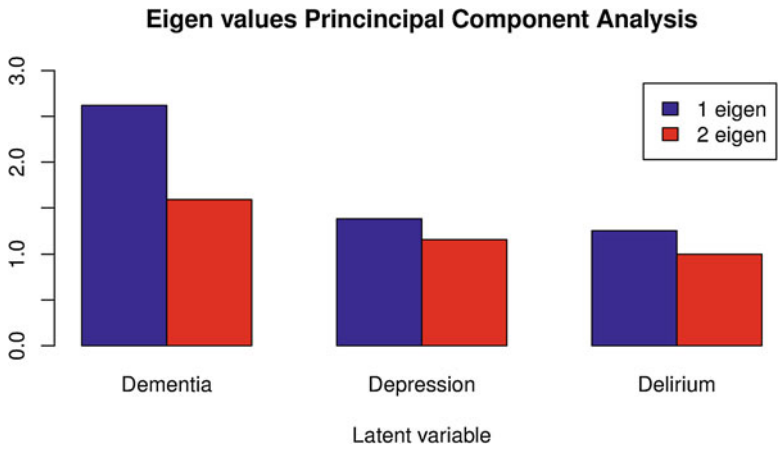

Fig. 19.3 Unidimensionality validation

19.4 3 *D* PLS Global Model Validation

The first step of a PATHMOX analysis consists of the specification of the global structural model describing the relationship between the variables of interest. In this section we present the main results obtained with the classical PLS-PM approach. We first discuss the validation of the outer model, and then we analyze the inner model. For sake of interpretation we just present the results regarding the three latent variables: *dementia*, *depression* and *delirium*

19.4.1 Outer Model

All constructs are specified as formative as said previously. In Fig. (19.3) we can verify the multidimensionality of each constructs. For all constructs, the first two eigenvalues are similar in magnitude, which suggest that the corresponding indicators describe different aspects of the latent variables and show clear evidence that they are not unidimensional.

In Fig. (19.4) we present the weights,[1] expression of the contribution of each manifest variable to the construct it is suppose to measure.

The most important weights are:

[1]The Bar-chart reveal that there are some weights with a value close to zero. In any case, as we have all formative constructs we can't delete any one of them due to each manifest variable contribute to measure a different facet of the latent variables.

Table 19.1 Indicators per latent variable

LV	MV	Item
MMSE	$mmse_3$	What month of the year is this?
	$mmse_8$	What city are we in?
	$mmse_{11}$	I am going to say 3 words. After I have said all three, I want you to repeat them
	$mmse_{12}$	Spell the word "world"
	$mmse_{16}$	Repeat the following phrase, "no ifs, ands or buts"
	$mmse_{17}$	Take this paper in your right/left hand, fold the paper in half and put it on the floor
	$mmse_{19}$	Copy this design
HDS	hds_2	Prefrontal subscale
	hds_6	Denomination subscale
	hds_7	Comprehension subscale Verbal
	hds_{11}	Reading subscale
	hds_{12}	Recent memory subscale
	hds_{18}	Motor subscale
	hds_{20}	Writing subscale
	hds_{22}	Similarities subscale
Depression	$corn_5$	Agitation; restlessness, hand wringing, hair pulling
	$corn_8$	Loss of interest; less involved in usual activities
	$corn_{15}$	Early morning awakening; earlier than usual for this individual
	$corn_{19}$	Mood congruent delusions; delusions of poverty, illness or loss
Delirium	del_4	Disorganized thinking
	del_5	Altered level of consciousness
	del_7	Memory impairment
	del_8	Perceptual disturbances
	del_{10-11}	Psychomotor agitation & Psychomotor retardation

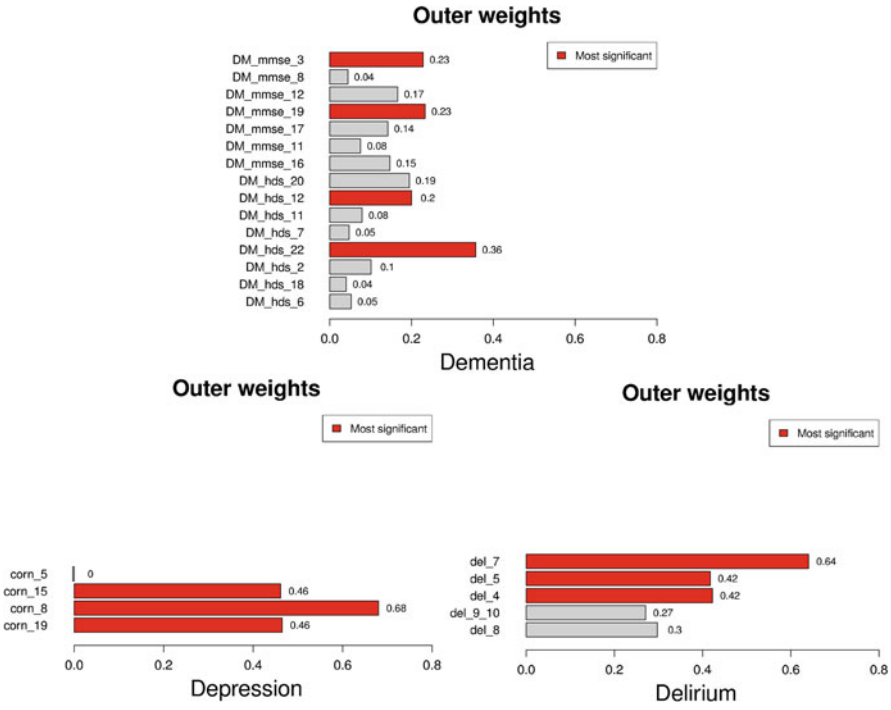

Fig. 19.4 Bar-chart of the outer weights, we represent in block the most important indicator per latent variable

- *Dementia*: *DM_mmse*_3 (What month of the year is this?), *DM_mmse*_19 (Copy this design) and *DM_hds*_12 (Recent memory sub-scale) and *DM_hds*_22 (Similarities sub-scale);
- *Depression*: *corn*_8 (Loss of interest; less involved in usual activities), *corn*_15 (Early morning awakening; earlier than usual for this individual), *corn*_19 (Mood congruent delusions; delusions of poverty).
- *Delirium*: *del*_7(Memory impairment), *del*_4 (Disorganized thinking) and *del*_5 (Altered level of consciousness)

19.4.2 Discriminant Validity

Table (19.2) shows the correlations between manifest variables and constructs. It should be noted that, as we would expect, the manifest variables (MV) are more correlated with their own constructs (LV) than with the others.

Table 19.2 Correlations between the manifest variables and the latent constructs

LV	MV	Dementia	Depression	Delirium
	DM_mmse_3	**0.56**	0.01	0.31
	DM_mmse_8	**0.21**	−0.07	0.15
	DM_mmse_12	**0.44**	0.01	0.19
	DM_mmse_19	**0.55**	0.21	0.30
	DM_mmse_17	**0.36**	0.16	0.19
	DM_mmse_11	**0.26**	−0.06	0.12
	DM_mmse_16	**0.43**	−0.08	0.14
Dementia	DM_hds_20	**0.48**	−0.02	0.19
	DM_hds_12	**0.54**	−0.08	0.28
	DM_hds_11	**0.30**	0.09	0.12
	DM_hds_7	**0.23**	−0.03	0.14
	DM_hds_22	**0.71**	−0.03	0.44
	DM_hds_2	**0.14**	0.13	0.09
	DM_hds_18	**0.27**	−0.03	0.14
	DM_hds_6	**0.16**	0.04	0.17
	corn_5	0.01	**0.29**	0.06
	corn_15	−0.03	**0.55**	0.12
Depression	corn_8	0.05	**0.68**	0.13
	corn_19	0.06	**0.60**	0.11
	del_7	0.37	0.09	**0.67**
	del_5	0.22	0.11	**0.43**
Delirium	del_4	0.26	0.05	**0.46**
	del_9_10	0.15	0.12	**0.33**
	del_8	0.15	0.15	**0.34**

19.4.3 Inner Model

The inner model validation is presented in Fig. (19.5):

We can see that when we analyze the *depression* construct, the effect of *dementia* on *depression* is very low, path coeff. = 0.04 (found not significant) and the R^2 is practically zero. When we consider *delirium* construct we find an important effect of *dementia* on *delirium* (path coeff. = 0.52) whereas the effect of *depression* on *delirium* is lower (path coeff = 0.17); in this case the R^2 is 0.31. Both coefficients are significant.

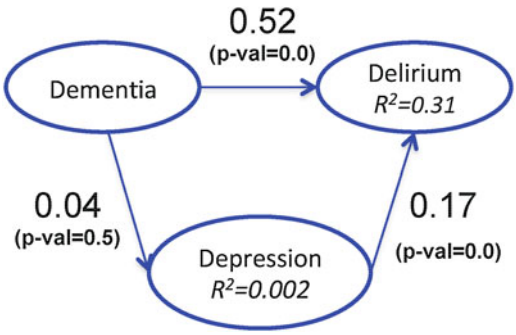

Fig. 19.5 Path diagram of *3D* latent variables relationship. In parentheses, we show the significance's *p*-value of the path coefficients

Table 19.3 *F*-block and *F*-coefficients results

F-block test			
Constructs	Statistic	P-value	Significance
Depression	2.19	0.11	No
Delirium	**3.79**	**0.01**	Yes
F-coefficient test			
Path coeff.	Statistic	P-value	Significance
Intercept on delirium	1.56	0.21	No
Dementia on delirium	2.04	0.15	No
Depression on delirium	**8.01**	**0.00**	Yes

19.5 PATHMOX Tree

We can now investigate by a tree analysis whether or not the global PLS-PM model is valid for the whole population. Our analysis suggest that the sample can be split into two subsamples, each with a distinct PLS-PM. The tree-structure obtained by PATHMOX is given in Fig. (19.6).

The partition is obtained by the segmentation variable *duration of hospitalizations* with *F*-global statistic of *1.91* and a corresponding *p*-value equal to .048. The root node is split in two children nodes: the **node two** with *61* patients with a duration of hospitalization less than one year (*shorter term hospitalization*) and the **node three** with *95* patients with a duration of hospitalization more than one year (*longer term hospitalization*).

19.5.1 Extended PATHMOX Analysis

As discussed above, the new extended version of PATHMOX also provides very useful aid to interpretation through the *F*-block and *F*-coefficient tests. The results of these tests as applied to our data are given in Table (19.3):

Note that the *F*-block test identifies the *delirium* construct as the one responsible for the difference of the inner models at the two children nodes (*F*-block statistic

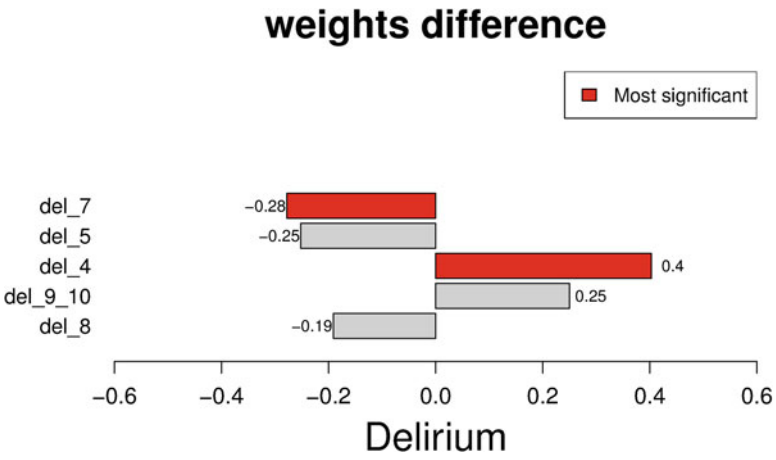

Fig. 19.6 PATHMOX's Segmentation Tree

PLS model Node 2: shorter term hospitalization patients **PLS model Node 3**: longer term hospitalization patients

Fig. 19.7 Path diagram of the two terminal nodes identified by PATHMOX. From left two the right we find the path diagram of node 2: *shorter term hospitalization patients* and the path diagram of node 3: *longer term hospitalization patients*

= 3.79, $p = .01$). The F-coefficient test identifies that it is the path coefficient that links *depression* to *delirium* significantly different across the two children nodes (F-coefficient statistic = 8.01, $p < .01$). Hence we can conclude that there is existence of model heterogeneity at the inner model level due to the different relationship between *depression* and *delirium* in the two detected segments. Further details are contained in the following Fig. (19.7) representing the inner models at the two terminal nodes.

For the shorter hospitalization node, we can see there is a large difference of the path coefficient *depression* on *delirium*: *0.60* and significant ($p < .01$) for shorter term hospitalization patients, but very small (−0.05) and not significant for the longer term hospitalization patients.

Now we turn to the measurement models. The F-invariance test is highly significant, with a Chi-square statistic of 93.39 with *50* degrees of freedom ($p < .01$). This implies that the measurement models are different in the two terminal nodes and the meaning of every latent variable is specific to its segment. Hence, the three latent variables defined in these two segments are not directly comparable, since

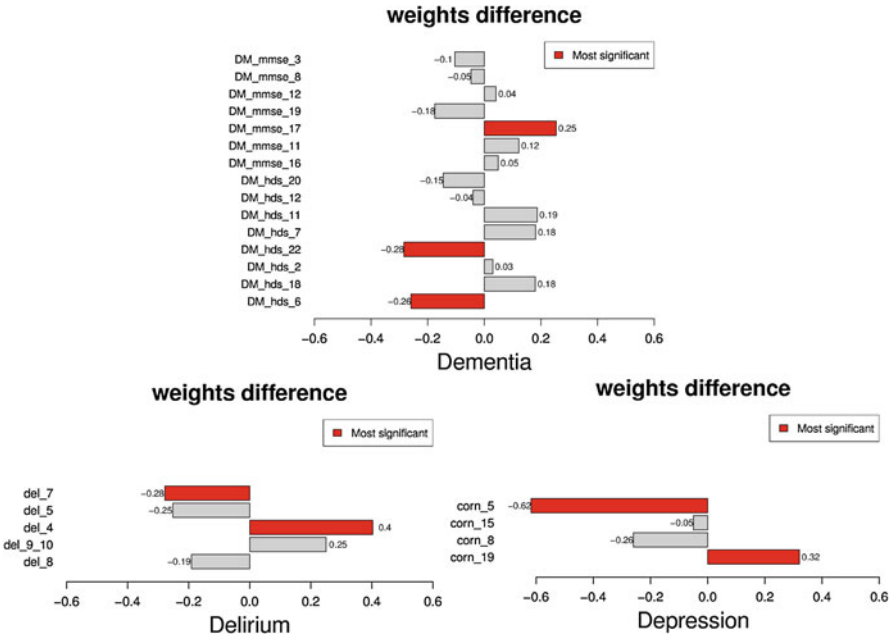

Fig. 19.8 Bar-chart of the weights difference of the two PATHMOX's terminal nodes

we found a significant p-value of the invariance test. This can be illustrated from the three bar-charts of Fig. (19.8), which graph the difference of the weights of two PLS-PM model.The bars in red indicate greater differences. We can see that for the measurement of *dementia* the greater difference are due to: *DM_mmse*_19 (Copy this design), *DM_hds*_22 (Similarities subscale) and *DM_hds*_6 (Denomination subscale); for the measurement of *depression* the greater difference are due to: *corn*_5 (Agitation; restlessness, hand wringing, hair pulling) and *corn*_19 (Mood congruent delusions; delusions of poverty, illness or loss); for the measurement of *delirium* the greater difference are due to: *del*_4 (Disorganized thinking) and *del*_7 (Memory impairment).

In conclusion, this work demonstrates the ability of PLS-PM to investigate the relationship between Dementia, Delirium, and Depression, and the usefulness of PATHMOX as a tool for identifying heterogeneity as regards these relationships.

Granted that the invariance measurement test works well and is suitable for detecting differences in the terminal nodes at measurement levels, this test is a global criterion: we know if the weights of the terminal nodes are the same or not, but we do not know which nodes or weights are responsible for the difference. Thus, an interesting work would consider the possibility of extending the same logic of the F-block and the F-coefficient test to a measurement model.

Further research would be the comparison of the segments found using the PATHMOX approach with others approaches that allows to consider heterogeneity in PLS-PM: REBUS (Esposito Vinzi et al. 2008) and FIMIX (Ringle et al. 2005), among others, to analyze similarities and differences.

Acknowledgements The data analyzed in this paper were collected with funding from the Canadian Institutes of Health Research (IAO69519), Canadian Institute of Aging & Institute of Gender and Health (CRG-82953) and the Alzheimer Society of Canada and the Canadian Nurses Foundation (07-91). Data were used with permission by J. McCusker and M. G. Cole.

References

Alexopoulos, G.S., Abrams, R.C., Young, R.C., et al.: Cornell scale for depression in dementia. Biol. Psychiatry **23**(3), 271–284 (1988)

Aluja, T., Lamberti, G., Sanchez, G.: Extending the pathmox approach to detect which constructs differentiate segments. In: Abdi, H., Chin, W.W., Esposito Vinzi, V., Russolillo, G., Trinchera, L. (eds.) New Perspectives in Partial Least Squares and Related Methods, pp. 269–280. Springer, New York (2013a)

Aluja, T., Lamberti, G., Sanchez, G.: Modelling with heterogeneity. In: Proceedings of SIS 2013, Singapore (2013b)

Chow, G.: Tests of equality between sets of coefficients in two linear regressions. Econometric **28**(3), 591–605 (1960)

Dastoor, D., Cole, M.: Age related patterns of decline in dementia as measured by the hierarchic dementia scale (hds). Am. J. Alzheimers Dis. Other Dement. **3**(6), 29–35 (1988)

Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M.: REBUS-PLS: a response-based procedure for detecting unit segments in PLS path modelling. Appl. Stoch. Models Bus. Ind. **24**(5), 439–458 (2008). John Wiley

Folstein, M.F., Folstein, S.E., McHugh, P.R.: Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. **12**(3), 189–198 (1975)

Lebart, L., Morineau, A., Fenelon, J.P.: Traitement des données statistiques, Dunod, Paris (1979)

Lohmoller, J.-B.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)

Pompei, P., Foreman, M., Cassel, C.K., Alessi, C., Cox, D.: Detecting delirium among hospitalized older patients. JAMA Intern. Med. **155**(3), 301–307 (1995)

Ringle, C.M., Wende, S., Will, A.: Customer segmentation with FIMIX-PLS. In: Aluja, T., Casanovas, J., Esposito, V., Morineau, A., Tenenhaus, M. (eds.) Proceedings of the PLS'05 International Symposium, SPAD Test&Go, pp. 507–514 (2005)

Sanchez, G.: PATHMOX approach: segmentation trees in partial least squares path modeling. Unpublished doctoral dissertation, Universitat Politecnica de Catalunya, Catalonia (2009)

Voyer, P., Monette, J., Champoux, N., Ciampi, A., et al.: Prevalence and incidence of delirium in long-term care. Int. J. Geriatr. Psychiatry **26**, 1152–1161 (2011)

Wold, H.: Soft modeling: the basic design and some extensions. In: Joreskog, K.G., Wold, H. (eds.) Systems Under Indirect Observation: Causality, Structure, Prediction, vol. 2, pp. 1e54. North Holland, Amsterdam (1982)

Chapter 20
Multi-group Invariance Testing: An Illustrative Comparison of PLS Permutation and Covariance-Based SEM Invariance Analysis

Wynne W. Chin, Annette M. Mills, Douglas J. Steel, and Andrew Schwarz

Abstract This paper provides a didactic example of how to conduct multi-group invariance testing distribution-free multi-group permutation procedure used in conjunction with Partial Least Squares (PLS).To address the likelihood that methods such as covariance-based SEM (CBSEM) with chi-square difference testing can enable group effects that mask noninvariance at lower levels of analysis problem, a variant of CBSEM invariance testing that focuses the evaluation on one parameter at a time (i.e. *single parameter invariance testing*) is proposed. Using a theoretical model from the field of Information Systems, with three exogenous constructs (routinization, infusion, and faithfulness of appropriation) predicting the endogenous construct of deep usage, the results show both techniques yield similar outcomes for the measurement and structural paths. The results enable greater confidence in the permutation-based procedure with PLS. The pros and cons of both techniques are also discussed.

Keywords Multi-group Invariance Testing • Permutation Analysis • PLS • Covariance Based SEM

W.W. Chin (✉)
Department of Decision and Information Systems, C. T. Bauer College of Business, University of Houston, Houston, TX 77204–6021, USA
e-mail: wchin@uh.edu

A.M. Mills
Department of Accounting and Information Systems, College of Business and Economics, University of Canterbury, Ilam Christchurch 8140, New Zealand
e-mail: annette.mills@canterbury.ac.nz

D.J. Steel
Department of Management Information Systems, School of Business, University of Houston-Clear Lake, Houston, TX 77058, USA
e-mail: steel@uhcl.edu

A. Schwarz
Louisiana State University, Baton Rouge LA, USA
e-mail: aschwarz@lsu.edu

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_20

20.1 Introduction

Multi-group invariance (MGI) testing is a technique that allows researchers to determine whether parameters of a measurement model and/or the structural model are equivalent (i.e. invariant) across two or more groups (Breckler 1990; Byrne 2010). For the measurement model, invariance testing indicates whether the items used mean the same thing to respondents from different groups or populations (Cheung and Rensvold 2002). If invariance cannot be established, it would be difficult to determine if the differences observed are due to true differences or to different psychometric responses to the items. For the structural model, MGI testing indicates whether the structural paths are equivalent across groups. MGI testing also provides a particularly strong test of the validity of the measurement model and replicability of the structural model across settings.

Covariance-based SEM (CBSEM) using chi-squared difference testing is the most common approach used to examine model invariance. However there is also the distribution- free multi-group permutation procedure used in conjunction with Partial Least Squares (PLS) (Chin 2003; Chin and Dibbern 2010). While some studies using PLS-based approaches have samples that are suited to covariance-based invariance testing, in many other situations the sample size may be too small or the data distribution may violate the assumptions of CBSEM. This paper provides a didactic introduction to how one performs both CBSEM and PLS invariance testing. Using a theoretical model with three exogenous constructs (routinization, infusion, and faithfulness of appropriation) predicting the endogenous construct of deep usage, this study shows convergence of both techniques in terms of both the measurement and structural paths. The pros and cons of the two procedures are also discussed.

20.2 Multi-group Invariance Testing

Establishing the equivalence of measures is critical for research across many disciplines including psychology, marketing, and information systems (Bagozzi and Foxall 1995; Doll et al. 2004; Steenkamp and Baumgartner 1998; Malhotra and Sharma 2008) that rely on latent constructs and comparison analyses. For example, without measurement equivalence, conclusions based on measurement scales, such as the meaning and interpretation of the latent constructs or determining differences or equivalences across populations, at best may be ambiguous, or worse, invalid (Steenkamp and Baumgartner 1998; Malhotra and Sharma 2008).

Multi-group invariance testing is therefore important for many reasons. It is most often used to establish the reliability of measurement scales across groups such as the Kirton Adaption-Innovation inventory (KAI) in psychology and end-user computing satisfaction in information systems (Bagozzi and Foxall 1995; Doll

et al. 2004) and for cross-validation (Steenkamp and Baumgartner 1998; Byrne 1993). It is also used for making comparisons within a study, whether this is to assess theoretical differences between subgroups of the same population (Saeed and Abdinnour-Helm 2008), across populations in the case of multicultural research (Teo et al. 2009), and the equality of (or changes in) responses over time in the case of longitudinal studies (Vandenberg and Lance 2000) to determine if samples taken from different sources can be combined into a single dataset (Steenkamp and Baumgartner 1998).

Despite its importance for validating models across groups and theory testing, MGI testing is relatively uncommon. This may be due to several reasons such as the relative unfamiliarity of researchers with different techniques for MGI testing, the methodological complexities involved in MGI testing, and the relatively large sample sizes needed for CBSEM MGI testing (Steenkamp and Baumgartner 1998; Malhotra and Sharma 2008). For studies whose datasets violate the parametric assumptions of CBSEM, many researchers now rely on distribution-free techniques such as Partial Least Squares (PLS). However, the options for MGI testing in conjunction with techniques such as PLS have been limited to date with many relying on relatively naïve approaches for making group comparisons (Chin and Dibbern 2010). To address this gap this paper describes a distribution-free permutation procedure that can be used with Partial Least Squares for multi-group analysis and contrasts it to CBSEM MGI testing.

For studies that do conform to the parametric assumptions of CBSEM, there are well-established techniques such as multi-group confirmatory factor analysis for conducting MGI testing (Byrne 2010). Many follow traditional procedures that begin with a global test of invariance in which sets of parameters (e.g. all factor loadings, factor covariances, and/or structural paths) are constrained to be equal across the groups. This is followed by a logically ordered series of increasingly restrictive models as each test provides evidence of invariance (Byrne 2010). However a major limitation is that this approach may yield conflicting results where equivalences across groups are demonstrated at one level but rejected at another level of analysis. For example, it is possible for invariance to be suggested at the factor unit level when all loadings are constrained to be equal for that factor, yet individual factor loadings can be found to be noninvariant. One reason is that within a set of items, a group of invariant items may compensate or mask the noninvariance of a single item. In addition to the issue of sets of parameters masking the assessment of a single parameter, questions arise as to which set(s) of parameters (e.g. factor loadings, factor covariances, means, structural paths, error variances/covariances) should be tested, how they should be combined, and what is an appropriate order for conducting the tests. While these decisions may be determined in part by the model and hypotheses being tested, different combinations and test sequences coupled with the practice of testing increasingly restrictive models can also lead to different conclusions regarding equivalences across groups.

To address these limitations and reduce the complexity involved in MGI testing, this study proposes a simplified procedure for identifying the constrained model and sequencing CBSEM tests of multi-group invariance. Instead of combining

one or more sets of parameters in a single test round, individual parameters (e.g. single factor loading, factor variance, or structural path) within the set of interest are constrained *one at a time*. We refer to this approach to MGI testing as *single parameter invariance testing* (*SPIT*). Since this procedure does not evaluate increasingly restrictive models, it addresses the inconsistencies that can arise when non-equivalences are masked by group effects or the sequencing of the model tests. This approach may also yield a more exacting test of invariance due to its ability to more consistently identify instances of noninvariance at the level of the individual parameter.

Using a theoretical model with three predictors (i.e., routinization, infusion, and faithfulness of appropriation) linked to the post-adoption use of Information Systems (IS), this paper illustrates how multi-group invariance testing can be implemented using the two procedures above – a distribution-free permutation procedure for PLS analysis and single parameter invariance testing for use with CBSEM analysis. The results of both procedures in terms of the measurement and structural paths are compared, and the pros and cons of each procedure discussed.

20.2.1 Traditional CBSEM Approach to MGI Testing

Multi-group invariance testing using covariance-based SEM is the most common approach used to establish measurement and structural equivalence of the model paths across groups (Byrne 2010; Doll et al. 2004; Malhotra and Sharma 2008). This approach often begins with an examination of the measurement model and estimation of the least restrictive (unconstrained) model for each group in the set, followed by the same unconstrained model for all the groups as a whole (i.e. the configural model). Equality constraints are then applied to sets of parameters across the groups (e.g. factor loadings, factor variances/covariances, means, error variances, structural paths) (Byrne 2010). Depending on the model and hypotheses being assessed different sets of parameters (or a combination thereof) are constrained in a hierarchical manner, yielding a nested set of increasingly restrictive models for testing. For example, tests for measurement invariance will often begin by constraining multiple elements for the entire model or at the factor-level (e.g. all the factor loadings) (Doll et al. 2004). The chi-square difference test is then used to compare the model fit of the configural model with that of the constrained model; statistically significant differences indicate that the model is noninvariant. According to Byrne (2010) usually tests of individual parameters (e.g. a single factor or factor loadings) are only conducted if model invariance is rejected. If the constraints at the model level are noninvariant, all constraints are removed and a series of tests of constraints at the factor level are performed. Those factors tested and found noninvariant are then subjected to item level testing. A similar procedure is used to assess the equivalence of the structural model.

A key disadvantage with beginning a series of multi-group invariance tests with a globally constrained set of parameters (e.g. all factor loadings, factor variances,

and/or structural paths) is that multiple constraints may mask non-invariance at a lower level of analysis (e.g. for an individual parameter) by confounding the model estimates and hence the estimation of the statistical significance of the changes of model fit between the models. Essentially, sets of parameters that appear invariant as a whole may include individual parameters (e.g. a single factor loading) that are noninvariant, but whose identification is masked by the group effect.

To address these limitations, this study eschews the standard practice of setting a block of parameters such as all the measurement model factor loadings to equality across data groups. Instead, we recommend a simplified procedure for invariance testing in which single parameters (e.g. factor loading or structural path) at a given level of invariance (e.g. metric invariance, structural invariance) are constrained *one at a time* in each round of tests. In other words for each round of invariance testing the constraint applied in the previous round is removed and the next parameter in the test sequence constrained. Hence, the constrained model in each test round differs from the configural model by one constraint only or one degree of freedom minimizing the types of errors discussed above and permitting more precise isolation of noninvariant parameters.

We refer to this simplified approach in which constraints are applied to one parameter at a time, as *single parameter invariance testing* (*SPIT*) and distinguish this approach from *omnibus tests of a given level of invariance* (e.g. model or factor level) in which the full set of parameters related to that level of invariance are constrained all at once. We also distinguish this approach from what we refer to as *forward stepwise tests of invariance* in which sets of parameters are constrained in an additive manner yielding a series of increasingly restrictive models. This would be the logical next step after running SPIT for all parameters; the parameter with the smallest non-significant chi-square difference is then chosen first for equality constraint. This is followed by another round of SPIT and the smallest non-significant chi-square difference chosen for the next equality constraint with that one parameter constrained. This may ultimately culminate in an *omnibus test of invariance* with all items constrained. Alternatively, one may use *backward stepwise tests of invariance* using a Lagrange Multiplier (LM) test to determine which of the set of constraints should be released first followed by another LM test (Bentler 1992).

20.2.2 PLS Permutation-Based Approach to MGI Testing

The Partial Least Squares (PLS) approach has been popularized among researchers in part because the sample size requirements are much smaller for complex models than required for covariance-based techniques, and there are fewer assumptions on data properties such as normality and heterogeneity. It is also considered a more appropriate choice when the emphasis is on prediction. However, to date the procedures used for multi-group comparisons have been relatively naive (Chin and Dibbern 2010) being focused on discussions of the magnitude of differences

between estimates or on t-test statistics when assessing differences (Keil et al. 2000). However, such techniques can be problematic if sample sizes are dissimilar or the data is not normally distributed (Chin and Dibbern 2010; Sarstedt et al. 2011). As an alternative test of equivalence, this paper provides an example of a distribution-free permutation procedure for performing multi-group comparisons with PLS and illustrates how this can be applied to multi-group invariance testing.

The procedure for permutation testing based on random assignment as described by Edgington (1987) and Good (2000) and outlined in Chin and Dibbern (2010) is as follows:

1. A test statistic is computed for the data (e.g., contrasting experimental treatment/control or non-experimental groupings).
2. The data are permuted (divided or rearranged) repeatedly in a manner consistent with the random assignment procedure. With two or more samples, all observations are combined into a single large sample before being rearranged. The test statistic is computed for each of the resulting data permutations.
3. These data permutations, including the one representing the obtained results, constitute the reference set for determining significance.
4. The proportion of data permutations in the reference set that have test statistic values greater than or equal to (or, for certain test statistics, less than or equal to) the value for the experimentally obtained results is the p-value (significance or probability value). For example, if the original test statistic is greater than 95 % of the random values, then the null hypothesis can be rejected at $p < .05$.

As discussed by Chin and Dibbern (2010) this procedure is considered especially suitable for small samples where the assumptions for parametric testing are not fully satisfied (Good 2000), but may also be applied to large samples where the dataset does not comply with the assumed distribution Noreen (1989).

20.3 Multi-group Invariance Testing: An Illustration

To illustrate the two procedures outlined above, this study examines a model derived from the discipline of information systems (IS) that links three exogenous variables representing types of post-adoption use (i.e., routinization, infusion, and faithfulness of appropriation (FOA)) to deep use of an IS.

20.3.1 The Research Model

For firms to maximize the potential returns on their IS investments, the technologies they have invested in must be used in ways that go beyond the initial acceptance towards fully utilizing the systems in ways that support and enhance task, job and/or organizational goals. In other words, when individuals engage in using systems

more deeply to support their goals (i.e. deep usage), performance improvements are likely (Chin and Marcolin 2001). This has led to an increased focus on post-adoption use types such as continued use (Bhattacherjee 2001), routinization (Schwarz 2003), infusion (Schwarz 2003; Sundaram et al. 2007), extended use (Hsieh et al. 2011) and innovation with IT (Wang and Hsieh 2006).

Prior research further suggests that individual engagement in different types of post-adoption use may lead to or impact other forms of use (Chin and Marcolin 2001; Saga and Zmud 1994; Wang and Hsieh 2006). For example, given the temporal distinctions between use types researchers have shown that use of more features of a system (i.e. extended use) may lead individuals to using systems in innovative ways (i.e. emergent use) that support task performance (Wang and Hsieh 2006). Similarly, Sundaram et al. (2007) suggested that the more a person uses a system (i.e. frequency of use), the more likely they will engage in use types reflected in the integration of the technology into how they do their work (i.e. routinization), and use the system in ways that enhance their productivity (i.e. infusion). Hence, it is suggested that frequency of use precedes routinization which in turn may precede infusion (Sundaram et al. 2007). Altogether, these findings are consistent with prior work which suggests that understanding use types such as deep usage, that is, the extent to which an individual will use different features of a technology to support their work or tasks (Schwarz 2003) can begin with examining and integrating different types of use as antecedents (Chin and Marcolin 2001). This study therefore focuses on three use types which, over time, are likely to lead to individuals using systems more deeply to support their work goals; these are routinization, infusion, and faithfulness of appropriation.

When users adopt and use a system in ways that were envisioned or expected by the organization (e.g. faithfulness of appropriation, routinization, infusion) it is likely that over time they will find other ways of using the system, thereby making greater and more extensive use of the system features to support their work. Persons who use a technology deeply are therefore expected to go beyond basic features and procedures that were prescribed by the organization to utilize system features more fully and more intensely, in ways that will help them to do their job or task well and enhance their performance (Chin and Marcolin 2001; Hsieh et al. 2011). Hence, it is suggested that use behaviors such as routinization, infusion and faithfulness of appropriation are likely to lead to deeper uses of a system that in turn support and enhance how individuals perform in their jobs or tasks (See Fig. 20.1).

20.4 Research Design

Before proceeding to MGI testing, the sufficiency of the measures and of the structural model is assessed. Equivalence of the measurement and structural paths in the proposed research model (Fig. 20.1) is then evaluated using data gathered from two distinct groups in a multi-group test of invariance.

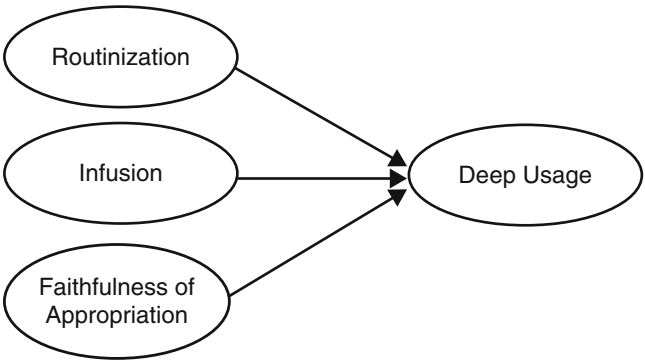

Fig. 20.1 Research model

20.4.1 The Sample

A review of the literature shows many studies use subgroups of a dataset rather than independent samples to assess invariance across groups. For example, in the IS discipline except for cross-cultural studies (Teo et al. 2009; Deng et al. 2008) many use a single demographic such as age, gender, application type and computing experience (Doll et al. 2004; Saeed and Abdinnour-Helm 2008; Lai and Li 2005) to segment a dataset into subgroups for MGI testing. Compared to using distinct and independent samples, subgroups of the same population (or different populations that are demographically similar) provide a weaker context for assessing multi-group invariance as the datasets will have many characteristics in common. On the other hand, studies that use datasets from distinct populations provide stronger evidence for invariance (Teo et al. 2009; Deng et al. 2008).

For this study, data was drawn from two independent and distinct organizations. SGA is a US state government agency which had implemented an electronic document management system 3 years prior to the survey; 111 persons completed the survey. FGA is a large US federal agency which had implemented a core accounting and financial module (from SAP R/3) 4 months prior to the survey; 268 persons completed the survey. Although both organizations are in the public-sector, other elements besides organization type and technology context distinguished the groups. Demographically, SGA had fewer females (51.4 %) than FGA (60.3 %). Respondents from SGA were also younger (56.6 % were 35 years or under compared with 19.6 % at FGA) and had shorter organizational tenure with 26.7 % having been with SGA for more than 5 years compared with 77.6 % at FGA.

20.4.2 Measures

Constructs were measured using validated scales. Infusion and routinization were assessed using 4 items each, from prior studies (Schwarz 2003; Sundaram et al. 2007); faithfulness of appropriation (FOA) used 5 items taken from Chin et al. (1997); and deep usage used 3 items from Schwarz (2003). Responses were captured using 7-point Likert scales anchored (1) Strongly Agree and (7) Strongly Disagree.

Since CBSEM tests of invariance assume the data is normally distributed, skewness, kurtosis and multivariate kurtosis were evaluated using SPSS 20.0. The results showed that across the samples the values for skewness ranged from −1.089 to 0.938, and for kurtosis from 0.697 to 1.202. Based on the rule of thumb that skewness $>$ 3 and kurtosis $>$ 10 represent extreme non-normality these results suggest the distribution of the individual variables is univariate normal (Byrne 2010). However even if the data are univariate normal, it is not necessarily the case that the set of variables 'as a whole' are normally distributed – that is, they are multivariate normal. Mardia (1970)'s normalized estimate of multivariate kurtosis was therefore evaluated. The results show critical ratio (*cr*) statistics of 18.968 and 42.303 for SGA and FGA respectively indicating the data as a whole was non-normal.

Next, the measurement model was examined in terms of convergent and discriminant validity to determine whether the measures perform sufficiently across the datasets. Using PLS-Graph, the results show the factor loadings for both samples were above the recommended threshold of 0.70, ranging from 0.704 to 0.966 (Chin 2010). Also composite reliability (CR) ranged from 0.894 to 0.957 and average variance extracted (AVE) from 0.738 to 0.834, exceeding the recommended thresholds of 0.60 for CR and 0.50 for AVE (Chin 2010).

For discriminant validity to be demonstrated items should load more highly on their own construct than other constructs (Chin 2010). One approach for assessing discriminant validity is to determine whether the average variance extracted exceeds the squared correlations among the constructs. The results show that for both samples the AVE exceeded the squared correlations, suggesting adequate discriminant validity.

Evaluation of the structural model shows that the models account for 0.417 and 0.319 of the variance observed for SGA and FGA, respectively. Structural paths linking routinization and infusion to deep usage were significant (at $p < .05$) for both samples; however, faithfulness of appropriation (FOA) was not significant with respect to deep usage.

20.5 Invariance Testing: Analysis and Results

This section demonstrates, and reports the results of, the two proposed procedures for assessing MGI that is, a variant of multi-group invariance testing with CBSEM which evaluates one parameter at a time (i.e. single parameter invariance testing), and a permutation procedure using PLS.

20.5.1 Multi-group Invariance Testing with Covariance-Based SEM (with AMOS 20.0)

As with prior work (Byrne 2010), the procedure described here for multi-group invariance testing begins with an estimate of the model for each group followed by an estimate of the configural model. We then apply the SPIT procedure where individual parameters (e.g. factor loading) are constrained one at a time and each form of the constrained model is compared with the configural model. Equivalency of the constrained model is then assessed using the chi-square difference test to determine whether the constrained parameter is invariant. This procedure, as applied to measurement and structural invariance testing of the conceptual model proposed in Fig. 20.1, is detailed below.

As with prior research, invariance testing began with a test of the measurement model. Although invariance testing can begin with or focus on different sets or combinations of parameters (e.g. factor loadings, means, factor variances/covariances, error variances) tests of the measurement model typically begin with factor loadings (Byrne 2010; Bollen 1989). For the purposes of demonstrating the proposed method, the focus likewise was on factor loadings.

In this study, the starting model was the same for both groups; hence the same model was estimated for each group separately. Tests began with an estimation of the least restrictive model in which factor variances and error loadings are fixed to 1 for identification, while the parameters to be estimated (i.e. factor loadings and factor covariances) are not constrained (see Fig. 20.2). Since none of the factor loadings are fixed in the unconstrained model this provides the same baseline for comparison across each test. In other words, the only difference between the baseline and any variant of the restricted model is the constrained element (in this case, the constrained factor loading). This also permits comparability with the PLS analysis which also sets the factor variance to a value of 1.

The results of model fitting of the unrestricted baseline model for SGA are $\chi^2 = 243.409$, CFI = 0.919, RMSEA = 0.116; and for FGA, $\chi^2 = 309.552$, CFI = 0.936, RMSEA = 0.090. The model fit for the group as a whole was also estimated, yielding the following fit indices for the configural model: $\chi^2 = 552.961$, CFI = 0.930, RMSEA = 0.070. While the fit indices are reasonable, there may be alternative models with better fit indices. However, since it is not the aim of this paper to posit a model with the best fit, it is deemed appropriate to proceed with invariance testing.

In this study the measurement model consists of 4 constructs with 16 indicators altogether. Equality constraints were then imposed on each factor loading one at a time in a logically ordered manner. For example, in this study beginning with the leftmost construct in the model (see Fig. 20.2) we constrained the first factor loading in the first construct and worked systematically through each factor loading from left to right until all 16 factor loadings had been evaluated. It is important to note that since only one constraint is applied at a time it does not matter which parameter is constrained first or in which order the constraints are applied. However,

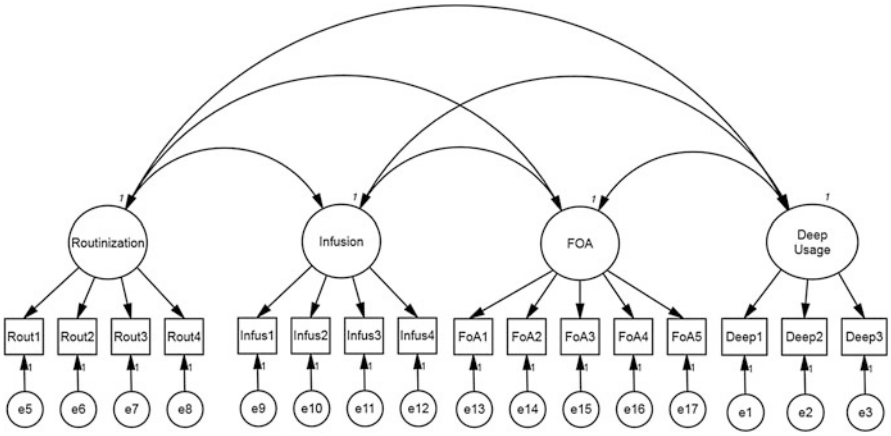

Fig. 20.2 Measurement model using covariance-based SEM (with AMOS 20.0)

having a logical sequence for constraining individual parameters can help ensure that parameters are not omitted from the test sequence.

Using this procedure, 16 alternative models were identified and a chi-square difference test applied to each model separately to determine whether the constrained parameter was invariant across the groups. For example, for Model M1 the factor loading for the first item Rout1 was constrained; the degrees of freedom (df) for the constrained model = 197 (compared with $df = 196$ for the baseline model); hence $\Delta df = 1$. Estimation of M1 yielded a χ^2 statistic of 553.076 and a $\Delta\chi^2$ of 0.115 (i.e. $\Delta\chi^2 = 553.076 - 552.961 = 0.115$); the p-value associated with the $\Delta\chi^2$ of 0.115 and $\Delta df = 1$ was determined from the χ^2 distribution; thus $p = .735$. The non-significant p-value for M1 suggests that Rout1 is invariant. The tests and evaluation sequence were then repeated for each alternative model that is, M2. . .M16 (see Table 20.1). Altogether the results (Table 20.1) show all the measures are invariant (at $p < .05$) with the exception of Infus3 (where $p = .049$).

The masking effect of setting invariances simultaneously on factor or model level loadings can be shown in this situation. For example, if we set all the factor loadings equal, the $\Delta\chi^2$ of 15.087 with $\Delta df = 16$ represents a p-value of .518 leading to a conclusion that all items are invariant. At this point, a researcher would conclude that all single item loadings are invariant in contrast to the SPIT outcome of noninvariant for Infus3.

Next, the structural model was evaluated. The approach used for structural invariance testing is similar to that for the measurement model where each structural path is constrained separately. Since the configural model comprises three paths, three alternative models are identified. The results (Table 20.2) suggest all three structural paths are invariant.

Table 20.1 Measurement model estimation: comparison of CBSEM (AMOS) and PLS results

		CBSEM analysis ($\Delta df = 1$)			PLS analyses				
	Manifest	Baseline (group) model: $\chi^2 = 552.961$				n = 1000	n = 2000	n = 5000	
Model#	variables	$\Delta\chi^2$	*p-value*	Sig.	Difference	*p-value*	*p-value*	*p-value*	Sig.
M1	Rout1	0.115	0.735	No	0.022	0.346	0.315	0.324	No
M2	Rout2	0.820	0.365	No	0.010	0.482	0.494	0.472	No
M3	Rout3	0.961	0.327	No	0.016	0.738	0.736	0.744	No
M4	Rout4	0.080	0.777	No	0.025	0.397	0.423	0.426	No
M5	Infus1	0.297	0.586	No	0.001	0.969	0.967	0.963	No
M6	Infus2	0.232	0.630	No	0.013	0.473	0.489	0.493	No
M7	Infus3	3.879	0.049	**Yes**	0.190	0.039	0.026	0.030	**Yes**
M8	Infus4	2.326	0.127	No	0.077	0.103	0.105	0.111	No
M9	FoA1	0.055	0.815	No	0.048	0.546	0.534	0.537	No
M10	FoA2	0.172	0.678	No	0.052	0.159	0.149	0.155	No
M11	FoA3	0.201	0.654	No	0.012	0.647	0.653	0.669	No
M12	FoA4	0.005	0.944	No	0.020	0.385	0.372	0.377	No
M13	FoA5	0.078	0.780	No	0.003	0.910	0.915	0.916	No
M14	Deep1	0.686	0.408	No	0.013	0.647	0.679	0.666	No
M15	Deep2	0.039	0.843	No	0.075	0.143	0.115	0.125	No
M16	Deep3	0.687	0.407	No	0.057	0.193	0.188	0.184	No

Table 20.2 Structural model estimation: comparison of CBSEM (AMOS) and PLS results

	CBSEM analysis ($\Delta df = 1$)			PLS analyses				
	Baseline (group) model: $\chi^2 = 552.961$				n = 1000	n = 2000	n = 5000	
	$\Delta\chi^2$	*p-value*	Sig.	Difference	*p-value*	*p-value*	*p-value*	Sig.
Routinization → Deep usage	0.417	0.518	No	0.111	0.256	0.266	0.266	No
Infusion → Deep usage	0.070	0.791	No	0.018	0.880	0.872	0.879	No
Faithfulness of appropriation (FOA) → Deep usage	1.607	0.205	No	0.176	0.103	0.095	0.103	No

20.5.2 Permutation-Based Multi-group Invariance Testing with PLS

Following the procedure outlined earlier, three permutation analyses were conducted using 1000, 2000 and 5000 permutations respectively. The proportion of data permutations in the reference set that have test statistic values greater than or equal to (or for some statistics, less than or equal to) the value for the obtained results determines the *p*-value or the significance of the difference (Chin and Dibbern 2010).

The results (Table 20.1) showed the group parameter differences for the measurement loadings ranged from 0.001 to 0.190. With the exception of Infus3 which returned significant *p*-values (i.e. $p < .05$) of .039, .026 and .030, the *p*-values for each permutation analysis ranged from .103 to .969, .105 to .967, and .111 to .963 for 1000, 2000 and 5000 permutations respectively. The non-significant values returned for all the indicators except Infus3 suggest partial invariance for the measurement model and matches those from the CBSEM analysis.

Next, the structural paths were evaluated. As with the procedures used to evaluate the measurement model, the original parameter differences between groups are compared to the permutated data sets. Original path differences of 0.111, 0.018 and 0.176 were obtained for the structural paths linking routinization, infusion, and faithfulness of appropriation to deep usage respectively. The results (Table 20.2) showed the p-values for the three permutation analyses (i.e. 1000, 2000 and 5000 permutations, respectively) for routinization (i.e. .256, .266, .266) infusion (.880, .872, .879) and faithfulness of appropriation (FOA) (.103, .095, .103) were non-significant demonstrating invariance of the structural paths.

Taken altogether, invariance testing of the measurement and structural paths using CBSEM and PLS supported procedures showed full convergence of the results. This outcome demonstrates the efficacy of the new PLS procedure as well as its usefulness particularly in cases where the datasets do not comply fully with the parametric assumptions for covariance-based SEM analysis.

20.6 Discussion and Conclusion

This paper provided a didactic introduction to how invariance testing can be conducted using a multi-group distribution-free permutation approach in conjunction with PLS. Attention was also given to the likelihood that common methods such as covariance-based invariance testing using chi-square difference testing, can enable group effects that mask noninvariance at lower levels of analysis, leading to contradictory findings and possible false conclusions. To address this issue, a variant of chi-square difference testing that focuses the testing procedure on evaluating one parameter (rather than sets of parameters) at a time, that is *single parameter invariance testing* was proposed and tested. By constraining and testing

only one parameter at a time for invariance, this approach allows for more precise identification of noninvariant parameters, and reduces the likelihood of noninvariant parameters being covered over by group effects that arise from constraining multiple parameters in each test round. Both procedures are demonstrated using example data from the field of information systems to test invariance of both the measurement and structural models.

Given the importance of invariance testing for knowledge contribution in fields that rely on latent constructs, this paper contributes in two key areas. Methodologically, the paper highlights a procedure that can be used in conjunction with PLS-based analyses for multi-group invariance testing. This addresses a key gap in the current pedagogy given the popularity of PLS-based analyses as the technique of choice in certain settings (e.g. when a dataset is characterized by small sample size or non-normal distribution) and the lack of procedures that enable researchers to test for invariance with PLS. Indeed, up until now researchers have relied on covariance-based methods to assess multi-group invariance or, if using a PLS-based procedure they opt not to demonstrate the invariance of their measures and structural paths or rely on relatively naive means to compare groups (Chin and Dibbern 2010). There is a clear need for more suitable methods for analyzing invariance in certain empirical settings. This paper therefore demonstrates a PLS-based approach to assessing multi-group invariance.

Turning to CBSEM techniques, this study introduces a simplified variant of CBSEM invariance testing, that is single parameter invariance testing (SPIT). This procedure represents a simpler yet potentially more exacting test of invariance due to its ability to identify instances of noninvariance at the level of the individual parameter. It will also help address the inconsistencies that can arise when non-equivalences are masked by group effects. Unlike common CBSEM techniques (Byrne 2010; Doll et al. 2004; Steenkamp and Baumgartner 1998) the current procedure recommends constraining one parameter at a time while all other non-fixed parameters are freely estimated. This means that the results of single parameter invariance testing are not affected by which parameter is constrained first or the sequencing of the tests.

A comparison of CBSEM and PLS showed full convergence of the results for both the measurement and structural models (Tables 20.1 and 20.2) enabling greater confidence in the efficacy of the new PLS procedure. For the measurement model, the CBSEM analysis suggested only one parameter (Infus3) was noninvariant. Given the data was multivariate non-normal concerns could be raised about the validity of these findings (Byrne 2010). However, the same pattern of findings was also identified by the PLS analyses suggesting that multivariate non-normality was not a major factor in the analyses and enabling greater confidence in the findings. This further suggests the usefulness of the PLS procedure as a way to validate the outcomes of CBSEM analyses where the data does not comply fully with the assumptions for such analyses.

For the research model, the results demonstrate invariance of the four indicators of systems use and their associated measures (i.e. routinization, infusion, faithfulness of appropriation and deep usage) across two independent samples. As far as

we are aware prior studies have not focused on enhancing the credibility of these measures. Establishing the invariance of key measures in IS research is important not only for enabling greater confidence in our understanding of how technologies are used by individuals in organizations, but also to enable meaningful comparisons across settings and the cumulative development of knowledge which relies heavily on the reliability of the latent variable measures used to capture key phenomenon. This paper represents a step in this direction. Empirically, the results demonstrate how engagement in one type of use can lead to deeper uses of information systems. The results also confirm prior research which suggests that certain types of use will impact other use types (Chin and Marcolin 2001) by providing evidence of these linkages in the context of post-adoption deep usage.

In summary, this paper provides a didactic example of a confirmatory test of measurement and structural multi-group invariance in the context of post-adoption use. It introduced a new approach to CBSEM invariance testing focusing on single parameter invariance testing. The paper also demonstrated a procedure for conducting multi-group invariance testing using a distribution-free permutation approach with PLS. The results showed convergence of the findings across both procedures. While this enables greater confidence in the permutation procedure as recommended by Chin and Dibbern (2010) it would be useful to compare the results for both procedures across varying levels of non-normality to see how the PLS method performs. Also, the number of individual parameters being tested was rather small. For more complex models, an adjustment for Type I error would be needed. In our case, with an alpha setting of 0.05, the 19 individual tests conducted in this paper would suggest on average one significant finding which was indeed what we found. To compensate for potential Type I error, we would suggest a Bonferroni or Sidàk type correction with a commensurate stricter alpha level for single parameter invariance testing.

References

Bagozzi, R.P., Foxall, G.R.: Construct validity and generalizability of the Kirton adaption-innovation inventory. Eur. J. Personal. **9**, 185–206 (1995)

Bentler, P.M.: EQS: Structural Equations Program Manual. BMDP Statistical Software, Los Angeles (1992)

Bhattacherjee, A.: Understanding information systems continuance: an expectation-confirmation model. MIS Q. **25**, 351–370 (2001)

Bollen, K.A.: Structural Equations with Latent Variables. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section. John Wiley, New York (1989)

Breckler, S.J.: Applications of covariance structure modeling in psychology: cause for concern? Psychol. Bull. **107**, 260–273 (1990)

Byrne, B.M.: The Maslach Burnout inventory: testing for factorial validity and invariance across elementary, intermediate and secondary teachers. J. Occup. Organ. Psychol. **66**, 197–212 (1993)

Byrne, B.M.: Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming. Lawrence Erlbaum Associates, Mahwah (2010)

Cheung, G.W., Rensvold, R.B.: Evaluating Goodness-of-fit Indexes for Testing Measurement Invariance. Lawrence Erlbaum Associates, Hillsdale, NJ (2002)

Chin, W.W.: A permutation procedure for multi-group comparison of PLS models. Invited presentation. In: Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (eds.) PLS and Related Methods, PLS'03 International Symposium – "Focus on Customers", Lisbon, pp. 33–43 (2003)

Chin, W.W.: How to write up and report PLS analyses. In: Esposito Vinzi, V., Chin, W.W., Hensler, J., Wang, H. (eds.) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics, pp. 655–690. Springer, Berlin/Heidelberg (2010)

Chin, W.W., Dibbern, J.: An introduction to a permutation based procedure for multi-group PLS analysis: results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between Germany and the USA. In: Esposito Vinzi, V., Chin, W.W., Hensler, J., Wang, H. (eds.) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics, pp. 171–193. Springer, Berlin/Heidelberg (2010)

Chin, W.W., Marcolin, B.L.: The future of diffusion research. Data Base Adv. Inf. Syst. **32**, 8–12(2001)

Chin, W.W., Gopal, A., Salisbury, W.D.: Advancing the theory of adaptive structuration: the development of a scale to measure faithfulness of appropriation. Inf. Syst. Res. **8**, 342–367 (1997)

Deng, X.D., Doll, W.J., Al-Gahtani, S.S., Larsen, T.J., Pearson, J.M., Raghunathan, T.S.: A cross-cultural analysis of the end-user computing satisfaction instrument: a multi-group invariance analysis. Inf. Manag. **45**, 211–220 (2008)

Doll, W.J., Deng, X.D., Raghunathan, T.S., Torkzadeh, G., Xia, W.D.: The meaning and measurement of user satisfaction: a multigroup invariance analysis of the end-user computing satisfaction instrument. J. Manag. Inf. Syst. **21**, 227–262 (2004)

Edgington, E.S.: Randomization Tests. Marcel Dekker, New York (1987)

Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edn. Springer, New York (2000)

Hsieh, J.J., Rai, A., Xu, S.X.: Extracting business value from IT: a sensemaking perspective of post-adoptive use. Manag. Sci. **57**, 2018–2039 (2011)

Keil, M., Tan, B.C.Y., Wei, K.-K., Saarinen, T., Tuunainen, V., Wassenaar, A.: A cross-cultural study on escalation of commitment behavior in software projects. MIS Q. **24**, 299–325 (2000)

Lai, V.S., Li, H.: Technology acceptance model for internet banking: an invariance analysis. Inf. Manag. **42**, 373–386 (2005)

Malhotra, M.K., Sharma, S.: Measurement equivalence using generalizability theory: an examination of manufacturing flexibility dimensions. Decis. Sci. **39**, 643–669 (2008)

Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika **57**, 519–530 (1970)

Noreen, E.W.: Computer Intensive Methods for Testing Hypotheses: An Introduction. John Wiley, New York (1989)

Saeed, K.A., Abdinnour-Helm, S.: Examining the effects of information system characteristics and perceived usefulness on post adoption usage of information systems. Inf. Manag. **45**, 376–386 (2008)

Saga, V.L., Zmud, R.W.: The nature and determinants of IT acceptance, routinization and infusion. In: Levine, L. (ed.) Diffusion, Transfer and Implementation of Information Technology, pp. 67–86. Elsevier Science B.V./North-Holland, Amsterdam (1994)

Sarstedt, M., Henseler, J., Ringle, C.M.: Multigroup analysis in partial least squares (PLS) path modeling: alternative methods and empirical results. In: Sarstedt, M., Schwaiger, M., Taylor, C.R. (eds.) Measurement and Research Methods in International Marketing. Advances in International Marketing, pp. 195–218. Emerald Group Publishing, Bingley (2011)

Schwarz, A.H.: Defining information technology acceptance: a human-centered, management-oriented perspective, University of Houston, Houston (2003)

Steenkamp, J.E.M., Baumgartner, H.: Assessing measurement invariance in cross-national consumer research. J. Consum. Res. **25**, 78–90 (1998)

Sundaram, S., Schwarz, A., Jones, E., Chin, W.W.: Technology use on the front line: how information technology enhances individual performance. J. Acad. Mark. Sci. **35**, 101–112 (2007)

Teo, T., Lee, C.B., Chai, C.S., Wong, S.L.: Assessing the intention to use technology among pre-service teachers in Singapore and Malaysia: a multigroup invariance analysis of the technology acceptance model (TAM). Comput. Educ. **53**, 1000–1009 (2009)

Vandenberg, R.J., Lance, C.E.: A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods **3**, 4–70 (2000)

Wang, W., Hsieh, P.: Beyond routine: symbolic adoption, extended use, and emergent use of complex information systems in the mandatory organizational context. In: Proceedings: International Conference of Information System, Milwaukee (2006)

Chapter 21
Brand Nostalgia and Consumers' Relationships to Luxury Brands: A Continuous and Categorical Moderated Mediation Approach

Aurélie Kessous, Fanny Magnoni, and Pierre Valette-Florence

Abstract This study investigates the role of nostalgia in the consumer-brand relationships in the luxury sector. Results indicate that the nostalgic luxury car brands (*vs*. futuristic luxury car brands) lead to stronger consumer-brand relationships. Moreover, brand nostalgia has a direct positive effect on brand attachment and separation distress. Brand attachment is also a partial mediator between brand nostalgia and separation distress. In addition, the influence of two moderating variables is examined. We show that past temporal orientation reinforces the relationship between (1) brand nostalgia and brand attachment, and between (2) brand nostalgia and separation distress. Finally, consumers' need for uniqueness reinforces the relationship between brand attachment and separation distress. On a methodological side, the study shows the ability of the PLS approach to handle higher order latent variables both in the context of continuous and categorical latent moderated mediation variables.

Keywords Luxury brands • Consumer-brand relationships • Nostalgia

21.1 Introduction

Nowadays, consumers need reassurance and feel emotional about the past. This retro trend especially prevails in the luxury sector, where brands play on the traditional and classical themes. Although marketers widely use nostalgia, no study has addressed to this date its impact in the luxury sector. Moreover, building a strong consumer-brand relationship is very important in order to make a business profitable

A. Kessous (✉)
CERGAM, Faculté d'Economie et de Gestion, Aix-Marseille Université, Marseille, France
e-mail: aurelie.kessous@univ-amu.fr

F. Magnoni • P. Valette-Florence
CERAG, IAE Grenoble, Université Grenoble Alpes, Grenoble, France
e-mail: Fanny.Magnoni@iae-grenoble.fr; Pierre.Valette-florence@iae-grenoble.fr

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_21

and increase customer lifetime value (Fournier 1998; Thomson et al. 2005). So, we may ask ourselves whether the effects of nostalgia are equally positive with socially-visible products from the luxury sector, imbued with symbolism and ostentation.

This chapter presents a comparison between consumer relationships with two luxury car brands: nostalgic luxury car brand (Mini) *vs.* futuristic luxury car brand (Infiniti). The main research question is, therefore, the following: How does the relationship to brands vary according to the perceived nostalgic *vs*. futuristic character of the luxury car brands? The relevance of this chapter is two-fold. On the one hand, it contributes to a better explanation of the nature of the links that consumers have with these two types of luxury car brands. On the other hand, it exemplifies the usefulness of the Partial Least Square (PLS) approach for handling both continuous and categorical moderated mediation variables.

Firstly, we briefly present the theoretical framework of research and the hypotheses. Secondly, the methodology for the collection and analysis of data is provided. Finally, we present the results of the study of a sample of 132 Mini owners and 123 Infiniti owners. In conclusion, we state the contributions, limits and research paths.

21.2 Conceptual Background and Hypotheses

In marketing, academics suggest multiple definitions of nostalgia. Holbrook and Schindler's definition (Holbrook and Schindler 1991, p. 330) is undoubtedly the most cited

> A preference (general liking, positive attitude, or favorable affect) toward objects (people, places, or things) that were more common (popular, fashionable, or widely circulated) when one was younger (in early adulthood, in adolescence, in childhood, or even before birth). According to this logic, nostalgic brands are defined brands that were popular in the past (and are still popular now), whereas the non-nostalgic brands as brands that are popular now (but were less so in the past or did not exist in the past) (Loveland et al. 2010).

Nostalgia is also well studied in the consumer-brand relationship literature (i.e., Fournier 1998). Two factors define nostalgic attachment: self-concept connection—which states the congruity between past, present, real or ideal self-image and those that he/she has of the brand—and nostalgic connection—which deals with a transfer of a person's remembrances of the brand.

Attachment refers to an emotional bond and comes from interpersonal relationships (Bowlby 1969). The recent study of Park et al. (2010) illustrates that two factors reflect brand attachment: brand-self connection and brand prominence. Brand-self connection refers to the consumer's degree of identification with a brand and expresses the incorporation of the brand into their self-concept (Fournier 1998; Escalas and Bettman 2003). Brand prominence can be considered as the salience of the cognitive and affective bond that links the brand with the self. As an attachment behavior, separation distress (i.e., emotional distress due to loss of proximity) is also strongly predicted by brand attachment (Thomson et al. 2005; Park et al. 2010). Separation distress refers to an emotional indicator of attachment inducing negative

feelings (e.g., anxiety, depression, loss of self). This positive influence of brand attachment on separation distress is also expected in our study. In addition, since nostalgic connection deals with a transfer of a person's remembrances of the brand (Fournier 1998), brand nostalgia should also positively impact separation distress.

Temporal orientation refers to cognitive involvement focused on one of the three time zones (i.e., past, present, future) that influences attitude and behavior. Research in psychology leads us to consider temporal orientation as a moderator of nostalgia. For instance, Sedikides et al. (2008) point out that nostalgia is a defense mechanism, protecting individuals from certain existential problems. Consequently, the influence of brand nostalgia on brand attachment and separation distress should be stronger when consumers tend to be highly past oriented.

If uniqueness is a specific dimension of luxury brands (Vigneron and Johnson 1999), consumers' need for uniqueness should be a relevant variable. According to Tian et al. (2001, p.172), consumers' need for uniqueness refers to

> individuals' pursuit of differentness relative to others that is achieved through the acquisition, utilization, and disposition of consumer goods for the purpose of developing and enhancing one's personal and social identity.

Consumers' need for uniqueness should influence the impact of brand attachment on separation distress. Indeed, we can suppose that separation distress will be stronger when consumers' need for uniqueness is high because in this case, possession of luxury brands highly contributes to develop the self-image and the feeling of being unique.

The mains points of the literature review are summarized in the following set of hypotheses and in Fig. 21.1.

- H1: Brand nostalgia has a direct positive effect on: (a) brand attachment; and (b) separation distress.
- H2: Brand attachment has a direct positive effect on separation distress.

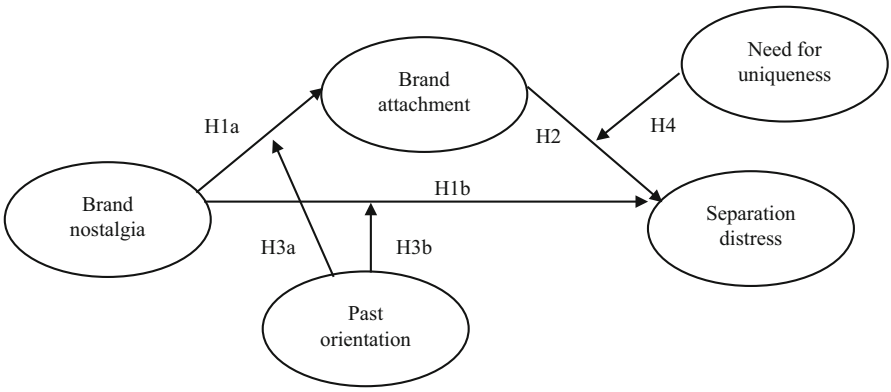

Fig. 21.1 Conceptual model

- H3: Past temporal orientation reinforces the relationship between: (a) brand nostalgia and brand attachment; and (b) brand nostalgia and separation distress.
- H4: Consumers' need for uniqueness reinforces the relationship between brand attachment and separation distress.

21.3 Methodology for Data Collection and Processing

We selected two luxury car brands: a nostalgic luxury car brand and a futuristic luxury car brand. The first one, the Mini by BMW, has an old and nostalgic connotation. It has a mythical history associated to the iconic 1959 Austin Mini model. It was redesigned in 2009 by BMW but has kept its British appeal of elegance and chic. The second one, "Infiniti" by Nissan, has, on the other hand, a new and futuristic connotation. It appeared in the US in 1989 and entered the French market in 2008. Performance, intuitive technologies and first-class comfort are the key words guiding the production supervised by Sebastian Vettel, the Formula 1 world champion. The scarcity of Infiniti car dealers in France enhances the rare, unique, and exclusive character of this elitist brand.

The questionnaires were distributed online between June and October 2013, with the support of two Mini dealerships in Marseilles, and three Infiniti dealerships in the South-east of France (Marseilles, Cannes, and Lyons). A filter question was used to select only respondents who were clients (ownership and/or purchasing of brand products). The final sample comprised 255 clients, who were quasi-equally distributed between the two brands. 132 Mini car owners and 123 Infiniti car owners to responded to the survey in full. The sample is male (78 % men) and relatively young (80 % under the age of 54). Then three manipulation checks were conducted.

1. We verified that the respondents had perceived Mini and Infiniti as prestigious brands. The degree of luxury associated with these brands was rated on a six-point Likert scale by indicating the importance of the characteristics "luxury" and "status" when purchasing the brand (Park et al. 2010), along with the items of the luxury scale of Vigneron and Johnson (2004). As expected, Mini and Infiniti are both perceived as luxury brands because means are fairly above 3.5 ($M_{\text{Mini}} = 4.76$; $M_{\text{Infiniti}} = 4.82$).
2. Nostalgic *vs.* futuristic orientation of the two brands was also tested. Respondents indicated the extent to which they viewed the brand as "retro" (−2) *vs.* "futuristic" (+2). Infiniti was perceived as significantly more "futuristic" than Mini ($F(2, 253) = 76.542$; $p = .001$; $M_{\text{Mini}} = 0.00$; $M_{\text{Infiniti}} = 1.14$).
3. Finally, we verified that there were no significant differences between the two brands in terms of subjective familiarity to control the possible effect of this variable. Three items from Brucks (1985) were used to evaluate subjective familiarity toward the brand (six-point Likert scale). As expected, Mini and Infiniti were perceived as similar ($F(2, 253) = 2.361$; $p > 0.05$; $M_{\text{Mini}} = 4.88$; $M_{\text{Infiniti}} = 4.66$).

This research used six-point Likert scales for all scales. Brand attachment and separation distress were measured with the scales proposed by Park et al. (2010). Brand self-connection and brand prominence (i.e., the two attachment components) were evaluated with four items (two items for brand self-connection and two items for brand prominence). Two items measured separation distress (scale of Park et al. 2010). For brand nostalgia, we selected the three-dimensional scale (i.e., personal memories, perceived brand oldness and historical memories) of Bartier (2013). We used twelve items (six items for personal memories, three items for perceived brand oldness and three items for historical memories). For past temporal orientation, we selected three items from Usunier and Valette-Florence (2007). Finally, we used eight items from Tian et al. (2001) to measure consumers' need for uniqueness.

21.4 Data Analysis and Test of Assumptions

A PLS approach has been selected because of its minimal demands on sample size and suitability to handle higher order latent constructs and violation of multivariate normality (Bagozzi and Yi 1994). Moreover, the present study relies on rather small sample sizes and the model is complex involving several second order reflective latent variables. In this research, the estimation of the different PLS models follows a two steps procedure.

First, although the measurement and structural models are simultaneously and iteratively estimated within the PLS approach, the reliability and validity of the measurement model should be firstly assessed. Once the adequacy of the construct measurements is verified, the structural relationships among the constructs and the quality of the overall model are then assessed (Fornell and Larcker 1981). The adequacy of the reflective measurement model can be assessed by looking at composite reliabilities, the convergent validity of the measures associated with individual constructs, and discriminant validity (Henseler et al. 2009). Results are displayed in Table 21.1. The causal model depicted in Fig. 21.1 encompasses first and second order latent variables. First order latent variables are modeled by means of reflective indicators whereas second order latent variables are conceptualized in a molecular way (i.e. a reflective relationship between the second order latent variables and their respective first order latent facets). All second order latent variables were measured via replicated indicators of all the first order latent variables they were connected with. The second order latent variables are respectively brand attachment, nostalgia and need for uniqueness.

As for the first order reflective latent variables, all the indicators of convergent validity and reliability are satisfied. As regards to the second order reflective latent variables, convergent validity and reliability are fairly good as well. Finally, a test of the discriminant validity (Fornell and Larcker 1981) shows that each first order latent variable shares more variance with its respective indicators than with the other latent variables it is correlated with.

Table 21.1 Convergent validity and reliability indices. Second order latent variables are in capitals

Latent variable	Convergent validity	Reliability
ATTACHMENT	0.781	0.834
Brand-self connection	0.891	0.942
Brand prominence	0.927	0.962
BRAND NOSTALGIA	0.568	0.740
Personal memories	0.728	0.941
Perceived brand oldness	0.818	0.931
Historical memories	0.682	0.865
Past orientation	0.803	0.924
NEED FOR UNIQUENESS	0.745	0.821
Creative choice counterconformity	0.698	0.920
Avoidance of similarity	0.710	0.918
Separation distress	0.870	0.931

Second, to assess the structural model a set of criteria should be verified. Although PLS does not provide any global goodness-of-fit indices as those used for covariance-based SEM, Tenenhaus et al. (2005) propose the geometric mean of the average communality (measurement model) as well as the average R^2(structural model), as an overall Goodness-of-Fit (GoF) measure for PLS. In this research, the absolute GoF value is 0.553, a value corresponding to an excellent adjustment according to Wetzels et al. (2009) (GoF higher than 0.36 are large). Moreover, in line with Henseler et al. (2009), the essential criterion is the coefficient of determination (R^2) of the endogenous latent variables. In our case, on average the R^2 is 48.1 % for the full causal estimated model.

A latent MANOVA and a step-down analysis were conducted. Since that from the outset the latent variables define a causal model, we decided to conduct a latent analysis of variance because it was necessary at that stage to delve deeper into the joint effects of the nostalgic *vs.* futuristic orientation of the luxury car brand (Mini *vs.* Infiniti) on the latent brand relationships variables encompassed by this research. One main advantage of analyzing variance at the latent level using a structural equations model is the ability to compare the strength of the effect between different dependent latent variables and to perform a step-down analysis at the latent level. When there is a causal relationships network among the dependent variables, step-down analyses provide useful information as to whether the mean difference in a dependent variable is due to the direct effect of the experimental manipulation or its dependence on other variables (Bagozzi and Yi 1989).

A step-down analysis proceeds into two sequential steps. The first stage begins with a latent MANOVA performed on all dependent variables. If the path estimates point to a rejection of equal means, then the next step consists of testing the dependent variables in the hypothesized causal network while partialling out all remaining dependent variables as covariates. As a result, the researcher can then assess the relative impact of the experimental manipulation, while taking into account the causal order between all the dependent latent variables.

Table 21.2 Latent MANOVA and step-down analysis results

		Latent MANOVA	Latent step-down analysis		
Concepts	Bootstrapped Path coefficients	Mini (1) *versus* Infiniti (0)	Mini (1) *versus* Infiniti (0)	Brand nostalgia	Attachment
Brand nostalgia	Direct effect	0.492*	0.492*		
	Total R^2	25.27 %	25.27%		
Attachment	Direct effect	0.140**	0.107**	0.508*	
	Indirect effect		0.250*		
	Total R^2	2.02 %	19.85 %	19.85 %	19.85 %
Separation distress	Direct effect	0.176**	0.0070	0.572*	0.357*
	Indirect effect		0.370*	0.181**	
	Total R^2	3.01 %	48.11 %	48.11 %	48.11 %

*: $p < .001$; **: $p < .05$

Table 21.2 shows the corresponding results. First, when a simple latent MANOVA is performed, results show that the nostalgic *vs.* futuristic orientation of the luxury car brand (Mini *vs.* Infiniti) has a significant influence on each of the dependent variables. In others words, our results highlight a positive effect of the nostalgic *vs.* futuristic orientation of the luxury car brand on brand nostalgia, brand attachment and separation distress.

Second, the results in Table 21.2 show the causal relationships between brand nostalgia, brand attachment and separation distress. As we can observe, brand nostalgia has a direct positive impact on brand attachment (0.508) and separation distress (0.572); supporting H2a and H2b. Brand nostalgia has also an indirect effect on separation distress (0.181) and brand attachment influences directly separation distress (0.357). These results support H3. Hence, brand attachment is a partial mediator between brand nostalgia and separation distress.

In addition, the latent Step-Down analysis permits to deeper examine the causal relationships. As we can see in Table 21.2, the direct influence of the nostalgic *vs.* futuristic orientation of the luxury car brand is no longer statistically significant in all. This means that all the effects are now due to the causal relationships between the dependent variables. This result seems both theoretically and managerially important. However, even if the nostalgic *vs.* futuristic orientation of the luxury car brand doesn't have any direct impact, it still has an important indirect effect. Two points deserve attention. First, all the indirect effects are now greater than when the nostalgic *vs.* futuristic orientation of the luxury car brand was solely taken into account (for brand attachment for example, 0.107 *vs.* 0.250). This means that the encompassed latent variables indirectly amplify the effect of the nostalgic *vs.* futuristic orientation of the luxury car brand. Second, there is once again an attenuation of the incidence of the brand orientation on the dependent variables. This indirect effect is indeed greater for separation distress (0.370) than brand attachment (0.250). Once again, this result puts the stress on the influence of brand nostalgia on separation distress, either directly or indirectly.

Finally, in order to study the joint effect of the two latent moderator variables, we relied on the normalized product indicator approach, hence following recent recommendations made by Henseler and Chin (2010) in the case of complex moderation investigations (we recall that we jointly model 3 moderations through either first or second order latent variables). Far and foremost, the past orientation positively moderates attachment ($\beta = 0.204$; $p = .001$). In other words, the past orientation reinforces the impact of brand nostalgia on brand attachment; supporting H4a. A similar effect arises as for the double moderation of the past orientation and the need for uniqueness on separation distress. Once again, this is the past orientation that has the greatest moderating incidence ($\beta = 0.194$; $p = .016$), compared to need for uniqueness ($\beta = 0.077$; $p = .046$). The past orientation and the need for uniqueness both reinforce the impact of either brand nostalgia or brand attachment on separation distress; supporting H4b and H5. However, one can notice that the moderating influence of the past orientation on brand nostalgia is almost three times higher than the moderating influence of the need for uniqueness on band attachment. In other words, brand nostalgia and past orientation seems the more important to predict separation distress.

21.5 Discussion

First, this study highlights the importance and relevance of the use of nostalgia in the luxury brand management. Brand nostalgia, brand attachment and separation distress are stronger for the nostalgic luxury car brand (*vs.* futuristic luxury car brands). A second main contribution is the moderating effect of past temporal orientation between brand nostalgia and brand attachment, and brand nostalgia and separation distress. Finally, on a methodological side, the study shows the ability of the PLS approach to handle higher order latent variables both in the context of continuous and categorical latent moderated mediation variables.

Nevertheless, some limitations should be noted. First, the research is focused on only one product category (i.e., automobile) and two luxury brands (i.e., Mini and Infiniti). Moreover, even though the two car brands are seen as luxury, more prestigious car brands, such as for instance Jaguar, could be investigated. Further research on different categories of products and luxury brands would be helpful to achieve a generalizability of the findings. On the methodological side, other recent approaches could be investigated as well. In that spirit, and taking into account the relative small sample sizes, consistent PLS estimation (Dijkstra and Henseler 2015), could be worth relying on. Moreover, formally testing the differences of parameter estimates between the two luxury brands by means of a generalized structured component analysis (GSCA), Hwang and Takane (2004) could give additional insights on a theoretical level and hence prove to be very useful.

References

Bagozzi, R.P., Yi, Y.: On the use of structural equation models in experimental designs. J. Mark. Res. **26**, 271–284 (1989)

Bagozzi, R.P., Yi, Y.: Advanced topics in structural equation models. In: Bagozzi, R.P. (ed.) Advanced Methods of Marketing Research, pp. 1–52. Blackwell, Oxford (1994)

Bartier, A.L.: An initial step towards conceptualization and measurement of brand nostalgia. In: Proceedings of the European Marketing Academy Conference, 42nd Annual Conference, Istanbul (2013)

Bowlby, J.: Attachment and Loss, Volume 1: Attachment. Basic Books, New York (1969)

Brucks, M.: The effects of product class knowledge on information search behaviour. J. Consum. Res. **12**, 1–16 (1985)

Dijkstra, T., Henseler, J.: Consistent partial least squares path modeling. MIS Q. **39**, 297–316 (2015)

Escalas, J.E., Bettman, J.R.: You are what they eat: the influence of reference groups on consumer connections to brands. J. Consum. Psychol. **13**, 339–348 (2003)

Fornell, C., Larcker, D.F.: Evaluating structural equation models with unobservable variables and measurement error. J. Mark. Res. **18**, 39–50 (1981)

Fournier, S.: Consumers and their brands: developing relationship theory in consumer research. J. Consum. Res. **24**, 343–373 (1998)

Henseler, J., Chin, W.: A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. Struct. Equ. Model. **17**, 82–109 (2010)

Henseler, J., Ringle, C.M., Sinkowics, R.R.: The use of partial least squares path modeling in international marketing. Adv. Int. Mark. **20**, 277–319 (2009)

Holbrook, M.B., Schindler, R.M.: Echoes of the dear departed past: some work in progress on nostalgia. Adv. Consum. Res. **18**, 330–333 (1991)

Hwang, H., Takane, Y.: Generalized structured component analysis. Psychometrika **69**, 81–99 (2004)

Loveland, K.E., Smeesters, D., Mandel, N.: Still preoccupied with 1995: the need to belong and preference for nostalgic products. J. Consum. Res. **37**, 393–408 (2010)

Park, W.C., MacInnis, D., Priester, J., Eisingerich, A.B., Iacobucci, D.: Brand attachment and brand attitude strength: conceptual and empirical differentiation of two critical brand equity drivers. J. Mark. **74**, 1–17 (2010)

Sedikides, C., Wildschut, T., Gaertner, L., Routledge, C.: Nostalgia as an enabler of self continuity. In: Sani, F. (ed.) Self Continuity: Individual and Collective Perspectives, pp. 227–239. Psychology Press, New York (2008)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)

Thomson, M., MacInnis, D., Park, W.C.: The ties that bind: measuring the strength of consumer's emotional attachments to brands. J. Consum. Psychol. **15**, 77–91 (2005)

Tian, K.T., Bearden, W.O., Hunter, G.L.: Consumers' need for uniqueness: scale development and validation. J. Consum. Res. **28**, 50–66 (2001)

Usunier, J.C., Valette-Florence, P.: The time styles scale, a review of developments and replications over 15 years. Time Soc. **16**, 333–366 (2007)

Vigneron, F., Johnson, L.W.: A review and a conceptual framework of prestige-seeking consumer behavior. Acad. Mark. Sci. Rev. **3**, 1–17 (1999)

Vigneron, F., Johnson, L.W.: Measuring perceptions of brand luxury. J. Brand Manag. **11**, 484–506 (2004)

Wetzels, M., Odekerken-Schroder, G., Van Oppen, C.: Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. MIS Q. **33**, 177–195 (2009)

Chapter 22
A Partial Least Squares Algorithm Handling Ordinal Variables

Gabriele Cantaluppi and Giuseppe Boari

Abstract The partial least squares (PLS) is a popular path modeling technique commonly used in social sciences. The traditional PLS algorithm deals with variables measured on interval scales while data are often collected on ordinal scales. A reformulation of the algorithm, named Ordinal PLS (OrdPLS), is introduced, which properly deals with ordinal variables. Some simulation results show that the proposed technique seems to perform better than the traditional PLS algorithm applied to ordinal data as they were metric, in particular when the number of categories of the items in the questionnaire is small (4 or 5) which is typical in the most common practical situations.

Keywords Partial least squares path modeling (PLS-PM) • Robust Methods • Ordinal Variables

22.1 Introduction

Partial Least Squares (PLS) path modeling is largely used in socio-economic studies where path analysis is performed with reference to structural equation models with latent variables. It often happens that data are measured on ordinal scales; a typical example concerns customer satisfaction surveys, where responses given to a questionnaire are on Likert type scales assuming a unique common finite set of possible categories. In several research and applied works, averages, linear transformations, covariances and Pearson correlations are computed on the ordinal variables coming from surveys. This practice can be theoretically justified by invoking the *pragmatic* approach to statistical measurement (Hand 2009). A more accurate way would require to adopt appropriate procedures in order to handle manifest indicators of

G. Cantaluppi (✉) • G. Boari
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy
e-mail: gabriele.cantaluppi@unicatt.it; giuseppe.boari@unicatt.it

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5_22

the ordinal type (Stevens 1946). Within the LISREL covariance-based framework, several approaches are suggested; in particular, Jöreskog (2005) and Bollen (1989) make the assumption that to each manifest indicator there corresponds an underlying continuous indicator.

In order to deal with ordinal variables within Partial Least Squares path modelling, Jakobowicz and Derquenne (2007) use a procedure based on generalized linear models, while (Russolillo 2012) and (Nappo 2009) use Optimal Scaling and Alternating Least Squares. As observed by Russolillo (2012) in the procedure by Jakobowicz and Derquenne (2007) a value is assigned to measure the impact of each explanatory variable on each category of the response, while the researcher may be interested in measuring the impact of each explanatory variable on the response as a whole. The same issue characterizes the techniques illustrated by Lohmöller (1989). The present proposal goes in this direction. Wold's (1979) PLS algorithm is presented in matrix form, starting from the covariance matrix of row data. This allows us to deal with variables of the ordinal type in a manner analogous to the covariance based procedures, according to Thurstone's (1959) scaling, which assumes the presence of a continuous underlying variable for each ordinal indicator (Sect. 22.2). The polychoric correlation matrix can be defined; it is used in Sect. 22.3 to obtain parameter estimates within the PLS framework. In Sect. 22.4 simulation results give evidence that the proposed solution is particularly appropriate in all situations with a low number of scale points. This is the most common situation encountered in questionnaire analysis, where items are usually measured on at most 4 or 5 alternative points.

22.2 The Model

A structural equation model with latent variables consists of two sets of equations: the inner model, describing the path of the relationships among the latent variables, and the measurement model or outer model, representing the relationships linking unobservable latent variables to appropriate corresponding manifest variables.

The inner model is represented by the linear relations

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{exog} \\ \mathbf{Y}_{endog} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \boldsymbol{\Gamma} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{exog} \\ \mathbf{Y}_{endog} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\zeta} \end{bmatrix} = \mathbf{DY} + \boldsymbol{\nu} \tag{22.1}$$

where $\mathbf{Y}_{exog}$ and $\mathbf{Y}_{endog}$ are vectors of n exogenous and m endogenous latent random variables, defining the vector $\mathbf{Y} = [Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_{n+m}]'$; $\boldsymbol{\zeta}$ is a vector of m error components. $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are respectively $(m \times n)$ and $(m \times m)$ matrices containing the structural parameters. Matrix $\mathbf{B}$ is assumed to be lower triangular and the predictor specification, $E(\zeta_j | Y_1, \ldots, Y_{n+j-1}) = 0, j = 1, \ldots, m$, is made.

The measurement model describes the relation between each latent variable Y_j in $\mathbf{Y}$ and a single block of p_j manifest indicators, $X_{jh}, h = 1, \ldots, p_j$, elements of the $(p \times 1)$ vector random variable $\mathbf{X}$. Different types of measurement models are

available, the reflective, the formative and the MIMIC (Esposito Vinzi et al. 2010; Tenenhaus et al. 2005). We consider measurement models of the reflective type

$$\mathbf{X} = \boldsymbol{\Lambda}\mathbf{Y} + \boldsymbol{\varepsilon}. \tag{22.2}$$

In presence of ordinal indicators, we assume that to the set of ordinal variables $\mathbf{X}$ there corresponds a p-dimensional unobservable continuous indicator $\mathbf{X}^*$, represented on an interval scale with a multinormal distribution function (Jöreskog 2005; Bollen 1989; Bollen and Maydeu-Olivares 2007). Each observed ordinal indicator X_{jh} can assume I categories and is related to the corresponding continuous indicator X^*_{jh} by means of a non linear monotone function

$$X_{jh} = \left\{1 \text{ if } X^*_{jh} \le a_{jh,1};\quad 2 \text{ if } a_{jh,1} < X^*_{jh} \le a_{jh,2}; \ldots;\quad I_{jh} \text{ if } a_{jh,I_{jh}-1} < X^*_{jh}\right\} \tag{22.3}$$

where $a_{jh,1}, \ldots, a_{jh,I_{jh}-1}$ are marginal threshold values defined as $a_{jh,i} = \Phi^{-1}(F_{jh}(i))$, $i = 1, \ldots, I_{jh} - 1$, with $\Phi(\cdot)$ the distribution function of a standard Normal random variable and $F_{jh}(\cdot)$ the empirical distribution function of X_{jh} (Jöreskog 2005); $I_{jh} \le I$ denotes the number of categories effectively used by the respondents. For each pair of ordinal categorical variables, (X_h, X_k), the polychoric correlation is defined as the maximum likelihood estimate of the Pearson correlation between the corresponding underlying Normal variables (X^*_h, X^*_k), Drasgow (1986). Then, a polychoric correlation matrix can be obtained which will be used in the PLS algorithm.

In presence of manifest indicators of the ordinal type, we therefore suggest a slightly modified version of model (22.1)–(22.2), where manifest variables $\mathbf{X}$, in (22.2), are in a certain sense 'replaced' by underlying unobservable continuous indicators $\mathbf{X}^*$

$$\mathbf{X} \leftarrow \mathbf{X}^* = \boldsymbol{\Lambda}\mathbf{Y} + \boldsymbol{\varepsilon}. \tag{22.4}$$

We do not write explicitly the dependence between $\mathbf{X}$ and $\mathbf{X}^*$, since for subject $s = 1, 2, \ldots, N$ the actual score x^*_{ks} for each indicator X^*_k cannot be identified; we only assume that it belongs to the interval defined by the threshold values in (22.3) having as image the observed category x_{ks}. It will be possible to obtain point estimates for the parameters in $\mathbf{D}$ and $\boldsymbol{\Lambda}$, while only estimates of the threshold values will be directly predicted with regard to the scores of the latent variables $Y_j, j = 1, \ldots, n + m$.

22.3 The Ordinal PLS Algorithm

Wold's PLS algorithm consists of a first iterative phase giving as result the weights which allow to define latent variables scores $\hat{\mathbf{Y}}$ as *composites*, linear combinations

of their respective manifest indicators. All unknown parameters in the model are then estimated by applying OLS to the linear multiple regression sub-problems into which the inner and outer models can be decomposed. In the algorithm presented below, we will update the weights defining the *composites* $\hat{\mathbf{Y}}$ according to the Mode A method, which is appropriate with reflective measurement models (Esposito Vinzi et al. 2010; Tenenhaus et al. 2005).

A square matrix $\mathbf{T}$, of order $(n+m)$, indicating the structural relationships among latent variables in the inner model can be defined, whose generic element t_{jk} is given unit value if the endogenous Y_j is directly linked to Y_k; t_{jk} is null otherwise. $\mathbf{T}$ is the indicator matrix of $\mathbf{D}$ by having set to 0 the elements on the main diagonal.

Wold (1979) presented two versions of the PLS algorithm; the first one starts from raw data, while the second considers cross-products of manifest variables. He suggested in Wold (1982) to implement computer programs starting from product moments, see Rönkkö (2014) for an implementation in R. We propose a procedure based on the polychoric correlation matrix across observed ordinal manifest variables $\mathbf{X}$, which corresponds to the Pearson correlation matrix across the underlying continuous indicators $\mathbf{X}^*$.

The algorithm,[1] see Fig. 22.1, starts from the polychoric correlation matrix of manifest variables $\boldsymbol{\Sigma}_{XX}$ (covariance matrix if raw data are all on interval scales).

The procedure begins with an arbitrary choice of starting values for the weights corresponding to the latent composites. All weights are collected in a matrix $\mathbf{W}$.

Following a Gauss Seidel procedure, the following quantities are computed at each iteration: the covariance matrix of outer estimates $\hat{\mathbf{Y}} = \mathbf{XW}$ of composites (22.5), the 'standardizing'[2] weights (22.6), the correlation matrix of outer estimates of composites (or covariance if 'standardizing' weights are not computed) (22.7), the cross covariances between manifest indicators and inner estimates $\mathbf{Z} = \mathbf{XW}\boldsymbol{\Upsilon}$ of composites defined according to Wold's centroid scheme (22.8)–(22.9), the cross covariances between manifest indicators and outer estimates of composites (22.10). $\mathbf{C}_j^{(s)}$ and $\pm$ are defined in order to update the outer weights according to the Mode A method (22.12). The symbol $*$ in (22.6), (22.8), (22.11) denotes the Hadamard product; $\boldsymbol{\chi}$ in (22.11) is the indicator function returning 1 for elements in its matrix argument different from 0.

Observe that in this way the simultaneous updating of all composite weights is performed, according to the PLS algorithm by Lohmöller (1989). The resulting weights, defining each composite, are normalised in the sense that they sum up to 1, according to the "New Mode A" Regularized Generalized Canonical Correlation Analysis (RGCCA) by Tenenhaus and Tenenhaus (2011).[3]

[1] A detailed description of the matrix PLS algorithm by Lohmöller (1989) Table 2.2, is reported in Cantaluppi (2012).

[2] We use 'standardizing' to denote weights normalized in order to obtain standardized composites.

[3] The 'standardizing' operation in (22.6) is made according to Wold's algorithm, but is not required by "new Mode A". In the latter case set ${}_s\mathbf{W}_{k-1}^{(s)} = \mathbf{W}_{k-1}^{(s)}$ in (22.7), (22.8), (22.9), (22.10), and (22.11).

Input: $\boldsymbol{\Sigma}_{XX}$ (in presence of only ordinal items the polychoric correlation matrix), $\mathbf{T}$

Initialization: Choose $n+m$ arbitrary vectors of weights $\mathbf{w}_j^{(0)} = \left[0,\ldots,w_{j1}^{(0)},\ldots,w_{jp_j}^{(0)},0,\ldots,0\right]'$, $j = 1,2,\ldots,n+m$, each summing up to 1 and build the matrix $\mathbf{W}_{-1}^{(0)} = \left[\mathbf{w}_1^{(0)},\ldots,\mathbf{w}_j^{(0)},\ldots,\mathbf{w}_{n+m}^{(0)}\right]$.

For $s = 0,1,\ldots$ (until convergence)

For $k = 0,1,\ldots,m$: Set $\mathbf{W}_{\text{TEMP}} = \mathbf{W}_{-1}^{(s)}$ and compute:

$$\boldsymbol{\Sigma}_{\hat{Y}\hat{Y},k}^{(s)} = \mathbf{W}_{k-1}^{(s)'}\boldsymbol{\Sigma}_{XX}\mathbf{W}_{k-1}^{(s)} \tag{22.5}$$

$${}_s\mathbf{W}_{k-1}^{(s)} = \mathbf{W}_{k-1}^{(s)}\left[\boldsymbol{\Sigma}_{\hat{Y}\hat{Y},k-1}^{(s)} * \mathbf{I}\right]^{-1/2} \tag{22.6}$$

$$\boldsymbol{\Sigma}_{\hat{Y}\hat{Y},k}^{(s)} = {}_s\mathbf{W}_{k-1}^{(s)'}\boldsymbol{\Sigma}_{XX}\,{}_s\mathbf{W}_{k-1}^{(s)} \tag{22.7}$$

$$\boldsymbol{\Upsilon}_k = (\mathbf{T}+\mathbf{T}') * sign\left(\boldsymbol{\Sigma}_{\hat{Y}\hat{Y},k}^{(s)}\right) \tag{22.8}$$

$$\boldsymbol{\Sigma}_{XZ,k}^{(s)} = \boldsymbol{\Sigma}_{XX}\,{}_s\mathbf{W}_{k-1}^{(s)}\boldsymbol{\Upsilon}_k \tag{22.9}$$

$$\boldsymbol{\Sigma}_{X\hat{Y},k}^{(s)} = \boldsymbol{\Sigma}_{XX}\,{}_s\mathbf{W}_{k-1}^{(s)} \tag{22.10}$$

$$\mathbf{C}_k^{(s)} = \boldsymbol{\chi}_{{}_s\mathbf{W}_{k-1}^{(s)}} * \boldsymbol{\Sigma}_{XZ,k}^{(s)} \quad \text{and} \quad \pm = sign\left\{\mathbf{1}_p'\left[sign\left(\boldsymbol{\chi}_{{}_s\mathbf{W}_{k-1}^{(s)}} * \boldsymbol{\Sigma}_{X\hat{Y},k}^{(s)}\right)\right]\right\} \tag{22.11}$$

$$\mathbf{W}_k^{(s+1)} = \mathbf{C}_k^{(s)}[diag(\mathbf{1}_{p+q}'\mathbf{C}_k^{(s)})]^{-1}diag(\pm) \tag{22.12}$$

$$\mathbf{W}_k^{(s+1)} = \mathbf{W}_{k-1}^{(s)}\begin{bmatrix}\mathbf{O}_{n+k,n+k} & \mathbf{O}_{n+k,m-k}\\ \mathbf{O}_{m-k,n+k} & \mathbf{I}_{m-k,m-k}\end{bmatrix} + \mathbf{W}_k^{(s+1)}\begin{bmatrix}\mathbf{I}_{n+k,n+k} & \mathbf{O}_{n+k,m-k}\\ \mathbf{O}_{m-k,n+k} & \mathbf{O}_{m-k,m-k}\end{bmatrix} \tag{22.13}$$

$$\begin{cases}\text{If } k < m \text{ set } \mathbf{W}_k^{(s)} = \mathbf{W}_k^{(s+1)}\\ \text{If } k = m \text{ set } \mathbf{W}_{-1}^{(s+1)} = \mathbf{W}_m^{(s+1)} \text{ and proceed with next } s\end{cases} \tag{22.14}$$

End
End

Fig. 22.1 Flow chart for Wold's, Tenenhaus and Tenenhaus' and Lohmöller's matrix PLS algorithms

Relationship (22.13), where $\mathbf{O}$ and $\mathbf{I}$ are null and identity matrices, allows Gauss Seidel procedure to be executed. In this way[4] only weights pertaining to the first $n+k$ composites are updated and the weights of the latent composite $\hat{Y}_{n+k+1}^{(s+1)}$ are obtained at step $k+1$, based on the weights of $\hat{Y}_1^{(s+1)},\ldots,\hat{Y}_{n+k}^{(s+1)}$ and $\hat{Y}_{n+k+2}^{(s)},\ldots,\hat{Y}_{n+m}^{(s)}$. To implement the Lohmöller's algorithm only step $k=0$ has to be performed in the internal loop and one can proceed with next step s by setting $\mathbf{W}_{-1}^{(s+1)} = \mathbf{W}_0^{(s+1)}$.

Convergence is achieved when $||\mathbf{W}_{-1}^{(s+1)} - \mathbf{W}_{\text{TEMP}}|| < \varepsilon$, the tolerance level. According to Wold's algorithm final 'standardizing' weights have to be obtained. This is not required by the PLS algorithm for RGGCA, where weights are normalized.

[4]The first n composites are exogenous. Their weights are computed at step 0 of the internal cycle.

Composite values can be computed if raw data are available and OLS regressions carried out to estimate model parameters. The inner regression models describe the relationship between each endogenous composite $\hat{Y}_j, j = n+1, \ldots, n+m$ and a subset of $\left\{\hat{Y}_1, \ldots, \hat{Y}_{j-1}\right\}$ defined according to the jth row of matrix $\mathbf{D}$. OLS estimates correspond to ${}_{Rj}\boldsymbol{\Sigma}^{-1}_{\hat{Y}\hat{Y}\,j}\,\boldsymbol{\Sigma}_{\hat{Y}\hat{Y}}$ where ${}_{Rj}\boldsymbol{\Sigma}_{\hat{Y}\hat{Y}}$ is the matrix defined by extracting from $\boldsymbol{\Sigma}_{\hat{Y}\hat{Y}}$ the rows and columns pertaining to the covariates of $\hat{Y}_j$ and ${}_j\boldsymbol{\Sigma}_{\hat{Y}\hat{Y}}$ is the vector containing the correlations between $\hat{Y}_j$ and its covariates. OLS estimates in the outer model are given by the elements in (22.10), when composite variables are standardized.

By applying transformation (22.3) to the ordinal variables, threshold values related to the underlying standard normal variables X^*_{jh} are available. A point estimate of the composite $\hat{Y}_j$ cannot be determined in presence of ordinal variables for the generic subject. We can only establish an interval of possible values conditional on the threshold values pertaining the continuous indicators X^*_{jh} that underlie each ordinal manifest variable; a 'category' indication follows for $\hat{Y}_j$ from this interval, according to one of the 3 estimation methods presented in Cantaluppi (2012), Sect. 6.3.

From now on we will refer to the proposed algorithm for ordinal manifest variables with Ordinal Partial Least Squares (OrdPLS).

Optimality criteria described in Tenenhaus and Tenenhaus (2011) and Russolillo (2012) are referred, for OrdPLS, to model (22.1), (22.4), defined among the continuous indicators $\mathbf{X}^*$ underlying the ordinal $\mathbf{X}$. Also the causal predictive properties, which characterize PLS, have to be referred for OrdPLS to the $\mathbf{X}^*$ variables. The use of the polychoric correlation matrix with 'standardizing' weights is consistent with the METRIC 1 option performing the standardization of all manifest indicators (Lohmöller 1989; Tenenhaus et al. 2005).

In Schneeweiss (1993) it is shown that parameter estimates obtained with the PLS algorithm are negatively biased for the inner model. These estimates are related to the covariances across latent composites, but OrdPLS is based on the analysis of the polychoric correlation matrix. When the number of categories is sufficiently high (8 or 9) polychoric correlation values are close to Pearson's ones; while they are usually larger than Pearson's ones in presence of items with a low number of categories (Coenders et al. 1997). For this reason we can expect a possible reduction of the negative bias of the inner model estimates. A positive bias in the outer model parameter estimates corresponds to the reduction in the bias of the inner model parameter estimates for OrdPLS (Fornell and Cha 1994).

Scale reliability can be assessed by having recourse to methods based on the polychoric correlation matrix for Cronbach's α (Zumbo et al. 2007). Dillon-Goldstein's rho (Chin 1998) will be inflated due to the positive bias in the outer model parameter estimates.

22.4 Simulation Results

Some simulations[5] were performed to analyze the behavior of the procedure, in particular when item scales have a low number of points. The OrdPLS methodology was implemented in R[6]; procedures by Fox (2010) and Revelle (2012) are used to compute polychoric correlation matrices, with minor changes to allow polychoric correlations to be computed when the number of categories is larger than 8. Simulations from the model

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1, \qquad \eta_2 = \beta_{21}\eta_1 + \gamma_{22}\xi_2 + \gamma_{23}\xi_3 + \zeta_2, \qquad \eta_3 = \beta_{32}\eta_2 + \zeta_3$$

were considered. Measurement models of the reflective type were assumed, with 3 manifest ordinal reflective indicators for each latent variable $X_{ih} = {}_X\lambda_{ih}\xi_i + \varepsilon_{ih}$, $Y_{ih} = {}_Y\lambda_{ih}\eta_i + \delta_{ih}, i = 1, 2, 3, h = 1, 2, 3$.

Latent exogenous variables ξ_i were generated according both to the standard Normal distribution for all ξ_i variables (first simulation design considering symmetric Normal distributions for the latent variables) and Beta distributions with parameters $(\alpha = 11, \beta = 2)$ for ξ_1, $(\alpha = 16, \beta = 3)$ for ξ_2, $(\alpha = 54, \beta = 7)$ for ξ_3 which were then standardized (second simulation design which takes into account the presence of skew distributions). Theoretical skewness indices -0.96, -0.80 and -0.60 correspond to the three Beta distributions. The model parameters were fixed to $\gamma_{11} = 0.9$, $\gamma_{22} = 0.5$, $\gamma_{23} = 0.6$, $\beta_{21} = 0.5$ and $\beta_{32} = 0.6$. The λ coefficients were set to $0.8, 0.9, 0.95$ in each measurement model. Error components were generated from Normal distributions. Both the variances of the error components ζ_i in the inner model and those pertaining errors in the measurement models were set to values ensuring the latent and manifest variables to have unit variance.

Manifest variables X_{ih} and Y_{ih} were rescaled according to the rule ${}_{SCALED}X_{ih} = \frac{X_{ih} - \min(X_{ih})}{\max(X_{ih}) - \min(X_{ih}) + 0.01} \cdot npoints + 0.5$ with extrema computed over the sample realizations, being *npoints* the desired number of points common to all items. Values were then rounded to obtain integer responses, corresponding to conventional ordinal variables.

Simulations were performed by considering $4, 5, 7$ and 9 categories in the scales. 500 replications for each instance, each with 250 observations were made.

We expected results from PLS applied to ordinal data, as they were of the interval type, and OrdPLS to be quite similar in presence of 9 categories, since in this case polychoric correlations are close to Pearson ones.

To compare the performance of the two procedures we considered the empirical distributions of the inner model parameter estimate biases, see Table 22.1. Results are reported only for the first simulation design with 4 points and Normal

[5] See Cantaluppi (2012) for an application of the OrdPLS methodology to the well-known ECSI data set (Tenenhaus et al. 2005).

[6] The R package matrixpls, independently implemented by Rönkkö (2014), also performs PLS starting from covariance matrices.

Table 22.1 Bias distribution of the inner model parameter estimates (4 points, Normal distribution) obtained with PLS and OrdPLS and distribution of the ratio between absolute values of the biases: percentage points, mean and standard deviation

	5 %	10 %	25 %	50 %	75 %	90 %	95 %	mean	sd
PLS									
γ_{11}	−0.166	−0.158	−0.144	−0.125	−0.107	−0.094	−0.087	−0.126	0.025
γ_{22}	−0.128	−0.118	−0.095	−0.067	−0.039	−0.019	−0.004	−0.068	0.039
γ_{23}	−0.147	−0.131	−0.110	−0.084	−0.056	−0.033	−0.021	−0.083	0.038
β_{21}	−0.131	−0.119	−0.098	−0.072	−0.046	−0.023	−0.010	−0.072	0.038
β_{32}	−0.164	−0.149	−0.115	−0.083	−0.050	−0.022	−0.006	−0.084	0.049
OrdPLS									
γ_{11}	−0.111	−0.103	−0.087	−0.070	−0.052	−0.039	−0.027	−0.070	0.025
γ_{22}	−0.101	−0.090	−0.065	−0.036	−0.004	0.019	0.035	−0.035	0.042
γ_{23}	−0.111	−0.095	−0.072	−0.044	−0.014	0.009	0.023	−0.044	0.042
β_{21}	−0.103	−0.091	−0.067	−0.039	−0.011	0.016	0.031	−0.039	0.042
β_{32}	−0.138	−0.111	−0.077	−0.044	−0.010	0.020	0.036	−0.046	0.052
Ratio of absolute biases OrdPLS over PLS								geometric mean	
γ_{11}	0.329	0.392	0.465	0.557	0.613	0.666	0.693	0.522	
γ_{22}	0.073	0.166	0.376	0.594	0.755	1.090	3.803	0.531	
γ_{23}	0.113	0.182	0.385	0.577	0.697	0.792	0.982	0.483	
β_{21}	0.100	0.207	0.414	0.621	0.747	0.914	2.559	0.543	
β_{32}	0.112	0.244	0.436	0.606	0.736	0.911	3.437	0.575	

distribution, see Cantaluppi (2012) for more detailed results. Estimates obtained with the PLS algorithm are negatively biased. Only for scales with 5, 7 and 9 categories we observed about 5 % of the trials with a small or negligible positive bias for Normal distributed latent variables. The negative bias gets more evident with decreasing number of scale points. The behavior is common both to Normal and Beta situations. With OrdPLS about 10 % of the simulations always present positive bias. Most percentage points of the bias distribution obtained with the OrdPLS procedure are closer to 0 than with PLS. Averages biases are again closer to 0 with the OrdPLS algorithm. Percentage points for the two estimation procedures in case of a 9 point scale are very close, as well as average values; in this case polychoric and Pearson correlations give similar values.

The ratio between the absolute biases observed in each trial with OrdPLS and PLS was also considered, in order to better compare the two procedures. The distribution of the ratios is shown in the third sections of Table 22.1 giving evidence that over 90 % of the trials have an absolute bias of OrdPLS lower than PLS, when scales are characterized by 4 points. By comparing the 5 % and 95 % percentage points for the distributions of ratios of absolute biases in case of the Normal assumption with 4 point scales, we can observe the better behavior of OrdPLS: for parameter γ_{22} we have 5 % and 95 % percentiles of absolute ratios equal to 0.0728 and 3.8032. According to the latter value 5 % of the trials have an absolute bias in OrdPLS estimates larger more than 3.8 times that of PLS. The former value shows

how 5 % of the trials have an absolute bias of PLS larger more than $1/0.0728 = 13.7$ times than OrdPLS.

Geometric means have been computed to summarize ratios between absolute biases of OrdPLS and PLS and in all situations (except for γ_{11}, 9 points, Beta distribution) they are lower than 1. Their values increase with increasing number of scale points and get close to 1 in presence of scales with 9 points and skew Beta distribution of the latent variables.

In Sect. 22.3 we reminded that to the reduction in the bias attained by OrdPLS, pertaining the inner model parameter estimates, there corresponds an increase in the bias of the outer model parameter estimates. The bias is evident in Fig. 22.2 which reports Box & Whiskers plots for the distribution of the bias of the inner and outer model coefficients estimates from their theoretical values and the distribution of the weights under the Normal assumption for scales with 4 points. According to the Box & Whiskers Plots, OrdPLS estimates of normalized weights, which sum up to one and give information about the strength of the relationship between each composite and its manifest indicators, are characterized by a lower interquartile range.

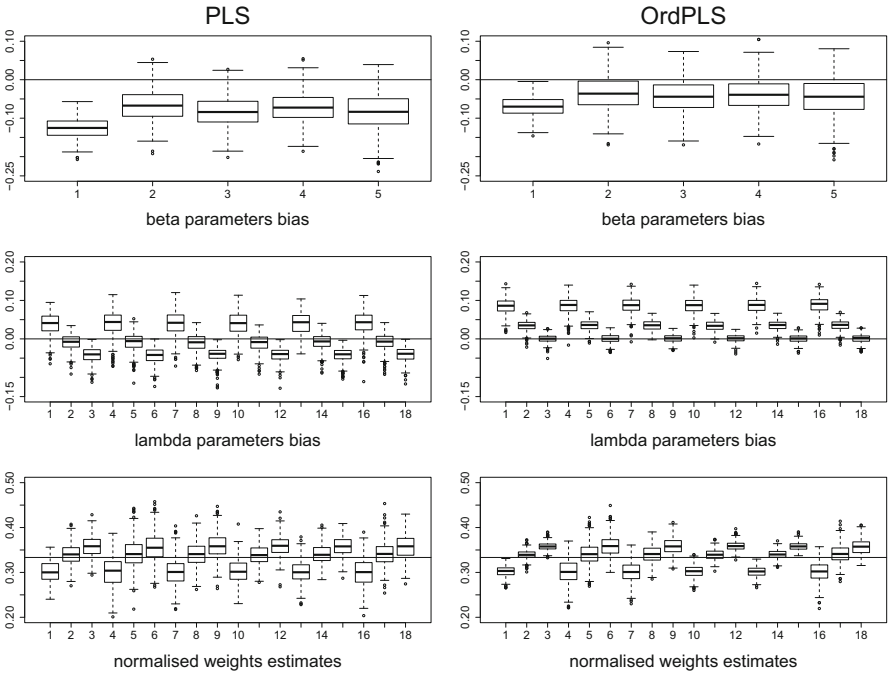

Fig. 22.2 Parameter estimates bias and weights distribution (4 points, normal distribution)

22.5 Conclusion

The Ordinal PLS (OrdPLS) algorithm dealing with variables on ordinal scales has been presented. It applies to unobservable underlying continuous indicators, assumed to generate the observed ordinal categorical variables. It is based on the use of the polychoric correlation matrix and shows better performance than the traditional PLS algorithm in presence of ordinal scales with a small number of point alternatives, by reducing the bias of the inner model parameter estimates.

A basic feature of PLS is the so-called soft modeling, requiring no distributional assumptions on the variables appearing in the structural equation model. With the OrdPLS algorithm the continuous variables underlying the categorical manifest indicators are considered multinormally distributed. This can appear a strong assumption but, as observed in Bartolomew (1996), every distribution can be obtained as a transformation of the Normal one, which can thus suit most situations. For instance, in presence of a manifest variable with a distribution skew to the left, points on the right side of the scale will have higher frequencies and the underlying continuous indicator should also be skew to the left; however, the transformation considered, see (22.3), will work anyway since it assigns larger intervals to the classes defined by the thresholds to which the highest points in the scale correspond.

Polychoric correlations are expected to overestimate real correlations when scales present some kind of skewness. This can be regarded as a positive feature for the OrdPLS algorithm when compared to the PLS algorithm applied to row data. It can represent a possible correction of the negative bias with regard to the estimates of the inner model parameters. The gain in the bias reduction is less evident for scales with a high number of categories, for which polychoric correlation values are closer to Pearson's correlations. In these cases ordinal scales can be considered as they were of the interval type, possibly according to the so-called pragmatic approach to measurement (Hand 2009).

Increasing the number of the points of the scale can help the performance of the traditional PLS algorithm when the scale is interpreted as continuous, but, as it often happens, in presence of skew distributions many points of the scale are characterized by low response frequencies, since the number of points that respondents effectively use is quite restricted. Thus the administered scale actually corresponds to a scale with a lower number of points and OrdPLS can anyway be useful in these situations.

A feature of the PLS predictive approach is that it gives direct estimation of latent scores. The OrdPLS algorithm allows only thresholds to be estimated for each composite, from which a 'category' indication for the latent variable follows according to one of the 3 estimation methods presented in Cantaluppi (2012).

Simulations have been carried out to evaluate the properties of the algorithm also in presence of skew distributions for latent variables. A reduction of the bias of the inner model parameter estimates obtained with the traditional PLS algorithm was observed. Results show also how the distributions of the weights obtained with OrdPLS have lower variability. Further research will consider a more detailed analysis of the causal predictive properties of OrdPLS and a comparison with the Optimal Scaling techniques proposed within the PLS framework by Russolillo (2012) and Nappo (2009).

References

Bartolomew, D.: The Statistical Approach to Social Measurement. Academic Press, San Diego (1996)

Bollen, K.: Structural Equations with Latent Variables. John Wiley, New York (1989)

Bollen, K., Maydeu-Olivares, A.: A Polychoric Instrumental Variable (PIV) Estimator for Structural Equation Models with Categorical Variables. Psychometrika **72**, 309–326 (2007)

Cantaluppi, G.: A Partial Least Squares Algorithm Handling Ordinal Variables also in Presence of a Small Number of Categories. Quaderno di Dipartimento, Università Cattolica del Sacro Cuore, Milano (2012). http://arxiv.org/pdf/1212.5049v1

Chin, W.: The Partial Least Squares Approach for Structural Equation Modeling. In: Marcoulides, G. (ed.) Modern Methods for Business Research, pp. 295–336. Lawrence Erlbaum Associates, London (1998)

Coenders, G., Satorra, A., Saris, W.: Alternative Approaches to Structural Modeling of Ordinal Data: a Monte Carlo Study. Struct. Equ. Model. **4**(4), 261–282 (1997)

Drasgow, F.: Polychoric and polyserial correlations. In: Kotz, S., Johnson, N. (eds.) The Encyclopedia of Statistics, vol. 7, pp. 68–74. John Wiley, New York (1986)

Esposito Vinzi, V., Trinchera, L., Amato, S.: PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement. In: Esposito Vinzi V. et al. (ed.) Handbook of Partial Least Squares, pp. 47–82. Springer-Verlag, Berlin/New York (2010)

Fornell, C., Cha, J.: Partial Least Squares. In: Bagozzi, R. (ed.) Advanced Methods of Marketing Research, pp. 52–78. Blackwell, Cambridge (1994)

Fox, J.: Polycor: Polychoric and Polyserial Correlations (2010). http://CRAN.R-project.org/package=polycor. R package version 0.7-8

Hand, D.J.: Measurement Theory and Practice: The World Through Quantification. John Wiley, New York (2009)

Jakobowicz, E., Derquenne, C.: A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy. Comput. Stat. Data Anal. **51**, 3666–3678 (2007)

Jöreskog, K.: Structural Equation Modeling with Ordinal Variables using LISREL. Scientific Software Internat. Inc. (2005). http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf

Lohmöller, J.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)

Nappo, D.: SEM with ordinal manifest variables. An Alternating Least Squares Approach. Ph.D. thesis, Università degli Studi di Napoli Federico II (2009)

Revelle, W.: Psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston (2012). http://personality-project.org/r/psych.manual.pdf. R package version 1.2.8

Rönkkö, M.: Matrixpls: Matrix-based Partial Least Squares Estimation (2014). https://github.com/mronkko/matrixpls. R package version 0.3.0

Russolillo, G.: Non-Metric Partial Least Squares. Electron. J. Stat. **6**, 1641–1669 (2012)

Schneeweiss, H.: Consistency at Large in Models with Latent Variables. In: Haagen et al. K. (ed.) Statistical Modelling and Latent Variables, pp. 299–320. Elsevier, Amsterdam/New York (1993)

Stevens, S.: On the Theory of Scales of Measurement. Science **103**, 677–680 (1946)

Tenenhaus, A., Tenenhaus, M.: Regularized Generalized Canonical Correlation Analysis. Psychometrika **76**, 257–284 (2011)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)

Thurstone, L.: The Measurement of Values. University of Chicago Press, Chicago (1959)

Wold, H.: Model construction and evaluation when theoretical knowledge is scarce: an example of the use of Partial Least Squares. Cahier 79.06 du Département d'économétrie, Faculté des Sciences Économiques et Sociales, Université de Genève, Genève (1979)
Wold, H.: Soft modeling: the basic design and some extensions. In: Jöreskog, K.G., Wold H. (eds.) Systems Under Indirect Observations, Part II, pp. 1–54. North-Holland, Amsterdam (1982)
Zumbo, B., Gadermann, A., Zeisser, C.: Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. J. Mod. Appl. Stat. Methods **6**(1), 21–29 (2007)

Author Index

A
Abdi, H., 17, 19, 73–89, 94, 101, 104
Abdinnour-Helm, S., 269, 274
Afshin-Pour, B., 93–101
Alexopoulos, G.S., 258
Allen, G.I., 89
Aluja-Banet, T., 253–265
Amato, S., 66, 180–182
Amenta, P., 219
Anderson, E.W., 170, 172
Arnoldi, W.E., 37

B
Bagozzi, R.P., 268, 289, 290
Baik, J., 233
Bang, S., 205
Bartier, A.L., 289
Bartolomew, D., 304
Basset, G.W., 157, 161, 182
Bastien, P., 190
Bates, D., 4
Baumgartner, H., 268, 269, 281
Beaton, D., 73–89
Bécue-Bertaut, M., 75, 78
Bentler, P.M., 271
Béra, M., 77
Berrendero, J.R., 202
Bertram, L., 85
Bertrand, F., 239–250
Bettman, J.R., 286
Bhattacherjee, A., 273
Billard, L., 156
Binder, H., 57
Blalock, H., 30
Blanco-Fernndez, A., 156
Blazère, M., 227–236
Boari, G., 295–304
Bock, J., 50
Boente, G., 202
Bollen, K.A., 65, 155, 276, 296, 297
Bookstein, F., 79, 80
Boudon, R., 29, 30
Boulesteix, A.-L., 45–57
Bougeard, S., 213–225
Bowlby, J., 286
Boyd, S., 134
Breckler, S.J., 268
Breiman, L., 5
Bretherton, C.S., 79
Bro, R., 19, 116
Brucks, M., 288
Brusquet, L., 115–126
Bry, X., 141–154
Bühlmann, P., 3–14
Butler, N.A., 228, 233–235
Byrne, B.M., 268–270, 275, 276, 281

C
Cai, T.T., 202
Calhoun, V.D., 74
Cantaluppi, G., 295–304
Cantor, R.M., 75
Cha, J., 300
Chen, X., 130
Cheung, G.W., 268

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5

Chin, W.W., 61, 62, 68, 160, 180, 182, 220, 267–282, 292, 300
Chow, G., 254
Christophersen, N., 228, 231, 233–235
Churchill, N.W., 93–101
Ciampi, A., 103–113, 253–265
Coenders, G., 300
Cole, M., 258
Combettes, P.L., 134
Coppi, R., 156
Cornu, G., 141–154
Corral, N., 156
Cox, D.R., 190
Crambes, C., 202
Cunningham, E., 103–113

D
Dalzell, C., 201
D'Ambra, L., 60
Dastoor, D., 258
Davino, C., 157, 161, 169–183
de Carvalho, F.A.T., 156
de Hoog, F., 18, 19, 23, 24
De Jong, S., 19, 228, 235
De la Cruz, O., 77
De Leeuw, J., 133
de Souza, B.F., 51
Delaigle, A., 202–204
Demsar, J., 46
Deng, X.D., 274
Denham, M.C., 108, 228, 233–235
Derquenne, C., 296
Diamond, P., 156
Dibbern, J., 268, 269, 271, 272, 280–282
Diday, E., 156
Dijkstra, T., 292
Dolce, P., 59–69, 169–183
Doll, W.J., 268, 270, 274, 281
Dougherty, E.R., 46, 50
Drasgow, F., 297
Dray, S., 77, 79
Duchesnay, E., 129–138
Duncan, O.D., 29, 30
Dunlop, J., 73–89

E
Edgington, E.S., 272
Efron, B., 87, 107, 182
El Hadri, Z., 29–43
Eldén, L., 19
Escalas, J.E., 286
Escofier, B., 80, 81
Escoufier, Y., 77
Eslami, A., 213–225

F
Fan, J., 203, 205
Ferraty, F., 201
Filbey, F.M., 73–89
Fleiss, J.L., 137, 138
Flury, B.N., 220
Folstein, M.F., 258
Folstein, S.E., 258
Fornell, C., 170, 172, 182, 289, 300
Fournier, S., 286, 287
Fox, J., 301
Foxall, G.R., 268
Fraiman, R., 202
Frank, I.E., 227, 229, 233
Frank, L.E., 3
Freedman, D.A., 242
Friedman, J.H., 3, 227, 229, 233
Frouin, V., 129–138

G
Galanaud, D., 115–126
Gamboa, F., 227–236
Geisser, S., 68
Geladi, P., 3
Genin, E., 75
Gertheiss, J., 203
Goeman, J.J., 193
Golub, G.H., 24
Good, P., 272
Götz, O., 178
Gould, W., 182
Goutis, C., 228, 235
Greenacre, M.J., 76–78, 81
Grubb, A., 120
Gui, J., 190, 192
Guillemot, V., 129–138
Guo, P., 156, 158

H
Hadj-Selem, F., 129–138
Hair, J., 68
Hall, P., 5, 202–204
Hanafi, M., 29–43, 62
Hand, D.J., 295, 304
Harshman, R.A., 116
Hastie, T., 3, 4, 203, 205, 241
Hauser, R.M., 29
Heise, D.R., 29

Helland, I.S., 227, 229
Henseler, J., 178, 182, 289, 290, 292
Hestenes, M., 23
Hesterberg, T., 87
Hoerl, A.E., 5
Holbrook, M.B., 286
Holmes, S.P., 77
Horowitz, J.L., 202
Höskuldsson, A., 240
Hosseini-Nasab, M., 202
Hotelling, H., 66
Hox, J., 214
Hoyle, R.H., 31
Hsieh, J.J., 273
Hsieh, P., 273
Huang, L., 201–209
Huber, P.J., 4
Hudson, T.J., 74
Hwang, H., 89, 219, 292

J
Jacques, J., 204
Jakobowicz, E., 296
Jhun, M., 205
Jiang, C.R., 202
Johnson, L.W., 287, 288
Joober, R., 103–113
Jöreskog, K.G., 31, 37, 43, 161, 296, 297
Jorgensen, K., 220
Judd, C.M., 156, 157
Jung, S., 219

K
Keil, M., 272
Kennard, R.W., 5
Kessous, A., 285–292
Kiers, H.A.L., 214
Kim, C.-H., 104
Kim, K.J., 157
Kline, R.B., 29
Kneip, A., 202
Kocherginsky, M., 182
Koenker, R., 157, 161, 170, 179, 182
Kohavi, R., 241
Koksalan, D., 157
Konishi, S., 206
Kovacevic, N., 95
Kowalski, B.R., 3, 217
Krämer, N., 18, 62, 63, 94, 190, 227, 241, 245
Kriegsman, M., 73–89
Krishnan, A., 79, 87, 94
Krzanowski, W.J., 214

L
Labbe, A., 103–113
Lai, V.S., 274
Lamberti, G., 253–265
Lambert-Lacroix, S., 190
Lance, C.E., 269
Larcker, D.F., 182, 289
Lauro, C., 59–69
Lauro, N., 60
Le Floch, E., 74
Lebart, L., 76–78, 254, 256
Lechuga, G., 115–126
Lee, D., 190
Lee, W., 190
Lee, Y., 190
Legleye, S., 213–225
Leung, S.-W., 189–198
Li, G., 170, 175
Li, H., 190, 192, 274
Li, R., 203, 205
Li, Y., 170, 175
Lian, H., 203
Liang, K.Y., 214
Lima Neto, E.A., 156
Lin, Y., 205
Lindgren, F., 220
Lingjaerde, O.C., 228, 231, 233–235
Liu, H., 130
Liu, J., 74
Löfstedt, T., 129–138
Lohmöller, J.-B., 17, 60, 67, 170, 258, 296, 298, 300
Loisel, S., 17–27
Loubes, J.-M., 227–236
Loveland, K.E., 286
Love,W., 66
Lukic, A.S., 97

M
Machado, J., 179, 182
Mage, I., 220
Magnanensi, J., 239–250
Magnoni, F., 285–292
Malhotra, M.K., 268–270
Marcolin, B.L., 273, 282
Mardia, K.V., 275
Marino, M., 156
Martens, H., 227
Martin, E., 220
Marx, B., 142
Matsui, H., 206
Maumy-Bertrand, M., 239–250

Maydeu-Olivares, A., 297
McClelland, G.H., 156, 157
McIntosh, A.R., 79, 94
McLachlan, G., 4
Meda, S.A., 74
Meinshausen, N., 4, 5, 7–10, 14
Mevik, B.H., 104
Meyer, N., 239–250
Meyer-Lindenberg, A., 74
Michel, V., 136
Mika, S., 121
Mills, A.M., 267–282
Mitteroecker, P., 79
Moè, A., 162
Molinaro, A., 46, 50
Mortier, F., 141–154
Moskowitz, H., 157

N
Naes, T., 227
Nappo, D., 296, 304
Nelder, J., 143
Nesterov, Y., 132
Nguyen, D.V., 190, 192
Noreen, E.W., 272

O
Oishi, K., 85

P
Pagès, J., 75, 78
Palmgren, J., 190
Palumbo, F., 155–166
Parikh, N., 134
Park, P.J., 190
Park, W.C., 286, 288, 289
Parzen, M.I., 182
Pawitan, Y., 190, 193
Peel, D., 4
Perlbarg,V., 115–126
Pesquet, J.C., 134
Phatak, A., 18, 19, 23, 24, 235
Philippe, C., 129–138
Pinheiro, J., 4
Polis, G.A., 30
Pompei, P., 258
Preda, C., 203, 204
Putter, H., 194
Puybasset, L., 115–126

Q
Qannari, E.M., 213–225
Qin, Z., 136

R
Ramsay, J.O., 201
Rencher, A., 66
Rensvold, R.B., 268
Revelle, W., 301
Ringle, C.M., 265
Ripatti, S., 190
Rocke, D.M., 190, 192
Romano, R., 155–166
Rönkkö, M., 298
Rosipal, R., 18, 94, 190, 227
Rosseel, Y., 41
Russolillo, G., 60, 63, 171, 296, 300, 304

S
Saad, Y., 22, 26, 229
Saeed, K.A., 269, 274
Saga, V.L., 273
Samworth, R.J., 5
Sanchez, G., 254
Sanchez-Pinero, F., 30
Saporta, G., 203
Sarstedt, M., 178, 272
Schäfer, J., 136
Schindler, R.M., 286
Schmidt, M., 135
Schneeweiss, H., 300
Schork, N.J., 74
Schultz, M.H., 26
Schumacher, M., 57
Schwarz, A., 267–282
Schwarz, G., 205
Sedikides, C., 287
Sewall, W.H., 29
Sharma, S., 268–270
Sheehan, K., 104
Sheng, J., 74
Shine, R., 30
Shipley, B., 30
Sidaros, A., 122
Silverman, B.W., 201, 202
Slawski, M., 51
Steel, D.J., 267–282
Steenkamp, J.E.M., 268, 269, 281
Stevens, S., 296
Stewart, D., 66
Stiefel, E., 23
Stone, M., 68, 182

Strimmer, K., 136
Strother, S.C., 93–101
Sugiyama, M., 241, 245
Sundaram, S., 273, 275

T
Takane, Y., 17–27, 89, 219, 292
Tam, F., 97
Tanaka, H., 156, 158
Ten Berge, J.M.F., 214
Tenenhaus, A., 62, 63, 115–126, 129–138, 298, 300
Tenenhaus, M., 60, 62, 63, 156, 161, 162, 169, 170, 181, 241, 290, 297, 298, 300
Teo, T., 269, 274
Therneau, T.M., 190, 192
Thompson, P.M., 74
Thomson, M., 286
Thurstone, L., 296
Tian, K.T., 287, 289
Tibshirani, R.J., 4, 5, 107, 182, 203
Tishler, A., 79
Trottier, C., 141–154
Tsaur, R.C., 158
Tucker, L.R., 79, 80
Tutz, G., 203

U
Usunier, J.C., 289
Utikal, K.J., 202

V
Valette-Florence, P., 285–292
van de Geer, S., 3
Van De Vijver, M.J., 190
van den Berg, E., 135
van Houwelingen, H.C., 194
van Loan, C.F., 24
Vandenberg, R.J., 269
van't Veer, L.J., 193
Verron, T., 141–154
Vieu, P., 201
Vigneron, F., 287, 288
Vinod, H., 130
Vinzi, V.E., 3, 59–69, 169–183, 265, 297, 298
Visscher, P.M., 75
Vittadini, G., 60
Vounou, M., 74
Voyer, P., 257

W
Wang, H.F., 158, 201–209
Wang, J.L., 202
Wang, W., 273
Wangen, L.E., 217
Wedderburn, R.W.M., 143
Wegelin, J.A., 79
Wehrens, R., 104
Weiner, M.P., 74, 85
Welch, B.L., 245
Wetzels, M., 290
Williams, L.J., 77, 79, 101
Witten, D.M., 131, 133
Wold, H., 3, 17, 27, 31, 60, 61, 69, 93, 155, 169, 170, 217, 227, 229, 240, 258, 296, 298
Wold, S., 190, 227, 229, 240
Wolfle, L.M., 30
Wollenberg, A.L., 62, 66
Wright, S., 29

Y
Yi, Y., 289, 290
Yourganov, G., 94
Yu, B., 5
Yuan, M., 202, 205

Z
Zadeh, L.A., 156
Zapala, M.A., 74
Zeger, S.L., 214
Zhou, Y., 189–198
Zhu, Y., 189–198
Zmud, R.W., 273
Zou, H., 203, 205
Zumbo, B., 300

Subject Index

A
ADHD. *See* Attention deficit hyperactivity disorder (ADHD)
Allele, 75, 85, 106
Allele coding, 75, 106
Alzheimer, 75, 84–88, 138
Apolipoprotein E (ApoE), 75, 88
Attention deficit hyperactivity disorder (ADHD), 103–113
Axial diffusivity (L1), 121

B
Basis functions, 202–206, 209
Behavioral partial least square (PLS), 95
Bias, 50, 93–101, 103–113, 300–304
Big data, 3, 201
Bootstrap, 48, 53, 56, 67, 87, 88, 94, 96, 104, 106, 107, 109, 111, 112, 137, 138, 182, 239–250
Bootstrapped variance (BV), 95–99, 101
Bootstrapping (BOOT), 46–48, 51, 54, 55, 57, 242, 248
Bootstrap resampling methods, 47, 48, 53, 67, 95–97
Brain, 74, 87, 94–101, 116, 117, 120–123, 125, 136, 138
Brain imaging, 84, 87, 88, 94, 95, 97
Burt's stripe, 73–88

C
Canonical correlation analysis (CCA), 66, 67, 74, 130, 219, 220
Canonical covariance analysis, 79
Categorical variables, 75, 145, 149, 151, 221, 286, 292, 297, 304
CBSEM. *See* Covariance based SEM (CBSEM)
CCA. *See* Canonical correlation analysis (CCA)
Centroid, 63, 131, 161, 171, 176, 182, 298
Chemometrics, 155
Clustered variable selection, 138
Co-inertia analysis, 79, 219
Common method variance (CMV), 280
Conceptual diagram, 30
Confirmatory factor analysis (CFA), 269
Conjugate gradient (CG), 17, 23–27
Constrained least squares (CLS), 18–20
Constrained principal component analysis (CPCA), 18
Constraint, 6, 63, 76–79, 83, 94, 116–119, 131–134, 136–138, 158, 215, 217, 270, 271, 276
Correlation, 4, 9, 31, 65, 68, 69, 96, 99–101, 104–113, 131, 152, 153, 161, 171, 172, 175–177, 179, 204, 205, 220, 229, 230, 242, 261, 262, 275, 295–298, 300–302, 304
Correspondence analysis (CA), 75, 77–78, 80–85
Covariance, 3, 5, 6, 29–43, 79, 80, 94, 95, 101, 104, 116, 117, 130, 131, 133, 155, 157, 159, 160, 182, 203, 204, 227, 229, 232, 240, 268–271, 276, 280, 281, 295, 296, 298, 300, 301
Covariance based SEM (CBSEM), 155, 157, 160, 267–282, 290

H. Abdi et al. (eds.), *The Multiple Facets of Partial Least Squares and Related Methods*, Springer Proceedings in Mathematics & Statistics 173,
DOI 10.1007/978-3-319-40643-5

Cox model, 190, 192, 194
Cross-validation (CV), 19, 45–48, 51–57, 68, 95, 108, 136, 137, 149, 152, 191, 193–198, 205, 241, 269

D
Diffusion tensor imaging (DTI), 84
Dimension reduction methods, 142, 190
Discriminant analysis (DA), 115–126

E
Eigendecomposition. *see* Eigenvalue and eigenvector
Eigenvalue, 22, 76, 96, 98, 99, 134, 229, 232–235, 259
Eigenvector, 117, 121, 122, 125, 144, 145, 229–233, 235
Endogenous, 30–33, 35, 39, 42, 61, 62, 64–67, 69, 159, 161, 162, 171, 175, 176, 180–182, 254, 255, 268, 290, 296, 298, 300
Equivalent models, 20, 43, 62, 268
Escofier, 80–82
Escofier's coding, 80, 81
Escofier-style, 80–83
Escofier-style transform, 80–83
Estimation, 3–5, 8, 13, 14, 46–49, 51, 53, 56, 60, 61, 63, 64, 68, 94, 95, 101, 142, 143, 151, 153, 155, 157, 159–161, 170–173, 175, 176, 182, 201–203, 205, 209, 216, 228, 241, 245, 270, 271, 276–279, 289, 292, 300, 302, 304
Exogenous, 30–35, 37, 40, 61, 62, 64, 65, 67, 69, 159, 161, 162, 171, 254, 268, 272, 296, 299, 301

F
Factorial, 63, 131, 171, 176, 182
FDA. *See* Fisher discriminant analysis (FDA)
FIM. *See* Finite iterative methods (FIM)
Finite iterative methods (FIM), 29–43
Fisher discriminant analysis (FDA), 116–119, 121–125
Fractional anisotropy (FA), 84–86, 121–123, 125
Frailty, 190–192, 195
Functional magnetic resonance imaging (*f*MRI), 94, 95, 97
Functional principal component (FPC), 202, 206

G
Gene, 87, 88, 104–111, 136, 138, 189–198
Generalized singular value decomposition (GSVD), 76–79
General linear model (GLM), 142–144, 147, 151, 153, 190, 194, 296
Genetics, 74, 84, 85, 87, 105, 108, 113, 227
Genome-wide association (GWA), 74, 75
Genome-wide association studies (GWAS), 74
Genomic(s), 74, 83, 104, 190, 191, 193–195
GLM. *See* General linear model (GLM)
Gradient, 132, 133, 136
GSVD. *See* Generalized singular value decomposition (GSVD)

H
Heterogeneous data, 3–14, 75
Heterozygote, 75, 86–88
Homozygote, 86–88
Horst's scheme, 131

I
Imaging genetics, 73–89
Imaging genomic, 74
Independent components analysis, 74
Inference, 10, 111, 182
Interaction effects, 220
Inter-battery factor analysis, 79
Iterated normed gradient (ING), 144

J
Joint variation, 290, 292
Joreskog's method, 31, 37–43

K
Kernel, 97, 121, 202, 209
k-fold, 45–48, 205
Krylov subspace, 19, 228, 229, 236

L
Lanczos, 21, 22, 27
Lanczos bidiagonalization, 17, 20–22, 27
Latent variable (LV), 43, 60–69, 79–82, 84–88, 97, 104, 151, 155–157, 159–162, 165, 170–172, 174–183, 190, 214, 219, 240, 255–264, 282, 289–292, 295–298, 301–304
Leave one out (LOO), 46–48, 51, 108, 109, 121, 123, 124

Likert scales, 163, 275, 288, 289, 295
Linear discriminant analysis (LDA), 51, 52, 56
Linear regression, 18, 108, 142, 149, 172, 201–209, 228, 240, 298
Loading, 18, 65, 66, 95, 105–107, 159, 160, 163, 165, 170, 175, 179, 214–220, 223–225, 269–271, 275–277, 280
Logistic, 241
Loss function, 132, 136
L_1 regularization, 137
L_2 regularization, 137

M
Machine learning, 46, 49, 130
Magging, 4–10, 12–14
Magnetic resonance imaging (MRI), 94, 116, 117, 120, 121, 124
Manifest variable (MV), 60, 63, 155, 159, 161, 170, 171, 257–262, 278, 296–298, 300, 301, 304
Maximin, 4–11, 14
Maximum likelihood *(ML),* 159, 297
MCA. *See* Multiple correspondence analysis (MCA)
Mean diffusivity (MD), 121
Mean square error of prediction (MSEP), 108, 110
Measurement model, 61, 63–65, 67, 68, 157, 159, 160, 162, 163, 170, 176, 180–182, 256–258, 264, 265, 268, 270, 271, 273, 275–278, 280, 281, 289, 290, 296–298, 301
MFDA. *See* Multiway Fisher Discriminant Analysis (MFDA)
Microarray, 51, 56, 190
MiMoPLS. *See* Mixed-modality partial least squares (MiMoPLS)
Minimax, 7
Mixed data, 73–89
Mixed-modality partial least squares (MiMoPLS), 73–89
Mixture model, 4, 8
Mode A, 61–64, 69, 171, 175, 298
Mode B, 61–64, 69, 171, 175
Model generating approach, 190
Model identification, 142, 276, 277
Model selection, 150–151
Mode Q, 176, 182
Monte Carlo, 206
Moore-Penrose, 20
MRI. *See* Magnetic resonance imaging (MRI)
Multiblock, 61, 64, 130, 213–225
Multiblock PLS (mbPLS), 213–225
Multicollinearity, 61, 65, 66, 68–69
Multigroup, 213–225
Multigroup PLS (mgPLS), 213–225
Multiple correspondence analysis (MCA), 75, 78–81, 83
Multiple Factor Analysis, 75
Multivariate, 17, 59, 60, 74, 79, 94, 101, 103–113, 116, 143–144, 147, 153, 160, 191, 192, 202, 207, 214, 275, 281, 289
Multiway, 115–126
Multiway Fisher Discriminant Analysis (MFDA), 116–126

N
Nesterov smoothing, 132
Neuroimaging, 74, 75, 83, 93–101, 116
Nonlinear effects, 75
Number of components, 10, 11, 104, 108–110, 112, 113, 150, 219, 240–243, 245–247, 249

O
OLS. *See* Ordinary Least Square (OLS)
Optimism bias, 103–113
Optimization, 5, 61, 63, 64, 69, 76, 77, 117–119, 130, 131, 133, 137, 138, 156, 160, . 181, 220, 249
Ordinal, 81, 88, 295–304
Ordinary Least Square (OLS), 5, 18, 69, 160, 170, 172, 220, 227, 234, 298, 300
Orthogonalization, 22
Orthogonal polynomial, 228, 230–231, 233, 235, 236
Overfitting, 104, 190, 241

P
Partial least square correspondence analysis (PLSCA), 80–84
Partial Least Squares Correlation (PLSC), 75, 79–84, 86, 89, 94–96, 98, 99, 101
Partial Least Squares path modeling (PLS-PM), 60–67, 69, 155, 169, 170, 176–183, 254, 256–259, 263, 265, 296
Partial Least Squares Regression/PLS Regression (PLSR), 5, 191, 240, 241, 243–246, 250
Partial Possibilistic Regression Path Modeling (PPRPM), 155–166
Path analysis, 29, 30, 43, 295
Path coefficient, 30, 160–162, 164, 166, 171, 175, 176, 178, 254–257, 263, 264, 291
Path direction, 59–69

Path Model, 29–43, 156, 254–256
Path Modeling, 59–69, 89, 130, 155–166, 169–183, 220, 295
PATHMOX, 253–265
PCA. *See* Principal component analysis (PCA)
Permutation, 267–282
PLS1 algorithm, 18–21, 27
PLSC. *See* Partial least squares correlation (PLSC)
PLSCA. *See* Partial least square correspondence analysis (PLSCA)
PLS-PM. *See* Partial Least Squares path modeling (PLS-PM)
PLSR. *See* Partial Least Squares Regression/PLS Regression (PLSR)
Power, 45–57, 93–101
PPRPM. *See* Partial Possibilistic Regression Path Modeling (PPRPM)
Prediction, 3, 4, 7, 8, 12, 13, 45–50, 59–69, 136, 137, 150, 152, 159, 160, 162, 180, 190–195, 202, 205, 209, 220, 225, 227, 241, 271
Principal component analysis (PCA), 18, 68, 77–81, 83, 101, 116, 138, 144, 145, 149, 214, 232
Procrustes, 96
Projection, 83, 84, 94, 117, 119, 120, 133–135, 219, 229, 232, 236

Q
Questionnaire, 157, 162–166, 215, 220, 288, 296

R
Radial diffusivity (Lt), 121
Recursive path, 29–43
Regression, 3–5, 8, 10–13, 17, 18, 20, 22, 30, 67, 68, 74, 89, 156, 158, 160, 161, 170, 171, 173, 179, 180, 190, 202, 209, 220, 228, 236, 240, 242, 254, 255
Regularization, 89, 117, 121, 123, 131, 136, 137, 142, 144, 205, 209
Regularized canonical correlation analysis, 130
Regularized Generalized Canonical Correlation Analysis (RGCCA), 129–138, 298
Reliability, 101, 163, 165, 174, 268, 282, 289, 290, 300
Reproducibility, 96
Resampling, 45–57, 67, 93–101, 182
RGCCA. *See* Regularized Generalized Canonical Correlation Analysis (RGCCA)
Robustness, 7, 8, 13–14, 101, 246

S
Salience, 95, 96, 101, 286
Sample size, 3, 11–13, 101, 104, 109, 111, 113, 207, 216, 244, 245, 258, 268, 269, 271, 272, 289, 292
SCGLR. *See* Supervised Component Generalized Linear Regression (SCGLR)
SEMs. *See* Structural equation models (SEMs)
SGCCA. *See* Sparse GCCA (SGCCA)
Shrinkage, 206, 209, 231–236
SIMPLS, 51
Simulation, 8, 41, 46, 50, 95, 97, 98, 101, 104–109, 111, 151, 206–207, 241–243, 248, 250, 296, 301–303
Single nucleotide polymorphisms (SNPs), 74, 75, 84, 85, 87, 88, 103–113
Singular value decomposition (SVD), 75–79, 95, 228–229
Singular vector, 76, 77, 94
SNPs. *See* Single nucleotide polymorphisms (SNPs)
Sparse GCCA (SGCCA), 130, 131, 138
Spline, 202
Split-half, 95, 96, 101
Statistical learning, 52
Structural equation models (SEMs), 29, 155–157, 159, 166, 267–282, 290, 295, 296, 304
Structural model, 60–62, 64, 65, 68, 69, 157, 160, 162, 163, 166, 171, 173, 178, 180–183, 254, 256, 259, 268, 270, 273, 275, 277, 279, 281, 289, 290
Supervised Component Generalized Linear Regression (SCGLR), 141–154
Survey, 172, 215, 220, 274, 288, 295
SVD. *See* Singular value decomposition (SVD)

T
Task-PLS, 95, 99
TBI. *See* Traumatic brain injury (TBI)
Traumatic brain injury (TBI), 120–125
t-test, 46, 49–51, 56, 245, 246, 248, 249, 272
Tucker's inter-battery factor analysis, 79

V
Validity, 31, 261–262, 268, 275, 281, 289, 290
Variational optimization problem, 117
Voxels, 74, 85–88, 94, 95, 98, 116, 121–123

W
Wold's algorithm, 298, 299