Contributions to Statistics

Anna Maria Paganoni
Piercesare Secchi *Editors*

Advances in Complex Data Modeling and Computational Methods in Statistics

Springer

Contributions to Statistics

More information about this series at
http://www.springer.com/series/2912

Anna Maria Paganoni • Piercesare Secchi
Editors

Advances in Complex Data Modeling and Computational Methods in Statistics

Springer

Editors
Anna Maria Paganoni
Dipartimento di Matematica
Politecnico di Milano
Milano
Italy

Piercesare Secchi
Dipartimento di Matematica
Politecnico di Milano
Milano
Italy

ISSN 1431-1968
Contributions to Statistics
ISBN 978-3-319-11148-3 ISBN 978-3-319-11149-0 (eBook)
DOI 10.1007/978-3-319-11149-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014955400

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Statistics is rapidly changing. Pushed by data generated by new technologies, statisticians are asked to create new models and methods for exploring variability in mathematical settings that are often very far from the familiar Euclidean environment. Functions, manifold data, images and shapes, network graphs and trees are examples of object data that are becoming more and more common in statistical applications but for which the established multivariate theory—frequentist or Bayesian—shows its limitations and downfalls. Moreover the questions the statistical analysis should answer are also new. Through the exploration of large administrative databases, public health institutions seek real time epidemiological information; by means of the analysis of the humongous data generated by social networks, large companies expect to measure intangible quantities like their reputation or the sentiment of their stakeholders. In life science as well as in geophysics statisticians are asked to integrate the information coming from the observation of real data with the knowledge encapsulated in mathematical models often expressed in terms of differential equations with difficult boundary conditions and constraints. New methodological problems are posed which do not have a counterpart in multivariate analysis, like decoupling phase and amplitude variability for the analysis of functional data or developing viable models for statistical inference based on manifold data. Last but not least the exploration of big data and the need to effectively transfer the acquired knowledge to the larger public of decision makers requires the statistician to design new graphical and visualization tools, freed from the standard representations developed over the years for communications in print. The papers of this volume have been selected among those presented at the conference "*S.Co.2013*: Complex data modeling and computationally intensive methods for estimation and prediction" held at the Politecnico di Milano, September 9–12, 2013. Over the years the *S.Co.* conference became a forum for the discussion of new developments and applications of statistical methods and computational techniques for complex and high dimensional data: that of 2013 is its eighth edition, the first one being held in Venice in 1999.

The book is addressed to statisticians working at the forefront of the statistical analysis of complex and high dimensional data and offers a wide variety of statistical models, computer intensive methods and applications: network inference from the analysis of high dimensional data; new developments for bootstrapping complex data; regression analysis for measuring the downsize reputational risk; statistical methods for research on the human genome dynamics; inference in non-Euclidean settings and for shape data; Bayesian methods for reliability and the analysis of complex data; methodological issues in using administrative data for clinical and epidemiological research; regression models with differential regularization; geo-statistical methods for mobility analysis through mobile phone data exploration.

Most notably we included in the book a paper coauthored by Ph.D. students and post-docs at the laboratory for Modeling and Scientific Computing (MOX) of the Department of Mathematics at the Politecnico di Milano who were the driving force of the Barcamp event *Technology foresight and statistics for the future* in honour of the 150th anniversary of the Politecnico di Milano and which followed the official *S.Co.2013* conference. The Barcamp was the final event closing a competition challenging young EU statisticians to envision statistical models and methods that will have an impact on the development of technology in the next 25 years, before the 175th anniversary of Politecnico di Milano. Through new means like drama playing, videos, interactive games and discussions which challenged the traditional experience of a scientific meeting, the barcampers explored the future of big data analysis, computational statistics and data visualization.

Milano, Italy — Anna Maria Paganoni
Milano, Italy — Piercesare Secchi

Contents

Inferring Networks from High-Dimensional Data with Mixed Variables

Antonino Abbruzzo and Angelo M. Mineo

1 Introduction

Graphical models are useful to infer conditional independence relationships between random variables. The conditional independence relationships can be visualized as a network with a graph. Graphs are objects with two components: nodes and links. Nodes are in one-to-one correspondence with random variables and links represent relations between genes. If a link between two genes is absent this means that these two genes are conditional independent given the rest. Pairwise, local and global Markovian properties are the connections between graph theory and statistical modeling [1–3].

Applications of graphical models include among others the study of gene regulatory networks where expression levels of large number of genes are collected, simultaneously [4]. A microarray is a collection of microscopic DNA spots attached to a solid surface. Understanding how genes work together as a network could (1) hold the potential for new treatments and preventive measures in disease, (2) add a new level of complexity to scientists' knowledge of how DNA works to integrate and regulate cell functionality. Many of the works on trying of inferring gene regulatory networks have focus on penalized Gaussian graphical models. The idea is to penalize the maximum likelihood function, for example with the ℓ_1-norm, to produce sparse solutions. The main assumption of these models is that the networks are sparse, which means many of the variables are conditionally independent from the others. In this setting, Meinshausen and Bühlmann [5] proposed to select edges for each node in the graph by regressing the variable on all the other variables using

A. Abbruzzo (✉) • A.M. Mineo
Dipartimento Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze, Ed. 13, 90128 Palermo, Italy
e-mail: antonino.abbruzzo@unipa.it; angelo.mineo@unipa.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_1

ℓ_1 penalized regression. Penalized maximum likelihood approaches using the ℓ_1 penalty have been considered in [6, 7] where different algorithms for estimating sparse networks have been proposed. The most known algorithm to estimate sparse graphs is probably the graphical lasso (glasso) proposed by Friedman et al. [8]. These models cannot deal with high-dimensional data with mixed variables. However, the need of statistical tools to analyze and extract information from such data has become crucial. For example, the most recent task in DREAM8 challenge [9] is related to predict the response of Rheumatoid Arthritis patients to anti-TNF therapy based on genetics and clinical data.

In their seminal paper Lauritzen and Wermuth [10] introduced the problem of dealing with mixed variables. Recently, Hoff [11] proposed a semiparametric Bayesian copula graphical model to deal with mixed data (binary, ordinal and continuous). The semiparametric Bayesian copula graphical model uses the assumption of Gaussianity on the multivariate latent variables which are in one-to-one correspondance with the observed variables. Conditional dependence, regression coefficients and credible intervals can be obtained from the analysis. Moreover, copula Gaussian graphical models allow to impute missing data. However, the Bayesian copula approach is infeasible for higher-dimensional problems due to its computational complexity and problem of convergence to the proposal distribution.

In this paper, we present two classes of graphical models, namely strongly decomposable graphical models [12] and regression-type graphical models [13], which are classes of models that can be used for analyzing high-dimensional data with mixed variables. Assuming that the conditional distribution of a variable A given the rest depends on any realization of the remaining variables only through the conditional mean function, the regression models are useful to find the matrix weights which can be further employed to recover the network. The aim here are (1) to give some insight on the use of decomposable models for recovering graph structure; (2) to connect this model with the use of regression-type graphical lasso; (3) to provide a simulation study to compare graphical lasso, which is a penalized approach, to strongly decomposable graphical models.

The rest of this paper is organized as follows. In Sect. 2, we briefly recall the methodologies used to infer decomposable graphical models and regression-type graphs for mixed data. In Sect. 3 we show a simulation study in which we compare several type of graphs. In Sect. 4, we show an application of the methodology to a real dataset which contains mixed variables that are the expression level of genes collected in a microarray experiment and some clinical information of the patients.

2 Methodology

A graph is a couple $G = (V, E)$ where V is a finite set of nodes and $E \subset V \times V$ is a subset of ordered couples of V. Nodes are in one-to-one correspondence with random variables. Links represent interactions between the nodes. In this paper, we are interested in links which represent conditional independence between two

random variables given the rest. Suppose we have d discrete and q continuous nodes and write the sets of nodes as Δ and Γ, where $V = \{\Delta \cup \Gamma\}$. Let the corresponding random variables be $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = (X_1, \ldots, X_d)$ and $\mathbf{Y} = (Y_1, \ldots, Y_q)$, and a typical observation be $(\mathbf{x}, \mathbf{y})$. Here, $\mathbf{x}$ is a d-tuple containing the values of the discrete variables, and $\mathbf{y}$ is a real vector of length q. We will denote with $P(\mathbf{z})$ a joint probability distribution for the random variables $(\mathbf{X}, \mathbf{Y})$.

2.1 Decomposable Graphical Models for High-Dimensional Data

Finding a conditional independence graph from data is a task that requires the approximation of the joint probability distribution $P(\mathbf{z})$. A product approximation of $P(\mathbf{z})$ is defined to be a product of several of its component distributions of lower order. We consider the class of second-order distribution approximation, i.e.:

$$P_a(\mathbf{z}) = \prod_{i=1}^{p} P(z_i, z_{j(i)}), \qquad 0 \leq j(i) \leq p$$

where $(j_1, \ldots, j_p)$ is an unknown permutation of integers $(1, 2, \ldots, p)$, where $p = d + q$.

For discrete random variables, Chow and Liu [14] proved that the problem of finding the goodness of approximation between $P(\mathbf{x})$ and $P_a(\mathbf{x})$ considering the minimization of the closeness measure

$$I(P, P_a) = \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{P_a(\mathbf{x})},$$

where $\sum_{\mathbf{x}} P(\mathbf{x})$ is the sum over all levels of the discrete variables, is equivalent to maximizing the total branch weight $\sum_{i=1}^{p} I(x_i, x_{j(i)})$, where

$$I(x_i, x_{j(i)}) = \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)}) \log \left(\frac{P(x_i, x_{j(i)})}{P(x_i) P(x_{j(i)})} \right). \tag{1}$$

Calculating the total branch weight for each of the p^{p-2} trees would be computationally too expensive even for moderate p. Fortunately, several algorithms can be used to solve the problem of finding dependence tree of maximum weight, such as Kruskal's algorithm, Dijkstra's algorithm, Prim's algorithm. These algorithms start from a square weighted matrix p by p, where a weight for a couple of variables (X_i, X_j) is given by the mutual information $I(x_i, x_j)$. So, the problem is reduced to calculating $p(p-1)/2$ weights. Consider, now a real application where probability

distributions are not given explicitly. Let $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N$ be N independent samples of a finite discrete variable $\mathbf{x}$. Then, the mutual information can be estimated as follows:

$$\hat{I}(x_i, x_j) = \sum_{u,v} f_{uv}(i,j) \log \frac{f_{uv}(i,j)}{f_{u(i)} f_{v(j)}},$$

where $f_{uv(i,j)} = \frac{n_{uv}(i,j)}{\sum_{uv} n_{uv}(i,j)}$, and $n_{uv}(i,j)$ is the number of samples such that their ith and jth components assume the values of u and v, respectively. It can be shown that with this estimator we also maximize the likelihood for a dependence tree.

This procedure can be extended to data with both discrete and continuous random variables [12]. The distributional assumption is that random variables $\mathbf{Z}$ are conditionally Gaussian distributed, i.e. the distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is multivariate normal $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ so that both the conditional mean and covariance may depend on i. We refer to the homogenous or heterogenous case if $\boldsymbol{\Sigma}$ does or does not depend on i, respectively. More details on this conditional Gaussian distribution can be found in [10]. To apply the Kruskal's algorithm, in the mixed case, we need to find an estimator of the mutual information $I(x_u, y_v)$ between each couple of variables. For a couple of variables (X_u, Y_v) we can write the sample cell counts, means, and variances as $\{n_i, \bar{y}_v, s_i^{(v)}\}_{i=1,\ldots,|X_u|}$. An estimator of the mutual information, in the homogenous case, is

$$\hat{I}(x_u, y_v) = \frac{N}{2} \log\left(\frac{s_0}{s}\right),$$

where $s_0 = \sum_{k=1}^{N} (y_v^{(k)} - \bar{y}_v)^2 / N$ and $s = \sum_{i=1}^{|X_u|} n_i s_i / N$. There are $k_{x_u, y_v} = |X_u| - 1$ degree of freedom. In the heterogeneous case, an estimator of the mutual information is

$$\hat{I}(x_u, y_v) = \frac{N}{2} \log(s_0) - \frac{1}{2} \sum_{i=1,\ldots,|X_s|} n_i \log(s_i)$$

with $k_{x_u, y_v} = 2(|X_u| - 1)$ degrees of freedom.

Note that the algorithm will always stop when it has added the maximum number of edges, i.e. $p-1$ for an undirected tree. Edwards et al. [12] suggested to use either $\hat{I}^{AIC} = \hat{I}(x_i, x_j) - 2k_{x_i, x_j}$ or $\hat{I}^{BIC} = \hat{I}(x_i, x_j) - \log(n) k_{x_i, x_j}$, where k_{x_i, x_j} are the degrees of freedom, to avoid inclusion of links not supported by the data.

The class of tree graphical models can be too restrictive for real data problem. However, we can start from the best spanning tree and determine the best strongly decomposable graphical model. A strongly decomposable graphical model is a graphical model whose graph neither contains cycles of length more than three nor forbidden path. A path exists between nodes A and B if one can reach A from B in

a finite number of steps. A forbidden path is a path between two not adjacent discrete nodes which passes through continuous nodes. The distributional assumption is that the random variables are conditional Gaussian distributed. This procedure would be NP-hard without the following result.

If $M_0 \subset M_1$ are decomposable models differing by one edge $e = (v_i, v_j)$ only, then e is contained in one clique C of M_1 only, and the likelihood ratio test for M_0 versus M_1 can be performed as a test of $v_i \perp v_j | C_{\setminus\{v_i, v_j\}}$. These computations only involve the variables in C. It follows that for likelihood-based scores such as AIC or BIC, score differences can be calculated locally which is far more efficient then fitting both M_0 and M_1. This leads to considerable efficiency gains.

To summarize, strongly decomposable model is an important class of model that can be used to analyze mixed data. This class restricts the class of possible interaction models which would be to huge to be explored. Moreover, we have the important results that for strongly decomposable graphical models closed-form estimator exists.

2.2 Regression-Type Graphical Models

Recently, Edwards et al. [13] proposed to estimate stable graphical models with random forest in combination with stability selection using regression models. Their main idea is motivated by the following theorem.

Assume that, for all $j = 1, \ldots, p$ the conditional distribution of Z_j given $\{Z_h; h \neq j\}$ is depending on any realization $\{z_h; h \neq j\}$ only through the conditional mean function:

$$\mu_j(\{z_h; h \neq j\}) = E[Z_j | z_h; h \neq j].$$

Assume the conditional mean exists, then

$$Z_j \perp Z_i | \{Z_h; h \neq j, i\}$$

if and only if

$$\mu_j(\{z_h; h \neq j\}) = \mu_j(\{z_h; h \neq j, i\})$$

does not depend on z_i for all $\{z_h; h \neq j\}$.

Suppose the network is composed by variables some of which are predictors and some of which are response variables. We use this theorem to determine the weight importance of each predictor on the response variable. To establish the importance of each predictor regression coefficients need to be comparable, i.e. standardized regression coefficients need to be used. These coefficients can also be interpreted as elasticity, i.e. how much we can change the regressor, by attempting to exogenously change one of the predictor.

2.3 Simple Example

In this section, we show a simple example on simulated mixed data. The aim is to recover the graph in Fig. 1 with a decomposable graphical model and to evaluate the relative importance of each predictor to the regressors with regression-type graphical models. In particular, in Fig. 1 we represents five variables with some of them that are regressor variables. These variables are those one having at least an incoming link. Table 1 shows distributions, models and conditional means of each variable. Regression coefficients are given in Table 2.

To generate $N = 100$ independent samples with structure given in Fig. 1 and conditional mean and distribution given in Table 1, we consider the following procedure:

- generate Y_5 from a normal with mean zero and variance one. Then, calculate π_4 and π_3 and generate Y_4 and Y_3;
- calculate μ_2 and generate Y_2. Then, calculate μ_1 and generate Y_1;
- repeat the process 100 times.

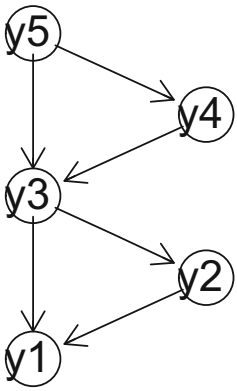

Fig. 1 Directed graph

Table 1 Model assumption for random variables represented in the DAG in Fig. 1

Distribution	Model	Conditional mean
Gaussian	$Y_1 \sim N(\mu_1, \sigma^2 = 1)$	$\mu_1 = \sum_{j=1}^{5} \beta_{j1} y_j$
Gaussian	$Y_2 \sim N(\mu_2, \sigma^2 = 1)$	$\mu_2 = \sum_{j=1}^{5} \beta_{j2} y_j$
Binomial	$Y_3 \sim Binom(1, \pi_3)$	$\pi_3 = \frac{\exp(\sum_{j=1}^{5} \beta_{j3} y_j)}{1+\exp(\sum_{j=1}^{5} \beta_{j3} y_j)}$
Binomial	$Y_4 \sim Binom(1, \pi_4)$	$\pi_4 = \frac{\exp(\sum_{j=1}^{5} \beta_{j4} y_j)}{1+\exp(\sum_{j=1}^{5} \beta_{j4} y_j)}$
Gaussian	$Y_5 \sim N(\mu_5, \sigma^2 = 1)$	$\mu_5 = 0$

There are three continuous Gaussian random variables and two binomial random variables. Regression coefficients are given in Table 2

Table 2 Regression coefficients

β	1	2	3	4	5
1	0	0	0	0	0
2	0.01	0	0	0	0
3	0.31	0.45	0	0	0
4	0	0	0.98	0	0
5	0	0	0.69	0.72	0

Table 3 Graph edge ordering and standardized regression coefficients

	V_i	V_j	AIC-ranking	SC
1	3	5	192.36	0.31
2	4	5	148.53	0.24
3	3	4	144.53	1.17
4	2	3	58.82	0.16
5	1	3	34.34	0.18
6	2	5	33.52	0.06
7	2	4	30.07	0.086
8	1	2	15.68	0.094
9	1	5	5.55	0.015
10	1	4	−1.60	0.07

Fig. 2 Recovered graph

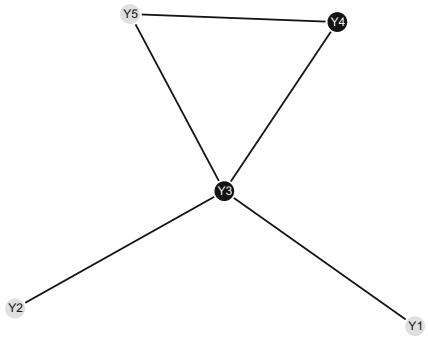

Table 3 shows the relative importance according to AIC ranking (AIC-ranking column) and the score calculated according to standard regression coefficients (SC column). There are ten possible links for an undirected graphical model. According to AIC ranking, the first link to be drawn in the tree is the link between variables Y_3 and Y_5. The selected strongly decomposable graphical model is shown in Fig. 2. It seems that ranking the links according to regression coefficients can give a more information on the relative importance of each link. In fact, from column SC in Table 3 we can see that regression-type graphical model would order the coefficients almost in the same order as the original coefficients.

3 Simulation Study

We perform a simulation study to compare the performance of graphical lasso to decomposable graphical models, in terms of recovering of the graph. The support recovery of the graph is evaluated by the following scores:

$$\text{PPR} = \frac{TP}{TP + FP}, \qquad \text{Sensitivity} = \frac{TP}{TP + FN},$$

and

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP \times TN)(FN \times FP)}},$$

where TP are the true positive, FP are the false positive, TN are the true negative and FN are the false negative. The larger the score value, the better the classification performance.

The "best" graph structures are estimated in terms of AIC (minForest-aic) and BIC (minForest-bic) for the decomposable graphical models. Whereas, for the graphical lasso we select the graph according to stability selection procedure [15].

We consider five models as follows:

- Model 1. A banded graph with bandwidth equal to 1;
- Model 2. A cluster graph where the number of cluster is about $p/20$ if $p > 40$ and 2 if $p \leq 40$. For cluster graph, the value 3/d is the probability that a pair of nodes has an edge in each cluster;
- Model 3. An hub graph where the number of hubs is about $p/20$ if $p > 40$ and 2 if $p \leq 40$;
- Model 4. A random graph where the probability that an edge is present between two nodes is $3/p$;
- Model 5. A scale-free graph where an edge is present with probability 0.9.

We use the function `huge.generator` of the R package `huge` to generate these graphical structures [16]. We keep the structure of the graph fixed and simulate $n = 100$ independence samples from a multivariate distribution with $\mu = 0$ and $\Sigma = K^{-1}$ where zero elements in K are absent links. For each model, we generate a sample of size $n = 100$ from a multivariate normal distribution We consider different values of $p = (10, 30, 50, 100, 150)$ and 100 replicates. We report the results for the support recovery of the precision matrix together with an example of the graph structures of each of the five models in Appendix.

The main conclusion which can be drawn from the results reported in the tables is that the strongly decomposable graphical model show, generally, comparable or better performance both in lower and high-dimensional case. We would expect minForest-bic have better results than minForest-aic but this doesn't appear in

our simulation study. The glasso-stars performs worse than minForest-aic and minForest-bic for banded graphs and hub graphs. This could be due to the particular structure of the graph and it should not be linked with the selection method. In other words, it seems to be a limitation of the glasso.

4 Analysis of Breast Cancer Data

In this section we analyze a breast cancer dataset. The data come from a study performed on 62 biopsies of breast cancer patients over 59 genes. These genes were identified using comparative genomic hybridization. Continuous measures of expression levels of those 59 genes were collected. In order to link gene amplification/deletion information to the aggressiveness of the tumors in this experiment, clinical information is available about each patient: age at diagnosis (AGE), follow-up time (Surv.Time), whether or not the patient died of breast cancer (C.Death), the grade of the tumor (C.Grade), the size of the tumor (Size.Lesion), and the Nottingham Prognostic Index (NPI). C.Death is a dichotomous variable, C.Grade is ordinal with three categories and NPI is a continuous index used to determine prognosis following surgery for breast cancer. NPI values are calculated using three pathological criteria: the size of the lesion; the number of involved lymph nodes; and the grade of the tumor. The complete dataset results in 62 units and 65 variables.

Our aim is to find a network which may underline some important relationships between the 65 variables. These variables comprise both gene expression levels and clinical variables. We use the package `gRapHD` [17] to analyse the breast cancer data. Firstly, the forest that minimizes the BIC is found by applying the function `minForest`. This result in a quite simple graph with at last 64 links. A more complex model can be found by applying the function `stepw`. This function performs a forward search strategy through strongly decomposable models starting from a given decomposable graphical model. At each step, the edge giving the greatest reduction in BIC is added. The process ends when no further improvement is possible.

Figure 3 shows the graph for the homogeneous strongly decomposable graphical model applied to the breast cancer data with starting point a minimum BIC forest with a link between C.Grade and C.Death. Black nodes indicate discrete variables while grey nodes represent continuous variables. The graph in Fig. 3 indicates that Gene 4 is the connection between Surv.Time, C.Death, C.Grade, NPI and Size.Lesion and the gene expression levels. Gene 4 separates two blocks of genes the one represented in the top part of Fig. 3 and the one represented in the bottom part of the same figure. The other most connected genes are Gene 12 and Gene 49 with 8 and 9 nodes, respectively. C.grade and Size.Lesion are linked to NPI as we expected and there is a short path between NPI and Survival time.

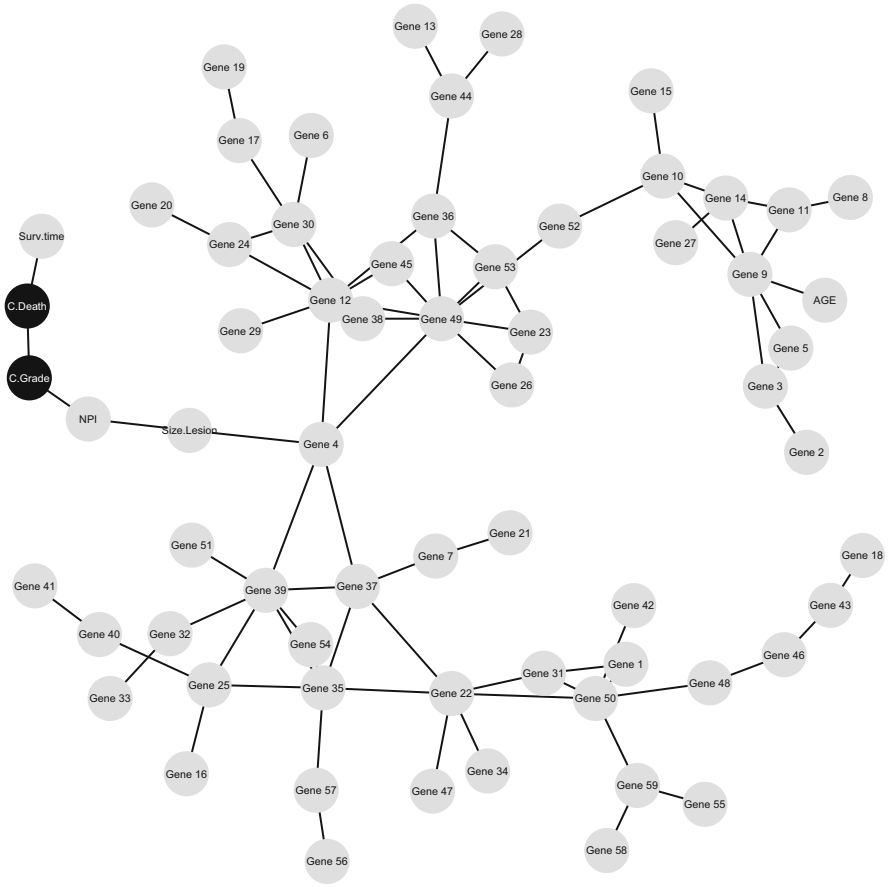

Fig. 3 Graph obtained by applying the homogeneous strongly decomposable graphical model to breast cancer data with starting point a minimum BIC forest with a link between C.Grade and C.Death. *Black dot nodes* indicate discrete variables while *circle grey nodes* represent continuous variables

5 Discussion

In this paper, we have explored a class of graphical models, the strongly decomposable graphical models, which can be used to infer networks for high-dimensional mixed data. Results from the simulation study shows comparable or better performance in terms of graphs recovering with respect to graphical lasso. There are some limitations. The first one is due to the assumption of decomposable models, namely neither cycle of length more than 3 nor forbidden path can be estimated. The second one is due to the distributional assumption. In fact, the conditional Gaussian distribution cannot take into account dependence of a continuous variable to a discrete one. So, careful attention should be paid during the analysis of real

data. In the real data analysis, in which mixed data are present, we have shown that a relation between gene expression levels and clinical conditions of the patients seems to be present. We have not dealt with parameter estimation which is indeed another challenge task for high-dimensional data. To conclude, the main advantages of using strongly decomposable graphical models we have illustrated in this paper are: (1) their feasibility for high-dimensional setting; (2) the facility to communicate the results by showing the graph; (3) the possibility to catch patters in terms of clustering, hubs, important variables from the conditional independent graph. Moreover, regression-type graphical models can give some insight on the ordering of importance for some of the regressors.

Appendix

See Fig. 4, Tables 4, 5, 6, 7, and 8

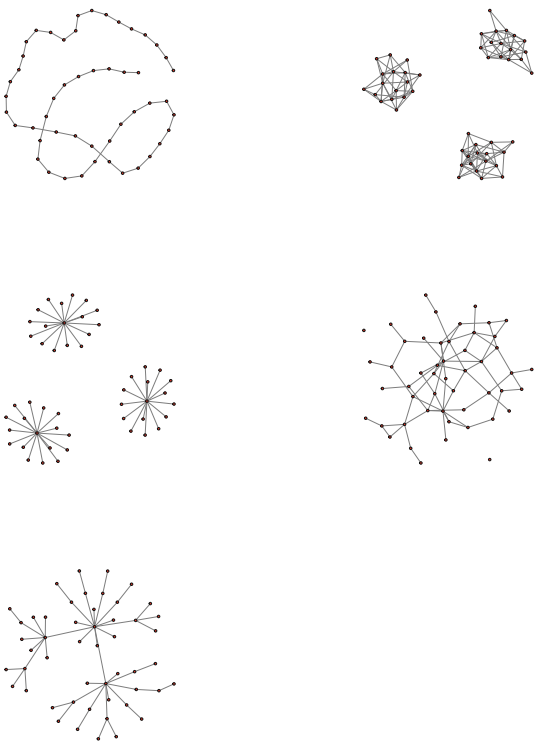

Fig. 4 Model structures from which we generate data. These graphs are described as models in this section and are named banded graph, cluster hub, random and scale free. All the graphs are sparse

Table 4 Model 1—banded graph

	glasso-stars	*minForest-aic*	*minForest-bic*
p	**PPV**		
10	88.88 (23.81)	98.38 (4.14)	96.18 (5.77)
30	67.55 (5.31)	96.72 (3.27)	94.06 (4.35)
50	56.04 (3.86)	95.69 (2.71)	93.16 (3.47)
100	40.99 (1.92)	94.23 (1.92)	91.35 (2.52)
150	32.23 (1.25)	93.42 (1.86)	90.53 (2.22)
	Sensitivity		
10	60.56 (35.57)	98.67 (3.63)	99.00 (3.20)
30	99.31 (1.70)	97.03 (3.02)	97.24 (2.90)
50	99.49 (1.10)	95.84 (2.61)	96.00 (2.44)
100	99.55 (0.62)	94.40 (1.86)	94.63 (1.79)
150	99.56 (0.50)	93.57 (1.84)	93.83 (1.70)
	MCC		
10	66.87 (27.41)	98.14 (4.77)	96.91 (5.02)
30	80.38 (3.63)	96.65 (3.32)	95.30 (3.58)
50	73.38 (2.84)	95.58 (2.76)	94.33 (2.86)
100	62.91 (1.55)	94.20 (1.92)	92.82 (2.05)
150	55.82 (1.12)	93.41 (1.87)	92.06 (1.88)

Table 5 Model 2—cluster

	glasso-stars	*minForest-aic*	*minForest-bic*
p	**PPV**		
10	90.50 (29.04)	90.98 (3.64)	96.73 (4.57)
30	74.29 (5.11)	79.99 (5.98)	77.77 (5.64)
50	67.43 (3.68)	82.37 (4.26)	78.55 (4.47)
100	52.52 (2.20)	73.99 (3.94)	71.14 (3.73)
150	43.84 (1.51)	74.20 (2.89)	71.60 (2.95)
	Sensitivity		
10	10.75 (6.90)	56.25 (12.48)	93.00 (12.23)
30	36.73 (5.72)	27.24 (2.35)	32.92 (3.72)
50	54.39 (4.02)	31.27 (1.73)	35.45 (2.52)
100	50.04 (2.31)	25.57 (1.38)	28.85 (1.63)
150	52.63 (1.98)	26.56 (1.12)	29.36 (1.40)
	MCC		
10	23.11 (10.26)	57.90 (11.32)	91.23 (12.05)
30	44.68 (4.55)	40.46 (4.19)	43.78 (4.53)
50	56.35 (3.13)	47.71 (2.86)	49.41 (3.00)
100	48.30 (1.96)	41.71 (2.41)	43.35 (2.28)
150	45.79 (1.48)	43.25 (1.83)	44.61 (1.89)

Table 6 Model 3—hub

	glasso-stars	*minForest-aic*	*minForest-bic*
p	**PPV**		
10	89.54 (16.62)	88.14 (3.84)	89.08 (8.42)
30	65.20 (6.42)	84.73 (5.74)	75.94 (6.05)
50	51.22 (3.92)	79.59 (4.61)	71.52 (4.78)
100	35.98 (2.18)	75.73 (4.36)	66.75 (3.56)
150	28.38 (1.29)	73.95 (3.42)	65.48 (2.87)
	Sensitivity		
10	74.12 (31.99)	98.88 (3.60)	99.25 (2.98)
30	90.50 (6.28)	88.57 (5.81)	90.61 (5.46)
50	91.00 (4.51)	83.64 (4.90)	85.77 (4.78)
100	92.56 (3.20)	79.60 (4.70)	82.34 (4.61)
150	92.75 (2.53)	78.09 (3.65)	80.29 (3.90)
	MCC		
10	75.89 (22.22)	91.84 (4.31)	92.53 (6.18)
30	74.85 (5.05)	85.68 (6.13)	81.63 (5.46)
50	66.66 (3.56)	80.84 (4.92)	77.36 (4.55)
100	56.53 (1.97)	77.19 (4.61)	73.57 (3.88)
150	50.36 (1.56)	75.67 (3.57)	72.11 (3.20)

Table 7 Model 4—random[spaced parenth - dt]

	glasso-stars	*minForest-aic*	*minForest-bic*
p	**PPV**		
10	92.40 (22.25)	95.61 (7.02)	90.15 (8.96)
30	68.28 (5.92)	77.75 (5.37)	79.31 (6.02)
50	62.82 (3.90)	81.79 (4.41)	79.96 (5.06)
100	48.74 (2.66)	74.87 (3.84)	71.91 (4.05)
150	36.80 (1.71)	62.68 (3.80)	60.62 (3.82)
	Sensitivity		
10	21.19 (10.09)	54.25 (4.34)	57.31 (5.69)
30	83.94 (5.82)	69.09 (5.26)	72.00 (6.35)
50	80.86 (4.76)	61.88 (3.31)	62.88 (3.41)
100	73.56 (3.68)	46.54 (2.33)	47.39 (2.44)
150	65.00 (3.17)	38.41 (2.32)	39.19 (2.28)
	MCC		
10	35.57 (11.24)	63.06 (7.41)	61.32 (6.98)
30	73.40 (4.74)	71.24 (5.64)	73.61 (5.74)
50	69.43 (3.73)	69.78 (4.00)	69.48 (4.16)
100	58.25 (2.71)	58.00 (3.06)	57.29 (3.11)
150	47.43 (2.09)	48.21 (3.02)	47.84 (2.98)

Table 8 Model 5—scale free

	glasso-stars	*minForest-aic*	*minForest-bic*
p	**PPV**		
10	92.46 (23.94)	95.51 (6.41)	88.63 (9.22)
30	68.44 (7.04)	72.87 (8.67)	62.23 (6.45)
50	48.81 (4.95)	54.98 (7.87)	47.40 (4.47)
100	27.23 (2.55)	33.57 (5.44)	30.78 (3.49)
150	17.75 (1.78)	23.86 (4.06)	22.97 (3.12)
	Sensitivity		
10	41.00 (24.96)	95.89 (6.25)	96.56 (5.63)
30	68.93 (9.54)	73.97 (8.84)	76.83 (9.18)
50	61.47 (7.60)	55.67 (8.14)	58.88 (8.98)
100	51.39 (6.03)	33.93 (5.65)	37.40 (6.51)
150	45.70 (5.99)	24.06 (4.11)	27.48 (5.00)
	MCC		
10	55.44 (22.77)	94.61 (7.79)	90.39 (8.27)
30	66.28 (7.36)	71.50 (9.36)	66.65 (7.61)
50	52.59 (5.66)	53.45 (8.33)	50.59 (6.40)
100	35.67 (3.78)	32.39 (5.65)	32.43 (4.83)
150	26.98 (3.24)	22.93 (4.14)	24.01 (4.00)

References

1. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
2. Edwards, D.: Introduction to Graphical Modelling. Springer, New York (2000)
3. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, New York (2009)
4. De Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. **9**, 67–103 (2002)
5. Meinshausen, N., Buhlmann, P.: High-dimensional graphs and variable selection with the lasso. Ann. Stat. **34**, 1436–1462 (2006)
6. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**, 19–35 (2007)
7. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Sparse inverse covariance estimation with the graphical lasso. J. Mach. Learn. Res. **9**, 485–516 (2008)
8. Friedman, J., Hastie, T., Tibshirani, R.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Biostatistics **9**, 432–441(2008)
9. DREAM8: Rheumatoid Arthritis Responder Challenge. https://www.synapse.org. (2014). Accessed 11 Feb 2014
10. Lauritzen, S.L., Wermuth, N.: Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Stat. **17**, 31–57 (1989)
11. Hoff, P.D.: Extending the rank likelihood for semiparametric copula estimation. Ann. Appl. Stat. **1**, 265–283 (2007)
12. Edwards, D., de Abreu, G., Labouriau, R.: Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinform. **11**, 18 (2010)
13. Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., Reinhardt, J.D.: Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. Comput. Stat. Data Anal. **64**, 132–152 (2013)

14. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**, 462–467 (1968)
15. Wasserman, L., Roeder, K. Liu, H.: Stability approach to regularization selection (stars) for high dimensional graphical models. Adv. Neural Inf. Process. Syst. **23**, 1432–1440 (2010)
16. Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.: Huge: High-dimensional Undirected Graph Estimation R Package. http://CRAN.R-project.org/package=huge (2014). Accessed 11 Feb 2014
17. Abreu, G.C.G., Edwards, D., Labouriau, R.: High-dimensional graphical model search with the gRapHD R package. J. Stat. Softw. **37**, 1–18 (2010)

Rounding Non-integer Weights in Bootstrapping Non-iid Samples: Actual Problem or Harmless Practice?

Federico Andreis and Fulvia Mecatti

1 Introduction and Motivation

Bootstrap method is a popular tool for numerically estimating estimators accuracy, constructing confidence intervals and determining p-values. In order to provide reliable results, feasible adaptations of the basic bootstrap algorithm are required to account for the non-iid nature of the sample data when they are collected from finite populations by means of without-replacement complex sample designs. Recent literature suggests the use of a weighting system in the resampling and/or the estimation procedure [1, 3, 12]. Weights can be familiar sample weights such as the inverse of unit inclusion probabilities as well as Hajek type or depending on particular probability distribution, normally defined as ratios. Integer weights would guarantee analytical properties of both the bootstrap procedure and of the final bootstrap estimates (see Sect. 2 for details). However, this is seldom the case in the applications. As a consequence integer-valued weights are an *ideal* conceptual situation rarely, if not never, occurring in practice. Two suggestions recur in the literature to deal with non-integer weights: (1) randomization and (2) systematical rounding. Randomization would require a further step added on top of the boostrap algorithm, thus affecting its computational efficiency. This step can be avoided by rounding non-integer weights to the nearest integer, according to some systematical

F. Andreis (✉)
DEMM - Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano, Via Conservatorio 7, 20100 Milano, Italy
e-mail: federico.andreis@unimi.it

F. Mecatti
Dipartimento di Sociologia e Ricerca Sociale, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, U7, 20126 Milano, Italy
e-mail: fulvia.mecatti@unimib.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_2

rule [2, 6]. Both solutions affect the Bootstrap process as well as the final boostrap estimates to an unknown extent; moreover, they violate basic Bootstrap principles such as the *mimicking* principle and the *plug-in* approach as it will be explained in Sect. 2 (see also [9] for an extended discussion of Bootstrap principles).

In this work we concentrate on rounding for dealing with non-integer weights as the computationally preferable option, our main aim being to produce empirical evidence of the extent of its effects. As a first investigation of the magnitude of the rounding effect on the Bootstrap process, we have focused on the estimation of the variance of the Horvitz-Thompson (HT) estimator for the mean of a study variable through an extended simulation exercise.

In Sect. 2 we describe the Bootstrap algorithm we have focused on, applying to samples from a finite population under a general without replacement design, and discuss the two most used rounding methods suggested in the literature in order to obtain an integer weights system. In Sect. 3 we outline the design of the simulation study, with a detailed description of the computational and numerical expedients implemented in order to simulate comparable scenarios. Particular attention has been given to the construction of *ideal*, though scarcely realistic, integer weights scenarios to be used as a benchmark for comparison with more realistic non-integer weights scenario where some sort of rounding would be unavoidable. Results from the simulation study are presented in Sect. 4 and discussed in Sect. 5. Section "Conclusions" concludes the work with some final remarks.

2 Bootstrapping from Non-iid Sample Data

We focus on samples of fixed size n selected without replacement from a finite population $U = \{1, \ldots, k, \ldots, N\}$ under a general random design where each population unit is assigned a specific probability π_k to be included in the sample. We adopt the popular approach based on weighting each sampled unit by the inverse of its own inclusion probability π_k^{-1} in order to produce an empirical population U^* where to perform resampling [10]. The pseudo-population U^*, usually named the *Bootstrap population* [4, 5, 8], is intended to mimick the *parent population* U according to fundamental Bootstrap principles such as the mimicking and the plug-in [9]. According to the same principles, the resampling plan should mimick the original sampling design. As a consequence, the creation of U^*, the resampling and the final estimates produced by the Bootstrap process depend on the weights π_k^{-1}, whether integer or to be rounded.

2.1 Bootstrap Algorithm

Let s denotes the original sample, fixed at its observed values. In the following, the popular star-notation * will be used to denote Bootstrap objects and quantities.

A non-iid Bootstrap algorithm, according to the Bootstrap population approach and basic mimicking principle, is composed by the following four steps:

1. *Bootstrap Population step*: construct U^* by replicating (a chosen number) d_k times each sampled unit $k \in s$;
2. *Resampling step*: select a Bootstrap sample s^* from U^* under a resampling plan mimicking the original design and with the same size $n^* = n$;
3. *Replication step*: on the Bootstrap sample s^*, compute a replicate of the original estimate computed on the original sample s, i.e. if $t = t(s)$ is the study estimate, thus $t^* = t(s^*)$ defines its Bootstrap replication;
4. *Bootstrap Distribution step*: iterate steps 2. and 3. a number B of times, chosen sufficiently large, resulting in the Bootstrap distribution $t_1^*, \ldots, t_b^*, \ldots, t_B^*$.

Owing to the mimicking principle, the Bootstrap distribution is assumed as a Monte Carlo simulation of the original estimator distribution, to be used for variance estimation and for producing p-values and confidence intervals.

Still according to the mimicking principle, a natural choice at step 1. for the frequency d_k of each sampled unit $k \in s$ in U^*, is $d_k = \pi_k^{-1}$ [10], although such a choice is unlikely to produce integers number, as discussed in the following subsection.

The actual construction of the set U^* at step 1. can be avoided, with significant computational advantages, by generating n values from a suitable probability distribution [3]. This method is based on the proven equivalence between the above bootstrapping scheme and weighting the units $k \in s$ by B random draws from a suitable probability function depending on the original sampling plan to be mimicked by the Bootstrap process [12]. We then exploit this technique for our simulative exercise.

2.2 Rounding Problem

As pointed out above, the weights d_k are usually not integer: for this to happen, in fact, $\pi_k \in \mathbf{Q}^+$shoud hold $\forall k \in s$. As a consequence, in most practical cases, some device to recover integer weights $\tilde{d}_k$ substituting $d_k = \pi_k^{-1}$ needs to be adopted in order to actually implement the Bootstrap algorithm previously outlined. Notice that this would be needed as well for implementing any of the non-iid bootstrap algorithms based on a weighting system such as the ones cited in the Introduction. The main methodologies suggested in the literature are based either on randomization or systematic rounding. The first option might compromise the computational efficiency of the entire process since it requires a further randomization step on top of the Bootstrap algorithm to perform n random trials in order to select a working Bootstrap population in a class of possible 2^n candidates [2]. In this paper two popular devices to obtain integer weights have been considered:

1. by means of nearest integer rule [2, 6];
2. by means of a Binomial distribution [5, 10].

The first approach is a systematic rule, whereas the second is a randomization procedure (thus, computationally more demanding).

Nearest integer rounding (1) is the simplest of the two methods: the weights are rounded according to the rule $\tilde{d}_k = \lfloor d_k + 0.5 \rfloor$, where $\lfloor \quad \rfloor$ denotes the *floor* function, i.e. $\lfloor a \rfloor = \max\{b \in \mathbf{N} : b \leq a\}$; this method is computationally efficient since it does not require any additional randomization step on top of the Bootstrap algorithm.

Rounding by means of a Binomial distribution (2) requires to obtain a realisation of a Bernoulli random variable for each $k \in s$ and to adjust the weights according to the rule

$$\tilde{d}_k = \begin{cases} \lfloor d_k \rfloor & \text{with probability } 1 - (d_k - \lfloor d_k \rfloor) \\ \lfloor d_k \rfloor + 1 & \text{with probability } d_k - \lfloor d_k \rfloor. \end{cases}$$

The special case of a symmetric Binomial distribution has also been considered, where each probability is kept constant and equal to 0.5. This approach (that can be traced back to the first proposal by Chao and Lo [5]) seemed to be the most natural (and simple) when applying a randomization procedure, and for this reason we have included it in our simulation. However, this special case will not be further discussed and no specific results regarding it will be shown in Sect. 4, for, besides its simplicity, no particularly interesting conclusions could be derived from it.

2.3 Sampling Schemes

Two largely applied selection schemes have been considered, both with fixed size and without replacement (see for instance [1] for details):

1. Simple Random Sampling (SRS) with constant inclusion probability $\pi_k = n/N$ for every population unit $k = 1, \ldots, N$;
2. Conditional Poisson Sampling (CPS), also known as Maximum Entropy, with unequal inclusion probability (exactly) proportional to a positive auxiliary variable x assumed totally known, i.e. $\pi_k = nx_k/X$ where $X = \sum_{k=1}^{N} x_k$ is the population auxiliary total. In a typical practical example the auxiliary variable x represents a measure of size of population units so that this kind of sampling is usually referred to as (inclusion) probability-proportional-to-size (πPS). It is well known that πPS designs might be significantly more efficient than equal probability sampling as the relation between the study and the auxiliary variables approaches proportionality. CPS is an easy to implement, sequential method for selecting πPS samples of fixed size.

Notice that, for the case of SRS, rounding affects but a single value, i.e. N/n which is the same for all units in U, and a unique control total, the Bootstrap population size $N^* = \sum_{k \in s} \tilde{d}_k$, generally different from the actual population

size N for non integer $d_k = N/n$. In the case of CPS the effect might be more severe, since the need for rounding may occur for multiple values, namely from 1 to n Bootstrap weights $\tilde{d}_k$. Moreover, in this case rounding would affect two control totals, both the Bootstrap population size N^* and the total of the auxiliary variable X^* which can severely depart from the corresponding actual population counts N and X, thus compromising the aim for U^* of mimicking U.

In order to build toy examples in both cases, numerical methods have been developed to find suitable sets of inclusion probabilities leading to integer weights, to be used as ideal situations in which the rounding problem (detailed in Sect. 2.2) does not arise. These toy examples will constitute the benchmark cases in the simulation study, for comparisons with the realistic practical cases of rounded weights. The methods we developed to produce them are outlined in Sect. 3.2.

3 Simulation Design

Desirable estimators' properties conceptually granted by the Bootstrap procedure are influenced by the rounding method employed to obtain integer weights. In order to evaluate the extent to which Bootstrap-based inference is affected and how, an extended simulation exercise is set up: a population variable y is generated from a Gamma distribution, allowing to investigate various scenarios, in which the level of variability of y (as measured by its coefficient of variation cv_y) varies and different departures from the ideal situation of integer weights (where no rounding is needed) are considered. These departures are induced by increasingly stronger perturbations on the population set of inclusion probabilities π_k. This would result in a negligible alteration of the population size in the SRS or in an additive noise randomly distributed over the set $\{\pi_k, k = 1, \ldots, N)$ in the CPS.

As a first, simple attempt at investigating the impact of each of the two rounding methods presented in Sect. 2.2, we focus on the Horvitz-Thompson estimator for the mean of y, denoted $\bar{y}_{HT}$. A set of Monte Carlo indicators is provided for each simulated scenario with the purpose of investigating the rounding problem under the following respects:

1. the mimicking principle: by evaluating distances between the nominal and the post-rounding characteristics of the Bootstrap algorithm, particularly on known population totals and size as compared to Bootstrap populations counts;
2. basic Bootstrap algorithm properties: particularly the *Bootstrap unbiasedness* [2] as measured by Monte Carlo expectation over $\bar{y}_{HT}$ estimates computed on both the original sample and the collection of Bootstrap samples;
3. inferential properties of the final Bootstrap estimates for the variance of the HT estimator, such as biasedness and stability as measured by Monte Carlo relative bias (RB) and relative root mean square error (RRMSE).

The following Monte Carlo measures, pertaining to each of the previous four points, are computed:

1. *population totals percentage relative bias (RB)*

$$\begin{aligned} \%RB_N &:= \frac{E[E^*(N^*)-N]}{N}100 \\ \%RB_X &:= \frac{E[E^*(X^*)-X]}{X}100; \end{aligned} \tag{1}$$

2. *Bootstrap unbiasedness*

$$BIAS_{boot} := E[E^*(\bar{y}^*_{HT}) - \bar{y}_{HT}]; \tag{2}$$

3. *percentage relative bias and root mean squared error (RRMSE) of the final variance Bootstrap estimate*

$$\begin{aligned} \%RB_V &:= \frac{E[V^*(\bar{y}^*_{HT})-V(\bar{y}_{HT})]}{V(\bar{y}_{HT})}100 \\ \%RRMSE_V &:= \sqrt{\frac{E\{[V^*(\bar{y}^*_{HT})-V(\bar{y}_{HT})]^2\}}{V(\bar{y}_{HT})}}100 \end{aligned} \tag{3}$$

where $N^* = \sum_{k\in s} \tilde{d}_k$, $X^* = \sum_{k\in s} \tilde{d}_k x_k$ and the expected values E are taken over the simulation runs, while E^* over the Bootstrap replications; $V(\bar{y}_{HT})$ denotes the estimator variance (as to be estimated via Bootstrap).

All the computations are carried out in *R* 3.0.2 [11]; the packages *sampling* and *BiasedUrn* [7] are respectively employed to select unequal-probability samples and to generate values from a non-standard distribution involved in the CPS case, the Fisher's Noncentral Multi-hypergeometric distribution (FNMHyg). Indeed, as shown in Ranalli and Mecatti [12], resampling from a CPS design is equivalent to resampling directly from the sample s by weighting each observation by random variates from a FNMHyg distribution (with proper choices of the parameters), thus avoiding the actual construction of the Bootstrap population U^*, which yields a sensitive computational benefit.

3.1 Simulated Scenarios

In order to simulate an assorted range of realistic cases, an asymmetric, positive study variable y, following a Gamma distribution with shape parameter $a > 0$ and scale parameter $\theta > 0$ is chosen as population variable. It follows that $E(y) = a\theta$, $Var(y) = a\theta^2$ and $cv_Y = \sqrt{a\theta^2}/a\theta = 1/\sqrt{a}$. The mean of y, as the population parameter to be estimated, is set to a constant value across all scenarios, specifically, $E(y) = 10$, the population size is $N = 600$ and the total X of the auxiliary variable is set equal to 1,000.

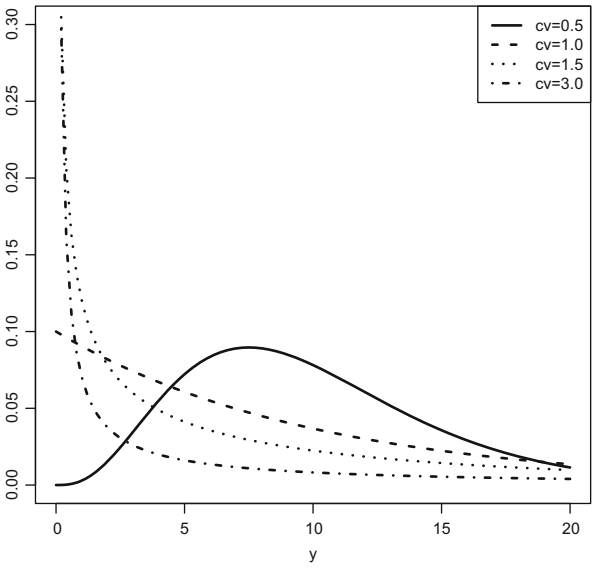

Fig. 1 Gamma densities used in the simulation

Table 1 Simulation parameters and scenarios

	SRS	CPS
n	{15, 30}	{15, 30}
cv_y	{0.5, 1.0, 1.5, 3.0}	{0.5, 1.0, 1.5, 3.0}
Departures	6	4

Figure 1 depicts the Gamma densities with fixed mean $E(y) = 10$, resulting at the different cv_y levels in Table 1, used in the simulation.

For both SRS and CPS, each scenario is characterized by the three simulation parameters:

1. sample size (n);
2. population variability (cv_y);
3. level of departure from the ideal integer-weights case, as simulated by the suitable toy example

as summarized in Table 1.

In order to investigate point 1, two sample sizes have been chosen, resulting in sampling fractions of exactly 5 % and 2.5 % for CPS and approximately so for SRS, where they are kept within a tight range, specifically 5 ± 0.2 % and 2.5 ± 0.1 %. Notice that larger sample sizes were not included due to computational limitations of the packages employed to implement the Bootstrap procedure. Specifically, the *BiasedUrn* package imposes a working maximum of $n = 32$ in order to retain both a good level of accuracy in random variates generation and an acceptable execution time. Notice that the core routines for generating from the FNMHyg distribution are written in C and appear to be reliable, so this seems not be a language-dependent

issue, rather more general, due to the computational burden of dealing with such a complex distribution. Possible workarounds, to overcome the $n = 32$ limitation, are currently under consideration.

Regarding point 2, suitable values of the parameters of the Gamma distribution are selected so to obtain a range of cv_y leading from low to high population variability, in order to investigate the performance of the bootstrapping procedure under different variability situations. This is accomplished by setting $a = cv_y^{-2}, \theta = cv_Y^2 E(y)$.

Finally, for what concerns point 3, departures from the ideal integer-weights case are obtained inducing an increasing level of perturbation on the inclusion probabilities $\pi_k, k \in U$. For SRS, this is achieved by means of small systematic modifications to the population size N, leading to the slight departure from the stated sample fractions described above, together with suitable element-wise deletions from U in order to retain the chosen population mean and (relative) variability dictated by each scenario. The aim of this perturbation is to force the constant inclusion probabilities to have inverses (the d_k) with decimal places either below (0.33), above (0.67) or exactly equal to 0.5, so to provide different situations where the two rounding methods might, at least theoretically, perform differently. A total of six modifications to N (three upwards and three downwards) are considered to this end for each sample size: $\{\pm 5, \pm 8, \pm 10\}$ when $n = 15$, and $\{\pm 10, \pm 15, \pm 20\}$ when $n = 30$.

For what concerns CPS, an additive random noise on an increasing fraction of the π_ks is introduced. This procedure can be summarized as follows:

1. set the fraction p of inclusion probabilities to be perturbated (in this simulation study, $p \in \{25\,\%, 50\,\%, 75\,\%, 100\,\%\}$);
2. randomly extract a fraction p of the population units;
3. perturbate the extracted units' inclusion probabilities by adding to half of them (randomly chosen) a fixed term δ while subtracting it from the remaining half. We choose δ so to ensure that the new (perturbated) inclusion probabilities $\tilde{\pi}_k$ lie in $(0, 1)$; moreover, adding and subtracting the same quantity grants that $\sum_{k \in U} \tilde{\pi}_k = n = \sum_{k \in U} \pi_k$.

In this way we obtain a new set of population inclusion probabilities that leads to non-integer weights and differs from the ideal integer-weights case in a way that can be measured by means of any distance (e.g., the Euclidean distance) calculated between the vector of the original π_k and the vector of the perturbated $\tilde{\pi}_k$; any such distance will be increasing in p. The population values $\{x_k, k = 1, \ldots, N\}$ of the auxiliary variable are then re-computed on the basis of the perturbed inclusion probabilities $\tilde{\pi}_k$ in order to retain the population total X.

This simulation settings yield a total of 48 scenarios for SRS and 32 for CPS, intended to explore the effect of rounding practice by varying sample size, population variability and extent of rounding; 45,000 simulation runs for each scenario, with each bootstrapping step being replicated $B = 2,500$ times have been performed. The Monte Carlo error on the mean and variance of $\bar{y}_{HT}$, are kept, respectively, under 0.5 % and 3.0 %, as measured against the known true values.

3.2 Toy Examples Probabilities

In order to construct the ideal situation of integer-weights for both simulated sampling designs SRS and CPS, we need to obtain suitable inclusion probabilities, given the sample size n and the population size N, that is, we are looking for some probabilities q_k such that $q_k^{-1} \in \mathbf{N}^+, \forall k$ and $\sum_{k=1}^{N} q_k = n$. We describe two possible solutions that are suitable to our purpose.

I $q_k = q, \forall k$

This assumption allows to immediately obtain the inclusion probabilities for the SRS toy example, upon setting $\pi_k = q = \frac{n}{N}$ and under the constraint that $N \bmod n = 0$.

II $q_k \in \{q_1, q_2\}$, i.e., two distinct values are allowed
$\sum_{k=1}^{N} q_k = \sum_{i=1}^{N_1} q_1 + \sum_{i=1}^{N_2} q_2 = N_1 q_1 + N_2 q_2$.

Setting this quantity equal to n and requiring that $N_1 + N_2 = N$ and $q_i^{-1}, N_i \in \mathbf{N}^+, i = 1, 2$, we obtain the following system:

$$\begin{cases} N_1 q_1 + N_2 q_2 = n \\ N_1 + N_2 = N. \end{cases}$$

To find q_1, q_2, N_1, N_2 that respect the aforementioned constraints is a problem that does not admit an analytical solution: numerical methods are therefore required. One might decide, given N and n, to solve the system for N_1 and N_2,

$$\begin{cases} N_1 = \frac{Nq_2 - n}{q_2 - q_1} \\ N_2 = \frac{n - Nq_1}{q_2 - q_1} \end{cases}$$

and then check, by *brute force*, which pairs of values among a (finite) set of proposals $\{\frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{M}\} \in \mathbf{Q}^+$ constitute a suitable choice for $\{q_1, q_2\}$, i.e. yield integer N_1, N_2 (for simplicity, we willingly choose not to let any of the k units have inclusion probability equal to 1). This method can yield multiple acceptable pairs to be used to build a toy example for CPS; more than one pair can be employed, collectively, in order to obtain a scenario with more than two distinct inclusion probabilities. In fact, we employ this procedure to obtain four different pairs, which we use in the simulation scenarios for CPS, leading to a total of eight distinct inclusion probabilities that satisfy the requirement of having integer inverses and lead to the desired sample size n.

Table 2 summarizes the values of N_i and π_i, obtained with the method described above, that have been used in the simulation study when $n = 30$ (for the case $n = 15$, the same probabilities, divided by two, have been employed). Clearly, for SRS we have $i = 1$, i.e. all the inclusion probabilities are equal, while for CPS we selected, as already said, eight distinct values, thus we have $i = 1, \ldots, 8$.

Table 2 Inclusion probabilities for the toy examples

	SRS	CPS
N_i	600	75
π_i	1/20	{1/11,1/12,1/14,1/15, 1/30,1/35,1/60,1/110}

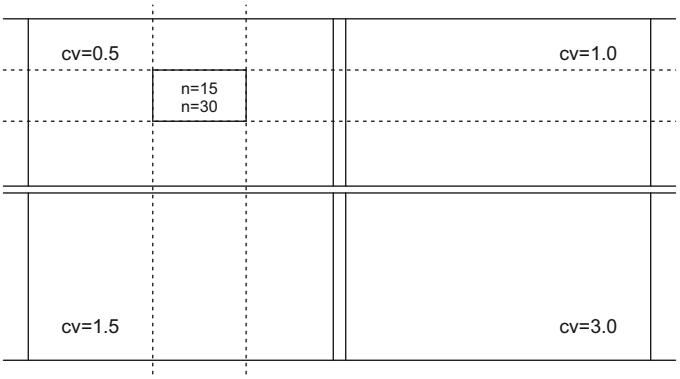

Fig. 2 How to read the tables of simulation results

4 Empirical Results

This section contains the results of the simulation study with respect to the Monte Carlo quantities of interest defined in Sect. 3. The results are presented separately for SRS (Tables 3 and 4 in Sect. 4.1) and CPS (Tables 5 and 6 in Sect. 4.2). Within each case two tables are provided, that collect the outcomes for systematic (Tables 3 and 5) and randomization-based rounding methods (Tables 4 and 6), as presented in Sect. 2.2. The tables are structured to contain results from all the scenarios as follows: each main quarter collects the results for a different level of population variability, while each cell contains two rows, one for each considered sample size, as shown in Fig. 2.

Each row header labels the specific quantity reported, while each column header indicates the level of departure from the reference situation, i.e. the toy examples with integer weights, as described in Sect. 3.1.

4.1 Equal-Probability Sampling (SRS)

Tables 3 and 4 contain the results of the simulation study for, respectively, systematic rounding and randomization-based (binomial) rounding under the SRS design. The column headers report the decimal places of the weights d_k in the integer case (0.00) and for three out of the six perturbations we have considered

Table 3 SRS—systematic rounding

$d_k - \lfloor d_k \rfloor$	0.00	0.33	0.50	0.67	0.00	0.33	0.50	0.67
$\%RB_N$	0	−0.83	1.15	0.82	0	−0.83	1.15	0.82
	0	−1.64	−2.44	1.61	0	−1.64	−2.44	1.61
$BIAS_{boot}$	0	−0.08	0.11	0.08	0	−0.08	0.11	0.08
	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
$\%RB_V$	−6.75	−8.36	−4.48	−4.72	−6.61	−8.48	−4.46	−4.74
	−3.27	−3.53	−3.19	−2.74	−2.82	−3.35	−3.32	−3.09
$\%RRMSE_V$	56.23	55.61	57.35	57.15	131.44	128.63	134.36	133.52
	27.93	27.80	28.05	28.35	65.93	65.60	65.77	66.50
$\%RB_N$	0	−0.83	1.15	0.82	0	−0.83	1.15	0.82
	0	−1.64	−2.44	1.61	0	−1.64	−2.44	1.61
$BIAS_{boot}$	0	−0.08	0.11	0.08	0	−0.09	0.12	0.08
	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
$\%RB_V$	−6.83	−8.20	−3.72	−5.02	−6.87	−7.5	−3.42	−4.36
	−2.95	−3.06	−3.15	−3.14	−3.35	−3.16	−2.90	−2.51
$\%RRMSE_V$	273.72	271.64	281.85	279.99	1,104.56	1,093.16	1,134.93	1,132.11
	137.68	138.04	138.46	138.40	554.33	561.11	560.21	559.98

Table 4 SRS—randomization-based rounding

$d_k - \lfloor d_k \rfloor$	0.00	0.33	0.50	0.67	0.00	0.33	0.50	0.67
$\%RB_N$	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
$BIAS_{boot}$	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	0.01
	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
$\%RB_V$	−6.75	−6.72	−6.56	−6.20	−6.61	−6.85	−6.57	−6.22
	−3.27	−3.09	−2.64	−2.46	−2.82	−3.04	−3.03	−2.95
$\%RRMSE_V$	56.23	56.29	56.69	56.58	131.44	130.45	132.10	131.92
	27.93	27.96	28.29	28.49	65.93	66.06	66.08	66.79
$\%RB_N$	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
$BIAS_{boot}$	0	<0.01	<0.01	<0.01	0	<0.01	<0.01	<0.01
	0	<0.01	<0.01	<0.01	0	0.01	<0.01	<0.01
$\%RB_V$	−6.83	−6.62	−5.89	−6.64	−6.87	−5.84	−5.54	−5.94
	−2.95	−2.79	−2.85	−3.01	−3.35	−2.95	−2.60	−2.51
$\%RRMSE_V$	273.72	275.99	276.03	275.77	1,104.56	1,113.50	1,112.43	1,114.28
	137.68	138.72	139.39	139.14	554.33	564.03	564.03	561.24

(0.33, 0.50, 0.67), as described in Sect. 3.1; specifically, we show only the upward deviations from $N = 600$, since no difference in absolute value has been observed with respect to the quantities of interest when considering the downward ones.

Table 5 CPS–systematic rounding

perturbation	25 %	50 %	75 %	100 %	25 %	50 %	75 %	100 %
$\%RB_N$	0.01	0.40	−0.46	0.36	0.09	0.11	0.44	0.42
	−0.13	−0.30	−0.36	−0.26	0.04	−0.10	0.08	−0.04
$\%RB_X$	0.16	0.36	0.60	0.58	0.17	0.36	0.46	0.65
	−0.08	−0.12	−0.18	−0.32	−0.07	−0.11	−0.13	−0.17
$BIAS_{boot}$	0.12	0.07	0.06	0.10	0.10	0.14	0.08	0.08
	0.16	0.18	0.19	0.21	0.22	0.19	0.25	0.19
$\%RB_V$	1.15	1.21	1.22	1.30	1.11	1.35	1.26	1.32
	1.57	1.83	1.96	1.84	1.18	1.68	1.90	1.85
$\%RRMSE_V$	1.23	1.34	1.39	1.58	1.53	2.28	1.94	2.27
	1.87	2.66	3.04	2.88	1.58	2.98	3.67	3.59
$\%RB_N$	0.23	−0.02	0.31	0.35	0.27	0.06	0.38	0.53
	−0.14	0.05	−0.29	−0.11	−0.11	−0.14	−0.24	−0.23
$\%RB_X$	0.14	0.36	0.49	0.71	0.19	0.29	0.41	0.65
	−0.03	−0.20	−0.17	−0.30	−0.03	−0.20	−0.21	−0.19
$BIAS_{boot}$	0.15	0.18	0.13	0.24	0.23	0.12	0.13	0.18
	0.16	0.05	0.12	0.30	0.32	<0.01	0.37	0.24
$\%RB_V$	1.17	1.33	1.22	1.08	1.38	1.58	1.47	1.44
	1.58	1.92	1.94	1.65	1.86	2.16	1.65	1.77
$\%RRMSE_V$	2.29	3.02	2.28	2.04	3.73	7.92	5.48	4.69
	3.24	6.08	4.99	3.45	7.37	13.93	5.38	6.45

4.2 Unequal-Probability Sampling (CPS)

Tables 5 and 6 contain the results of the simulation study for, respectively, systematic rounding and randomization-based (binomial) rounding under the CPS design. The column headers report the extent of the perturbation induced on the inclusion probabilities, from 25 % to 100 %, as described in Sect. 3.1. Both Tables 5 and 6 do not report results concerning the benchmark case of all integer weights, i.e., $\tilde{d}_k \equiv d_k = \pi_k^{-1}, \forall k \in s$, which are in fact redundant since (1) the bootstrapping algorithm outlined in Sect. 2.1 grants that $X^* = X$ holds when the weights d_k are integers, hence always yielding $\%RB_X = 0$, as expected; and (2) the quantities $\%RB_N$ and $BIAS_{boot}$ [defined, respectively, in Eqs. (1) and (2)] resulted negligible under every scenario. Moreover, unlike for Tables 3 and 4 in the previous subsection, in both Tables 5 and 6 the results concerning relative bias ($\%RB_V$) and relative root mean square error ($\%RRMSE_V$) of the final Bootstrap estimate of $V(\bar{y}_{HT})$, are now expressed as ratios between the perturbated scenario and the reference case, showing in this way the relative trend of the rounding effects as the departure from the ideal integer-case increases.

Table 6 CPS—randomization-based rounding

perturbation	25 %	50 %	75 %	100 %	25 %	50 %	75 %	100 %
$\%RB_N$	−0.05	0.15	0.24	0.15	0.09	−0.18	0.43	0.15
	0.28	0.54	0.91	1.20	0.67	0.33	1.00	1.33
$\%RB_X$	0.07	−0.03	0.23	0.3	0.16	−0.07	0.41	0.27
	0.52	1.07	1.61	1.8	0.78	0.47	1.14	1.8
$BIAS_{boot}$	0.12	0.09	0.07	0.10	0.12	0.13	0.09	0.10
	0.15	0.18	0.15	0.14	0.22	0.21	0.23	0.15
$\%RB_V$	1.15	1.20	1.22	1.30	1.11	1.35	1.26	1.31
	1.58	1.84	1.97	1.85	1.18	1.68	1.90	1.85
$\%RRMSE_V$	1.23	1.33	1.38	1.58	1.53	2.28	1.94	2.27
	1.88	2.67	3.05	2.90	1.58	2.98	3.67	3.60
$\%RB_N$	−0.02	−0.16	0.12	0.02	0.29	−0.06	0.18	0.23
	0.22	0.84	0.36	1.51	−0.07	0.71	1.28	1.06
$\%RB_X$	−0.2	0.11	0.22	0.21	0.19	0.13	0.12	0.24
	0.51	0.96	0.7	1.96	−0.02	0.61	1.27	1.63
$BIAS_{boot}$	0.16	0.19	0.16	0.26	0.19	0.10	0.12	0.21
	0.16	0.02	0.08	0.30	0.31	−0.03	0.40	0.26
$\%RB_V$	1.17	1.33	1.22	1.08	1.38	1.58	1.47	1.44
	1.58	1.93	1.95	1.65	1.87	2.16	1.64	1.77
$\%RRMSE_V$	2.29	3.02	2.27	2.04	3.73	7.92	5.48	4.69
	3.24	6.08	5.01	3.45	7.38	13.93	5.36	6.44

5 Discussion

In what follows, we discuss the most interesting conclusions that could be derived from the simulation study described in Sect. 3, whose results were presented in Tables 3, 4, 5, and 6. Once again, we separate the two sampling designs with equal and unequal inclusion probabilities (respectively, SRS and CPS, defined in Sect. 2.3), in order to better highlight the most significant differences we found in their respect regarding the Monte Carlo quantities defined in Eqs. (1)–(3).

SRS—Tables 3 and 4

- $\%RB_N$
  - **integer-weights case:** as expected, no bias was detected on the estimation of N by N^*;
  - **systematic rounding:** (small) bias was found, that worsen proportionally with sample size. This makes sense due to the nature of the sampling design, since it shows that more observations lead to a propagation of error on the estimation

of N by N^*, irrespective of population variability (all inclusion probabilities are equal). The worst bias has been observed with decimal places 0.50 (which constitutes, in this case, the maximum extent of rounding);
  - **randomization-based rounding:** the bias on estimation of N by N^* was found to be negligible, irrespective of population variability, sample size and extent of rounding. Randomization seems, then, to grant an unbiased estimation of the population size N. This effect might be, however, expected to increase with higher sample sizes.

- $BIAS_{boot}$
  - **integer-weights case:** as expected according to theory, Bootstrap unbiasedness is granted;
  - **systematic rounding:** negligible bias was found, that reduces even more with sample size. Bootstrap unbiasedness seems to be attained, regardless of population variability and extent of rounding;
  - **randomization-based rounding:** randomization helps attaining Bootstrap unbiasedness even faster than systematic rounding. $BIAS_{boot}$ appears to be negligible irrespective of sample size, population variability and extent of rounding; such results is in line with the theory (see, e.g., [5]).

- $\%RB_V$
  - **integer-weights case:** a moderate level of relative bias of the Bootstrap variance estimator is present, regardless of population variability; this bias has been observed to be always negative, i.e. leading to an under-estimation of the quantity of interest. It would appear, however, that some improvement is observed proportionally to sample size;
  - **systematic rounding:** we observe an overall moderate $\%RB_V$ that improves when sample size increases, regardless of population variability and extent of rounding, aligning with the results of the reference case;
  - **randomization-based rounding:** the results are in line with those obtained in the systematic rounding case.

- $\%RRMSE_V$
  - **integer-weights case:** $\%RRMSE_V$ strongly increases, as might be expected, with population variability, and decreases proportionally with sample size;
  - **systematic rounding:** the results are comparable to those obtained in the reference case;
  - **randomization-based rounding:** the results are comparable to those obtained in the reference case.

Overall, there seems to be no relevant effect of rounding when dealing with SRS, neither with respect to rounding method (systematic vs randomization) nor to the extent of rounding itself ($d_k - \lfloor d_k \rfloor$ assigned to $0.33, 0.50$ or 0.67). The only significant difference seems to concern the mimicking of population size N by N^*, for which the systematic approach yields a small bias; this, however, does not

seem to affect the other quantities relevant to rounding. Systematic rounding would therefore seem to constitute the preferable option, due to the observed equivalence with the randomization-based approach in terms of simulation results, and being computationally less demanding.

In conclusion, the simulation study shows a negligible rounding effect for SRS, where a single weight $d_k = \pi_k^{-1} = N/n$ and a single population count $N = \sum_{k \in s} d_k$ are concerned.

CPS—Tables 5 and 6

- $\%RB_N$
  - **systematic rounding:** an overall negligible bias in estimating N by N^* was found, irrespective of population variability, sample size and extent of perturbation. The worsening effect for increasing sample size observed in SRS does not appear here, maybe due to the random nature of the perturbation, possibly compensating for a systematic propagation effect induced by the rounding;
  - **randomization-based rounding:** bias, if overall negligible, seems to exhibit some sort of dependence on sample size (analogous to the one observed with SRS) and extent of perturbation; it appears that the higher the amount of rounding needed, the higher the bias in estimation of N becomes (still, always under a maximum in absolute value of 1.51 %).
- $\%RB_X$
  - **systematic rounding:** we observe an overall negligible bias, irrespective of population variability, sample size and extent of perturbation. There seems, however, to be some sort of improvement proportional to sample size on the estimation of the population total X by X^*;
  - **randomization-based rounding:** we find overall small bias, irrespective of population variability. Larger sample sizes tend to yield greater bias, which might be due to some rounding error propagation effect. Moreover, it would appear that bias in estimation of X is proportionally related to extent of rounding, meaning that it appears to lead to worse departures from the reference situation of integer-weights (to a maximum percentage relative bias of about 2 %).
- $BIAS_{boot}$
  - **systematic rounding:** the bias shows to be negligible under all scenarios, irrespective of population variability and extent of rounding. It does seem, however, to grow with sample size when variability is low ($cv_y \in \{0.5, 1.0\}$);
  - **randomization-based rounding:** the randomization procedure leads to conclusions analogous to those for the systematic rounding: we observe overall

negligible bias regardless of population variability and extent of rounding. It seems, again, to be growing with sample size when variability is low.

- $\%RB_V$
  - As observed in the SRS case, $\%RB_V$ grows steadily with population variability and has negative sign, indicating a persistent under-estimation of the true variance of $\bar{y}_{HT}$;
  - **systematic rounding:** the ratio of $\%RB_V$ for systematic rounding over $\%RB_V$ for the reference situation is consistently higher than one, suggesting a significant increase in relative bias of variance estimation when rounding is needed: the effect is particularly evident with respect to the larger sample size considered ($n = 30$), where all the ratios are higher, leading to double bias in one scenario. Figure 3 visually depicts this behaviour: on the x-axes we represent the extent of perturbation (and, thus, of the needed rounding), while on the y-axes $\%RB_V^{systematic}/\%RB_V^{reference}$ is reported;
  - **randomization-based rounding:** with respect to $\%RB_V$, randomization performs almost identically as compared to the systematic rounding under every

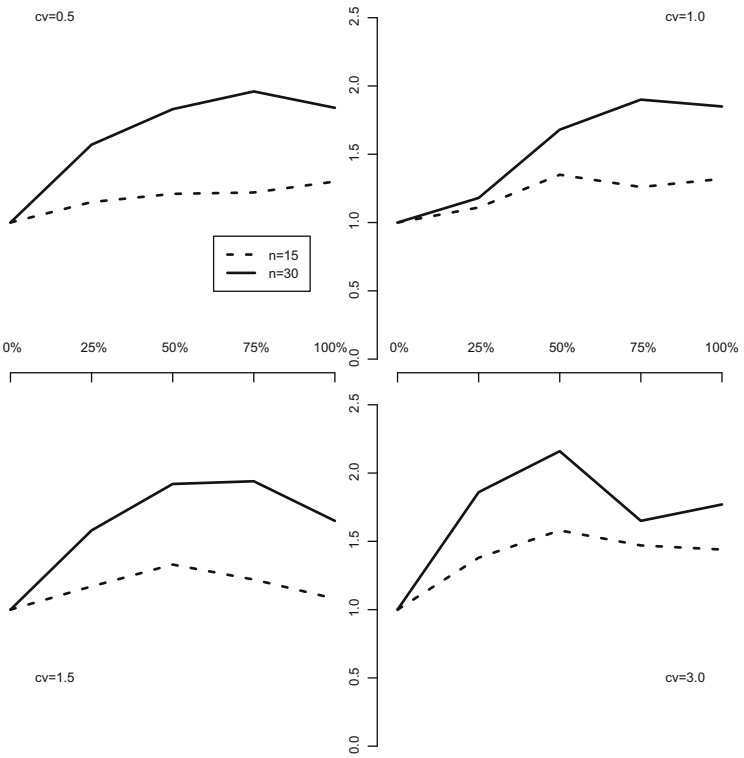

Fig. 3 Systematic rounding vs no rounding $\%RB_V$ ratios—perturbation level on the x-axes

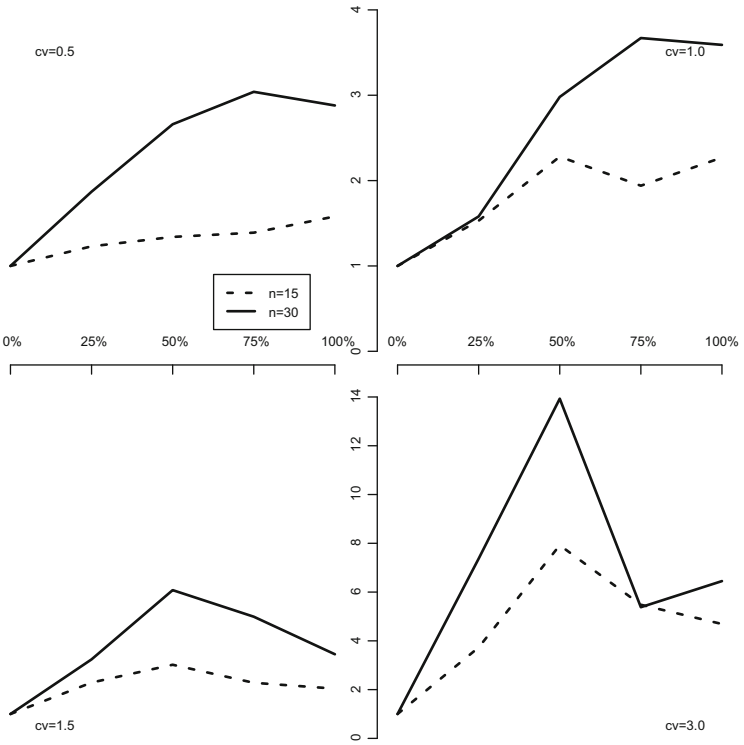

Fig. 4 Systematic rounding vs no rounding $\%RRMSE_V$ ratios—perturbation level on the x-axes

scenario. Notice that this is in agreement with other simulation studies [2, 6] thus confirming that randomization can be avoided in favour of a computationally more efficient systematic rounding.

- $\%RRMSE_V$
  - Again, as observed in the SRS case, $\%RRMSE_V$ grows steadily with population variability, indicating, as might be expected, increasing instability of the Bootstrap variance estimate as the population variability increases;
  - **systematic rounding:** the effect of systematic rounding becomes particularly evident here, and increases substantially with sample size and population variability. $\%RRMSE_V$ reaches values as high as 14 times higher when rounding has to be applied, even in presence of a moderate (50 %) departure from the reference case when $n = 30$ and $cv_y = 3.0$. There appears not to be a general recognizabile pattern related to the extent of perturbation, seemingly indicating an overall severe rounding effect. Figure 4 depicts this behaviour: on the x-axes we represent the extent of perturbation, while on the y-axes we report the value of the ratio $\%RRMSE_V^{systematic} / \%RRMSE_V^{reference}$;

– **randomization-based rounding:** with respect to $\%RRMSE_V$, as was the case for $\%RB_V$, randomization performs almost identically as compared to the systematic rounding under every scenario.

Overall, we observe a significantly different scenery as compared to the previous SRS equal-probability case. Although slight and negligible with respect to the Bootstrap algorithm characteristics, i.e. the ability of matching population totals and first-moment, the rounding effect is remarkable on the inferential properties of the final Bootstrap estimate. Specifically, $\%RB_V$ and $\%RRMSE_V$ reveal both rounding approaches (systematic and randomization) to have a severe impact that critically inflates bias and instability of the Bootstrap variance estimation. As a consequence, empirical evidence clearly indicates the practice of rounding under unequal-probability CPS as critically affecting the Bootstrap process. A reason for this, conflicting with the equal-probability SRS case, is the fact that when unequal inclusion probabilities are involved, the rounding affects a larger set of crucial Bootstrap quantities, namely 1 to n weights $d_k = \pi_k^{-1}, k \in s$ and two population counts, the size N^* and the auxiliary total X^*.

Conclusions

In this paper we have investigated the effect of the popular practice of rounding non-integer weights as it is usually the case in bootstrapping complex samples from finite populations. In particular, we have focused on analyzing its impact on the properties of non-iid Bootstrap methodology for estimating the variance of the Horvitz-Thompson estimator for the mean. An extended simulation study, aimed at providing evidence under various experimental conditions, has been set up. Numerical devices, specifically developed to obtain benchmark scenarios to which compare the results of rounding have been described.

The simulation results clearly indicate that using rounding (both randomization-based and systematic) in bootstrapping samples under a SRS-equal probabilities design might be considered a harmless practice, at least when considering a simple linear estimator such as $\bar{y}_{HT}$. Indeed, no significant effects have been detected with respect to violations of the mimicking principle, to basic Bootstrap algorithm properties nor to inferential properties of the final Bootstrap estimates.

Vice versa, under the more complex CPS-varying probabilities design, some relevant issues arise. While basic properties of the Bootstrap algorithm result only marginally affected, a severe effect was detected for what concerns the properties of the Bootstrap estimates for the variance of $\bar{y}_{HT}$ finally provided by the algorithm. Particularly, the relative bias in matching the population totals N and X by, respectively, N^* and X^*, seems to be acceptable and Bootstrap unbiasedness appears as satisfied. On the other hand,

(continued)

relative bias and relative root mean square error for the variance estimate, regardless of the employed rounding method, clearly show to suffer an overall significant increase (meaning greater instability) as compared to the benchmark ideal integer-weights scenarios. This leads to the conclusion that, even in the presence of a simple estimator such as $\bar{y}_{HT}$, rounding induces an actual problem in bootstrapping non-iid samples when the inclusion probabilities are not all equal.

This suggests more research on the topic. In particular we deem worth of attention: (1) the investigation of the rounding effect when dealing with more complex estimators such as semi-linear or non-linear; (2) further investigation of the rounding effect also contemplating Bootstrap confidence intervals; and (3) the developing of alternative bootstrapping algorithms possibly not requiring integer weights as an alternative to existing methods based on rounding on a routine basis as discussed in this paper. This could be prompted by the innovative framework proposed in Ranalli and Mecatti [12].

References

1. Antál, E., Tillé, Y.: A direct bootstrap method for complex sampling designs from a finite population. J. Am. Stat. Assoc. **106**, 534–543 (2011)
2. Barbiero, A., Mecatti, F.: Bootstrap algorithms for variance estimation in πPS sampling. In: Mantovan, P., Secchi, P. (eds.) Complex Data Modeling and Computationally Intensive Statistical Methods. Springer, Berlin (2009)
3. Beaumont, J.-F., Patak, Z.: On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. Int. Stat. Rev. **80**(1), 127–148 (2012)
4. Booth, J.G., Butler, R.W., Hall, P.: Bootstrap methods for finite populations. J. Am. Stat. Assoc. **89**, 1282–1289 (1994)
5. Chao, M.T., Lo, A.Y.: A Bootstrap method for finite population. Sankhya **47(A)**, 399–405 (1985)
6. Chauvet, G.: Méthodes de Bootstrap en population finie. Ph.D. Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Université de Rennes 2. Available at http://tel.archives-ouvertes.fr/docs/00/26/76/89/PDF/thesechauvet.pdf (2007)
7. Fog, A.: Biased Urn Theory. R package vignette. http://cran.r-project.org/web/packages/BiasedUrn/vignettes/UrnTheory.pdf (2013)
8. Gross, S.: Median estimation in sample Surveys. In: Proceedings of SSRM. American Statistical Society, pp. 181–184 (1980)
9. Hall, P.: The Bootstrap and Edgeworth Expansion. Springer, New York (1992)
10. Holmberg, A.: A Bootstrap approach to probability proportional to size sampling. In: Proceedings of Section on Survery Research Methods. American Statistical Association, pp. 181–184 (1998)
11. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org (2013)
12. Ranalli, M.G., Mecatti, F.: Comparing recent approaches for bootstrapping sample survey data: a first step towards a unified approach. In: Proceedings of Section on Survery Research Methods. American Statistical Association, pp. 4088–4099 (2012)

Measuring Downsize Reputational Risk in the Oil & Gas Industry

Marika Arena, Giovanni Azzone, Antonio Conte, Piercesare Secchi, and Simone Vantini

1 Introduction

The issue of reputational risk has always attracted much attention from academics and practitioners, since reputation is generally considered a critical asset for a company [1, 2]. It influences the behavior of the company's stakeholders from different points of view. A "favorable" reputation draws attention of qualified staff, helps the company to retain customers, contributes to legitimate its operations with public authorities and policy makers and stimulates the shareholders to invest in a company [3]. Any damage to corporate reputation, on the other hand, could have severe consequences, such as loss of current or future customers, loss of employees or managers within the organization, reduction in current or future business partners and increase in financial funding cost [4].

From this perspective, reputation has to be adequately managed as other company's assets, which leads to the need of dealing with the problem of reputational risk [5–7]. Not by chance, in recent years, the issue of the reputational risk management moved high on top managers' agenda and a survey done by Economist Intelligence Unit with 269 senior executives shows that it is considered one of the most significant threat to business success [36].

M. Arena (✉) • G. Azzone • A. Conte
Dipartimento di Ingegneria Gestionale, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: marika.arena@polimi.it; giovanni.azzone@polimi.it; antonio1.conte@polimi.it

P. Secchi • S. Vantini
MOX – Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: piercesare.secchi@polimi.it; simone.vantini@polimi.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_3

Generally speaking, reputation can be conceived as the system of stakeholders' perceptions and expectations towards the corporation [8], and reputational risk can be defined as the risk of having this system of perceptions/expectations altered or damaged. From this perspective, the ability of a company of handling a crisis depends not only on the quality and timeliness of its decision making, but also on how its actions are perceived by its stakeholders [9, 10].

However, perceptions and expectations are hardly measurable, leading to the need of defining an ad hoc approach to measure reputational risk. Only few studies have addressed this specific problem and they can be distinguished into two main streams: qualitative and quantitative studies. Qualitative studies mainly focus on the identification of the core components of the reputation (i.e. reputational drivers) and try and associate performance indicators to each of them in order to identify and monitor potential risks. For instance, the Reputation Quotient model created by Fombrun et al. [11] defines six dimensions and 20 characteristics to measure reputational risk. Similarly, Rayner [8] identifies seven reputational drivers and 27 indicators that can somehow support the evaluation of each driver and its exposure to uncertainty. Quantitative studies, on the other hand, generally relies on the event study methodology and examine share market reaction to specific types of risk. Most of this research focuses on financial sector organizations, where quantitative models are commonly used for examining potential losses associated to operational risk.

In this paper, we follow the second stream of works and we aim to propose a quantitative approach to measure reputational risk, relying on the event study methodology. Compared to prior research, we do not limit the analysis to big losses/catastrophic risk, but we aim to explore events that have a limited impact from an operational perspective (e.g. protest) but could have a considerable reputational effect. From this point of view, we do not focus on a specific risk category, but we take into consideration more generally different type of events that can potentially affect the system of perceptions and expectations of the system of stakeholders. In details, we propose a new model that aim to overcome the limitations of previous approaches by providing for each event a confidence interval for the estimated downsize impact and taking into account both the presence of price sensitive events and the share volatility. Empirical data used to test the model are derived from a leading multinational company, that competes in the Oil & Gas industry and is listed on NY Stock Exchange. This industry has been chosen because it is subject to growing pressures from the public opinion and media for its potential impact on the environment and the society, and the Deepwater Horizon explosion in 2010 rose even more attention on it. Hence, these companies are under scrutiny and reputational risk is an hot topic for them. In addition, the analysed company has been selected because of the possibility for the researchers to access the company's informants to ensure data triangulation (e.g. verify the database of events, access to the list of price sensitive events, verify the operational impacts associated to identified events).

The remaining of the paper is articulated as follows. Section 2 provides a review of prior works dealing with the issue of reputational risk, analyzing different approaches followed. Section 3 presents the data sets and introduces the event study

methodology. Section 4 presents the results of data analysis; and finally we draw some conclusions in "Conclusion" section.

2 Literature Review

Prior literature on quantitative assessment of reputational risk is limited, since this is a still young field of research, in particular if compared to different types of risk (e.g. credit and market risks). As mentioned above, prior papers on this issue tend to focus on specific industry sectors and specific types of risk (see Table 1 for a summary of the main contributions in the field).

A few works dealing with the quantitative assessment of reputational risk are settled in the financial sector (banks and insurance companies) and examine reputational impacts associated to operational losses [4,7,12–17]. The Basel framework defines operational losses as the "loss resulting from inadequate or failed internal

Table 1 Prior research on quantitative assessment of reputational risk

Authors	Industry	Type of events considered
Strachan et al. [22]	Non financial	Illegal allegations
Davidson and Worrell [23]	Non financial	Illegal allegations
Skantz et al. [24]	Non financial	Illegal allegations
Karpoff and Lott [25]	Non financial	Corporate frauds
Hamilton [30]	Non financial	Environmental news
Long and Rao [26]	Non financial	Illegal allegations
Lanoie and Laplante [32]	Non financial	Environmental news
Reichert et al. [27]	Non financial	Illegal allegations
Konar and Cohen [31]	Non financial	Environmental news
Palmrose et al. [28]	Non financial	Restatement announcements
Perry and De Fountnouvelle [4]	Financial	Operational losses
Cummins et al. [12]	Financial	Operational losses
Karpoff et al. [33]	Non financial	Environmental violations
Murphy et al. [29]	Non financial	Corporate misconduct
Cannas et al. [20]	Financial	Operational losses related to frauds
Carretta et al. [10]	Both	Corporate governance news
Gillet et al. [13]	Financial	Operational losses
Ruspantini and Sordi [21]	Financial	Operational losses related to frauds
Plunus et al. [19]	Financial	Operational losses
Soana [7]	Financial	Operational losses
Sturm [14]	Financial	Operational losses
Biell and Muller [15]	Financial	Operational losses
Fiordelisi et al. [16]	Financial	Operational losses
Fiordelisi et al. [17]	Financial	Operational losses

processes, people and systems or from external events"; this definition includes legal risk, but explicitly excludes strategic risk and reputational risk [18]. Hence, reputational risk can be measured by comparing the amount of the market value loss with the amount of the operational loss; when the market loss exceeds the operational loss, there is a reputational damage and this difference quantifies reputational risk. To this scope, the operational loss should be known and, for this motivation, most of these studies take into account operational losses that are quite large, for instance greater than 10 million euros [12, 13], in order to improve the quality of data (since larger losses are supposed to be carefully recorded). There are also few papers addressing both big (higher than 10 USD millions) and small operational losses, still setting a minimum threshold of 1 million USD [16,17].

In the same stream of research, but doing a step further, a few authors have gone beyond the measurement of reputational risk, assessing also the role of various determinants of reputational damage exploring issues such as the bank riskiness, profitability, level of intangible assets, firm capitalization, firm size, the entity of the operational losses and the business units that suffered from the operational losses [13,16,17]. These authors provide evidence that reputational damage increases in association to some company's characteristics (e.g. firm's profitability and size) and decreases in association to other company's characteristics (e.g. level of capital invested and intangible assets) [16, 17].

A second stream of references, again focused on the financial sector, comprises a few works that introduce some modifications in the approach used to quantify reputational risk. In this stream of research, Plunus et al. [19] examine the bond market reaction to operational losses, considering debt market more suitable than share market to isolate the reputational damage. According to the authors, debt contracts should be less sensitive to pure operational effects, compared to shareholders equity, hence allowing to consider return effects as purely reputational. Cannas et al. [20] examine share market reactions to the announcement of operational losses that involve an internal fraud and provide an estimation of a reputational value at risk in order to quantify the economic capital needed to hedge associated reputational effects. Finally, Ruspantini and Sordi [21] evaluate the reputational risk impact arising from the customers' negative reaction to 20 internal fraud cases occurred in some UniCredit Group retail branches. The reputational risk impact (to be expressed in terms of business capability) is evaluated in terms of strength and length of the customers' reaction to the event: the assets under custody and management have been chosen to describe them.

A third stream of research investigates the issue of reputational risk in non financial companies. Compared to the contributions discussed above, these papers generally focus on specific types of risk.

A first rich group of contributions includes works that analyse the stock market reaction to illegal allegations. Some of these works date back to the 1980s and the 1990s [22–27]; whilst other are more recent [28,29] confirming the long-lasting attention of researchers to this specific type of risk. For instance, Karpoff and Lott [25] study the evidence of the reputational losses when the firms have criminal fraud charges and demonstrate that the corporate fraud announcements, which can be

actual or alleged, lead to a loss in firm's common share market value. Similarly, Murphy et al. [29] examine the impact of allegations of corporate misconduct on firms' profitability and market share volatility. The authors report significant declines in profitability and increased share return volatility in association to allegations of misconduct and the changes are found to be consistently greater for related-party offenses (e.g. those that damage clients). Palmrose et al. [28] examine the association between share price reactions to restatement announcements (where companies correct inaccurate, incomplete or misleading disclosures) and restatement characteristics and they find that frauds and restatement attributed to auditors are associated with more negative returns.

Focusing on a different type of events, Carretta et al. [10] analyse the impact of corporate governance press news on stock market returns before and after the news publication. Based on the analysis of the Italian market, the authors highlight how investors are influenced by rumors about corporate governance news and, before the news publication, they are only able to assess the type of corporate governance event that underlie the news; after the news publication, investors' behavior is influenced by the content and tone of the news, and no more by type of corporate governance event.

Finally, there is a limited number of papers that analyze the reaction of share market to environmental news. These studies show that firms suffer from a decline in market values following the announcement of adverse environmental news [30–32]. Doing a step further, Karpoff et al. [33] investigate the sizes of the fines, damage awards, remediation costs, and market value losses imposed on companies that violate environmental regulations and show that firms that violate environmental laws suffer significant losses.

Table 1 provides a picture of the state of the art literature in connection to quantitative evaluation of reputational risk, highlighting the industry where the study is carried out and the type of events considered.

3 Data Collection and Event Study Methodology

This section describes the data sets used in this research, the variables selection and the statistical model developed.

3.1 Data Collection

The empirical analysis is based on two data sets: (1) the "potentially reputational" events, and (2) the price sensitive events, both covering a timeframe of 10 years, between January, 1st 2003 and June, 30 2013.

Table 2 List of keywords

Dimension	Keyword
Environmental	Oil spill, spill
	Gas leak
	Blow out, explosion, fire
Social	Accident, fatality, injury
	Protest, complaint
Economic	Bribery, corruption, scandal
	Sabotage, bunkering, vandalism
	Business interruption
	Project cancellation, plant closure

The first data set comprises 67 "potentially reputational" events, that were identified by means of Lexis Nexis, through a keywords search on the All English news database (Table 2).

Keywords were chosen based on the analysis of internal and external documentation concerning the case company and a set of interviews performed with 12 key informants from the following areas: Health, Safety, and the Environment; Environmental Management; Continuous Improvement; Quality management; International negotiations; Stakeholders management; Operations; Planning and Control; Reserves and portfolio optimization. The documental analysis and the interviews allowed to achieve a better understanding of potential sources of reputational risk in the specific context, and selecting proper keywords.

The above keywords were used to retrieve the news referring to potentially reputational events. More specifically, we searched for events related to the above categories, specifically referred to the case company, and with an operational impact relatively limited (i.e. not exceeding 1 million euros).

The search results were ordered by relevance, and they were cross-checked based on the article's title and summary, for eliminating the events not related with the analyzed company or with a potential reputational event. The event date was recorded as the first time when the news about an event appeared on the press, and all the news related to the same event where grouped together. Then, descriptive information were used to categorize the events based on the following dimensions:

- Event category, that defines the type of event according to the following classification: oil spill, gas leak, blow out, flaring, sabotage, fire, accident, plant closure, business interruption, project cancellation, business misconduct, complain and protests;
- Injury/fatality, that specifies whether the event involves an injury or a fatality;
- Pollution, that specifies whether the event involves environmental pollution;
- Geographical area, that refer to countries and continents, where the event took place.

Finally, the data set was validated by the company's informants, in order to double check that the identified events were all related to the organization and their impact in terms of operational costs was not significant (i.e. not exceeding 1 million

euros), the latter condition meaning that a reduction of the market share value could be almost entirely attributed to reputational dynamics.

Compared to prior research, the process of construction of the first data set is more articulated due to the choice of not focusing on a specific type of risk or big operational losses, where ad hoc databases are already available (e.g. OpVar database, OpVantage First for the financial sector).

The second data set comprises the price sensitive events, that consists in a list of investor communications, mainly related to the company's financial performance, potentially affecting the market share value (e.g. publication of financial reports, dividends distribution, communications related to relevant business choices). This data set is a public archive, that was provided by the company itself.

3.2 Event Study Methodology

This section presents the approach adopted to analyze the share market reaction to each one of the 67 events previously identified. Our approach relies on the concept of abnormal return and the basic idea of our approach is to look—in proximity of each event—at the percentage variation of the share market value from the dynamics it would have had in the days after the event if the event had not happened.

To measure the variation imputable to the event is of primary importance to capture in a suitable neighborhood of the event the dependence between the share value and both the market and the reference sector. Indeed, by exploiting this dependence it is possible to decouple that part of variation imputable to the market and/or sector from the variation directly imputable to the company performance. This latter variation is indeed the relevant one from a reputational perspective being the one possibly affected by the reputational event.

The most common approach presented in the literature to estimate the abnormal return of a given event (e.g., [4]) is to fit a linear model relying on the data referring to the τ_1 days before the event and then look at the prediction residuals the τ_2 days after the event. More in detail, the daily log-return of the share value $\log S_t/S_{t-1}$ is used as the response variable, one or more proxies of the reference market are used as regressors, and finally the amplitudes τ_1 and τ_2 of the fitting and of the prediction windows, respectively, are kept fixed across the events. The logarithmic abnormal return is finally computed as the sum of the residuals of the fitted model over the prediction window after the event. This approach has been used for instance in the literature to estimate the impacts of financial frauds on the market value of a bank.

Compared to prior literature, we introduced two main modifications. First, in order to take into account the industry specificities of an oil company, we moved from a univariate regression model to a bivariate one by introducing a proxy of the Oil & Gas industry. In detail, we directly model—as the response variable of the linear model—the company share value S_t at NYSE at day t and not the daily log-return of the share value $\log S_t/S_{t-1}$); and the Dow Jones index DJ_t at day (a proxy of the market) and the oil price per barrel OIL_t at NYSE at day t (a proxy

of the Oil & Gas industry) as regressors of the linear model. Alternative choices for both the response variable (e.g., the daily log-return of the share value) and the regressors (e.g., the average value of the share market values of the eight major companies operating in the Oil & Gas sector and indexed at the NYSE) have been tested. All candidates models have been compared in terms of residual diagnostic, model fitting, and significance and collinearity of regressors, ending up with the model detailed above.

Second, we explicitly took into consideration the existence of other price sensitive events, that consists in those events that are known anyhow to possibly affect the share dynamics. These events are characterized by: a magnitude of their respective impacts comparable with the magnitude of the potential impacts of the reputational events under study; and by being at least as frequent as the reputational events themselves. Hence, during the estimation of the abnormal return of a reputational event their presence cannot be ignored. Neglecting them might strongly bias the estimation of the impact due to the reputational event. This is for instance particularly evident when a positive price sensitive event (e.g., release of a positive semester report) occurs in close proximity of a negative reputational event (e.g., a fatality). Indeed in this case we would have an underestimate of the impact of the reputational event or even observe an unrealistic positive abnormal return due to the event. To overcome this issue we moved from a fixed window perspective to an adaptive window one. In details for the each event we identify: the starting point of the left window with the date of the last (either reputational or price sensitive) event occurred before the reputational event currently under investigation; and the ending point of the right window with the date of the first (either reputational or price sensitive) event occurred after the reputational event under investigation. This adaptive approach leads to window sizes τ_1 and τ_2 that will become specific of the reputational event under study. The counterpart of having reduced the possible bias—having neither the left fitting window nor the right prediction window affected by confounding events—is that the impact of each reputational event will be estimated with a different accuracy: high accuracy for "lucky" reputational events far from its closest price sensitive or reputational event; and low accuracy for "unlucky" reputational events which are very close to its closest price sensitive or reputational event.

This latter fact leads to a third major change with respect to the approach presented in the literature. The idea proposed in the literature of computing the abnormal return from the predictive residuals after the event date just allows a point estimate of it without any insight on the reliability of the estimates. We thus propose a new model able to provide an interval estimate of the impact. In detail, we overcome the concept of a left fitting window for fitting the model and right prediction window for computing the disagreement with respect to the fitted model, and move towards the definition of a unique model defined before and after the event including in its definition a possible discontinuity in correspondence of the event date. This has been done by introducing among the regressors a dummy variable distinguishing between the days before the event (i.e., dummy set to zero on those days) and after the event (i.e., dummy set to one on those latter days). Our final

proposal for modeling the share market value of the Oil & Gas company in a window of the *i*th reputational event is thus:

$$S_t = \beta_0^i + \beta_{Loss}^i \, \mathbf{1}_{\{t \geq t_i\}}(t) + \beta_{DJ}^i \, DJ_t + \beta_{OIL}^i \, OIL_t + \varepsilon_t,$$

with $t \in \{t_{Previous\ Event}, \ \ldots, \ t_{Next\ Event-1}\}$, t_i the date of the i-th reputational event, and:

- S_t the share value at NYSE at day t (i.e., the response variable acting as proxy of the economic value of the company);
- DJ_t the value of the Dow Jones Index at day t (i.e., the regressor acting as proxy of the market);
- OIL_t the price of oil per barrel at NYSE at day t (i.e., the regressor acting as proxy of the Oil & Gas sector);
- $\mathbf{1}_{\{t \geq t_i\}}(t)$ the dummy variable modeling a jump at day t_i (the date of the i-th reputational event). This quantity is set to zero for $t < t_i$ and to one for $t \geq t_i$;
- Finally, ε_t is the zero-mean error term.

The regression coefficients β_{DJ}^i and β_{OIL}^i describe the influence of market and sector, respectively, on the dynamics of the share market value in proximity of the event date. The coefficient β_{Loss}^i is associated instead to a possible jump in the model at day t_i and it can be thus interpreted as the expected absolute loss imputable to the *i*th reputational event.

The coefficient can be estimated by fitting the model to the data of the i-th window through ordinary least square. Thus, the coefficient $\widehat{\beta_{Loss}^i}$ will be our point estimate of the absolute impact. Indeed we have that:

$$\widehat{\beta_{Loss}^i} = \widehat{S_{t_i no\ event}} - \widehat{S_{t_i event}}$$

is the expected difference between the share market value at day t_i if the event had or had not happened. These two quantities are immediately computable by plug in of the value of the Dow Jones Index and of the price of oil per barrel at day t_i in the estimated model and by switching off and on the dummy variable respectively.

In detail, to make the 67 impacts comparable across time and describable in terms of abnormal return we focus on relative impacts. Our estimate of the abnormal return will be indeed the relative loss:

$$\frac{\widehat{S_{t_i no\ event}} - \widehat{S_{t_i event}}}{\widehat{S_{t_i event}}}$$

that is the ratio between the estimated absolute loss $\widehat{S_{t_i no\ event}} - \widehat{S_{t_i event}}$ and the expected share market value $\widehat{S_{t_i no\ event}}$ at day t_i if the event had not happened. Note that the estimated relative loss is a rational function of the OLS estimates of the coefficient parameters. This allows also to build approximated confidence intervals whose amplitude will depend on both the amplitude window (i.e., isolated

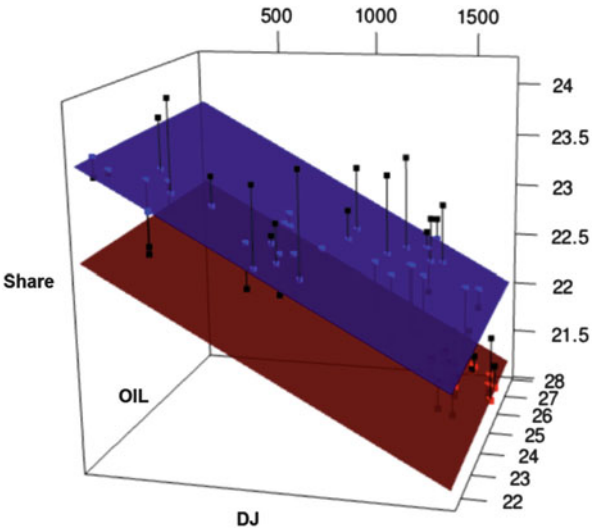

Fig. 1 Visual representation of the estimated model for the business interruption in Nigeria on July 18, 2008

reputational events will be typically associated to smaller confidence intervals while reputational events surrounded by other events will be typically associated to larger confidence intervals) and on the local volatility of the share value (i.e., reputational events occurring in periods of large volatility will be typically associated to larger confidence intervals while reputational events occurring in periods of reduced volatility will be typically associated to larger confidence intervals). In detail, the confidence intervals for the 67 relative losses will computed relying on a second order Taylor expansion and the Chebychev inequality.

As an illustrative example, in Fig. 1 we report a visual representation of the estimated model for a particular event (i.e., July 18, 2008, business interruption in Nigeria). The axes X, Y, and Z refer to the Dow Jones Index, to the price of oil per barrel, and to the company share market value, respectively. The blue plane represents the share dynamics estimated before the event and the red plane the share dynamics after the event. Black points represent the daily values of the triplets (DJ_t, OIL_t, S_t). The ones projected on the blue plane are the ones associated to the days before the events while ones projected on the red plane the ones associated to the days after the events. Finally the vertical displacement between the two planes correspond to the estimated absolute loss.

4 Results

In Fig. 2 we report along time the estimated impacts of the 67 reputational events identified for the Oil & Gas company under investigation with their corresponding confidence intervals. Focusing on events having significant impact at 10 % level, we find a strong excess of negative events (30 %) with respect to an ideal scenario

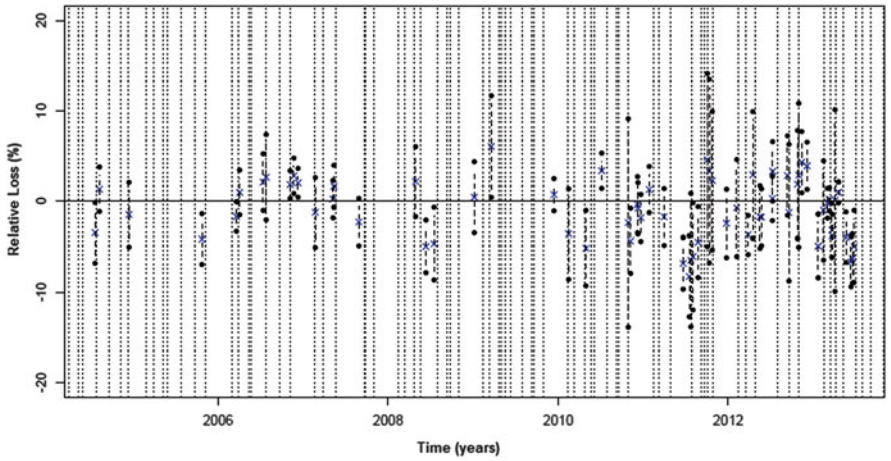

Fig. 2 Confidence intervals along time of the estimated impacts of the 67 reputational events identified. *Vertical dashed lines* are drawn in correspondence of price sensitive events

in which reputational events have no negative impact on share market value. On the contrary, we do not find any significant excess on the number of significantly positive event (9 %) which is indeed in line with the expected value we would have observed an a scenario in which reputational events have no positive impact on share market value. This prominence of significantly negative impacts over the positive ones is an evidence of the existence of a downsize effects of reputational events (at least for some of them). This evidence leads to a future area of investigation in whose focus will be the identification of both those types of events which the company is reputationally exposed and those type which do not significantly impact the company share market value.

The dashed vertical lines reported in Fig. 2 have been drawn in correspondence of events which were price sensitive for the company. It is clear from the picture that while the temporal frequency of the price sensitive events has remained unchanged along time, on the contrary the frequency of reputational events has grown significantly in the last years. This could be related to the increasing sensitivity of traditional and new information and communication media to the economic, social, and environmental performance of companies operating in the Oil & Gas industry. This trend points out the fact that that nowadays the presence of neighboring price sensitive and reputational events cannot be neglected in the estimation of the impact of an event. As a comparison, in Fig. 3 we compare the estimated impacts of the 67 reputational events taking into account (bottom panel) or not taking into account (top panel) the price sensitive or reputational events occurring in proximity of the reputational event under investigation. In details, in the bottom panel, the 67 impacts as estimated by the model we propose based on adaptive windows, and in top panel in the 67 impacts we would have obtained if we had kept the window fixed as proposed in the literature (i.e., a windows [−30;

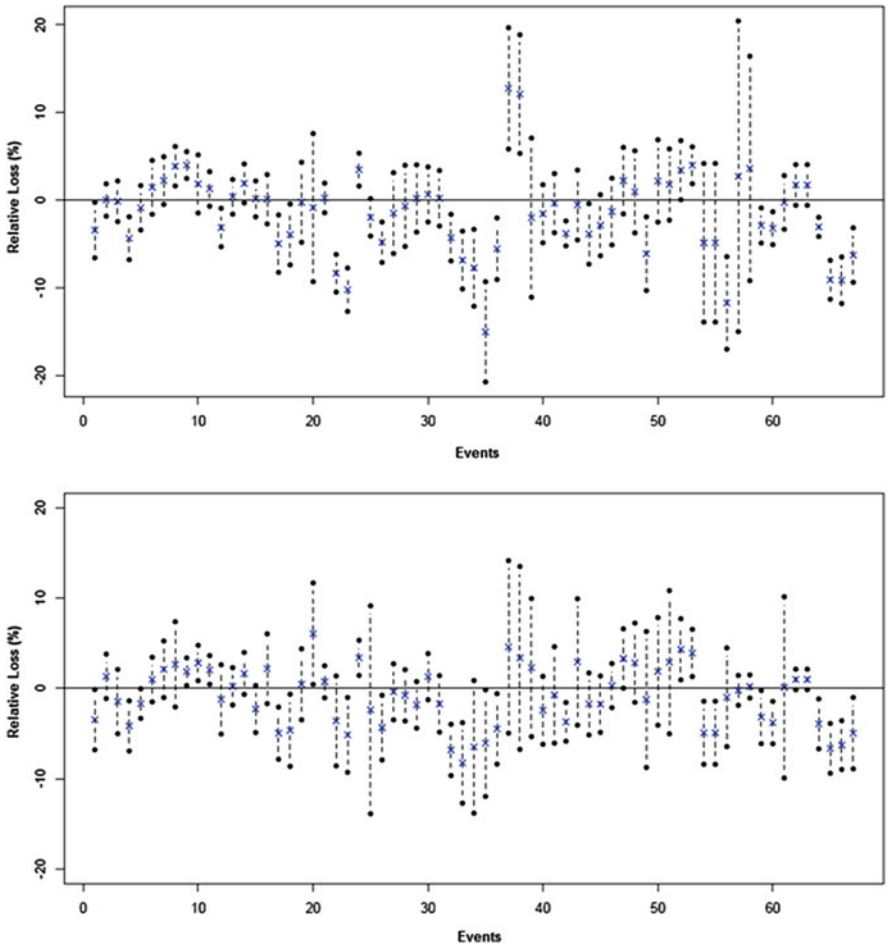

Fig. 3 Estimated impacts of the 67 reputational events taking into account (*bottom panels*) or not taking into account (*top panel*) the price sensitive or reputational events occurring in proximity of the reputational event under investigation

+15] around the event date). The 37th and the 38th events (i.e., an oil spill in Nigeria on October 4, 2012 and a business misconduct in Nigeria on October 12, 2011, respectively) come out as paradigmatic, dramatically pointing out the effect of neglecting neighboring events. Indeed, when the neighboring events are not taken into account, the two events are wrongly identified as events providing an important upside of the company.

It is worth noticing that this process mostly rely on data publicly available. This makes the approach possibly fruitful in the direction of a strategic comparison with direct competitors, possibly able to unveil areas of reputational exposure common to the Oil & Gas industry or very peculiar of the company itself.

Conclusions

So far, in the literature, the issue of reputational risk has been explored only partially, with a prevalence of qualitative studies, whilst quantitative approaches have been used mainly in connection to specific types of reputational risks and/or in a specific industry sectors (e.g. frauds and financial service companies).

Though this balance can be easily explained considering the intangible nature of reputation, that deals with stakeholders' perceptions and expectations, the possibility of applying quantitative approaches to reputational risks appears to be particularly interesting in the light of the close relationship between corporate reputation and a company's performance in financial markets.

Moving from these considerations, this paper aimed to present a new model developed to quantitatively measure reputational risk, using an event study methodology. The proposed model is specifically aimed to evaluate different types of reputational risks (related to the environmental, social and economic sphere), where the operational cost associated to the event is relatively limited. This choice is new compared to prior research, that instead focused on specific types of risks—e.g. violation of environmental regulation, frauds . . . —associated to events with a significant operational impact (i.e. big operational losses). The rationale at the basis of this choice grounds on the growing sensitivity of a broader range of stakeholders, including the public opinion, towards corporate behavior, whereby even events that do not imply significant operational cost can actually harm the corporate reputation since they indicate the failure of a company to address its stakeholders' expectations [34, 35].

From a modeling perspective the major innovations introduced by this work are the introduction of adaptive windows (based on the neighboring price sensitive or reputational events) and the direct modeling of the possible impacts of reputational events (based on the share dynamics both before and after the event under study). Compared to the previous literature, these two innovations jointly allow to obtained confidence intervals for the estimated impacts based on both the local volatility of the share and the distance of the event to the two closest neighboring events, giving a clear insight on the reliability of the corresponding estimates.

Finally, we discuss possible paths for future research. This paper focused on the identification of the events that determine a reaction in the share market, refining the methodology for the estimation of the loss. Stemming from this result, a first possible development consists in the exploration of the probability of occurrence of different events, allowing to achieve a better understanding of prospective dynamics associated to reputational risk. This is

(continued)

particularly interesting in the view of a predictive use of the model although the low number of reputational events (i.e., 67) poses some limits to the use of our results for the estimation of the impact of future reputational events.

Second, future studies could analyze the determinants at the basis of the observed impacts, for instance exploring if different risk categories (environmental, social and economic risk) or location of the event may determine a different reaction of the share market. In this same stream, the interaction of different determinants could be studied—e.g. considering if a certain risk category in a specific location is more or less dangerous from a reputational point of view compared to other possible combinations. From a managerial perspective this analysis could be particularly interesting, because it could offer managers and decision-makers within the organization a tool to decide where to concentrate preventive actions and control options.

Finally, from an empirical perspective, the comparison of different companies competing in the same industry could be particularly relevant. In this way, in fact, it would be possible to understand how different systems of stakeholders influence different organizations and impact on their financial performances.

References

1. Roberts, P.W., Dowling, G.R.: Corporate reputation and sustained superior financial performance. Strateg. Manag. J. **23**, 1077–1093 (2002)
2. Duhé, S.C.: Good management, sound finances, and social responsibility: two decades of US corporate insider perspectives on reputation and the bottom line. Public Relat. Rev. **35**, 77–78 (2009)
3. Chun, R.: Corporate reputation: meaning and measurement. Int. J. Manag. Rev. **7**, 91–109 (2005)
4. Perry, J., De Fontnouvelle, P.: Measuring Reputational Risk: The Market Reaction to Operational Loss Announcements. SSRN eLibrary. Available at SSRN: http://ssrn.com/abstract=861364 or http://dx.doi.org/10.2139/ssrn.861364 (2005)
5. Waddock, S.A., Graves, S.B.: The corporate social performance. Strateg. Manag. J. **8**, 303–319 (1997)
6. Burke, R.J., Martin, G., Cooper, C.L.: Corporate reputation: managing opportunities and threats. Gower Publishing, Famhman (2011)
7. Soana, M.G.: Reputazione e rischio reputazionale nelle banche. EIF-e.Book, Venezia (2012)
8. Rayner, J.: Managing Reputational Risk: Curbing Threats, Leveraging Opportunities. Wiley, Chichester (2004)
9. Atkins, D., Drennan, L., Bates, I.: Reputational Risk: A Question of Trust. Global Professional Publishing, London (2006)
10. Carretta, A., Farina, V., Martelli, D., Fiordelisi, F., Schwizer, P.: The impact of corporate governance press news on stock market returns. Eur. Financ. Manag. **17**, 100–119 (2011)
11. Fombrun, C.J., Gardberg, N.A., Server, J.M.: The reputation quotient: a multi-stakeholder measure of corporate reputation. J. Brand. Manag. **7**, 241–255 (2000)

12. Cummins, J.D., Lewis, C.M., Wei, R.: The market value impact of operational loss events for US banks and insurers. J. Bank. Financ. **30**, 2605–2634 (2006)
13. Gillet, R., Hübner, G., Plunus, S.: Operational risk and reputation in the financial industry. J. Bank. Financ. **34**, 224–235 (2010)
14. Sturm, P.: Operational and reputational risk in the European banking industry: the market reaction to operational risk events. J. Econ. Behav. Organ. **85**, 191–206 (2013)
15. Biell, L., Muller, A.: Sudden crash or long torture: the timing of market reactions to operational loss events. J. Bank. Financ. **37**, 2628–2638 (2013)
16. Fiordelisi, F., Schwizer, P., Soana, M.G.: The determinants of reputational risk in the banking sector. J. Bank. Financ. **37**, 1359–1371 (2013)
17. Fiordelisi, F., Schwizer, P., Soana, M.G.: Reputational losses and operational risk in banking. Eur. J. Financ. **20**, 105–124 (2014)
18. Basel Committee. Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards. A Revised Framework. Comprehensive Version, June (2006)
19. Plunus, S., Gillet, R., Hübner, G.: Reputational damage of operational losses on the bond market: evidence from the financial industry. Int. Rev. Financ. Anal. **24**, 66–73 (2012)
20. Cannas, G., Masala, G., Micocci, M.: Quantifying reputational effects for publicly traded financial institutions. J. Financ. Transform. **27**, 76–81 (2009)
21. Ruspantini, D., Sordi, A.: The reputational risk impact of internal frauds on bank customers: a case study on UniCredit Group. Working paper series. 5. Unicredit & Universities (2011)
22. Strachan, J.L., Smith, D.B., Beedles, W.L.: The price reaction to (alleged) corporate crime. Financ. Rev. **18**, 121–132 (1983)
23. Davidson, W.N., Worrell, D.L.: The impact of announcements of corporate illegalities on shareholder returns. Acad. Manag. J. **31**, 195–200 (1988)
24. Skantz, T., Cloninger, D.O., Strickland, T.H.: Price-fixing and shareholder returns: an empirical study. Financ. Rev. **1**, 153–163 (1990)
25. Karpoff, J.M., Lott, J.R.: The reputational penalty firms bear from committing criminal fraud. J. Law. Econ. **36**, 757–802 (1993)
26. Long, D., Rao, S.: The wealth effects of unethical business behavior. J. Econ. Financ. **19**, 65–73 (1995)
27. Reichert, A.K., Lockett, M., Rao, R.P.: The impact of illegal business practice on shareholder returns. Financ. Rev. **1**, 67–85 (1996)
28. Palmrose, Z.V., Richardson, V.J., Scholz, S.: Determinants of market reactions to restatement announcements. J. Account. Econ. **37**, 59–89 (2004)
29. Murphy, D.L., Shrieves, R.E., Tibbs, S.L.: Understanding the penalties associated with corporate misconduct: an empirical examination of earnings and risk. J. Financ. Quant. Anal. **44**, 55–83 (2009)
30. Hamilton, T.: Pollution as news: media and stock market reaction to the toxics release inventory data. J. Environ. Econ. Manag. **28**, 98–113 (1995)
31. Konar, S., Cohen, M.A.: Information as regulation: the effect of community right to know laws on toxic emissions. J. Environ. Econ. Manag. **32**, 109–124 (1997)
32. Lanoie, P., Laplante, B.: The market response to environmental incidents in Canada: a theoretical and empirical analysis. South. Econ. J. **60**, 657–672 (1994)
33. Karpoff, J.M., Lott, J.R., Wehrly, E.W.: The reputational penalties for environmental violations: empirical evidence. J. Law Econ. **48**, 653–675 (2005)
34. Porter, M.E., Kramer, M.R.: Strategy and society. Harv. Bus. Rev. **84**, 78–92 (2006)
35. Godfrey, P.C., Merrill, C.B., Hansen, J.M.: The relationship between corporate social responsibility and shareholder value: an empirical test of the risk management hypothesis. Strateg. Manag. J. **30**, 425–445 (2009)
36. Economist Intelligence Unit: Reputation: Risk of Risks. The Economist (2005)

BarCamp: Technology Foresight and Statistics for the Future

Laura Azzimonti, Marzia A. Cremona, Andrea Ghiglietti, Francesca Ieva, Alessandra Menafoglio, Alessia Pini, and Paolo Zanini

1 Introduction and Motivations

In the last two decades a drastic renewal has occurred in Statistics and in all the fields that involve Statistics. New requirements have arisen from new kinds of data, while technology has increased the ability of exploration and computation using massive amounts of information. As a result, more and more disciplines have started making an intensive use of statistical methods, driving the development of novel tools and providing new questions to be answered in a wide range of new application settings. In general, the production and the communication of results have changed in an extremely rapid way.

The aim of this paper is to introduce the BarCamp as an innovative way of producing and communicating statistical knowledge. For this purpose, we propose an algorithm to organize a scientific BarCamp and we describe it in detail in Sect. 2. In Sect. 3 we describe the BarCamp held at Politecnico di Milano and we discuss the vision of Statistics for the next 25 years emerged during the event. Finally, some conclusive observations are drawn in the section "Conclusions".

L. Azzimonti • M.A. Cremona • A. Ghiglietti • A. Menafoglio • A. Pini • P. Zanini
MOX – Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: laura.azzimonti@polimi.it; marziaangela.cremona@polimi.it; andrea.ghiglietti@polimi.it; alessandra.menafoglio@polimi.it; alessia.pini@polimi.it; paolo.zanini@polimi.it

F. Ieva (✉)
Department of Mathematics "Federigo Enriques", Università degli Studi di Milano, via Saldini 50, 20133 Milano, Italy
e-mail: francesca.ieva@unimi.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_4

2 The Method: BarCamp as a New Way of Making and Sharing Science

BarCamp is quite a new type of event for the scientific and technological community, organized for the first time in the US a decade ago. In particular, the first BarCamp took place in Palo Alto, California, on August 19–21, 2005. No official definition of BarCamp is available, though the most common way to describe it is "user-generated unconference". *User-generated* means that, although a generic organization is set and a theme is chosen before the event starts, users decide the main problems to be discussed, according to their interests and knowledge. *Unconference* means that a BarCamp is not structured as a typical conference with speaker names and talk session topics established in advance. In a BarCamp everyone can be a speaker rising a subject, commenting on a topic previously discussed or even just asking a question. Therefore, the themes faced in a BarCamp can be several and various, although they usually concern technology and the Web. BarCamps are mainly organized online and, despite their youth, they are very popular also outside the US. Indeed, according to the official BarCamp wiki [4], in 2013 six BarCamp were held in Africa, 15 in Asia and 15 in Europe.

When the main topic of a BarCamp concerns a popular subject, it is surely simple to stimulate a debate. However, the organization of a BarCamp on a more technical argument could be interesting, too. In Statistics, for example, sharing knowledge and ideas is very convenient in facing the new challenges provided by technological advances. We thus propose an algorithm to organize a BarCamp on a scientific topic. This algorithm has to be intended as a proposal based on our direct experience in the organization of the BarCamp held in Milan [3].

First of all, the organizers decide the main theme of the event. It should not be too restrictive and specific but it should identify a scientific area, possibly original and rapidly expanding. At the same time a competition (say c_0) is launched: candidates are invited to describe the specific topic they wish to cover and, above all, the way to present it to the audience in order to stimulate a debate among the participants of the BarCamp. Then the organizers establish n winners, say $w_1, \ldots, w_n$, and single out p topics to be assigned to $t_1, \ldots, t_p$ round tables. For each table t_i, suitable material upon which to enhance the open discussion is prepared and possibly shared in advance. If p is large enough, the BarCamp may last k days $d_1, \ldots, d_k$. Finally, a logistics set $\mathscr{L}$, satisfying all the setting properties that characterize a BarCamp, is arranged. For example, the area where the BarCamp takes place is opportunely organized, and a website offering place to preliminary discussions is created. Since a BarCamp is not a conventional conference, an open space with an informal setting is preferable. Moreover, specific corners with one or more computers where participants can browse the internet may help to create an informal environment. Music, leisure and social activities are planned, too. The detailed algorithm we propose for the organization of a successful BarCamp is described in Table 1.

Once the event is ended it is very important to gather the most interesting ideas emerged during discussions. For example, a short video about the BarCamp could

Table 1 The algorithm we propose for the organization of a scientific BarCamp, based on our experience with the BarCamp held in Milan

The BarCamp Algorithm
Initialization.
Launch a competition c_0. Create a captivating logo. Arrange a web site and a social network page.
Step 1.
Establish the winners $w_1, \ldots, w_n$ of competition c_0. Establish the p topics to be debated in $t_1, \ldots, t_p$ tables and assign a color to each topic. Characterize the logistics set $\mathscr{L}$.
Step 2.

for(day $d_i = d_1, \ldots, d_k$){
if($d_i == d_1$){Organize an initial activity to allow participants to know each other.}
Organize a session where $l \leq n$ winners present their ideas and stimulate a discussion among the participants.
Organize p_i parallel sessions (with p=$\sum_{i=1}^{k} p_i$). In each session, the topics of the round tables $t_{h+1}, \ldots, t_{h+p_i}$ are deepened ($h = \sum_{j=0}^{i-1} p_j$ and $p_0 = 0$). Participants are encouraged to bring some presentations, photos or videos they prepared.
Collect ideas emerged during discussions and associate to each contribution the color of the corresponding topics.
Organize some leisure and social activities (e.g, sport, a concert etc. . .).}

be produced and we recommend to share online all the material collected during the event. A website, possibly with a forum, may represent the most appropriate and interactive way to reach this goal.

3 Case Study: BarCamp–Technology Foresight and Statistics for the Future

In this section we present a 1D (one day) implementation of the algorithm described in Table 1. The real case of interest is the BarCamp held in Milan, in honor of the 150th anniversary of Politecnico di Milano [3]. Figure 1 shows the logo created for the event. The BarCamp was an event related to S.Co. 2013 [26] conference on "Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction". This BarCamp aimed at discussing the vision of Statistics for the next 25 years and was entitled "Technology Foresight and Statistics for the Future".

Fig. 1 The logo of the BarCamp held in honor of 150 years of Politecnico di Milano

We set the BarCamp in an open space, (i.e., the Agorà of Politecnico di Milano), on the compact one-day time interval ($k = 1$) of September 12, 2013. The detailed structure of the day is reported in Appendix.

During the months preceding the event, a competition c_0 was launched in order to challenge young (less than 33 years old) statisticians to envision statistical models and methods that will have an impact on the development of technology in the next 25 years (before the 175th anniversary of Politecnico di Milano). Competitors were asked to submit an essay describing their vision of Statistics for technology of the future, together with a description of the way chosen to present their ideas, in case of winning the contest. We did not restrict either the area or the way participants could present their work, in order to stimulate new approaches to the study and the communication of statistical knowledge and results. We selected $n = 4$ proposals, that are listed below in alphabetical order:

- w_1 *"Statistics: Pushing or Pulled by Technology?"*—Antonio Canale, Università di Torino, Italia;
- w_2 *"Statistics and new Technologies: Challenges and new Opportunities"*—Davide Pigoli, University of Warwick, United Kingdom;
- w_3 *"LEMMA"*—Ivan Vujacic, University of Groningen, Netherlands; Antonino Abbruzzo, Università di Palermo, Italia; Javier Gonzaléz, University of Groningen, Netherlands; Giulia Marcon, Università L.Bocconi, Milano, Italia;
- w_4 *"Statistics: an Important Player in the Big Data Age"*—Diwei Zhou, University of Brighton, United Kingdom.

The authors of these contributions were invited to participate to the S.Co. 2013 conference. They received travel support, conference registration and lodging. The winners of the competition led the main streams of the BarCamp, by sharing their ideas and stimulating new perspectives on Statistics and technology.

Simultaneously with the choice of winners, we created a Facebook page to ease the diffusion of the initiative and to share and collect ideas. Discussions on this page highlighted a variety of topics to be treated during the BarCamp. According to the themes emerged as most interesting from the social media and to the winners' proposals, we defined $p = 3$ main topics and we associated a color to each one: Big Data (Green), Computational Statistics (Orange) and Visualizing Data (Blue). Three round tables, t_1, t_2, t_3 were set up on these main topics. Table t_1 focused on Big Data and the new technologies characterizing the fields of application of Statistics. Table t_2 was about statistical computing techniques, which are relevant to a proper handling of complex data. Finally, table t_3 concentrated on the visualization tools

that are fundamental to an appropriate preliminary investigation of complex data and effective communication results.

The logistics set $\mathscr{L}$ was characterized in order to support the BarCamp spirit. At the very beginning of the day a leisure activity was organized to let participants know each other. Then, a huge panel containing the three keywords *Big Data*, *Computational Statistics* and *Visualizing Data* was placed in the center of the open space. During the day, this conceptual map was enriched with all the words, concepts, ideas and suggestions arisen from discussions, colored according to the topic they were related to. The evolution of the discussion along the day was recorded taking one photo per minute of the map. A poster session for the participants contributions was planned, too. Finally, the event day was enriched with music and refreshments for lunch and dinner, sport activities and a final concert in the evening.

On September 12, early in the morning, the event started with registrations and the mixing game as opening activity. It consisted of a game where participants were invited to perform 3-min-long interviews to ten attendees about their scientific activity. In doing so, they assigned each speaker an index of similarity for possible future interactions and/or collaborations.

The winners w_4, w_3 gave their contributions in the morning, while w_2, w_1 presented their ideas in the afternoon. All the contributes focused on the role of Statistics at the present time. The core idea was that new data and modern technologies are themselves vehicles of new modeling challenges. In such a setting, the statistician must evolve and acquire new skills, such as computational ones, or problem-specific background (it is the case especially in applications to medicine, biology and industry). Even though all the winners touched common topics in their discussions, they focused on different points using their own style. For example, Diwei Zhou proposed a very interactive debate dividing people into groups and proposing a sequence of 5-min brainstorming within groups on specific questions, to be further compared and discussed between groups. Questions focused on the main features and challenges of Big Data and on the key role of Statistics in this new era of science. On the other hand, the "Lemma team" composed and played a drama about different aspects of the statistical approach to data analysis. Among these we cite: multidisciplinarity, diffusion of the statistical culture, computational issues, interaction with other disciplines, communication of results and the importance of theoretical knowledge (especially when dealing with a huge amount of not structured data). Finally, both Antonio Canale and Davide Pigoli chose an interactive presentation to discuss the differences between Statistics and computer science. Moreover, the issues related to the dissemination and teaching of Statistics arose. Furthermore, all the winners dealt with the ways Statistics could act as a leading discipline, not only as the servant of all sciences. They also mentioned the influence of technology (in terms of computing capability) in pushing statistical development, finally underlining the ethical issues to be faced in statistical studies and communication.

3.1 t_1: Big Data and Technology

Table t_1 focused on the issues arising from managing, mining and analyzing new types of data that are commonly called Big Data. One of the most stimulating challenges Statistics will face in the next 25 years is represented by the improvement of the current techniques, that need to be modified and adapted to Big Data. Although the term Big Data seems to refer only to the size of the information collected, this expression has a wider meaning and it is especially related to the complexity of the data.

The 3 Vs The features characterizing Big Data are nowadays recognized to be *Volume*, *Variety* and *Velocity*.

- *Volume* indicates all the issues related to the techniques used to handle a great quantity of data. In fact, first problems with Big Data arise at the moment of collecting information. New tools are needed to store and mine a huge and still increasing amount of data, which nowadays is easily available thanks to new technologies. Moreover, Statistics needs new theoretical methods to improve the inferential analysis in problems concerning a big number of variables;
- *Variety* points out the motley nature of Big Data. Statistics has the important task of developing new theoretical tools for modeling data characterized by different input sources and various information content. This is highly challenging since most of the current statistical techniques can be applied only to specific and not composite data;
- *Velocity* indicates the high frequency of data collection and updating in a unit of time. New technologies are nowadays able to get a continuous data stream that requires real time analysis. All the issues related to this aspect concern data storage, data analysis and the need to provide results in the shortest possible time.

There is another important feature related to Big Data called *Viability*. When a big amount of data is collected it is necessary to filter through this information, by selecting the factors that are most likely to provide evidence for prediction and discrimination. This pre-processing phase is a crucial point to enable the statistical analysis of Big Data.

Big Data. Is it Worth? The discussion of table t_1 mainly involved the strategies that should be adopted to develop and potentiate the analysis of Big Data—in particular management and mining. A great importance was given to the pre-processing phase, that assumes an extremely relevant role when data are big and complex. Many contributions and comments warned that a huge amount of data does not always correspond to a huge amount of information concerning the problem under study [6]. They highlighted how data are not always really informative, in particular when the underlying planning is not well realized. This problem is considerable especially when computational cost to collect and manage data is high, but the inferential tools to extract valid information from them are very poor. The growth in computational power has to be supported and, at the same time,

has to fulfill the inferential and descriptive requirements imposed by the new data. Statistics is going to face this trade-off in the future. Its central role in the scientific method is strictly connected to the extent to which it will be able to identify the right cut-off point in such a trade-off, as emphasized by the winner w_4, Diwei Zhou.

Big Data in Social Networks, Medicine and Industry The discussants brought some examples of Big Data in social life behaviors and healthcare. They also treated applications arisen from recent technologies employed in the industrial context.

The main example coming from the social media framework was Facebook. With more than 950 million users who spend on average 6.5 h a month on the platform, Facebook has at its disposal an incredibly huge amount of data (see, for instance, [10]). This data potentially include a great quantity of useful information for the company. To extract this knowledge, the analyst must be able to deal with Big Data. The discussion moved over the technology adopted to manage this massive quantity of data. In particular, Facebook uses a software platform called Hadoop to process and analyze the data streaming through the web. In addition, Facebook developed two new software platforms, Corona and Prisma, to improve Hadoop scalability and increase usable memory (see [11, 12]).

The main examples in clinical context involved genomics and the integration of different health databanks (see e.g. [2, 17]). Even if these two examples are both characterized by high dimensionality, they show different features. Genomic data are frequently homogeneous and their high dimensionality results in a number of variables much higher than the number of available observations. In health databanks, a great heterogeneity is added to the huge amount of records, images, texts and information measured on the statistical units (usually, the patients) over time. The multidisciplinarity, essential when dealing with data in clinical context, generates the necessity of a new professional figure, namely the Data Scientist [18]. Indeed, Data Scientist is not just a mathematician, but also a computer scientist with expertise in a specific field of knowledge or application [13].

The last example discussed involved an industrial application. A participant presentation focused on the various problems in the development of a statistical process monitoring tool for manufacturing processes [15]. In that example, the test case of machine health monitoring in waterjet cutting was discussed to highlight current methodologies and open issues.

3.2 t_2: Computational Statistics

Table t_2 on Computational Statistics focused the discussion on all the challenges brought to light by computing in statistical modeling. Specifically, the main topics discussed during the round table are programming, parallel computing, data storage and data assimilation/integration. All these subjects were born only few decades ago and they rapidly grew up in the present Big Data era, stimulated by new technologies. They share multidisciplinarity as a common feature.

Multidisciplinarity is the fundamental ingredient in all the main themes in Computational Statistics. It is indeed necessary to mix expertise from different fields, such as informatics and numerical analysis but also biology and bioinformatics, to upgrade state-of-the-art techniques. A deep knowledge of informatics and numerical analysis helps statisticians in improving the statistical methodology both from a theoretical and a computational point of view. Moreover, when the multidisciplinarity is produced by the nature of the analyzed data, it is essential for statisticians to develop the skill of interacting with experts from different fields, using a common language to collect and share information.

The main examples of multidisciplinary fields, introduced by the winner w_3 (the Lemma Team) and discussed during the round table can be grouped in *programming*, *scientific computing* and *biomedicine and genomics*.

Programming The discussion focused on how Statistics has been influenced in the last years by the increase of computational capacity, as pointed out by the winner w_1, Antonio Canale: nowadays statisticians are dealing with huge amount of complex data, so efficient programming languages and algorithms are necessary. Interpreted languages (such as R) are going to be replaced by or integrated with compiled ones, such as C, C++ or Java, (see e.g. [7]) and computations can be performed in parallel on huge clusters (see e.g. [25]). During the discussion the need of developing strong computer science skills was pointed out. Specifically it would be necessary to teach advanced computational courses in mathematical and statistical undergraduate and graduate programs, and to stimulate more interactions between computer scientists and the statistical community.

Scientific Computing Together with programming, a good knowledge of scientific computing and numerical analysis is very useful for the analysis of complex and high dimensional data. Numerical analysis methods can be used indeed to speed up computations, efficiently solve a linear system, accurately discretize curves or surfaces and integrate quantities of interest. Statisticians can exploit numerical analysis techniques for the analysis of images or for speed up MCMC computation. Moreover in the last years new branches of numerical analysis, devoted to inverse problems, parameter estimation and uncertainty quantification, have emerged. The interaction between this new community and the statistical one will provide stimulating problems to be addressed and new efficient tools to solve them.

Biomedicine and Genomics Biomedical disciplines mix together different scientific approaches and integrate methods from medicine and biology with strategies that are typical of mathematics and computer science. Traditional medical models are gradually giving way to personalized medicine. This change creates new challenges that need different expertise to be faced. Next Generation Sequencing (NGS) techniques have revolutionized the genomic field, enabling scientists to directly study genetic and epigenetic processes. Statisticians should develop new methods or adapt existing ones to pre-process and analyze these new types of NGS data, by collaborating closely with bioinformaticians and biologists, with the purpose to unveil the complexity of the genome (see e.g. [9]). Moreover,

genomic and epigenomic knowledge must be combined within each other as well as integrated with clinical information, by using efficient models and statistical techniques.

Many researchers with different backgrounds and interests contributed with their ideas to the discussion. All of them agreed to conclude that a deep knowledge and an intense use of computational tools in Statistics will provide big improvements in forefront statistical research fields.

3.3 t_3: Visualizing Data

Table t_3 focused on open problems related to Data Visualization and, generally, to statistical communication. As highlighted by the winner w_2, Davide Pigoli, graphical communication has become a key aspect in the statistical research process. Thanks to the advances in technology, a huge quantity of data is now available to a wide public. As a consequence, the development of communication tools able to transfer the statistical knowledge, concerning data or results, to a public (composed by statisticians or uninitiated people) is an extremely important theme to be faced by the statistical community. A key aspect of this topic is the graphical communication. This stands at the frontier of two disciplines, Statistics and Information Visualization (InfoVis) [14]. The latter is the set of disciplines whose aim is to study the interactive visual representation of abstract data in order to strengthen human cognition. In the Big Data era, it is important to bear in mind that "information is not knowledge" (Einstein): great effort must be put in extracting and transmitting the statistical information hidden below the data. Therefore, statistical communication needs specific and effective graphical tools able to precisely convey complex messages.

Motivated by several examples, the discussion mainly developed around three key aspects: *curse of dimensionality*, *dynamic graphics* and *subjectivity and interpretation*.

Curse of Dimensionality Recent advances in technology provided scientists with increasingly complex data, for instance high-dimensional and functional data. A relatively large part of the literature had already been devoted to address several challenging problems arisen in dealing with these types of data, such as the problem of inference (see, e.g., [24, 27] and references therein). Nevertheless, still little attention had been paid to the problem of properly visualizing complex data through effective graphs.

Figure 2 shows an example of functional data visualization discussed during the round table. Each panel shows a different representation of box-office revenues in a year. The data refer to Italian revenues in 2012 [23], the right panel to USA revenues in 2008 [8]. The purpose of such charts is to show when the movies came out, how much they earned, and how long they lasted in cinemas. The curves representation shown in the top left panel is the most common way to visualize functional data in the statistical literature. The steamgraph (bottom left panel) is

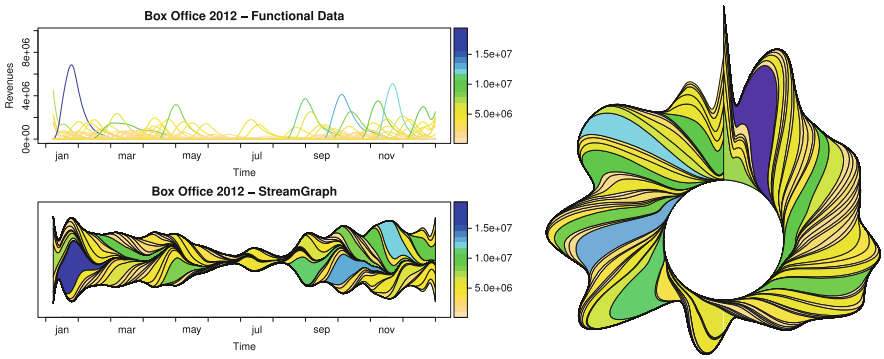

Fig. 2 Three different representations of box-office data. Plot of the curves (*top left*); steamgraph (*bottom left*); circular steamgraph (*right*)

instead commonly employed in the InfoVis approach. The latter presents a better visual impact, which however could prevent from the correct comparison among the represented quantities, which is immediate with curve representation. Nonetheless, the simplicity of the curves representation is likely to have a lower chance to be noticed and remembered [16]. Indeed, embellished charts are perceived as more attractive, more enjoyable and easier to remember [5]. They draw audience attention to the issue, by stimulating their curiosity. However, an excessive beautification can lead to bias: the chart presented in the right panel of Fig. 2 is an example of the extreme distortion that can be produced by further embellishing a steamgraph. The circularity, added to provide a natural representation of the year and to further increase the chart visual impact, leads to a hardly readable result: movies on the edge looks thinner, and such an artifact makes hard to evaluate and compare the actual revenues.

In this context, the classical paradigm "simple is better", along with the data-ink ratio rule [28], appears to be replaced by the claim "sufficient is better", which advocates for a fair balance between simplicity and visual impact. However, distortion is key to understand the problem of curse of dimensionality in graphical communication. The statement "Do not disturb my circles", attributed to Archimedes, expresses the discussion stream: the representation of high-dimensional object in 2D implies a bias that a single viewpoint cannot avoid. Therefore, to create new methods able to effectively represent complex data, the actors of the visualization play, among others statisticians and InfoVis people, should not be in contrast. The best results in visual communication will be instead reached through a marriage between Statistics and info graphics. None of the players strictly needs the other, but they are likely to play better together.

Dynamical Graphics Dynamic and interactive graphics constitute powerful tools to overcome the curse of dimensionality when representing complex and high dimensional data. In fact, dynamics allows to explore dimensions via time-varying frames, while interactive graphical tools allow the users to choose their own

viewpoint and to explore hidden relationships. In the Big Data era one can readily find instances of intrinsically dynamic data, such as real-time data, social network activities data, space-time data or georeferenced functional data [21, 22]. In this context, dynamic and interactive graphics can be the key to produce knowledge from data, through an active participation of the user to the cognitive process. New graphical standard should be established to ease this process, together with new softwares to enhance graphical production and dissemination.

A remarkable attempt of real-space graphical communication "Immaterials: Light painting WiFi" was made by Timo Arnall, Jørn Knutsen and Einar Sneve Martinussen at the Oslo School of Architecture and Design [1,20]. They represented the strength of the WiFi signal in the open spaces surrounding the University of Oslo. Data were displayed precisely in their 3D space locations. This project, making the invisible visible, opens new perspectives for the representation of other kind of real-space real-time data, as pollution or radiation data [1].

The need of representing data and knowledge in new ways induces the development of new ways of communicating statistical analysis and results, in addition to printed journals. Electronic journals together with electronic supplements of printed articles are moving onward and upward, while classical publications are becoming too restrictive for presenting statistical analysis and results. In this sense, *video-articles* and *e-articles*, which are commonly employed as means of InfoVis communication (e.g., [19]), are likely to be the future of statistical communication.

Subjectivity and Interpretation Discussants finally underlined that the statistical community needs to master the psychology of visualization. This is one of the keys to face the ethical issues mentioned by the winner w_2 Davide Pigoli as "visualization risk". Indeed, embellishments, dynamic plot and visual effects, if misused, may be as impressive as misleading.

Objective rules of an ideal visual display, together with new standards, should be established to prevent ethical issues related to the subjectivity of communication. Ethic, interpretation and misleading charts were the keywords of the last part of the discussion, stimulated by the two examples reported in Fig. 3, which are inspired by

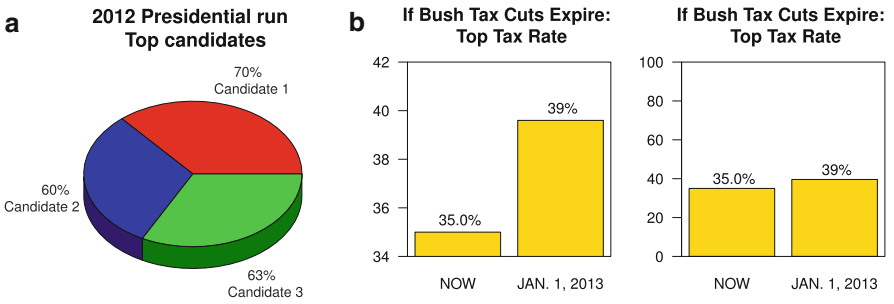

Fig. 3 Examples of misleading charts. (**a**) wrong pie chart (the sum of the percentages is not 100); (**b**) bar-plot giving a misleading message, due to the scale on the y axis (*left*) and correct version of the same bar-plot (*right*)

two real graphical communications. While in the first case (top panel), the graphic is completely wrong, but it does not provide a misleading message, in the second case (bottom panel) the y axis setting is responsible for a wrong perceived message. Looking at the barplot on the left, the y axis showing only values between 34 % and 42 % makes the tax rate seem to have an extremely sharp increase. The correct barplot on the right clearly shows that this rise is not so remarkable. The marriage between Statistics and info graphics proposed before would help to avoid such errors, providing a more clear and reliable graphical communication.

Conclusions

In this paper we described the BarCamp as an innovative way of creating and sharing statistical knowledge. We described the experiment held at Politecnico di Milano, as a solution to discuss the vision of Statistics for the next 25 years.

The main outcome produced by the BarCamp was the creation of a network among participants (especially among young statisticians attenders), to share materials and ideas, and to discuss research topics and related issues. The algorithm of a successful BarCamp proposed in Sect. 2 could become a new paradigm, characterized by vitality, velocity and multidisciplinarity, to create and communicate statistical thinking.

The BarCamp was a great opportunity both for the organizers and the participants to experience new ways of producing and communicating Statistics, by having a confrontation with other statisticians. For this reason, we stimulated discussions even after the end of the event. We received several feedbacks that underlined both the successful and the not fully satisfying aspects of our BarCamp. Participants enjoyed the main topics of discussions. Moreover, the informal setting positively influenced the engagement of all attendees. Participants were also positively impressed by the active discussions emerged in the round tables, as well as by the topics solicited by the winners of the competition. They highlighted the importance of writing and sharing the discussions results and they also gave some suggestions for a more effective communication of results.

In conclusion, the BarCamp was a really interesting experience, both for organizers and attenders. All the details (shared material, scientific contents, photos and videos) of the BarCamp held at Politecnico di Milano can be found on the website http://www1.mate.polimi.it/barcamp2013/ [3].

Acknowledgements BarCamp is one of the activities planned for celebrating the 150th anniversary of Politecnico di Milano. The authors wish to thank the organizers of S.Co. Conference and the Department of Mathematics of Politecnico di Milano.

Appendix: The structure of the BarCamp Held at Politecnico di Milano

```
streams = list("Visualizing Data", "Technology for the
          future and Big Data", "Computational Statistics");
// Take part to the facebook discussions and add new streams!!!
sport = list("Football", "Volleyball");

switch(time){

    case [9.30 - 10:30]: activity= "Opening and Registration";
    case [10.30 - 11.00]: activity= "Welcome activities";
    case [11.00 - 13.00]: activity= "Camp";

// the winners of the BarCamp competition lead the discussion
// Vujacic I. and Zhou D.
    case [13.00 - 15.00]: activity= "Lunch and Posters";
    case [15.00 - 17.30]: activity= "Streams";
print(streams);

// the winners of the BarCamp competition lead the discussion
// Canale A. and Pigoli D.

    case [17.30]: activity= "Closing";

// the BarCamp goes on with free discussion, sports and leisure

    case [17.30 - 20.00]: activity= "Sport/leisure activities";

print(sport);

    case [19.30 - 21.00]: activity= "Dinner";
    case [20.30 - 23.00]: activity= "Concert";
}

> print(streams);

> [1] "Visualizing Data"
//Do designers do it better?

> [2] "Technology for the future and big data"
//Remote Sensing
//Statistical Process Control
//How big data are improving and transforming healthcare
//Big data analysis in genomics

> [4] "Computational Statistics"
//Challenges of computing in statistical modeling
//Parallel computing
//Numerical issues
//MCMC
```

```
> print(sport);
> [1] "Football"
> [2] "Volleyball"
```

References

1. Arnall, T., Knutsen, J., Martinussen, E.S.: Immaterials: light painting WiFi. Significance **10**(4), 38–39 (2013) doi:10.1111/j.1740-9713.2012.00623.x
2. Barbieri, P., Grieco, N., Ieva, F., Paganoni, A.M., Secchi, P.: Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region. In: Mantovan, P., Secchi, P. (eds.) Complex Data Modeling and Computationally Intensive Statistical Methods. Series Contribution to Statistics, pp. 41–56. Springer, Milan (2010). ISBN 978-88-470-1386-5
3. BarCamp - Technology Foresights and Statistics for the Future. Agora' di Architettura, Politecnico di Milano, Milano, 12 September 2013. http://www1.mate.polimi.it/barcamp2013/ (2014). Accessed 5 Jan 2014
4. BarCamp. http://barcamp.org/w/page/402984/FrontPage (2014)
5. Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., Brooks, C.: Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2573–2582 (2010)
6. Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**(5), 662–679 (2012). doi:10.1080/1369118X.2012.678878
7. Chambers, J.: Software for Data Analysis, Programming with R. Springer, New York (2008)
8. Champkin, J.: 2008 movies, by Graham Wills. Significance **9**, 36–37 (2012). doi:10.1111/j.1740-9713.2012.00607.x
9. Datta, S., Datta, S., Kim, S., Chakraborty, S., Gill, R.S.: Statistical analyses of next generation sequence data: a partial overview. J. Proteom. Bioinform. **3**(6), 183–190 (2010)
10. Facebook is collecting your data – 500 terabytes a day. http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/ (2014)
11. Facebook pushes the limits of Hadoop. http://www.infoworld.com/article/2616022/big-data/facebook-pushes-the-limits-of-hadoop.html (2014)
12. Facebook Tackles (Really) Big Data With "Project Prism". http://www.wired.com/2012/08/facebook-prism/ (2014)
13. Foster, P., Fawcet, T.: Data science and its relationship to big data and data driven decision making. Big Data **1**(1) (2013). doi:10.1089/big.2013.1508
14. Gelman, A., Unwin, A.: Infovis and statistical graphics: different goals, different looks. J. Comput. Graph. Stat. **22**(1), 2–28 (2013). doi:10.1080/10618600.2012.761137
15. Grasso, M., Goletti, M., Annoni, M., Colosimo, B.M.: A new approach for online health assessment of abrasive waterjet cutting systems. Int. J. Abrasive Technol **6**(2), 158–181 (2013)
16. Hullman, J., Adar, E., Shah, P.: Benefitting infovis with visual difficulties. IEEE Trans. Visual. Comput. Graph. **17**(12), 2213–2222 (2011)
17. Ieva, F., Paganoni, A.M., Secchi, P.: Mining administrative health databases for epidemiological purposes: a case study on acute myocardial infarctions diagnoses. In: Pesarin, F., Torelli, N. (eds.) Advances in Theoretical and Applied Statistics, chap. 38, pp. 417–426. Springer, Berlin, Heidelberg (2013)
18. Liddy, E.D., Stanton, J., Mueller, K., Farnham, S.: Educating the next generation of data scientist. Big Data **1**(1) (2013). doi:10.1089/big.2013.1510
19. Martinussen, E.S.: Immaterials: Light painting WiFi, YOUrban. http://yourban.no/2011/02/22/immaterials-light-painting-wifi/ (2011)

20. Martinussen, E.S.: Making material of the networked city. In: Design Innovation for the Built Environment – Research by Design and the Renovation of Practice. Taylor & Francis, Routledge (2011)
21. Menafoglio, A., Secchi, P., Dalla Rosa, M.: A universal kriging predictor for spatially dependent functional data of a hilbert space. Electron. J. Stat. **10**, 2209–2240 (2013)
22. Menafoglio, A., Guadagnini, A., Secchi P.: A kriging approach based on aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stoch. Environ. Res. Risk. A (2014). doi:10.1007/s00477-014-0849-8
23. Mymovies: http://mymovies.it/ (2012)
24. Pini, A., Vantini, S.: The interval testing procedure: inference for functional data controlling the family wise error rate on intervals. Tech. Rep. MOX, **13**/2013 (2013)
25. Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., Mansmann, U.: State of the art in parallel computing with R. J. Stat. Softw. **31**, 1 (2009)
26. SCo - Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction. Politecnico di Milano, Milano, 9–11 September 2013 http://mox.polimi.it/sco2013/
27. Secchi, P., Stamm, A., Vantini, S.: Inference for the mean of large p small n data: A finite-sample high-dimensional generalization of Hotelling's theorem. Electron. J. Stat. **7**, 2005–2031 (2013) doi:10.1214/13-EJS833
28. Tufte, E.: The Visual Display of Quantitative Information. Graphics Press, Cheshire (1983)
29. Wills, G.: Visualizing Time: Designing Graphical Representations for Statistical Data. Springer, New York (2012)

Using Statistics to Shed Light on the Dynamics of the Human Genome: A Review

Francesca Chiaromonte and Kateryna D. Makova

1 Introduction

In this article we give an overview of some recent human genomics studies conducted by an interdisciplinary group of computational biologists, experimental biochemists and statisticians at The Pennsylvania State University. We showcase them as instances of statistics' critical role for producing scientific insight in contemporary genomics research.

The studies under consideration investigate various aspects and facets of the dynamics of the human genome. Several processes of mutagenesis contribute to changing the nuclear DNA. Some among them are both relatively simple in terms of their nature and localization, and relatively abundant. One example is nucleotide substitutions, where one nucleotide (A, C, G or T) is replaced by another. Another

F. Chiaromonte (✉)
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

Center for Medical Genomics, The Pennsylvania State University, University Park, PA 16802, USA

Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: chiaro@stat.psu.edu

K.D. Makova
Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

Center for Medical Genomics, The Pennsylvania State University, University Park, PA 16802, USA

Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: kdm16@psu.edu

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_5

example is insertions and deletions of short sequences (e.g., sequences comprising 30 or fewer nucleotides). Yet another instance of simple and frequent change is represented by repeat number alterations at microsatellite loci. A microsatellite comprises repetitions of a short motif; say, the di-nucleotide AT repeated five times to form ATATATATAT. The number of motif repetitions can increase or decrease very easily through a sort of replication hick-up called DNA strand slippage. More complex types of change that affect broader regions of the DNA include large insertions and deletions (e.g. insertions of interspersed transposable elements of varying sizes); gene duplication, loss and copy number variation; rearrangements of entire portions of the genome, etc.

As reference genomes for an increasing number of species, and then individual genomes for an increasing number of humans, became available over the last 10–15 years, the work by our group and others has produced mounting evidence that mutagenic processes don't act uniformly across the nuclear DNA and interact with one another (e.g. Chiaromonte et al. 2001; Hardison et al. 2003; Yang et al. 2004; [1]; see also Hodgkinson et al. 2012 for a review). The questions we attempted to answer concern how the action of single mutagenic processes, as well as their concerted action, unfold along the genome and are affected by the local "landscape".

With the term landscape we refer to a variety of composition, location and biochemical features of the nuclear DNA. They range from something as simple as GC content (the prevalence of G and C over A and T nucleotides), to distance to telomeres (repetitive sequences located at the tips of chromosomes) or centromeres (the constricted regions of chromosomes where spindle fibers attach during cell division), to measures of the propensity to recombine in the male and female germlines or to be transcribed to produce RNA and thus proteins, to signatures of the so-called chromatin environment, etc. In the studies reviewed here, abundances of certain families of interspersed transposable elements—which proxy transposition activity—are considered as part of the genomic landscape for simpler mutagenic processes such as nucleotide substitutions or small insertions and deletions. In other studies by our group though they are considered as the signatures of another type of process changing the DNA (see above) and themselves investigated in relation to the genomic landscape [2].

Questions about mutagenic processes and their genomic landscape are important because of the light they shed onto core mechanisms and trends in the evolution of genomes, but also because of their biomedical implications. Nucleotide substitutions, small insertions and deletions, as well as repeat number changes at microsatellite loci, have all been known to cause a large number of human genetic diseases.

2 The Data

Since we are interested in the processes that change DNA and in their landscape correlates, the first step in creating our data is to identify a subset of the human genome where these processes can be observed minimizing the effects of selection; that is, a subgenome comprising, to the best of our knowledge, only neutral DNA. One way to achieve this is to focus on so-called ancestral repeats. These are particular sequences that inserted in the genome of our ancestors long ago and can still be traced scattered across the human genome and the genomes of other species sharing those ancestors. Even though some of these sequences acquired a function over time, most of them never did—so by and large change unfolds in them without positive or negative consequences for adaptation. An alternative and complementary way to identify a neutral subgenome is to remove from the human genome all sequences that are currently annotated as having a known or putative function (e.g., sequences coding for proteins within genes and sequences that, proximal or distal to genes, participate in regulating their transcription levels) as well as all sequences classified as repetitive—including the ancestral repeats considered above. In our studies we often refer to the resulting subgenomes as AR (for Ancestral Repeats) and NCNR or NFNR (for Non-Coding Non-Repetitive, or Non-Functional Non-Repetitive). These "models" of neutral behavior are by no means universally accepted, but have been used repeatedly and with success over the years.

The next step is to consider genomic alignments restricted to a neutral subgenome. Comparing aligned genomes is what allows us to locate change events such as nucleotide substitutions, small insertions and deletions, and repeat number changes at microsatellites. Notably, the genomes comprised in the alignments determine the evolutionary radius at which an analysis is performed. In our studies we focus mostly on alignments of the reference human genome with the references genomes of other primates (e.g., chimpanzee, orangutan, rhesus macaque) and on alignments of multiple human genomes.

Once events due to various mutagenic processes are recorded within a neutral subgenome, one can proceed to compute their rates. This requires defining intervals along the human genome to perform the calculation. One approach, albeit certainly not the only one, is to create a partition of the human genome in non-overlapping windows of a given size—most but not all of the studies reviewed in this article utilize such a partition. Considering events that fall in the neutral portions of each such window, one can then employ appropriate models to produce rate estimates (e.g., the models proposed by [3] or [4] for substitutions; the equation proposed by [5] for mutability in microsatellite repeat numbers). In the same windows, we also retrieve or derive our genomic landscape features from publicly available data (e.g., genome annotations, large consortia studies, published information from different types of high-throughput experiments). Notably, the window size determines the scale at which the analysis is performed.

The scale determined by the choice of window size and the evolutionary radius determined by the choice of aligned genomes, are critical parameters. Characteristics and interdependencies of mutagenic processes, as well as their associations with the genomic landscape, may change with these parameters. There are good practical reasons to focus on some ranges of scales and radii; for instance, very small windows may bring to the fore patterns that are "averaged out" at larger scales, but will decrease the accuracy of rate estimation and force us to eliminate from the analysis landscape variables generated by low resolution experimental techniques. Similarly, some patterns may concern only the most recent segments of our evolutionary history, but a very small evolutionary radius will complicate the identification of certain events for technical reasons (e.g., substitution rates estimated from human-chimpanzee alignments are more affected by ancient polymorphisms than those estimated from human-macaque alignments [6]). Within viable ranges though, repeating analyses at varying parameter values can produce important insights, separating recurrent patterns from patterns that are specific to some genomic scales and evolutionary radii.

It is important to stress that acquiring and processing alignments and data from large public repositories to produce usable mutation rates and landscape features is in itself a delicate and complicated task; a lot of time and effort is spent to generate quality data even before any statistical analysis begins.

3 One Mutation Rate at a Time: Regression Analyses

Several of the studies we conducted over the years investigated the dependence of individual mutagenic processes on genomic landscape features. These analyses were performed utilizing the traditional regression toolkit. Statistical units are windows, on each of which we consider values for a mutation rate (the response) and multiple genomic features (the candidate predictors). Using windows with sizes ranging from 0.1 to 10 Mb (megabases) we have tens of thousands to a few hundred windows available to investigate a candidate predictor pool of $\sim$10 to over 50 features—depending on the study. We are therefore not faced with a so-called large-p-small-n setting. However, window size does create a trade off between autocorrelation and over-fitting. Windows of smaller size are abundant, but behave less like independent statistical units and may have less accurate mutation rate estimates (see above). In contrast, windows of larger size are too few to reliably parse the effects of dozens of predictors and may average out some critical signals, but do not show autocorrelations and have more accurate mutation rate estimates. We often struck a balance using a 1 Mb scale for our main analysis (see also [7]) but reporting also secondary analyses at varying window sizes and discussing robust vs. scale-specific outcomes.

While traditional regression tools can be applied straightforwardly, the analysis pipeline must be rigorous and combine them systematically and wisely because the data are complex. We usually start with several filtering, preprocessing and

transformation steps to deal with anomalous loci, regularize response and predictor distributions, and reduce the candidate predictors at the outset if warranted. Next, we employ rigorous variable selection and model building procedures combining best-subset type algorithms with multiple testing adjustments and variance inflation factors screens to account for sizeable interdependencies in the predictor pool. We also use multiple diagnostics tools throughout the model building iterations—including residual-based model and autocorrelation diagnostics, and case diagnostics for outliers and influence. To weigh the role of each individual predictor retained in the final models we often use the Relative Contribution to Variance Explained (RCVE), a coefficient similar to the partial R^2.

Our results indicate that features of the genomic landscape explain a substantial share of the local variability in mutation rates; ∼30 % for insertions and deletions, as high as ∼50 % for nucleotide substitutions and ∼80 % for nucleotide substitutions at particular sites called CpG sites [8, 9].

Some landscape features are very powerful predictors for insertions, deletions and nucleotide substitutions—suggesting mutational contexts and biochemical mechanism shared by multiple mutagenic processes. Among them GC content, which appears to have a bi-phasic effect on mutation rates, and distance from telomeres, with rates increasing non-linearly near telomeres (the final regression models contain significant quadratic terms; see for instance Fig. 1 adapted from [9]). Other strong shared predictors are positioning on the X chromosome, which decreases rates likely due to decreased replication errors (X undergoes fewer replications because it is present in only one copy in the male germline), and

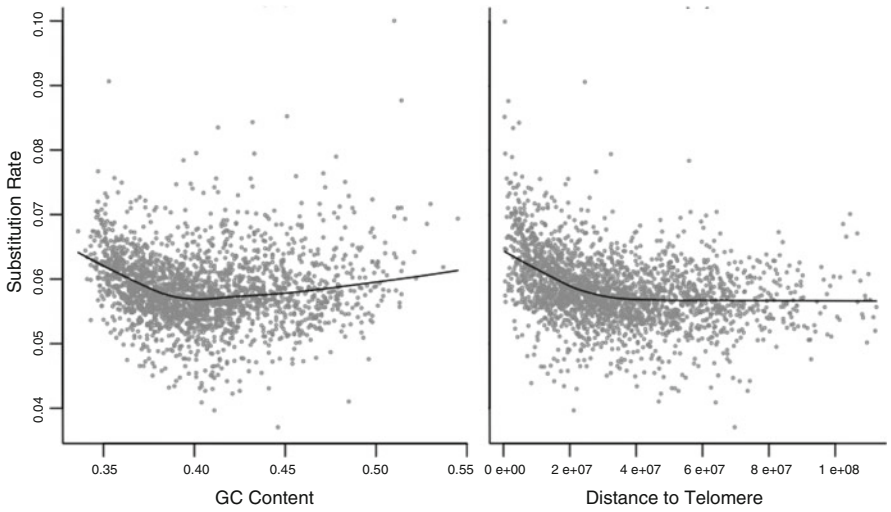

Fig. 1 Scatter plots of human-macaque substitution rates against human GC content (*left*) and distance to telomeres (*right*). The rates are computed in 1 Mb windows using the AR neutral subgenome. Lowess smoothers superimposed to the plots reveal curvature—consistent with significant quadratic terms in our regression models. Adapted from [9]

recombination in the male germline, which appears to have more direct and sizeable mutagenic effects than recombination in the female germline.

Interestingly though, the landscape correlates of different mutagenic processes also show notable differences—suggesting differences in the underlying mechanisms. For instance, even though insertions and deletions are often lumped together as "indels" in genomic studies, female recombination seems to play a role in increasing rates for insertions but not deletions, and in terms of transpositional context abundance of SINEs (Short INterspersed Elements) appears to increase insertion rates and decrease deletion rates, while abundance of LINEs (Long INterspersed Elements) appears to only increase deletions rates [8]. Another interesting finding was that propensity of a stretch of nuclear DNA towards nucleotide substitutions is at least partially conserved during evolution; substitution rates computed at orthologous regions from mouse-rat and dog-cow alignments are significant predictors of those computed from primate alignments. We also found that the bi-phasic relationship between substitution rates and GC content could be explained by different mutational patterns of CpG and non-CpG sites [9].

We also used regression analysis to investigate the determinants of mutability of microsatellite repeats, obtaining models that explain in excess of 90% of its variability [10]. However, the main drivers in such models are not genomic landscape features but intrinsic features of the microsatellite loci, such as the motif being repeated, its size and the number of repetitions (mutability in repeat number increases with the repeat number itself with a very distinct and informative pattern; see also below). Some landscape features though do appear to affect changes in microsatellite repeat numbers. For instance, microsatellites are most mutable on the Y chromosome and least mutable on the X chromosome. Microsatellite mutability is also increased by co-location with *Alu* repetitive elements (a particular subclass of SINEs). Recombination rates appear to gain a significant positive effect on microsatellite mutability as one reduces the window size—suggesting that the genomic landscape may in fact play a substantial role, but at scales smaller than those observed for other mutagenic processes. This remains an open question, as smaller scales could not be investigated with the data resolution currently available to us [10].

The pattern that links a microsatellite's propensity to add/remove repetitions of its motif to the number of repetitions itself has received much attention over the years. In fact, a long-lasting controversy concerns the existence of a threshold; short tandem repeats are hypothesized to undergo a transition when they reach a critical number of repetitions and "become" hyper-mutable microsatellites.

We explored this hypothesis with further in silico analyses using human polymorphism (as opposed to primate mutations) and with specialized in vitro experimental assays conducted by our collaborators in the Eckert group at The Pennsylvania State University Medical campus in Hershey [11]; (Ananda et al. 2013). For the in silico analyses here we did not utilize a system of non-overlapping windows; running a number of scripts and algorithms on reference and re-sequenced human genomes, we identified repetitive sequences, binned them based on their "typical" repeat number and computed the polymorphism incidence for each bin. We did this

also separating the sequences by motif size (mono-, di-, tri-, and tetra-nucleotide), and within motif sizes by the motif itself (e.g., an A mono-nucleotide, an AC di-nucleotide, etc.)—which allowed us to investigate the relationship between polymorphism incidence and repeat number taking into account the other intrinsic features.

In [11], where the re-sequencing data concerned 48 individuals and 10 specific genomic regions (from the HapMap-ENCODE resequencing and genotyping project [12] [13]), we found very strong evidence for a threshold. In [14], where the re-sequencing data concerned the whole genomes of 179 individuals from the 1000 Genomes Pilot 1 Project [15], we found that tandem repeats and microsatellites (i.e. repetitive sequences before and after the repeat number threshold) differ not only in their absolute polymorphism levels but also in the way these depend on repeat number. More specifically, we fitted segmented regression models on the logarithm of polymorphism incidences using the iterative algorithm in [16] as implemented in [17]. We observed that before the threshold these are very low and show a steep exponential growth, while after the threshold they are much higher but show a slower exponential growth (significant change in slope according to Davies test; see Fig. 2 adapted from [14]).

The thresholds we were able to infer from these biphasic regimes are 9, 5, 4 and 4 repeats, respectively, for mono-, di-, tri-, and tetra-nucleotide microsatellites. The inference for mono-nucleotide microsatellites had weaker statistical evidence

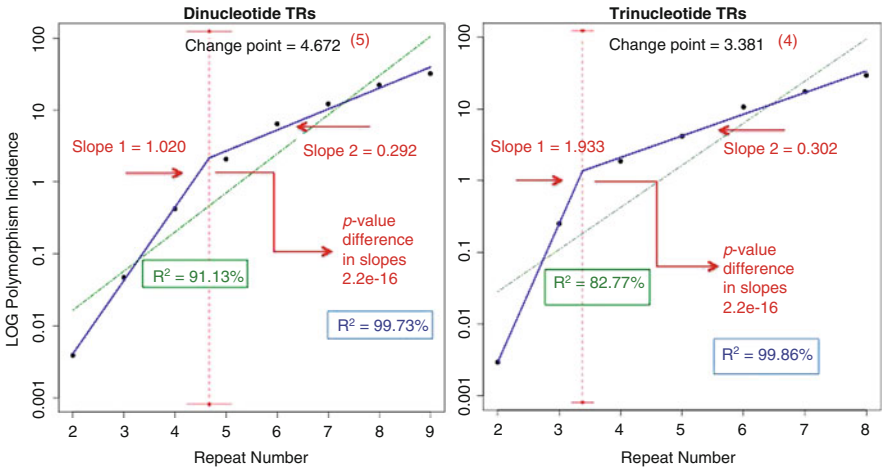

Fig. 2 Scatter plots of log polymorphism incidence against repeat number, with fits from regular and segmented regression for di- (*left*) and tri- (*right*) nucleotide tandem repeats. Annotations on the figure indicate R^2 values, location and 90% confidence interval for the change point, slope estimates before and after the change point and p-values for their difference. Similar information for mono- and tetra-nucleotide tandem repeats can be found in [14]. In each case the thresholds was defined as the smallest integer larger than the segmented regression change point. Adapted from [14]

than those for microsatellites with larger motif sizes—likely because current re-sequencing data do not allow us to produce reliable polymorphism measures for sequences comprising more than ten repetitions.

We conclude this section with a recent regression study [18] that utilizes again a large number of genomic features as predictors. This time though the responses are not mutation rates derived from primate comparisons or human re-sequencing information. We considered so-called common fragile sites—unstable regions of the DNA that are prone to gaps and breaks during replication and often host viral integrations and chromosomal rearrangements in cancer. A class of these sites can be induced through a special cellular treatment and were located genome-wide with an experimental assay, along with a measurement of fragility for each such site (breakage frequency) [19].

Among the sites identified in [19], we focused on a subset of 73 that are best characterized experimentally (see also Lukusa and Fryns 2008) and reside on autosomes (the non-sex chromosomes). The resolution of the assay employed in the genome-wide screen is fairly low and these sites, which are broadly distributed and cover approximately 15% of the autosomal genome, vary from less than 1 to ∼25 Mb in length. We formed a control set of non-fragile sites (124 regions covering ∼35 % of the autosomal genome and varying in length from ∼1.5 to 33 Mb), and for each of the fragile and non-fragile sites computed an array of 54 genomic landscape features.

Because of very strong interdependencies among some of the landscape features, we used hierarchical clustering based on Spearman's correlation to identify tight groups of predictors and select one (biologically meaningful) "representative" for each such group. After this preliminary variable selection, which allowed us to focus on 19 predictors, we used logistic regression on fragile and non-fragile sites, and standard regression on fragile sites alone, to investigate factors affecting, respectively, location and degree of fragility. The regression analysis pipeline was similar to the one described above (further variable selection and model building with best-subset type algorithms, multiple testing adjustments and variance inflation factors screens; iterative use of several diagnostics tools; quantification of the contribution of each individual predictor retained in the final models)—with the needed shifts from variance to deviance calculations for logistic regression runs.

We found that features of the genomic landscape are excellent predictors of both the location of fragile sites and their breakage frequency, allowing us to shed some light on the molecular mechanisms shaping DNA instability. The deviance explained when contrasting fragile and non-fragile sites with logistic regression is ∼77%, and the share of variability in breakage frequency at fragile sites explained with standard regression is in excess of 43%. Fragile sites reside predominantly in so-called G-negative chromosomal bands, which reflect chromatin structure and base composition patterns, and away from centromeres. They also appear to abound in *Alu* elements and to have high DNA flexibility. The breakage frequency of fragile sites, too, increases in G-negative chromosomal band and away from centromeres; these significant predictors are shared by the final logistic and standard regression models. Moreover, breakage frequency appears to increase when fragile sites

co-locate with evolutionarily conserved chromosomal breakpoints and decrease when fragile sites co-locate with CpG islands (these are regulatory regions with a distinctly high density of CG dinucleotides).

4 A Multivariate View: Principal Components and Canonical Correlations

After having analyzed nucleotide substitution, small insertion, small deletion, and microsatellite mutability rates individually as a function of genomic landscape features through regressions, we shifted to a multivariate perspective [1]. We went back to a partition of the human genome in non-overlapping windows, considering for each mutation rate estimates (based on primate comparisons) along with 15 genomic features. On these data, we used Principal Components (PC) to characterize the co-variation structure of mutation rates, and Canonical Correlations (CC) to characterize the multivariate associations between rates and genomic features. We implemented linear and non-linear versions of both analyses, employing Gaussian kernels for the latter. These analyses offered yet richer insights on the intricate web of interdependencies connecting different aspects of mutagenesis and the genomic landscape in which it unfolds.

One of our most robust findings, which we were able to replicate for different choices of evolutionary radius, genomic scale and neutral subgenome, is that nucleotide substitutions, small insertions and small deletions have a strong and positive linear covariation—their rates have similar standardized loading on the first linear PC. In contrast, microsatellite mutability varies orthogonally loading exclusively or almost exclusively on the second linear PC (see Fig. 3 adapted from [1]). For rates computed from human-orangutan comparisons in 1-Mb windows and using the AR subgenome, the first PC explained $\sim$54% and the second $\sim$25% of the overall data variability.

The first three linear CC pairs, which carry significant correlations ($\sim$0.73, $\sim$0.53 and $\sim$0.33, respectively), nuance the associations of co-varying substitution, insertion and deletion rates with combinations of genomic features. For instance, the first pair indicates that all three rates are elevated by high content of GC, SINE elements and protein-coding sequences. The second pair indicates that, orthogonally, substitutions may be elevated by high male recombination and proximity to telomeres. The third pair indicates that, orthogonally to the first two signals, deletion rates may be elevated by low content of GC, SINE elements and protein-coding sequences (a scenario opposite to that of the first pair) in combination with low female recombination, etc.

Notably, and in agreement with the regression analysis summarized above, at smaller genomic scales we found some evidence of association between microsatellite mutability and the other mutation rates, as well as the genomic landscape.

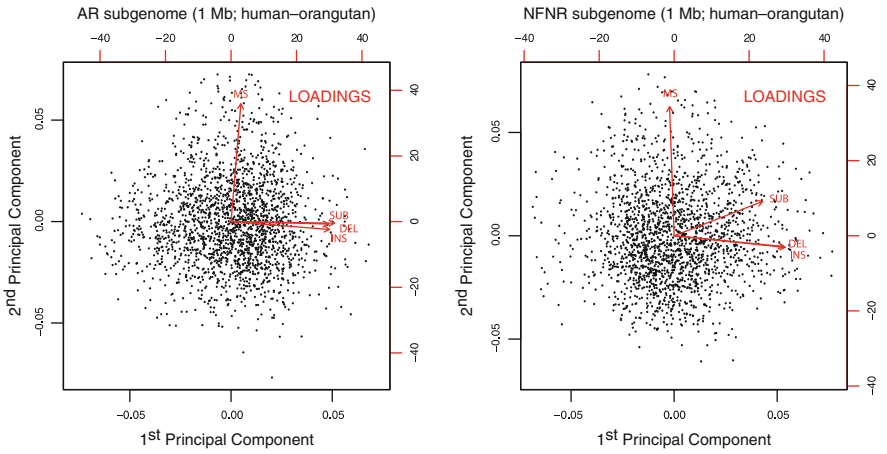

Fig. 3 Biplots of the first two linear PC components for human-orangutan mutation rates computed in 1 Mb windows using the AR (*left*) and the NFNR (*right*) neutral subgenome. The vectors labeled INS, DEL, SUB, and MS depict loadings for insertion, deletion, nucleotide substitution and mononucleotide microsatellite mutability rates, respectively. Adapted from [1]

Also the non-linear analyses offered interesting insights. Perhaps the most relevant, at least for the summary provided here, concerns localization. To understand what kernel PC could add in terms of interpretation, since this analysis does not produce loadings, we regressed the scores associated with its leading component on the scores from the first two linear PCs. This regression had a high R^2 (about 75%). However, when we looked at the genomic positions of the windows carrying the largest residuals, we found that loci where the non-linear leading signal is poorly recapitulated by linear ones correspond to extremely high mutation rates concentrated near the telomeres of autosomes, and extremely low mutation rates concentrated on chromosome X.

Linear and non-linear PC and CC analyses do not explicitly utilize the fact that our statistical units (the windows) are contiguously positioned along chromosomes. This evidence of a "geographical" characterization, which was also reminiscent of some landscape effects glanced through our previous regression analyses, naturally led us toward methodology capable of directly capturing and exploiting 1-dimensional spatial patterns in mutation rates.

5 The Geography of Genome Dynamics: Hidden Markov Models

Given the evidence that geography may be critical to genome dynamics, our the next step was to represent the observed variation and covariation in mutation rates as generated by "hidden" underlying mutational states that alternate along the genome

[20]. We performed some preliminary filtering and transformation steps, aimed at eliminating windows with scarce alignments or unreliable rate values (due to paucity of neutral subgenome within them) and at regularizing the marginal distributions of the rates. After these, we fitted Multivariate Gaussian Hidden Markov Models (HMMs) to our vector of four mutation rates (nucleotide substitutions, small insertions, small deletions and microsatellite repeat number changes) measured in non-overlapping windows along the genome.

HMMs have a long tradition in genomics and bioinformatics, where they have been used in algorithms to find and predict genes, and more recently to produce segmentations of the genome based on so-called epigenomic signatures (see for instance [21–23]). In full generality, an HMM is a tool for modeling an ordered sequence of observable measurements produced by an underlying process that alternates among unobservable ("hidden") states with a Markovian path dependence (if its order is 1, the state in each position depends only on the immediately preceding one). Based on the observable measurements one (1) infers model parameters (i.e. prior and transition probabilities for the hidden states, and state specific distributional parameters for the measurements) using an Expectation–Maximization type algorithm such as Baum–Welch; and (2) reconstructs the most likely sequence of hidden states using an algorithm such as Viterbi (see [24,25]).

The main differences between our application and prior applications of HMMs to genomic data are the simultaneous use of multiple continuous variables (the four rates) and the scale—ours is much larger than those previously used (individual bases or small intervals up to 100 bp). Notably, in addition to being the appropriate one to investigate some types of change according to past literature, and the viable one to compute some of the variables of interest, our much larger scale gave us the advantage of nimble computations (when using 1 Mb we had to handle only 2,500–2,700 windows) and excellent interpretability. We were indeed able to paint a broad-brush, meaningful picture of the geography of genome dynamics.

The Bayesian Information Criterion suggested the existence of six hidden mutational states underlying our observed sequence of 4-dimensional rate vectors. Moreover, six states were indeed sufficient to capture signals noted in our previous studies concerning microsatellite mutability and chromosome X, and using more states did not produce solutions with richer biological interpretations.

The six states are beautifully characterized in terms of rates profiles. They resonate with our findings in [1] and allow us to further nuance mutational patterns. We have an autosomal cold state for insertions, deletions, and substitutions (*Cold auto* in the following) that contrasts a hot state for all three processes (*Hot*). In addition we have a very cold state that covers only and almost all of chromosome X (*Cold X*), and two warm states where only insertions (*Ins warm*) or deletions and substitutions (*Del/sub warm*) appear enhanced. Finally, we have a state where microsatellite mutability—and no other process—is very active (*Microsat*).

Interestingly, about 15% of the genome is hot (in *Hot* or *Microsat*). About 45% is warm (in *Ins warm* or *Del/sub warm*), and about 40% is cold (in *Cold auto* or *Cold X*). Also interestingly, the contiguous segments of hot states tend to be short (median length 1–3 windows), those of warm states of medium length (median length 5–6

windows), and those of cold states even longer (median length 9 on autosomes; chromosome X is an almost entirely uninterrupted very cold segment). In other words, in terms of the underlying Markovian structure, "transitioning out" appears to be easier the hotter the state.

Concerning location, the short segments of *Hot* and the medium length segments of *Del/sub warm* tend to co-locate near the telomeres of autosomes. However, there are exceptions this trend. Interestingly, these exceptions can be traced back to what is known of the evolutionary history of human chromosomes; we found instances of hot segments near current centromeres that were near telomeres in ancestral genomes. Even though they "moved", these segments appear to still retain the mutational behavior typical of their previous environs. The medium length segments of *Ins warm* co-occur with *Hot* and *Del/sub warm* near the telomeres, but can also extend inward towards the center of chromosomes. The long segments of *Cold auto* tend to be removed from the telomeres, almost suggesting that a certain distance from the ends may be necessary for the DNA to "cool off" ($\geq$60 Mb on most autosomes; the short arms of many smaller ones, e.g. chromosomes 16, 17, 18, and 21, are almost entirely composed of hot or warm states). *Cold X* covers exclusively chromosome X. Finally, the enhanced microsatellite mutability state *Microsat* manifests itself in very short segments (mostly individual windows) interspersed across the genome, consistent with our previous observation that genomic landscape has little effect on microsatellite mutability at large scales (see Fig. 4 from [20]).

As in our previous studies, we also considered a long list of genomic landscape features to investigate their relationships with our mutational states. For each feature and each state we computed sign and significance of the association—the latter through a parametric null bootstrap in which we simulated a large number of genome "tilings" using the Multivariate Gaussian HMM parameter estimates, but unrelated to the actual rates on which the fit was performed.

We comment here on a few of the many possible interpretations afforded by this exercise. Contrasting the genomic landscape of the hot and cold autosomal states, the concauses of enhanced mutability in the GC-rich *Hot* appear to comprise enhanced male recombination (as associated with biased gene conversion or for its own mutagenic effects), aberrant repair near telomeres, the open chromatin architecture of these regions (which may make them more susceptible to change), and, given their high levels of transcription activity, transcription-mediated errors. Conversely, *Cold auto* inhabits GC-poor, low recombination inner regions of autosomes, with compact chromatin and lower transcription levels. Interestingly, the genomic landscape of *Microsat*, except for a few mild associations, appears by and large non-descript relative to genome baselines. This finding, too, is consistent with the notion that intrinsic microsatellite features (repeat number, motif size and motif composition) affect the mutability of these loci more than the genomic landscape—at least when observed at large scales.

In addition to genomic landscape features, we investigated the associations between our mutational states and some important functional annotations of the genome (also here, significance was assessed using a parametric null bootstrap). For instance, we considered annotated genes as well as regulatory elements annotations

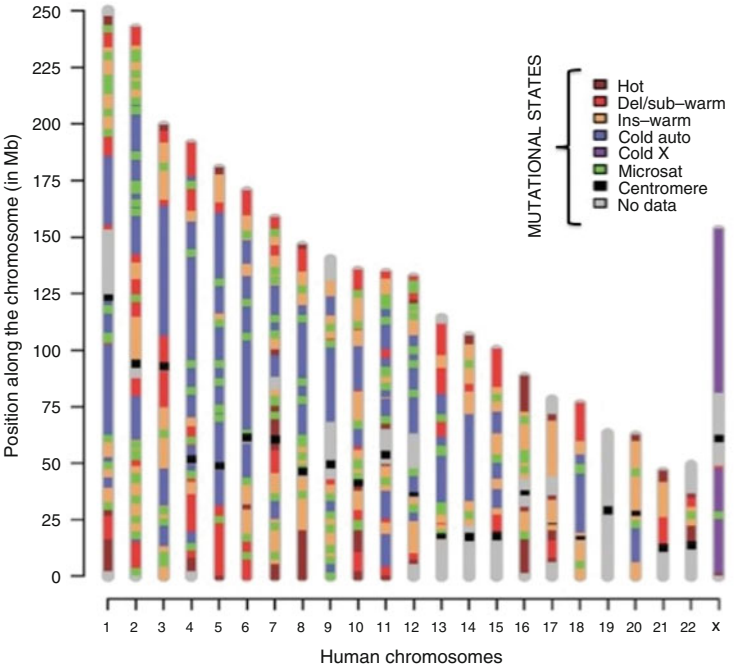

Fig. 4 Genomic locations of segments belonging to six mutational states obtained fitting Multivariate Gaussian HMMs on human-orangutan mutation rates computed in 1 Mb windows using the AR neutral subgenome. Bars represent human chromosomes, reported in scale and with positions indicated on the vertical axis (*gray regions* correspond to windows excluded from the analysis due to assembly gaps or data pre-processing filters; see [20] and its supplemental information for more details). Adapted from [20]

generated by the ENCODE project [23] (e.g. predicted binding locations for transcription factors, promoters in the transcription start sites of genes, enhancers). Among other things, we found that genes and functional elements governing the regulation of gene expression are enriched in *Hot*. This is a somewhat counterintuitive but very intriguing result; likely due to their open chromatin architecture, these regions host at the same time much of the function and much of the mutagenic action in the genome.

The geographical characterization of genome dynamics we obtained in this study is fairly robust to genomic scale (i.e. the choice of window size) as well as evolutionary radius. In particular, we could retrace its general outline repeating the analysis on rates computed within humans through data from the 1000 Genomes Project [15]. In addition to providing insights on the mechanisms of mutagenesis, our segmentations could have important practical applications. For instance, they could be used to benchmark signals in algorithms that predict the location of functional elements in the genome. These algorithms often assess whether a sequence has changed less or more than one might expect, and thus infer that it

may be under constraint (i.e. under negative selection) or usefully "accelerated" (i.e. under positive selection) because it has a function in the genome. In this approach, an expected baseline for change can be constructed taking into consideration the mutational state in which the sequence resides. This would allow one to create a local background that leverages information on multiple and interconnected types of change—and thus to improve algorithms performance by reducing both false negative and false positive rates.

A similar reasoning applies to algorithms that screen genomes for disease-related variants; when differences are detected between the genomic sequences of individuals affected by a disease and those of control individuals, knowing in what mutational state the differences occur can help evaluating the likelihood that they may indeed be relevant for the disease. Our segmentations, which have been made publicly available through the UCSC Genome Browser and the Galaxy Portal at Penn State, could therefore aid attempts to meaningfully mine biomedical data.

In this respect, of particular interest may be applications to cancer genomics. Our segmentations are based on neutral mutation rates measured at various genomic scales and evolutionary radii in the *germline*. They can certainly be informative when screening germline mutations that may predispose individuals towards one or more types of cancers. However, whether or how they may assist in the characterization of mutations and broader chromosomal changes that occur at the *somatic level* during oncogenesis, is a complex question that we have not yet fully explored at this point in time. Some studies support a positive relationship between the propensity of certain regions of the nuclear DNA to mutate in the germline, and their propensity to mutate somatically in cancer (e.g., [26]). However, and intriguingly, other lines of evidence point towards a negative relationship. For instance, while our study depicts a very "cold" chromosome X in terms of germline mutations, [27] found evidence of enhanced somatic mutations on this sex chromosome in cancer. Importantly, depressed germline mutation rates may to a large extent be due to reduced replication errors, since chromosome X spends less time than autosomes in the male germline (see [8,20,28] for more details).

Conclusions

The human genomics studies reviewed in this article demonstrate how established statistical techniques—ranging from regression, to multivariate analysis, to the modeling of latent structure—intelligently combined and rigorously applied to large and complex data sets generated by a variety of high-throughput experimental techniques, can help us address important scientific questions. They also demonstrate the power of, and need for, interdisciplinary collaborations to advance contemporary genome sciences. Importantly, these collaborations also serve as the breeding ground for novel developments in statistical methodology.

(continued)

One instance among many is the development of methodology for under-sampled data. In [18] we could utilize 73 fragile sites and a control group of 124 non-fragile sites—the sample sizes were therefore relatively small compared to the number of genomic features involved in the regressions; even after pre-selecting ∼20 predictors, we operated with 10 or less "observations" per feature under analysis. In other studies conducted by our group (e.g., [29]) the problem was even more marked, with the number of available observations smaller than the number of features of interest. For some types of genomic data, e.g. those generated by genome-wide transcription profiling technologies such as microarrays or RNA-seq, the difference between number of available observations and number of features can be of several orders of magnitude (tens or at most hundreds of individual samples vs. tens of thousands of transcripts).

We have been directly involved in the development of sufficient dimension reduction methodology applicable to under-sampled settings [30], as well as in the development of an "all-purpose" bootstrap-like approach that permits to artificially augment data with minimal distortions prior to the application of any statistical technique [31].

Acknowledgments We wish to thank G. Ananda, A. Fungtammasan, Y.D. Kelkar, E.M. Kvikstad, P. Kuruppumullage Don and S. Tyekucheva—the brilliant and hard working graduate students that took the lead and collaborated with each other in the studies reviewed in this article. We also wish to thank our collaborators in the Center for Medical Genomics of The Pennsylvania State University, in particular K. Eckert whose group performed experimental work critical for our studies of microsatellites and common fragile sites. Finally, we are in debt to a reviewer of this manuscript who offered useful and interesting comments on our work. Our research over the years has been supported by various sources; particularly important for the studies reviewed here were awards from the NSF (DBI 0965596) and the NIH (General Medical Sciences R01 GM087472-01).

References

1. Ananda, G., Chiaromonte, F., Makova, K.D.: A genome-wide view of mutation rate co-variation using multivariate analyses. Genome Biol. **12**(3), R27 (2011)
2. Kvikstad, E.M., Makova, K.D.: The (r)evolution of SINE vs LINE distributions in primate genomes: Sex chromosomes are important. Genome Res. **20**, 600–613 (2010)
3. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In: Munro, H.N. (ed.) Mammalian Protein Metabolism, pp. 21–123. Academic, New York (1969)
4. Hasegawa, M., Kishino, H., Yano, T.: Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**(2), 160–74 (1985)
5. Webster, M.T., Smith, N.G., Ellegren, H.: Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc. Nat. Acad. Sci. USA **99**, 8748–8753 (2002)
6. Li, W.H., Yi, S., Makova, K.D.: Male-driven evolution. Curr. Opinion Genetics Develop. **12**, 650–656 (2002)

7. Gaffney, D.J., Keightley, P.D.: The scale of mutational variation in the murid genome. Genome Res. **15**, 1086–1094 (2005)
8. Kvikstad, E.M., Tyekucheva, S., Chiaromonte, F., Makova, K.D.: A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS Comput. Biol. **3**(9)e176, 1772–1782 (2007)
9. Tyekucheva, S., Makova, K.D., Karro, J., Hardison, R.C., Miller, W., Chiaromonte, F.: Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. **9**(4), 76 (2008)
10. Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., Makova, K.: The genome-wide determinants of microsatellite evolution. Genome Res. **18**, 30–38 (2008)
11. Kelkar, Y.D., Strubczewski, N., Hile, S.E., Chiaromonte, F., Eckert, K.A., Makova, K.D.: What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol. Evolu. **2**, 620–635 (2010)
12. International HapMap Consortium: The International HapMap Project. Nature **426**(6968), 789–96 (2003)
13. International HapMap Consortium: A haplotype map of the human genome. Nature **437**(7063), 1299–320 (2005)
14. Ananda, G., Walsh, E., Jacob, K.D., Krasilnikova, M., Eckert, K.A., Chiaromonte, F., Makova, K.D.: Distinct mutational behaviors distinguish simple tandem repeats from microsatellites in the human genome. Genome Biol. Evolu. **5**(3), 606–620 (2012)
15. 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature **467**(7319), 1061–73 (2010)
16. Muggeo, V.: Estimating regression models with unknown break-points. Stat. Med. **22**(19), 3055–71 (2003)
17. Muggeo, V.: Segmented: an R package to fit regression models with broken-line relationships. R. News. **8**, 20–25 (2008). http://cran.r-project.org/doc/Rnews/
18. Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K.A., Makova, K.D.: A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Res. **22**, 993–1005 (2012)
19. Mrasek, K., Schoder, C., Teichmann, A.C., Behr, K., Franze, B., Wilhelm, K., Blaurock, N., Claussen, U., Liehr, T., Weise, A.: Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. Int. J. Oncol. **36**, 929–940 (2010)
20. Kuruppumullage, D.P., Ananda, G., Chiaromonte, F., Makova, K.D.: Segmenting the human genome based on states of neutral genetic divergence. Proc. Nat. Acad. Sci. USA **110**(36), 14699–14704 (2013)
21. Majoros, W.H., Pertea, M., Antonescu, C., Salzberg, S.L., Glimmer, M.: Exonomy and unveil: three ab initio eukaryotic gene finders. Nucleic Acids Res. **31**(13), 3601–3604 (2003)
22. Ernst, J., et al.: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature **473**(7345), 43–49 (2011)
23. Dunham, I., ENCODE Project Consortium, et al.: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
24. Taramasco, O., Bauer, S.: *R package RHmm*. http://CRAN.project.org/package=RHmm (2007)
25. Eddy, S.R.: What is a hidden Markov model? Nature Biotechnol. **22**(10), 1315–1316 (2004)
26. Hodgkinson, A., Chen, Y., Eyre-Walker, A.: The large scale distribution of somatic mutations in cancer. Hum. Mut. **33**(1), 136–143 (2012)
27. Davoli, et al.: Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape in the cancer genome. Cell **155**(4), 948–962 (2013)
28. Makova, K.D., Li, W.H.: Strong male-driven evolution of DNA sequences in humans and apes. Nature **416**(6881), 624–626 (2002)
29. Carrel, L., Park, C., Tyekucheva, S., Dunn, J., Chiaromonte, F., Makova, K.D.: Genomic environment predicts expression patterns on the human inactive X chromosome. PLoS Gen. **2**(9) e151, 1477–1486 (2006)

30. Cook, R.D., Li, B., Chiaromonte, F.: Dimension reduction in regression without matrix inversion. Biometrika **94**, 569–584 (2007)
31. Tyekucheva, S., Chiaromonte, F.: Augmenting the bootstrap to analyze high dimensional genomic data (invited discussion article). Test **17**, 1–18 (article) and 47–55 (rejoinder) (2008)
32. Chiaromonte F., Yang S., Elnitski L., Bing Yap V., Miller W., Hardison R.C.: Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. Proc. Natl. Acad. Sci. USA. **98**(25), 14503–14508 (2001)
33. Hardison R.C., Roskin K.M., Yang S., Diekhans M., Kent J.W., Weber R., Elnitski L., Li J., O'Connor M., Kolbe D., Schwartz S., Furey T.S., Whelan S., Goldman N., Smit A., Miller W., Chiaromonte F., Haussler D.: Co-variation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. Genome Res. **13**, 13–26 (2003)
34. Yang S., Smit A.F., Schwartz S., Chiaromonte F., Roskin K. M., Haussler D., Miller W., Hardison R.C.: Patterns of insertions and their covariation with substitutions in the rat, mouse and human genomes. Genome Res. **14**, 517–527 (2004)
35. Hodgkinson, A., Chen, Y., Eyre-Walker, A.: The large scale distribution of somatic mutations 534 in cancer. Hum. Mut. **33**(1), 136–143 (2012)
36. Lukusa T., Fryns J.P.: Human chromosome fragility. Biochim Biophys Acta. **1779**, 3–16 (2008)

Information Theory and Bayesian Reliability Analysis: Recent Advances

Nader Ebrahimi, Ehsan S. Soofi, and Refik Soyer

1 Introduction and Overview

As noted by Ebrahimi et al. [14] in a recent review, information theory provides measures for handling diverse problems in modelling and data analysis in a unified manner. Information theory statistics have been considered in reliability modelling and life data analysis; see [11, 12] for a review of such work. Information theory-based work in reliability can be grouped into three main areas as suggested in [12]. These include development of information functions for reliability analysis, information theory-based diagnostics and hypothesis tests for model building and measures that quantify the amount of information for prediction.

Since the seminal work of Lindley [20], information theory has played an important role in Bayesian statistics. The mutual information which is also known as Lindley's measure has been used by Bernardo [6] as the expected utility for the decision problem of reporting a probability distribution. It also has provided the foundation for the reference priors of Bernardo [7]. Other uses of Lindley's information have been in design problems; see for example [2] for a comprehensive review. An information processing rule has been defined in [31] using information measures and the Bayes rule has been shown as the optimal solution.

N. Ebrahimi
Division of Statistics, Northern Illinois University, DeKalb, IL, USA
e-mail: nader@math.niu.edu

E.S. Soofi
Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
e-mail: esoofi@uwm.edu.in

R. Soyer (✉)
Department of Decision Sciences, The George Washington University, Washington, DC, USA
e-mail: soyer@gwu.edu

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_6

As noted by Ebrahimi and Soofi [12] an area of Bayesian reliability analysis where information theory has been often used is the optimal design of life tests; see [9] and the references therein. Bayesian nonparametric entropy estimation and Bayesian estimation of information indices for lifetime modelling by Mazzuchi et al. [13, 21] have been another area of focus of information theoretic work.

In this paper we consider some recent advances in use of information theory in Bayesian reliability analysis. We present a range of information functions for reliability analysis, present their properties and discuss their use in addressing different issues in reliability. Our discussion focuses on use of Bayesian information measures in failure data analysis, prediction, assessment of reliability importance and optimal design of life tests. Below we present some preliminaries associated with Bayesian reliability analysis and introduce notation. Section 2 presents information measures such as mutual information that are used in reliability analysis and their properties. Parameter and predictive information concepts are considered and their properties are discussed in Sect. 3 with implications on Bayesian designs. Section 4 considers informativeness of observed failures and survivals from life tests and presents some new results for comparison. The notion of information importance is presented in Sect. 5 as an alternative measure of reliability importance of components of a system. Concluding remarks are given in the section "Conclusions".

1.1 Preliminaries

Reliability analysis deals with quantification of uncertainty about certain event(s) and making decisions. Typical issues of interest include: (i) if a component (or a system) performs its mission; (ii) if time to failure of a component (or a system) exceeds a specified (mission) time, and (iii) if mean time to failure exceeds a specified time. For example, if Y denotes lifetime of a component (or a system) then the event of interest is $Y > y$ where y is the specified mission time. The quantity

$$P(Y > y|\theta) = Prob\{Y > y|\theta\} \tag{1}$$

as a function of y is known as the reliability function, where θ is a parameter which can be a scalar or a vector. If $f(y|\theta)$ denotes the density function of the failure model for Y, then we can write the reliability function as

$$P(Y > y|\theta) = \int_y^\infty f(x|\theta)dx. \tag{2}$$

Decision problems that involve reliability assessment include design of life tests, system design via reliability optimization, and developing optimal maintenance strategies. In Bayesian reliability analysis, uncertainty about θ is described

probabilistically via the prior distribution $f(\theta)$. Prior to observing any data uncertainty statements about a future lifetime Y_ν is made using the prior predictive distribution

$$f(y_\nu) = \int f(y_\nu|\theta) f(\theta) d\theta. \tag{3}$$

Given n observations, $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ from $f(y|\theta)$, posterior inference for θ is obtained via the posterior distribution

$$f(\theta|\mathbf{y}) \propto f(\theta) f(\mathbf{y}|\theta). \tag{4}$$

The posterior predictive distribution for the future lifetime Y_ν is given by

$$f(y_\nu|\mathbf{y}) = \int f(y_\nu|\theta) f(\theta|\mathbf{y}) d\theta. \tag{5}$$

2 Information Functions for Reliability Analysis

Let Q be an unknown quantity of interest which can be a scalar or a vector. Q may be a parameter Θ such as failure rate or may represent a future outcome Y_ν such as component lifetime or $Q = (\Theta, Y_\nu)$. We denote the distribution of Q by F and its probability mass or density function by f.

The unpredictability of Q depends solely on the concentration of its distribution measured by an uncertainty function $\mathscr{U}(f)$. As pointed out by Ebrahimi et al. [14], two desirable properties of the uncertainty function are: (i) $\mathscr{U}(\cdot)$ is concave. (ii) $\mathscr{U}(f) \leq \mathscr{U}(f^*)$, where f^* is the pdf of the uniform distribution (the least concentrated model). An uncertainty function with these properties is Shannon entropy

$$H(Q) = H(f) = -\int f(q) \log f(q) dq. \tag{6}$$

The uncertainty about Q is measured by $H(Q)$ and $I(Q) = -H(Q)$ is information about Q; see [20].

Information provided by the data $\mathbf{y}$ about Q is measured by the entropy difference

$$\Delta H(\mathbf{y}; Q) = H(Q) - H(Q|\mathbf{y}), \tag{7}$$

where $H(Q|\mathbf{y})$ is obtained using the posterior distribution $f(q|\mathbf{y})$. In (7) $\Delta H(\mathbf{y}; Q)$ is referred to as *observed sample information* about Q and can be positive or negative.

The information discrepancy between $f(q|\mathbf{y})$ and $f(q)$ can also be measured by the Kullback–Leibler (KL) divergence

$$K[f(q|\mathbf{y}) : f(q)] = \int f(q|\mathbf{y}) \log \frac{f(q|\mathbf{y})}{f(q)} dq \geq 0, \tag{8}$$

where the equality holds if and only if $f(q|\mathbf{y}) = f(q)$ almost everywhere. The information discrepancy is a *relative entropy* which only detects changes between the prior and the posterior, without indicating which of the two distributions is more informative. It is invariant under all one-to-one transformations of Q.

Expected sample information measures are obtained by viewing the information measures as functions of data $\mathbf{y}$ and averaging them with respect to the distribution of $\mathbf{y}$. Conditional entropy of Q given $\mathbf{y}$ is defined as

$$\mathscr{H}(Q|\mathbf{y}) = E_{\mathbf{y}}\{H(Q|\mathbf{y})\} = \int H(Q|\mathbf{y}) f(\mathbf{y}) d\mathbf{y}. \tag{9}$$

The *conditional information* is then defined as $\mathscr{I}(Q|\mathbf{y}) = -\mathscr{H}(Q|\mathbf{y})$. It follows from the above that the expected sample information is given by

$$E_{\mathbf{y}}[\Delta H(\mathbf{y}; Q)] = H(Q) - \mathscr{H}(Q|\mathbf{y}) \geq 0, \tag{10}$$

which is always nonnegative.

The expected entropy difference and expected KL divergence provide the same measure, known as the *mutual information*

$$M(\mathbf{y}; Q) = E_{\mathbf{y}}\{\Delta H(\mathbf{y}; Q)\} = E_{\mathbf{y}}\{K[f(q|\mathbf{y}) : f(q)]\}. \tag{11}$$

Another representation of the mutual information, $M(\mathbf{y}; Q)$, is given by

$$M(\mathbf{y}; Q) = H(Q) - \mathscr{H}(Q|\mathbf{y}) = K[f(q, \mathbf{y}) : f(q) f(\mathbf{y})]. \tag{12}$$

These representations are in terms of the expected uncertainty reduction, and imply that the mutual information is symmetric in Q and $\mathbf{y}$. We note the following properties of mutual information:

1. $M(\mathbf{y}; Q) \geq 0$, where the equality holds if an only if Q and $\mathbf{y}$ are independent;
2. we can write $M(\mathbf{y}; Q)$ as

$$M(\mathbf{y}; Q) = H(Q) + H(\mathbf{y}) - H(Q, \mathbf{y});$$

3. the conditional mutual information is defined by $\mathscr{M}(\mathbf{y}; Q|S) = E_s[M(\mathbf{y}; Q|s)] \geq 0$, where the equality holds if an only if Q and $\mathbf{y}$ are conditionally independent;
4. $M(\mathbf{y}; Q)$ is invariant under one-to-one transformations of Q and $\mathbf{y}$.

For $Q = \Theta$, the expected sample information about the parameter, $M(\mathbf{y}; \Theta)$ is known as Lindley's measure; [20]. It is also referred to as the *parameter information*. Lindley's measure has been widely used in Bayesian optimal design. It was first considered by Stone [28] in the context of normal linear models. El-Sayyed [18] used Lindley's measure for information loss due to censoring in the exponential model and Polson [26] considered it in design of accelerated life tests.

3 Parameter and Predictive Information and Bayesian Designs

The predictive version of Lindley's measure is referred to as *predictive information*. For $Q = Y_\nu$, the expected information $M(\mathbf{y}; Y_\nu)$ is referred to as the predictive information; see for example, San Martini and Spezzaferri [27] and Amaral and Dunsmore [5]. Verdinelli et al. [29] proposed predictive information for optimal design of accelerated life tests with lognormal lifetimes.

Verdinelli [29] considered a linear combination of the parameter and predictive information as design criteria

$$U(\mathbf{Y}; \Theta, Y_\nu) = w_1 M(\mathbf{Y}; \Theta) + w_2 M(\mathbf{Y}; Y_\nu), \tag{13}$$

where $w_k \geq 0$, $k = 1, 2$ reflect the relative importance of the parameter and prediction. As noted by Ebrahimi et al. [15], since Θ and Y_ν are not independent quantities, $M(\mathbf{Y}; \Theta)$ and $M(\mathbf{Y}; Y_\nu)$ are not separable. The weights in the above do not take into account the dependence between the prediction and the parameter.

Taking the dependence between the parameter and prediction into account requires the joint information. Following [15], if we let $Q = (\Theta, Y_\nu)$ then the observed and expected information measures are given by $\Delta H[\mathbf{y}; (\Theta, Y_\nu)]$ and $M[\mathbf{Y}; (\Theta, Y_\nu)]$. The joint information measures enable us to explore the relationship between $M(\mathbf{Y}; \Theta)$ and $M(\mathbf{Y}; Y_\nu)$ as given by the following result.

Theorem 1 *Let $Y_1, Y_2, \cdots$ have distributions $f_{Y_i|\theta}$, $i = 1, 2, \cdots$ which, given θ, are conditionally independent, then*

1. $\Delta H(\mathbf{y}; \Theta) = \Delta H[\mathbf{y}; (\Theta, Y_\nu)]$;
2. $M(\mathbf{Y}; \Theta) = M[\mathbf{Y}; (\Theta, Y_\nu)]$;
3. $M(\mathbf{Y}; Y_\nu) \leq M(\mathbf{Y}; \Theta)$.

Proof of the theorem is given in [15]. From Part (1) of the Theorem, we note that information provided by the (observed) sample about the parameter is the same as joint information about the parameter and prediction. Part (2) of the theorem provides a broader interpretation of Lindley's information, namely expected information provided by the data about the parameter and for the prediction. The inequality in (3) is the Bayesian version of the information processing inequality of information theory. As suggested by Ebrahimi et al. [15], it may be referred to as the

Bayesian data processing inequality mapping the information flow $\mathbf{Y} \to \Theta \to Y_\nu$. We note that parts (2) and (3) of Theorem 1 are due to the decomposition:

$$M[\mathbf{Y}; (\Theta, Y_\nu)] = M(\mathbf{Y}; \Theta) + \mathscr{M}(\mathbf{Y}; Y_\nu|\Theta) = M(\mathbf{Y}; Y_\nu) + \mathscr{M}(\mathbf{Y}; \Theta|Y_\nu). \quad (14)$$

An immediate implication of Theorem 1 is that the design maximizing $M(\mathbf{Y}; \Theta)$ also maximizes sample information about the parameter and prediction jointly. However, such optimal design may not be optimal according to $M(\mathbf{Y}; Y_\nu)$. Similarly, the optimal design maximizing $M(\mathbf{Y}; Y_\nu)$ may not be optimal according to $M(\mathbf{Y}; \Theta)$.

When $Y_i,\ i = 1, 2, \cdots$ are not conditionally independent given θ, the information decomposition is given by

$$M[\mathbf{Y}; (\Theta, Y_\nu)] = M(\mathbf{Y}; \Theta) + \mathscr{M}(\mathbf{Y}; Y_\nu|\Theta) \geq M(\mathbf{Y}, ; \Theta) \quad (15)$$

where $\mathscr{M}(\mathbf{Y}; Y_\nu|\theta) > 0$ is the measure of conditional dependence which reduces to 0 in the conditional independent case. For the conditionally dependent case we have

$$M(\mathbf{Y}; Y_\nu) \leq M(\mathbf{Y}; \Theta) \Longleftrightarrow \mathscr{M}(\mathbf{Y}; \Theta|\mathbf{Y}_\nu) \geq \mathscr{M}(\mathbf{Y}; Y_\nu|\Theta). \quad (16)$$

We note that under strong conditional dependence the predictive information $M(\mathbf{Y}; Y_\nu)$ can dominate $M(\mathbf{Y}; \Theta)$, the parameter information.

4 Failures Versus Survivals

In a probe of the common belief that observing failures in life testing is always more informative than survivals, Abel and Singpurwalla [1] posed the following question:

> During the conduct of the test, what would you rather observe, a failure or a survival?

The answer to the question has practical implications. For example, if a failure is preferred for inference, then one may wait until a failure occurs or perhaps even induce a failure through an accelerated environment. Abel and Singpurwalla showed that the answer to the question depends on the inferential objective of the life test.

The authors considered an observation $y = y_0$ from the exponential model,

$$f(y|\theta) = \theta e^{-\theta y}$$

and assumed a gamma prior for θ with parameters α and β. They used Shannon entropy for measuring observed information utility

$$H(\Theta|y_0) = -E_{\Theta|y_0}[\log f(\theta|y_0)].$$

As the posterior distribution of Θ is gamma with $(\alpha + 1)$ and $(\beta + y_0)$ for the case of a failure at y_0 and with α and $(\beta + y_0)$ for the case of a survival at y_0, they were able to compare the gamma entropies.

The entropy of a gamma distribution with parameters a and b is given by

$$H_G(a) - log(b),$$

where

$$H_G(a) = log\Gamma(a) - (a-1)\psi(a) + a.$$

Since the scale parameter is the same, the comparison of the failure and the survival implies that

$$H_G(\alpha + 1) > H_G(\alpha). \tag{17}$$

Thus, for the failure rate Θ survival gives more information than failure.

However, if the objective is to make inference about the mean $\mu = 1/\Theta$, then the failure provides more information than the survival. Having the gamma prior on Θ implies an inverse gamma prior $1/\Theta$ and it can be shown that in this case the comparison gives

$$H_{IG}(\alpha + 1) < H_{IG}(\alpha). \tag{18}$$

It is important to note that the entropy is not invariant under transformations of Θ or the data. Thus, the comparison of information about a parameter depends on the parameterization of the lifetime model.

As pointed out by Ebrahimi et al. [16], findings by Abel and Singpurwalla raise some questions:

- Is the exponential case a counter example due to the memoriless property ?
- Can the result be generalized to other life models?
- What could possibly explain the preference for failures?
- What would you rather observe, a failure or a survival, for prediction of the lifetime of an untested item?

In order to address the above questions [16] considered a more general setup where $D_n = (y_1, \cdots, y_n)$ and $D_k = (y_1, \cdots, y_k, y^*_{k+1}, \cdots, y^*_n)$ denote the data provided by the failure and survival scenarios with y_i's and y^*_i's representing failure and survival times, respectively. The setup implies that the sufficient statistic for parameter Θ is the same for both scenarios, that is, $t_n(\mathbf{y}) = t_k(\mathbf{y})$. The corresponding likelihood functions for Θ are given by

$$\mathscr{L}(D_n|\theta) \propto \prod_{i=1}^{n} f(y_i|\theta), \quad \mathscr{L}(D_k|\theta) \propto \prod_{i=1}^{k} f(y_i|\theta) \prod_{i=k+1}^{n} S(y^*_i|\theta),$$

where $S(y|\theta) = P(Y > y|\theta)$ is the survival function.

As before, we let Q denote the unknown quantity of interest such as a parameter Θ, a function of the parameter such as $1/\Theta$, or the lifetime of an untested item Y_v with a distribution $f(\cdot)$. Using the Abel and Singpurwalla set up, the sample D_n is said to be more informative than D_k about Q whenever

$$H(Q|D_n) < H(Q|D_k). \tag{19}$$

The comparison is well-defined for improper priors for Θ as long as $f(q|D_n)$ and $f(q|D_k)$ are proper. With proper prior, the above is equivalent to the observed information criteria

$$\Delta H(D_n; Q) = H(Q) - H(Q|D_n) > H(Q) - H(Q|D_k) = \Delta H(D_k; Q). \tag{20}$$

Ebrahimi et al. [16] considered the class of models with survival function

$$S(y|\theta) = P(Y > y|\theta) = e^{-\theta\phi^{-1}(y)}, \tag{21}$$

where $Y = \phi(X)$, $S(x|\theta) = e^{-\theta x}$, and ϕ is an increasing function such that $\phi(0) = 0$ and $\phi(\infty) = \infty$. Since the survival function of X is exponential, the class of models is referred to as the time-transformed exponential (TTE) models and θ is referred to as the *proportional hazard* parameter; see [3]. Examples of lifetime models in the TTE family are given in Table 1. The sufficient statistics for the proportional hazard parameter under the two scenarios are the same

$$t_k(\mathbf{y}) = t_n(\mathbf{y}) = t_n = \sum_{i=1}^{n} \phi^{-1}(y_i).$$

Using the conjugate gamma prior for Θ with parameters α and β, denoted as $G(\alpha, \beta)$, the posterior distributions based on samples from models in the TTE family under both scenarios are gamma

$$f(\theta|D_s) = G(\alpha + n_s, \beta + t_n), n_s = k, n.$$

Table 1 Examples of time transformed exponential family

TTE model	$\phi(x)$
Exponential	x
Weibull	$x^{1/b}$
Linear failure rate	$\frac{1}{b}(a^2 + bx)^{1/2}$
Pareto Type I	ae^x
Pareto Type II	$e^x - 1$
Pareto Type IV	$(e^x - 1)^{1/a}$
Extreme value	$\log(1 + x)$

For the TTE family models, by the observed information criteria, we have:

1. for Θ

$$H_G(\alpha + n) > H_G(\alpha + k),$$

that is, survival is more informative than failure about the proportional hazard parameter Θ;

2. for $1/\Theta$

$$H_{IG}(\alpha + n) < H_{IG}(\alpha + k),$$

that is, failure is more informative than survival about the inverse parameter.

As noted by Abel and Singpurwalla [1], "The aim of life testing is to better predict the life lengths of untested items." In view of this, Ebrahimi et al. [16] investigated if the failure is more informative than the survival to predict future life lengths in the exponential model. Using the gamma prior $G(\alpha, \beta)$ in the exponential model the predictive distributions of Y_v will be Pareto. Thus, the posterior predictive entropies are given by

$$H(Y_v|n_s, t_n) = H_P(\alpha + n_s) + \log(\beta + t_n), \quad \alpha, \beta \geq 0, \quad n_s = k, n,$$

where

$$H_P(\alpha) = \frac{1}{\alpha} - \log \alpha + 1.$$

Since H_P is decreasingly ordered by α, we have

$$H_P(\alpha + n) < H_P(\alpha + k), \quad k < n. \tag{22}$$

In other words, for the exponential model, failure is more informative than survival about prediction of Y_v. Using the conjugate gamma prior for Θ, the result holds for most members of the TTE family for fixed values of other parameters a and b.

As shown by Ebrahimi et al. [16], a more general result can be obtained in comparing informativeness of failures and survivals about prediction by ordering entropies of predictive distributions. The important quantity for the stochastic ordering of the predictive distributions is

$$\Lambda(\theta) = \prod_{i=k+1}^{n} \lambda(y_i^*|\theta), \tag{23}$$

where $\lambda(y|\theta)$ is the hazard (failure) rate function of Y. The following result provides a comparison of the entropies of predictive distributions.

Theorem 2 *Given the definition of* $\Lambda(\Theta)$ *in (23)*

1. *if the predictive density function* $f(y_v|D_n)$ *is decreasing (increasing), then* D_n *is more (less) informative than* D_k *about the prediction of* Y_v*, if and only if*

$$\text{COV}\left[S(y_v|\Theta),\ \Lambda(\Theta)|D_k\right] < 0.$$

2. *if* θ *orders the survival function* $S(y|\theta)$*, then* $\text{COV}\left[S(y_v|\Theta),\ \Lambda(\Theta)|D_k\right] < 0$.

The terms $S(y_v|\Theta)$ and $\Lambda(\Theta)$ are functions of Θ and their covariance is obtained under the posterior distribution $f(\theta|D_k)$.

It is important to note that the Theorem 2 enables us to compare the entropies of predictive distributions for many lifetime models without a need to specify any prior distribution. The result is applicable to many of the TTE models. Also, all decreasing failure rate (DFR) distributions have decreasing density functions and mixtures of DFR distributions are also DFR. Thus, if the model $f(y|\theta)$ is DFR, then the predictive density is also DFR, and the decreasing condition in the result is satisfied. For example, the result holds for DFR models such as Pareto Type I, Pareto Type II, and Half-logistic, Weibull with $b \leq 1$, gamma with $a \leq 1$, and generalized-gamma with $ab \leq 1$, but it is not limited to the DFR models. It also applies to the IFR models: linear failure rate, extreme value, and models with non-monotone failure rates such as Half-Cauchy. Table 2 gives some examples where θ orders the survival function. When the conditions of the Theorem 2 do not hold, one can do the comparison directly by computing entropy of prediction under both scenarios.

Table 2 Examples where survival (S) or failure (F) is more informative for prediction

Model (more informative)	Density and support
Half-normal (F)	$f(y\|\theta) = \sqrt{\frac{2\theta}{\pi}} e^{-\frac{\theta}{2}y^2},\ y \geq 0$
Half-cauchy (F)	$f(y\|\theta) = \frac{2}{\pi\theta}\left(1 + \frac{y^2}{\theta^2}\right)^{-1},\ y \geq 0$
Half-logistic (F)	$f(y\|\theta) = \frac{\theta e^{-\theta y}}{(1+e^{-\theta y})^2}, y \geq 0$
Gamma ($a \leq 1$) (F)	$f(y\|a,\theta) = \frac{\theta^a}{\Gamma(a)} y^{a-1} e^{-\theta y},\ y \geq 0$
Generalized gamma ($ab \leq 1$) (F)	$f(y\|a,b,\theta) = \frac{b\theta^a}{\Gamma(a)} y^{ab-1} e^{-\theta y^b}, y \geq 0$
Generalized Pareto ($a > 0, , \theta \neq 1$) $\theta < 1$ (F) $\theta > 1$ (S)	$f(y\|\theta) = a\left(1 - \frac{\theta y}{a}\right)^{1/\theta - 1}, \begin{cases} y \geq 0, \theta < 0 \\ 0 < y \leq a/\theta, \\ \theta > 0 \end{cases}$
Power ($\theta < 1$, F) ($\theta > 1$, S)	$f(y\|\theta) = \theta y^{\theta-1},\ 0 < y \leq 1$
Beta ($\theta < 1$, S) ($\theta > 1$, F)	$f(y\|\theta) = \theta(1-y)^{\theta-1},\ 0 \leq y \leq 1$

Ebrahimi et al. [16] considered expected sample information as an alternative criterion to the observed information for a plausible explanation for the perception that failures are more informative. The expected information was measured by the conditional entropy $\mathscr{H}(Q|D_s)$ given by

$$\mathscr{H}(Q|D_k) = E_k\{H(Q|D_k)\} = \int H(Q|D_k)\, f_k(\mathbf{y}) d\mathbf{y}, \quad k = 0, 1, \cdots, n, \quad (24)$$

where E_k denotes averaging with respect to $f_k(\mathbf{y})$. Using the conditional entropy criteria and the conjugate gamma prior for Θ, the authors showed that for all members of the TTE family, failure is more informative than survival about prediction of a new lifetime Y_v, the proportional hazard parameter Θ, and its inverse $1/\Theta$. Furthermore, they showed that unlike the observed information measured by $-H(\Theta|D_k)$, the expected information measured by $-\mathscr{H}(\Theta|D_k)$ is increasing in the number of failures k.

This result may be interpreted as, on average, observing a failure is more informative than observing a survival. As noted by Kruskal [19], thinking in terms of averages is a tradition in statistics, and this provides a plausible explanation for the perception that failures are more informative.

5 Reliability Importance of System Components

Birnbaum [8] defined reliability importance of a component i for coherent systems as

$$I_i^B(t) = \frac{\partial \overline{F}(t)}{\partial \overline{F}_i(t)}, \quad (25)$$

where $\overline{F}(t)$ is the system reliability and $\overline{F}_i(t)$ is the component i's reliability at time t. Barlow and Proschan [4] introduced another measure of relative reliability importance as the conditional probability that the system's failure is caused by component i's failure. It can be shown that

$$I_i^{BP}(t) = \int_0^\infty I_i^B(t)\, dF_i(t)\, dt \quad (26)$$

where $F_i(t)$ is the distribution function for lifetime of component i and $\sum_i I_i^{BP}(t) = 1$. Alternative measures of reliability importance were proposed by Natvig [23, 24], and Armstrong [2]. More recent results are given in [25]. All these measures are in terms of contribution of a component to the reliability of the system.

Ebrahimi et al. [17] noted that reliability importance can be interpreted in terms of how knowledge of status of a component changes our knowledge of

the system. In other words, their interpretation is in terms of which component's status knowledge matters most in reducing our uncertainty about systems' status. The authors suggested an alternative notion of component importance in terms of information measures.

Consider a system that consists of n components $C_1, \cdots, C_n$ that collectively determine a random variable of interest Q for the system. Let Q_i be the corresponding random variable associated with $C_i, i = 1, \cdots, n$. More formally, let $Q_i = Q_i(C_i), \mathbf{Q} = (Q_1 \cdots Q_n)$ and define the *system structure function* as

$$Q = \phi(\mathbf{Q}), \phi : \Re^n \to \Re. \tag{27}$$

We note that Q_i can be the indicator variable of the state (failure/survival) or the life length of a component and Q is the respective random variable for the system.

The information notion of importance maps the expected utility of the component variable Q_i for prediction of the system variable Q in terms of the dependence implied by the joint distribution $F(q, q_i)$. More specifically, it is defined as follows.

Definition 1 The importance of component C_i is defined by the expected information utility of C_i for the system measured by the mutual information $M(Q; Q_i)$, provided that $F(q, q_i) \ll F_q(q)F_i(q_i)$. Component C_i is more important than the component C_j for the system if and only if $M(Q; Q_i) \geq M(Q; Q_j)$.

Under the information notion of importance component C_i is more important than the component C_j for the system if and only if

$$M(Q; Q_i) \geq M(Q; Q_j) \iff \mathscr{I}(Q|Q_i) \geq \mathscr{I}(Q|Q_j), \tag{28}$$

where $\mathscr{I}(Q|Q_i) = -\mathscr{H}(Q|Q_i)$ is the conditional information.

Let $T_1, \cdots, T_n$ denote independent random variables representing the life lengths of components $C_1, \cdots, C_n$ and T denote the life length of the system. The survival of system is defined as the survival up to a mission time τ. We define the binary variables for the states of the component and system as

$$Q_i(C_i) = X_i(\tau) = 1(0), \ \ if \ \ T_i > \tau \ \ (T_i \leqslant \tau)$$

and

$$Q = X(\tau) = \phi(X_1(\tau), \ldots, X_n(\tau)) = 1(0), \ \ if \ \ T > \tau \ \ (T \leqslant \tau),$$

where $\phi : \{0, 1\}^n \to \{0, 1\}$.

The marginal distributions of $X_i(\tau)$ and the conditional distribution of $X(\tau)$ given $X_i(\tau) = x_i$ are Bernoulli with parameters $p_i = p_i(\tau) = \bar{F}_i(\tau)$ and $p_{x|x_i} = p_{x|x_i}(\tau) = P(X(\tau) = x|X_i(\tau) = x_i)$ for $i = 1, \ldots n$, respectively. For each mission time τ, we obtain the conditional information $\mathscr{I}[(X(\tau)|X_i(\tau)]$ as

$$\mathscr{I}[(X(\tau)|X_i(\tau)] = p_i(\tau)I[(X(\tau)|X_i(\tau) = 1] + (1 - p_i(\tau))I[(X(t)|X_i(t) = 0]. \tag{29}$$

Note that this measure ranks the importance of components for a fixed mission time τ.

The following result by Ebrahimi et al. provides an ordering of the components for three types of systems at a fixed time point $t = \tau$.

Theorem 3 *Consider a system with n independent components such that $P(X_i(\tau) = 1) = p_i, i = 1, \cdots, n$.*

1. *if the system is series, then the component C_j is more important than the component C_i, for $p_j < p_i$, $i, j = 1, \cdots, n$;*
2. *if the system is parallel, then the component C_j is more important than the component C_i, for $p_j > p_i$, $i, j = 1, \cdots, n$;*
3. *for the k-out-of n system, the component C_j is more important than the component C_i, if $p_j < p_i$, $i, j = 1, \cdots, n$ and $P\left\{S_{(i)}(\boldsymbol{X}(\tau)) \geq k\right\} \geq \frac{1}{2}$, $i = 1, \cdots, n$, where*

$$S_{(i)}(\boldsymbol{v}) = \sum_{k \in \mathscr{N}_i} v_k, \quad \mathscr{N}_i = \{1, \cdots, i-1, i+1, \cdots, n\}$$

for a vector $\boldsymbol{v} = (v_1, \cdots, v_n)$.

The information measure (28) orders the components in the same way as the reliability importance index of Birnbaum [8]. The information analog of Barlow and Proschan's [4] measure of component's importance is the expected mutual information $E_i\left[M\left(X(\tau), X_i(\tau)\right)\right]$ where the expected value is with respect to the distribution of T_i. Thus, we can order the components by evaluating $E_i\left[\mathscr{I}\left(X(\tau)|X_i(\tau)\right)\right]$.

As shown in [17], stochastic ordering of life times of the components $T_1 \overset{st}{\leq} \ldots \overset{st}{\leq} T_n$ is sufficient for Theorem 3 to hold. The next example illustrates the implementation of the Theorem.

5.1 Example

Consider a system of two components with independent lifetimes.

1. *Bernoulli Distributions.* We can obtain the conditional information measures for the series and parallel systems as

$$\mathscr{I}\left(X(\tau)|X_i(\tau)\right) = p_i(\tau) I(X_j(\tau)) \quad \text{and} \quad \mathscr{I}\left(X(\tau)|X_i(\tau)\right) = (1 - p_i(\tau)) I(X_j(\tau)),$$

respectively. Thus, C_i is more (less) important for a series system than for a parallel system whenever $\tau >$ (<) median lifetime;

2. *Proportional Hazard Distributions.* Suppose that the components' lifetimes, $T_i, i = 1, 2$, have proportional hazard (PH) distributions

$$\bar{F}_i(\tau) = [\bar{F}_0(\tau)]^{\theta_i}, \ \tau \geq 0, \ \theta_i > 0.$$

For $\theta_1 > \theta_2$, we have $T_1 \overset{st}{\leq} T_2$. Thus for any mission time τ, $p_1(\tau) < p_2(\tau)$ and by Theorem 3 C_1 is more (less) important than C_2 for the series (parallel) system;

3. *TTE Family of Distributions.* Suppose that the components' lifetimes $T_i, i = 1, 2$, having TTE family of distributions as defined in Sect. 4. Note that the TTE model is a PH model and therefore we can conclude that for $\theta_1 > \theta_2$, $T_1 \overset{st}{\leq} T_2$. Thus, for any mission time τ, $p_1(\tau) < p_2(\tau)$ and C_1 is more (less) important than C_2 for the series (parallel) system.

The information based analog of Barlow-Proschan importance index can also be computed for TTE models. It can be shown that $E_i\left[\mathscr{I}\left(X(t)|X_i(t)\right)\right]$ for the series and parallel systems are functions of $\theta = \dfrac{\theta_i}{\theta_j}$, decreasing (increasing) for $\theta < (>)1$.

The notion of information importance is also applicable to component and system lifetimes as continuous random variables. As shown in [17], it is possible to develop results for convolutions and order statistics. For convolutions, results in entropy ordering can be used to obtain component importance ordering. Considering parallel and series systems with continuous lifetimes, the system lifetime being order statistics leads to singular distributions, where the mutual information is not well-defined. In order to alleviate this problem, a modification of the information importance index was used by the authors.

Ebrahimi et al. [17] also considered an entropy-based importance measure using the Maximal Data Information Prior (MDIP) criterion proposed by Zellner [30]. The MDIP criterion which was originally proposed for developing priors for Bayesian inference, provides the same importance ordering of the components as the mutual information.

Conclusions

Information measures play an important role in Bayesian reliability analysis. In this paper our focus was on Bayesian information measures in failure data analysis, prediction, assessment of reliability importance and optimal design of life tests. Other uses of these measures include failure model selection [21, 22], characterization of univariate and multivariate failure models and characterization of dependence for system reliability [13].

Computation of information functions can be quite challenging in many applications but decomposition type results given in [13] can be helpful

(continued)

for certain problems. Parametric and nonparametric Bayesian estimation of information functions and related indices (see [10]) are required for multivariate life models. Recent advances in Bayesian computing, especially, those in efficient Markov chain Monte Carlo methods can be exploited in many cases.

References

1. Abel, P.S., Singpurwalla, N.D.: To survive or to fail: that is the question. Am. Stat. **48**, 18–21 (1994)
2. Armstrong, M.J.: Joint reliability-importance of components. IEEE Trans. Reliab. **44**, 408–412 (1995)
3. Barlow, R.E., Hsiung, J.H.: Expected information from a life test experiment. Statistician **48**, 18–21 (1983)
4. Barlow, R.E., Proschan, F.: Importance of system components and fault tree event. Stoch. Process. Their Appl. **3**, 153–172 (1975)
5. Amaral-Turkman, M.A., Dunsmore, I.: Measures of information in the predictive distribution. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (eds.) Bayesian Statistics, vol. 2, pp. 603–612. Elsevier, Amsterdam (1985)
6. Bernardo, J.M.: Expected information as expected utility. Ann. Stat. **7**, 686–690 (1979)
7. Bernardo, J.M.: Reference posterior distribution for Bayesian inference. J. R. Stat. Soc. Series B **41**, 605–647 (1979)
8. Birnbaum, Z.W.: On the importance of different components in a multicomponent system. In: Krishnaiah, P.R. (eds.) Multivariate Analysis II, pp. 581–592. Academic Press, New York (1969)
9. Chaloner, K., Verdinelli, I.: Bayesian experimental design: a review. Stat. Sci. **10**, 273–304 (1995)
10. Dadpay, A.N., Soofi, E.S., Soyer, R.: Information measures for generalized gamma family. J. Econom. **138**, 568–585 (2007)
11. Ebrahimi, N., Soofi, E.S.: Recent developments in information theory and reliability analysis. In: Basu, A.P., Basu, S.K., Mukhopadhyay, S. (eds.) Frontiers in Reliability, pp. 125–132. World Scientific, New Jersey (1998)
12. Ebrahimi, N., Soofi, E.S.: Information functions for reliability. In: Soyer, R., Mazzuchi, T.A., Singpurwalla, N.D. (eds.) Mathematical Reliability: An Expository Perspective, pp. 127–159. Kluwer, New York (2004)
13. Ebrahimi, N., Soofi, E.S., Soyer, R.: Multivariate maximum entropy identification, transformation, and dependence. J. Multivar. Anal. **99**, 1217–1231 (2008)
14. Ebrahimi, N., Soofi, E.S., Soyer, R.: Information measures in perspective. Int. Stat. Rev. **78**, 383–412 (2010)
15. Ebrahimi, N., Soofi, E.S., Soyer, R.: On the sample information about parameter and prediction. Stat. Sci. **25**, 348–367 (2010)
16. Ebrahimi, N., Soofi, E.S., Soyer, R.: When are observed failures more informative than observed survivals ? Nav. Res. Logist. **60**, 102–110 (2013)
17. Ebrahimi, N., Jalali, N.Y., Soofi, E.S., Soyer, R.: Importance of components for a system. Econom. Rev. **33**, 395–420 (2014)
18. El-Sayyed, G.M.: Information and sampling from exponential distribution. Technometrics **11**, 41–45 (1969)

19. Kruskal, W.: Relative importance by averaging over orderings. Am. Stat. **41**, 6–10 (1987)
20. Lindley, D.V.: On a measure of information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956)
21. Mazzuchi, T.A., Soofi, E.S., Soyer, R.: Computations of maximum entropy Dirichlet for modeling lifetime data. Comput. Stat. Data Anal. **32**, 361–378 (2000)
22. Mazzuchi, T.A., Soofi, E.S., Soyer, R.: Bayes estimate and inference for entropy and information index of fit. Econom. Rev. **27**, 428–456 (2008)
23. Natvig, B.: A suggestion of a new measure of importance of system component. Stoch. Process. Their Appl. **9**, 319–330 (1979)
24. Natvig, B.: New light on measures of importance of system components. Scand. J. Stat. **12**, 43–54 (1985)
25. Natvig, B., Gsemyr, J.: New results on the Barlow–Proschan and Natvig measures of component importance in nonrepairable and repairable systems. Methodol. Comput. Appl. Probab. **11**, 603–620 (2009)
26. Polson, N.G.: A Bayesian perspective on the design of accelerated life tests. In: Basu, A.P. (eds.) Advances in Reliability, pp. 321–330. North Holland, Amsterdam (1993)
27. San Martini, A., Spezzaferri, F.: A predictive model selection criteria. J. R. Stat. Soc. Series B **46**, 296–303 (1984)
28. Stone, M.: Application of a measure of information to the design and comparison of regression experiments. Ann. Math. Stat. **29**, 55–70 (1959)
29. Verdinelli, I., Polson, N.G., Singpurwalla, N.D.: Shannon information and Bayesian design for prediction in accelerated life-testing. In: Barlow, R.E., Clarotti, C.A., Spizzichino, F. (eds.) Reliability and Decision Making, pp. 247–256. Chapman Hall, London (1993)
30. Zellner, A.: Maximal data information prior distributions. In: Aykac, A., Brumat, C. (eds.) New Developments in the Applications of Bayesian Methods, pp. 211–232. North-Holland, Amsterdam (1977)
31. Zellner, A.: Optimal information processing and Bayes' theorem. Am. Stat. **42**, 278–284 (1988)

(Semi-)Intrinsic Statistical Analysis on Non-Euclidean Spaces

Stephan F. Huckemann

1 Introduction

Often, problems concerning dynamics in biology, due to natural constraining conditions, equivalence relations and/or identifications, lead to data of statistical interest which come to lie on spaces with no Euclidean structure. For such data [38] coined the term *object data*. Frequently, the underlying spaces can be equipped with structures of non-flat Riemannian manifolds or with structures of stratified spaces made up from manifolds of varying dimensions. Exemplary in this paper we consider dynamics caused by growth, by simple rotational deformations and by human gait motion. In order to describe such dynamics, suitable descriptors are introduced which themselves live in stratified spaces.

In general, (*semi-*)*intrinsic statistical analysis* considers data on a topological space Q, links these data via a *linking distance* to Fréchet ρ-means in another space P that admits a manifold stratification and conducts statistical analysis on P. This analysis is *semi-intrinsic* if for the final statistical analysis, e.g. for reasons of modeling and computational simplicity, extrinsic methods rather than intrinsic ones are used. This approach which has been briefly described in [26] is introduced in detail in Sect. 2. In Sect. 3, inference on *geodesic PCs* is exemplary studied in case where a $\sqrt{n}$-Gaussian central limit theorem is assumed valid. Section 4 describes some curious phenomena when this is not valid. We start now with four motivational examples that will be taken up along the development of the methodology.

S.F. Huckemann (✉)
Felix Bernstein Institute for Mathematical Statistics in the Biosciences,
Georg-August-Universität Göttingen, Göttingen, Germany
e-mail: huckeman@math.uni-goettingen.de

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_7

Example 1 As a first example we consider the problem of studying the growth of Canadian black poplar leaves conducted in a joint research with the Department of Oecological Informatics, Biometry and Forest Growth at the University of Göttingen. Over several growing periods, leaf contours have been non-destructively digitized. The task is to describe growth and discriminate growth among leaves of the same tree of identical clones and over different genetic expressions.

For this scenario it turns out that geodesics in underlying shape spaces qualify well as such descriptors, cf. [24]. Two such shape spaces come to mind. If only specific landmarks are taken into account, Kendall's planar shape spaces based on landmarks (cf. [31]) seem canonical. In a separate effort, from these the entire contour could be reconstructed (cf. [19]). Alternatively one can consider the shape of the entire contour. Then modeling in the shape spaces of closed 2D contours based on angular direction by Zahn and Roskies [47] seems appropriate. These two spaces are Riemannian manifolds in a canonical way, cf. [34]. In particular Kendall's shape space of planar landmarks are the well known complex projective spaces. If we were to analyze Kendall shapes based on 3D landmarks or closed contours without a specified point (the stalk entry to the leaf is such a specified point) the underlying shape space would cease to be Riemannian manifolds while the former would still be a stratified space (cf. [22, 27, 33]).

Example 2 As a second example in a joint research with the Departments of Computer Science and Statistics at the Universities of North Carolina and Pittsburgh we consider the problem of estimating deformations of internal organs which is crucial for instance in radiation oncology, where one challenge consists in relocating an internal organ at therapy time while this organ has been carefully investigated at planning time. Frequently, this relocation involves not only the estimation of Euclidean motions but also the estimation of deformations, e.g. a cancerous prostate might be bent around a filled bladder.

For this task as an underlying model, skeletal representations (cf. [7, 43]) seem very appropriate. In 2D the intuition goes as to mark the medial skeleton by the loci where an inward "grassfire" originating on the contour eventually dies out. Equivalently, the medial skeleton is characterized as the set of points that are along with a ball fully contained in the object, where the ball is centered at such a point and touches the boundary at least at two points. The vectors from the balls' centers to the touching points are called *spokes*. In this vein, statistics of sufficiently well behaved 3D objects translates into statistics of 2D medial sheets along with their spokes. "Well behaved" in this context means that there may be only very mild boundary perturbations, since every protruding boundary kink results in a medial branch ending there. Thus two versions of the same object under small noise may have medial sheets with drastically different topological branching structures.

In order to treat realistically occurring boundary noise, Pizer et al. [42] have introduced *skeletal representations* (s-reps) with skeletal sheets as "medial as possible" while ideally granting a common branching topology over an entire sample of interest (for the mathematical challenges involved, cf. [10, 11]).

The fundamental deformations due to rotations, bendings and twistings cause the spokes directions to move on specific concentric small circles with a common axis for every fundamental rotational deformation. Thus the spaces of concentric small circles on the two-sphere qualify as descriptors of fundamental rotational deformations.

Example 3 The study of the third example originated from the AOOD Workshop 2010/11 at the Statistical and Applied Mathematical Sciences Institute (SAMSI) in North Carolina, cf. [44]. In their seminal paper [15] introduced *phylogenetic trees* to assess genetic mutation distances. From the mutations of a specific gene over a fixed set of different species, an optimal descendant tree is estimated which takes the role of a descriptor. Usually, different genes over the same set of species give different trees. Measuring tree distances by comparing interior edge lengths (corresponding to mutation distance) over equivalence classes of trees leads to a metric space which can be given the structure of a stratified space with flat manifold strata (orthants of varying dimensions) such that the entire space carries non-positive curvature, cf. [6]. Similar models are appropriate to model biological tree-like structures such as the artery tree of the brain (cf. [2, 44]) or of the respiratory tract (cf. [14]).

Example 4 In a joint research with the School of Rehabilitation Science and the Department of Kinesiology at McMaster University biomechanical human gait is analyzed statistically. In particular, the relative rotation of the lower leg (tibia) with respect to the thigh (femur) during a volunteer's gait is estimated by a standard protocol [12] which uses the spatial paths of markers placed on various locations of the upper and lower leg by a specialist, cf. [8]. For every gait cycle the non-Euclidean data object of interest is thus a (quasi-closed) curve in the space of three-dimensional rotations. Instead of entire curves as descriptors, one may use configurations of specific locations of critical events on the curve modulo initial rotations, as the latter convey effects of varying marker placement. Following the standard protocol, the rotations are parametrized by three Euler angles, where the first and dominant angle measures flexion/extension about a medial-lateral axis attached to the thigh. In order to obtain a landmark descriptor one may consider the following four critical gait events which are defined as pseudo-landmarks determined by the first angle. Locally extremal flexion/extension angles occur while the foot is on the ground (stance) at *heel contact* and *mid stance* as well as at *maximal bend* while the foot is in the air (swing); when the foot leaves the ground at *toe off* the angular velocity is maximal, cf. [45].

2 (Semi)-Intrinsic Statistical Analysis

2.1 The Setup

We begin with two topological spaces:

Q is the *data* space and P is the *descriptor space*.

Further, we assume that data and descriptors are linked via a continuous function

$$\rho : Q \times P \to [0, \infty)$$

called a *linking function* which takes the role of a distance between a datum and a descriptor. Finally, we assume that there is a continuous mapping $d : P \times P \to [0, \infty)$ vanishing on the diagonal $\{(p, p) : p \in P\}$.

Random variables on Q and P are mappings from an abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$ that are measurable w.r.t. the corresponding Borel σ-algebras.

Definition 2.1 We call such a tuple (ρ, d) a *uniform link* if for every $p \in P$ and $\epsilon > 0$ there is a $\delta = \delta(\epsilon, p) > 0$ such that

$$|\rho(x, p') - \rho(x, p)| < \epsilon \text{ for all } x \in Q, p' \in P \text{ with } d(p, p') < \delta.$$

Moreover, it is a *coercive link* if for every $p_0 \in P$

(i) for every $C > 0$ and sequence $p_n \in P$ with $d(p_0, p_n) \to \infty$ there is a sequence $M_n \to \infty$ with $\rho(x, p_n) > M_n$ for all $x \in Q$ with $\rho(x, p_0) < C$, and
(ii) if $d(p^*, p_n') \to \infty$ for some other sequence $p_n' \in M$ and $p^* \in M$ then $d(p_0, p_n') \to \infty$.

2.2 Fréchet ρ-Means

Definition 2.2 For random elements $X, X_1, X_2, \ldots$ on Q define the *set of population Fréchet ρ-means of X on P* by

$$E^{(\rho)}(X) = \operatorname*{argmin}_{\mu \in P} \mathbb{E}\big(\rho(X, \mu)^2\big) .$$

For $\omega \in \Omega$ denote by

$$E_n^{(\rho)}(\omega) = \operatorname*{argmin}_{\mu \in P} \sum_{j=1}^{n} \rho\big(X_j(\omega), \mu\big)^2$$

the *set of sample Fréchet ρ-means on P*.

Note that without further conditions, $E^{(\rho)} = \emptyset$ is possible. In most applications, in particular in those considered here, ρ and P are sufficiently well behaved so that existence of mean sets is not an issue.

By continuity of ρ, the mean sets are closed random sets. For our purpose here, we rely on the definition of *random closed sets* as introduced and studied by Choquet [9], Kendall [29] and Matheron [39]. For an overview over the well developed asymptotic theory of closed random sets in Banach spaces cf. [40]. As it seems,

a more general asymptotic theory for random closed sets in non-linear spaces as are of concern here, has received less attention in literature.

Fréchet ρ-means in case of a metric $\rho = d$ on $P = Q$ have been first introduced by Fréchet [16] and generalized to quasimetrics $\rho = d$ by Ziezold [48]. A quasimetric is a continuous and symmetric mapping $Q \times Q \to [0, \infty)$ vanishing on the diagonal that satisfies the triangle inequality. In case of $P = Q$ and $\rho = d$ satisfying the triangle inequality, it is a uniform coercive link. More generally if P and Q are compact then (ρ, d) is a uniform coercive link.

2.3 Examples of Fréchet ρ-Means

Intrinsic Means If $Q = P$ is a manifold or a stratified space and ρ a geodesic distance, Fréchet ρ-means are usually called *intrinsic means*. On Riemannian manifolds they have been first studied by Kobayashi and Nomizu [35] as *centers of gravity*.

Extrinsic and Residual Means If $P = Q$ is embedded in a Euclidean space $\mathbb{R}^D$ with Euclidean norm $\|\cdot\|$ and $\rho(x, y) = \|x - y\| = d(x, y)$ $(x, y \in Q)$ being the *extrinsic metric* (also called chordal distance), the corresponding Fréchet ρ-means are called *extrinsic means*. They have been introduced for manifolds as *mean locations* by Hendriks and Landsman [18]. If additionally $\Phi : \mathbb{R}^D \to Q$ is the orthogonal projection (which is locally well defined around Q wherever Q is locally Euclidean), then the Fréchet ρ-means with respect to the *residual link* $\rho(x, y) = \|d\Phi_x(y - x)\| = d(x, y)$ $(x, y \in Q)$ are called *residual means*. On spheres, residual means have been introduced by Jupp [28]. There, the residual link is globally well defined and in fact a metric.

Ziezold and Procrustean Means Suppose now that M is a Riemannian manifold embedded in a Euclidean space $\mathbb{R}^D$ with Euclidean norm $\|\cdot\|$ on which a discrete or a Lie group G acts from the left via

$$M \to M, x \overset{g}{\mapsto} gx \text{ for all } g \in G\,, \tag{1}$$

giving rise to the quotient

$$Q := M/G = \{[x] : x \in M\}, \quad [x] = \{gx : g \in G\},\ x \in M \tag{2}$$

such that the action is isometric with respect to the extrinsic distance, i.e.

$$\|gx - gy\| = \|x - y\| \quad \forall g \in G \text{ and } x, y \in M\,.$$

Then we have the following two linking functions $\rho = d$ on $P = Q$:

$$\begin{aligned} \rho_Z([x], [y]) &:= \inf_{g \in G} \|gx - y\|, && \text{the } \textit{Ziezold link} \\ \rho_P([x], [y]) &:= \inf_{g,h \in G} \|d\Phi_{gx}(hy - gx)\|, && \text{the } \textit{Procrustean link} \end{aligned}$$

giving rise to *Ziezold means* and *Procrustean means*. The quotient Q is often called a *shape space*. Note that the Ziezold link is a quasimetric, hence it is a uniform coercive link. The first type of mean has been introduced by Ziezold [49] for Kendall's shape spaces. Although also designed to tackle Kendall shapes, the notion of the *full Procrustes mean* by Gower [17] seems to have preceded these. On Kendall's shapes spaces the Ziezold link and the Procrustean link are metrics.

Kendall's Shape Spaces In view of Example 1, consider configurations of $k \in \mathbb{N}$ landmarks in a $m \in \mathbb{N}$ dimensional space represented by $m \times k$ matrices (cf. [30,31]). The *shape* of a landmark configuration is its equivalence class under common translation, scaling and rotation of all columns (these are the landmarks) of the matrix. Usually one normalizes for scaling and translation (e.g. via Helmertizing, cf. [13]) to obtain a matrix $x \in M$ where M is the unit sphere embedded in the Euclidean space $M(m, k-1)$ of $m \times (k-1)$ matrices on which $G = SO(m)$ acts by matrix multiplication from the left giving rise to Kendall's shape space $\Sigma_m^k := M/G$. For a short derivation cf. [23].

Geodesic Principal Components On a metric space (Q, τ) a rectifiable curve $\gamma : I \to Q$, $I \subset \mathbb{R}$, is called a *geodesic* if

(i) for all $t_0 \in I$ there are $t < t_0 < s$, $p = \gamma(t)$ and $q = \gamma(s)$ such that

$$\inf_{\substack{n \in \mathbb{N} \\ t = t_0 < \ldots < t_n = s}} \sum_{j=1}^{n} \tau\big(\gamma(t_{j-1}), \gamma(t_j)\big) = \tau(p, q)\,;$$

(ii) γ cannot be extended beyond I such that it satisfies (i).

Denote by P the space of point sets of geodesics on Q. If P can be given a topological structure such that

$$\rho : Q \times P \to [0, \infty), \quad (q, \gamma) \mapsto \rho(q, \gamma) := \inf_{t \in I} \tau\big(p, \gamma(t)\big)$$

is a linking function, the corresponding Fréchet ρ-means are called *first geodesic principal components* (1stGPCs). If there is a concept of orthogonality of geodesics, higher order principal components can be defined as in [27].

Concentric Small Circles For spoke data $z = (z_1, \ldots, z_k) \in (S^2)^k$ as in Example 2, here $S^2 := \{x \in \mathbb{R}^3 : \|x\| = 1\}$ is the two-sphere and $k \in \mathbb{N}$, consider the space P of k concentric small circles with the straightforward quotient topology induced by the Ziezold distance inherited from the extrinsic distance on $S^2 \times [0, \pi]^k$ as follows. Let

$$\delta(c, r) := \big\{z \in (S^2)^k : \langle c, z_j \rangle = r_j, j = 1, \ldots, k\big\}$$
$$\text{for } c \in S^2 \text{ and } r = (r_1, \ldots, r_k) \in [0, \pi]^k\,,$$

$$
\begin{aligned}
[\delta(c,r)] &:= \{\delta(c,r), \delta(-c,\pi-r)\} \\
&\qquad \text{where } \pi - r := (\pi - r_1, \ldots, \pi - r_k) \text{ and} \\
P &:= \left\{[\delta(c,r)] : c \in S^2, r \in [0,\pi]^k\right\} .
\end{aligned}
$$

Here, $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product. A linking function is given by the geodesic distance

$$
\rho\big(z, [\delta(c,r)]\big) := \sqrt{\sum_{j=1}^{k} \big(\arccos(\langle c, z_j \rangle) - r_j\big)^2} .
$$

Manifold Configurations Above, spoke data are given by configurations on manifolds. If the coordinate system is rotated, say, spoke configurations are rotated alike. More generally, consider the case where the underlying space is a manifold M on which a Lie group G acts via (1). In analogy to (2), with

$$
M^k := \{q = (q_1, \ldots, q_k) : q_j \in M, j = 1, \ldots, k\}, \quad [q] := \{gq : g \in G\},
$$

the space of k-landmark configurations on M and orbits of this action, respectively, this gives rise to the shape space

$$
M^k/G := \{[q] : q \in M^k \setminus N^o\} .
$$

Here, $N^o \subset M^k$ denotes a *singularity set* that may have to be removed in order to obtain a topological Hausdorff space in a suitable topology of the quotient. In case of Kendall's shape spaces, a minimal singularity set is the set of all configurations with all landmarks coinciding, where, however, the topology of the quotient is not the canonical quotient topology, cf. [23, 33].

In view of gait analysis in Example 4 consider the special case where $M = G = SO(3)$. Then, the quotient topology may be used with void singularity set $N^o = \emptyset$ giving a shape space with a Bookstein type structure

$$
\begin{aligned}
G^k/G &\cong G^{k-1} \\
[g_1, \ldots, g_k] &\leftrightarrow (g_1^{-1}g_2, \ldots, g_1^{-1}g_k)
\end{aligned} .
$$

Given a distance δ on G (e.g. the intrinsic distance due to a left-invariantly extended inner product of the tangent space T_eG at the identity element $e \in G$ or the extrinsic distance due to an embedding of G in a suitable Euclidean space), as a link for G^{k-1} one may use the canonical product distance on G^{k-1}

$$
\rho\big((g_1, \ldots, g_{k-1}), (g'_1, \ldots, g'_{k-1})\big) := \sqrt{\sum_{j=1}^{k-1} \delta(g_j, g'_j)^2} .
$$

2.4 Asymptotics of Fréchet ρ-Means

Definition 2.3 Let $E_n^{(\rho)}(\omega)$ ($\omega \in \Omega$) be a random closed set and $E^{(\rho)}$ a deterministic closed set in P. Then,

(ZC) $E_n^{(\rho)}(\omega)$ is a *strongly consistent estimator* of $E^{(\rho)}$ in the sense of *Ziezold* if a.s. for $\omega \in \Omega$

$$\bigcap_{n=1}^{\infty} \overline{\bigcup_{k=n}^{\infty} E_k^{(\rho)}(\omega)} \subset E^{(\rho)} ,$$

(BPC) $E_n^{(\rho)}(\omega)$ is a *strongly consistent estimator* of $E^{(\rho)}$ in the sense of *Bhattacharya-Patrangenaru* if $E^{(\rho)} \neq \emptyset$ and if for every $\epsilon > 0$ and a.s. for $\omega \in \Omega$ there is a number $n = n(\epsilon, \omega) > 0$ such that

$$\bigcup_{k=n}^{\infty} E_k^{(\rho)}(\omega) \subset \{p \in P : d(E^{(\rho)}, p) \leq \epsilon\} .$$

In linear spaces, usually convergence w.r.t. the Hausdorff distance is considered, cf. [40]. If curvature is involved, however, there may be convergence in the above sense but no longer convergence w.r.t. Hausdorff distance. For example, recall the uniform distribution on a sphere, the set of its intrinsic Fréchet means being that entire sphere. Every sample mean, however, is atomic with probability one (for the detailed construction cf. [4, Remark2.6.]).

Ziezold [48] introduced (ZC) and proved it for quasi-metrical means on separable (i.e. containing a dense countable subset) spaces. Bhattacharya and Patrangenaru [4] introduced (BPC) and proved it for metrical means on spaces that enjoy the stronger d-Heine–Borel property, i.e. that every d-bounded (A is d-bounded if there is a point $p \in A$ such that $d(p, p_n)$ is bounded for every sequence $p_n \in A$) closed set is compact. Both properties, have been called 'strong consistency' by their respective authors. We have the following generalization, cf. [24, Theorems A.3 and A.4].

Theorem 2.4 *Suppose that the data space Q is separable, ρ is a uniform link and that $\mathbb{E}\big(\rho(X, p)^2\big) < \infty$ for all $p \in P$. Then property (ZC) holds for the set of Fréchet ρ-means on P.*

If additionally $E^{(\rho)} \neq \emptyset$, P enjoys the d-Heine-Borel property and (ρ, d) is also a coercive link then property (PBC) holds for the set of Fréchet ρ-means on P.

In order to formulate a Gaussian central limit theorem, we require additional properties.

Assumption 2.5 *The population Fréchet-ρ mean is unique up to a discrete group action, i.e. there are a discrete group H acting on P and $\mu \in P$ such that $\{h\mu : h \in H\} = E^{(\rho)}$ and there is an open neighborhood U of μ in P that is a D-dimensional twice differentiable manifold, $D \in \mathbb{N}$.*

Definition 2.6 Under Assumption 2.5 we say that a P-valued estimator $\mu_n(\omega)$ of μ satisfies a *Gaussian $\sqrt{n}$-Central-Limit-Theorem* (CLT), if in any local chart (ϕ, U) near $\mu = \phi^{-1}(0)$ there is a random Gaussian $D \times D$ matrix $\mathcal{G}_\phi$ with zero mean and semi-definite covariance matrix Σ_ϕ such that

$$\sqrt{n}\big(\phi(\mu_n) - \phi(\mu)\big) \;\to\; \mathcal{G}_\phi$$

in distribution as $n \to \infty$.

In consequence of the "δ-method", for any other chart (ϕ', U) near $\mu = \phi'^{-1}(0)$ we have simply

$$\Sigma_{\phi'} = J(\phi' \circ \phi^{-1})_0 \Sigma_\phi J(\phi' \circ \phi^{-1})_0^T$$

where $J(\cdot)_0$ denotes the Jacobi-matrix of first derivatives at the origin. In fact, the argument for a central limit theorem rests on the "δ-method" from M-estimation technology, cf. [46]. To this end we require additional smoothness assumptions.

Assumption 2.7 *Under Assumption 2.5 assume further for every local chart (ϕ, U) near $\mu = \phi^{-1}(0)$ that*

$$\text{the mapping } x \mapsto \rho(X, \phi^{-1}(x))^2 \text{ is a.s. twice differentiable in } U \tag{3}$$

and that

$$\left.\begin{array}{rl} \mathbb{E}(grad_2\rho(X,\mu)^2) & \text{exists,} \\ Cov(grad_2\rho(X,\mu)^2) & \text{exists,} \\ \mathbb{E}(Hess_2\rho(X,\nu)^2) & \text{exists for } \nu \text{ near } \mu, \text{ is continuous at } \nu = \mu \\ & \text{and of full rank there,} \end{array}\right\} \tag{4}$$

where $grad_2\rho(q,\nu)^2$ and $Hess_2\rho(q,\nu)^2$ denote the gradient and the Hessian of the above mapping.

Obviously the validity of (4) is independent of the particular chart chosen. The following Theorem is a straightforward consequence of [23, Theorem 6]. It is a generalization of [5] who provided a proof for the case of $P = Q$ a manifold and ρ either the intrinsic or extrinsic (arising from an embedding) distance. Also in case of $P = Q$ a manifold and ρ the intrinsic distance, Kendall and Le [32] have derived intrigued versions for independent but non-identically distributed samples.

Theorem 2.8 *Under Assumptions 2.5 and 2.7 suppose that $E_n^{(\rho)}$ is a strongly consistent estimator in the sense of Bhattacharya–Patrangenaru of a Fréchet population ρ-mean set $E^{(\rho)} = \{h\mu : h \in H\}$, $\mu \in P$, unique up to a discrete group action H. Then for every measurable choice $\mu_n(\omega) \in E^{(\rho)}$ there is a random sequence $h_n(\omega) \in H$ such that $h_n(\omega)\mu_n(\omega)$ satisfies a Gaussian $\sqrt{n}$-CLT. In a suitable chart (ϕ, U) the corresponding matrix from Definition 2.6 is given by*

$$\Sigma_\phi = A_\phi Cov(grad_2\rho(X,\mu)^2) A_\phi^{-1} \text{ where } A_\phi = \mathbb{E}(Hess_2\rho(X,\mu)^2)\,.$$

Remark 2.9 The condition that A_ϕ be of full rank is not at all trivial. A consequence of A_ϕ failing to to so is discussed in Sect. 4. Also in Sect. 4 we see that (3) from Assumption 2.7 is not necessary for the validity of the Gaussian $\sqrt{n}$-CLT.

2.5 The Two-Sample Test

Two-sample tests for metrical means on shape spaces and on manifolds have been around for a while, e.g [13, 37, 41]. In the following we extend these to our context of Fréchet ρ-means.

Suppose that $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} Y$ and $Z_1, \ldots, Z_m \overset{i.i.d}{\sim} Z$ are independent samples on Q with unique Fréchet ρ-means μ_Y and μ_Z, respectively, on Q $(m, n > 0)$ and suppose X and Y are a.s. contained in $U \subset Q$, a twice differentiable Riemannian manifold with geodesic distance d. Under the null hypothesis we assume $\mu_Y = \mu_Z = \mu \in U$. For $p \in U$ let (ϕ_p, U) denote a chart with $\phi_p(p) = 0$. Moreover assume that ρ is uniform coercive and that Assumption 2.7 holds for a random variable X with $\mathbb{P}^X = \frac{n\mathbb{P}^Y + m\mathbb{P}^Z}{n+m}$ and that there is a constant $C > 0$ such that $\|\phi_p(X) - \phi_\mu(X)\|, \|\phi_p(Y) - \phi_\mu(Y)\| \leq C d(p, \mu)$ a.s. for p near μ. Then the classical Hotelling T^2 statistic $T^2(n, m)$ of $\phi_{\mu_{n+m}}(Y_1), \ldots, \phi_{\mu_{n+m}}(Y_n)$ and $\phi_{\mu_{n+m}}(Z_1), \ldots, \phi_{\mu_{n+m}}(Z_m)$ is well defined for $n + m$ sufficiently large where μ_{n+m} denotes a measurable selection of a pooled Fréchet ρ-sample mean (e.g. [1, Chapter 5]). In consequence of Theorem 2.8 we have the following, cf. [25, Theorem 10].

Theorem 2.9 (Two-Sample Test) *Under the above hypotheses for $n, m \to \infty$, $T^2(n, m)$ is asymptotically Hotelling T^2-distributed if either $n/m \to 1$ or if $Cov(\phi_\mu(X)) = Cov(\phi_\mu(Y))$.*

3 Application: Semi-Intrinsic Inference on the Mean Geodesic of Kendall Shapes

The space of geodesics $\Gamma(\Sigma_m^k)$ of Kendall's shape space Σ_m^k of m-dimensional configurations with k landmarks $(m < k)$ can be given the following structure of a double quotient yielding a stratified space (cf. [27, Theorem 5.3])

$$\Gamma(\Sigma_m^k) = O(2) \backslash O^H(m, k-1) / SO(m) \,.$$

Here, the two dimensional orthogonal group $O(2)$ acts from the right and the m-dimensional special orthogonal group $SO(m)$ act from the left on the following submanifold of an orthogonal Stiefel manifold

$$O^H(m, k-1) := \{(x, v) \in M(m, k-1)^2 : \\ \langle x, v \rangle = 0, \langle x, x \rangle = 1 = \langle v, v \rangle, x v^T = v x^T\}$$

where $\langle x, y\rangle = \operatorname{trace}(xy^T)$ denotes the Euclidean inner product. The canonical embedding $O^H(m, k-1)$ in $\mathbb{R}^{m(k-1)} \times R^{m(k-1)}$ with the extrinsic metric leads to a Ziezold link ρ_Z on $\Gamma(\Sigma_m^k) \times \Gamma(\Sigma_m^k)$. We have the following extension of [24, Theorem 3.1].

To this end recall that on a Riemannian manifold the *cut locus* $C(p)$ of p comprises all points q such that the extension of a length minimizing geodesic joining p with q is no longer minimizing beyond q. If $q \in C(p)$ or $p \in C(q)$ then p and q are *cut points*. E.g. on a sphere, antipodals are cut points.

Theorem 3.1 *The following hold*

(i) the Ziezold link ρ_Z is a metric on $\Gamma(\Sigma_m^k)$;
(ii) there is an open and dense set $U \subset \Gamma(\Sigma_m^k)$ that carries the structure of a Riemannian manifold such that ρ_Z^2 is twice differentiable on $U \times U$ except at cut points;
(iii) $U = \Gamma(\Sigma_2^k)$ can be chosen in (ii) in case of $m = 2$.

In case of $m = 2$ the Ziezold metric can be computed explicitly. An example using the two sample test for inference on the mean geodesic of leaf growth as in Example 1 of the Introduction can be found in [24].

4 Sticky and Smeary Limit Theorems

We conclude by surveying some peculiarities that come along with the non-Euclidean nature of sample spaces.

On the Circle Recall that the condition that $A_\phi = \mathbb{E}(\operatorname{Hess}_2 \rho(X, \mu)^2)$ be of full rank is among the Assumptions 2.7 ensuring a Gaussian $\sqrt{n}$-CLT in Theorem 2.8. For the intrinsic metric on the circle it turns out that this condition is necessary. Another condition was local twice differentiability a.s. of $x \to \rho^2\big(X, \phi^{-1}(x)\big)$. For the intrinsic metric this condition is violated whenever X has a non-vanishing density near the cut locus of an intrinsic mean. On the circle it turns out that this condition is not necessary, cf. Theorem 4.2 below.

Let S^1 be the unit circle which we represent by $[-\pi, \pi)$ with the endpoints identified. $\rho(x, y) = \min\{|x - y|, 2\pi - |x - y|\}$, $x, y \in [-\pi, \pi)$ denotes the intrinsic distance. The following three theorems are taken from [20]. The assertion of the first is actually a special case of a deeper result due to [36], that on any complete connected Riemann manifold, the cut locus of an intrinsic mean cannot carry mass at all, if the cut locus can be reached from the intrinsic mean by at least two different minimal geodesics.

Theorem 4.1 *Let X be a random variable on the circle S^1 with intrinsic mean $\mu = 0$. Then*

$$\mathbb{P}\{X = -\pi\} = 0\,,$$

i.e. there can be no point mass antipodal to an intrinsic mean.

If X restricted to some neighborhood of $-\pi$ features a continuous density f, then

$$f(-\pi) \leq \frac{1}{2\pi}\,.$$

Moreover, $\mu = 0$ is contained in a whole continuum of intrinsic means if there is $\epsilon > 0$ such that $f(x-\pi) = \frac{1}{2\pi} = f(\pi - x)$ for all $0 \leq x \leq \epsilon$.

Theorem 4.2 *Let X be a random variable on the circle S^1 with unique intrinsic mean $\mu = 0$ featuring a continuous density f near $-\pi$. Assume that $\mathbb{E}(X^2) = \sigma^2$ where X is viewed as taking values in $[-\pi, \pi)$ and that μ_n is an intrinsic sample mean. If $f(-\pi) < \frac{1}{2\pi}$ then*

$$\sqrt{n}\,\mu_n \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{\big(1 - 2\pi f(-\pi)\big)^2}\right).$$

If $f(-\pi) = \frac{1}{2\pi}$ suppose that there are $\delta > 0$ and $k \in \mathbb{N}$ such that f is $k-1$ times continuously differentiable in $(-\pi, \delta - \pi)$ and $(\pi - \delta, \pi)$ with $f^{(j)}(\pi-) = f^{(j)}(-\pi+) = 0$ for all $1 \leq j < k$, and that there are k-th order continuous directional derivatives with $0 \neq f^{(k)}(\pi-) = (-1)^k f^{(k)}(-\pi+) < \infty$. Then

$$\sqrt{n}\,\operatorname{sign}(\mu_n)\,|\mu_n|^{k+1} \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2\big((k+1)!\big)^2}{\big(2\pi f^{(k)}(-\pi+)\big)^2}\right).$$

Theorem 4.3 *Consider the distribution of X, decomposed into the part λ which is absolutely continuous w.r.t. Lebesgue measure, with density f, and the part η singular to Lebesgue measure. Let $S_1, \ldots, S_k$ be maximal arcs distinct up to their endpoints on which $f \leq \frac{1}{2\pi}$, assume that their interiors* $\operatorname{int} S_j$ *are all disjoint from* $\operatorname{supp}\eta$ *$(j = 1, \ldots, k)$ and that $\{x \in S^1 : f(x) = \frac{1}{2\pi}\}$ is a Lebesgue null-set. Then X has at most k intrinsic means and every* $\operatorname{int}\widetilde{S_j}$ *contains at most one candidate.*

Definition 4.4 We say that the limiting distribution of $\hat{\mu}_n$ is k-th order *smeary* if

$$n^{\frac{1}{2(k+1)}}\,\hat{\mu}_n$$

has a non-trivial limiting distribution.

In the first case of Theorem 4.2, $\hat{\mu}_n$ is 0-th order smeary as is always the case for means on Euclidean spaces. In the second case it is k-th order smeary. Below is an example of a circular von-Mises mixture that is 2nd order smeary.

Example 4.5 For a random variable X on S^1 following a bimodal von Mises mixture density

$$f(x) := \frac{1}{I(a,b,\kappa,\tau)}\left(a\, e^{\kappa\cos x} + b\, e^{-\tau\cos x}\right) \text{ with}$$
$$I(a,b,\kappa,\tau) := \int_{-\pi}^{\pi} (a\, e^{\kappa\cos x} + b\, e^{-\tau\cos x})\, dx$$

and suitable $a, b, \kappa, \tau > 0$ such that there is a major mode at 0 of height $f(0) > (2\pi)^{-1}$ and minor mode at $-\pi$ of height $f(-\pi) = (2\pi)^{-1}$, we have that X features a unique intrinsic mean at 0 (due to Theorem 4.3) which is approached by sample means μ_n with a rate of

$$n^{1/6}\mu_n \to Y \text{ in distribution, with } Y^3 \sim \mathcal{N}\left(0, \frac{9}{\pi^2}\frac{\int_{-\pi}^{\pi} x^2 f(x)\, dx}{(a\kappa e^{-\kappa} - b\tau e^{-\tau})^2}\right)$$

(here $k = 2$ in Theorem 4.2), much slower than $\sqrt{n}$ because $\frac{1}{2}\mathbb{E}(\mathrm{Hess}_2\rho(X,0)^2) = 1 - 2\pi f(-\pi) = 0$.

The Open Book is a model space for the space of phylogenetic trees as introduced by Billera et al. [6], cf. Example 3. More precisely, near a co-dimension-1 singularity the tree space is locally an open book. The d-dimensional open book is defined as

$$Q = S \cup \bigcup_{j=1}^{k} H_j^+$$

with the *spine* $S = \mathbb{R}^{d-1} = Q^0$ and $k \in \mathbb{N}$ *leaves* $H_j^+ = \mathbb{R}^{d-1} \times (0,\infty)$, $(j = 1, \ldots, k)$ $d \in \mathbb{N}$ and $k > 2$. The topological identification is given by $S \sim R^{d-1} \times \{0\}$ is detailed in [21]. For $j = 1, \ldots, k$ introduce the *folding maps*

$$F_j : Q \to \mathbb{R}^d, (x,t) \mapsto \begin{cases} (x,t) & \text{if } x \in H_j^+ \\ (x,-t) & \text{else} \end{cases}$$

and *folded moments*

$$m_j = \int_{\mathbb{R}^d} z\, d\mathbb{P}^{F_j \circ X}(z)\,.$$

Theorem 4.6 ([21]) *There is an index $j_0 \in \{1, \ldots, k\}$ such that $m_j < 0$ for all $j_0 \neq j \in \{1, \ldots, k\}$. Moreover, sample and population means are unique. If they are denoted by $\hat{\mu}_n$ and μ, respectively, with $Y_n = \sqrt{n}\big(F_{j_0}(\hat{\mu}_n) - F_{j_0}(\mu)\big)$, one of the following is true*

(i) $m_{j_0} > 0 \Leftrightarrow \mu \in H^+_{j_0}$ *and the limiting distribution of* Y_n *is a Gaussian on* $\mathbb{R}^{d-1} \times (0, \infty)$*;*
(ii) $m_{j_0} = 0 \Leftrightarrow$ *the limiting distribution of* Y_n *is supported on* $\mathbb{R}^{d-1} \times [0, \infty)$ *assuming* $\mathbb{R}^{d-1} \times \{0\}$ *with positive probability;*
(iii) $m_{j_0} < 0 \Leftrightarrow$ *the limiting distribution for* Y_n *is a Gaussian on* $\mathbb{R}^{d-1} \times \{0\}$.

Definition 4.7 We say that the Fréchet ρ-mean of a random variable X on Q *sticks* to a subset $P_0 \subset P$ if for all compactly supported random variables Y on Q independent of X, there is $C_Y > 0$

$$E^{(\rho)}(Z) \cap P_0 \neq \emptyset \quad \text{for all } \epsilon < C_Y \text{ and random variables } Z \sim \frac{\mathbb{P}^X + \epsilon \mathbb{P}^Y}{1+\epsilon} \text{ on } Q\,,$$

In case (iii) above in Theorem 4.6, μ *sticks* to the spine S. In particular, for suitable random $N \in \mathbb{N}$ we have a.s. that $\hat{\mu}_n \in S$ for all $n \geq N$.

More intrigued non-Gaussian limit theorems can be found in [3] which covers T_4, the space of trees with four leaves. The general picture is still open.

5 Outlook

In this article we have introduced and illustrated for some examples the concept of (semi)-intrinsic statistical analysis on stratified spaces. Although the approach is clear in its outline, many essential details still present challenging research topics, among others, conditions for uniqueness of Fréchet ρ-means, which are only fully resolved for the circle. Moreover, we have touched on the effect of degeneracy of $\mathbb{E}(\mathrm{Hess}_2\rho(X,\mu)^2)$ which may lead to arbitrary slow convergence rates of sample Fréchet ρ-means. Again, to date the precise picture is only known on circles.

In addition to stickiness phenomena, ongoing research on the *kale*, the cone $\mathcal{K} := [0, \infty) \times (\mathbb{R}/\alpha\mathbb{Z})$ of angle $\alpha > 2\pi$, another model space for phylogenetic trees, suggests that the failure of a.s. twice differentiability of $x \to \rho^2\big(X, \phi^{-1}(x)\big)$ (cf. Assumption 2.7) may destroy Gaussianity, also in the non-sticky case.

The following picture materializes: While rates may be lower on positive curvature spaces due to mass near cut loci of means, rates may be considerably higher on non-positive curvature spaces with non-continuous drops in curvature.

To date, these non-Euclidean effects have only been studied on a few spaces such as the circle, the open book and T_4. Obviously this is still considerably far from descriptor spaces of interest such as spaces of geodesics or of concentric circles as in the second example of the introduction. The general theory is open.

Acknowledgements The author is grateful for support by the Niedersachen Vorab of the Volkswagen foundation.

References

1. Anderson, T.: An Introduction to Multivariate Statistical Analysis, 3nd edn. Wiley, New York (2003)
2. Aydın, B., Pataki, G., Wang, H., Bullitt, E., Marron, J.: A principal component analysis for trees. Ann. Appl. Stat. **3**(4), 1597–1615 (2009)
3. Barden, D., Le, H., Owen, M.: Central limit theorems for Fréchet means in the space of phylogenetic trees. Electron. J. Probab. **18**(25), 1–25 (2013)
4. Bhattacharya, R.N., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds I. Ann. Stat. **31**(1), 1–29 (2003)
5. Bhattacharya, R.N., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds II. Ann. Stat. **33**(3), 1225–1259 (2005)
6. Billera, L., Holmes, S., Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27**(4), 733–767 (2001)
7. Blum, H., Nagel, R.N.: Shape description using weighted symmetric axis features. Pattern Recogn. **10**(3), 167–180 (1978)
8. Cappozzo, A., Croce, U.D., Leardini, A., Chiari, L.: Human movement analysis using stereophotogrammetry part 1: theoretical background. Gait Posture **21**, 186–196 (2005)
9. Choquet, G.: Theory of capacities. Annales de l'Institut de Fourier **5**, 131–295 (1954)
10. Damon, J.: Smoothness and geometry of boundaries associated to skeletal structures I: sufficient conditions for smoothness. Annales de l'institut Fourier **53**(6), 1941–1985 (2003)
11. Damon, J.: Global geometry of regions and boundaries via skeletal and medial integrals. Commun. Anal. Geom. **15**(2), 307–358 (2007)
12. Davis, R.B., Ounpuu, S., Tyburski, D., Gage, J.R.: A gait analysis data collection and reduction technique. Hum. Mov. Sci. **10**(5), 575–587 (1991)
13. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, Chichester (1998)
14. Feragen, A., Lauze, F., Lo, P., de Bruijne, M., Nielsen. M.: Geometries on spaces of treelike shapes. Computer Vision–ACCV 2010, pp. 160–173 (2011)
15. Fitch, W., Margoliash, E.: Construction of phylogenetic trees. Science **155**(760), 279–284 (1967)
16. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'Institut de Henri Poincaré **10**(4), 215–310 (1948)
17. Gower, J.C.: Generalized Procrustes analysis. Psychometrika **40**, 33–51 (1975)
18. Hendriks, H., Landsman, Z.: Asymptotic behaviour of sample mean location for manifolds. Stat. Probab. Lett. **26**, 169–178 (1996)
19. Henke, M., Huckemann, S., Kurth, W., Sloboda, B.: Growth modelling of leaf shapes exemplary at leaves of the populus x canadensis based on non-destructively digitized leaves. Silvica Fennica (2014, to appear)
20. Hotz, T., Huckemann, S.: Intrinsic means on the circle: Uniqueness, locus and asymptotics. Ann. Inst. Stat. Math. (2014, to appear)
21. Hotz, T., Huckemann, S., Le, H., Marron, J.S., Mattingly, J., Miller, E., Nolen, J., Owen, M., Patrangenaru, V., Skwerer, S.: Sticky central limit theorems on open books. Ann. Appl. Probab. **23**(6), 2238–2258 (2013)
22. Huckemann, S.: Dynamic shape analysis and comparison of leaf growth. arXiv:1002.0616v1 [stat.ME] (2010, preprint)
23. Huckemann, S.: Inference on 3D Procrustes means: tree boles growth, rank-deficient diffusion tensors and perturbation models. Scand. J. Stat. **38**(3), 424–446 (2011)
24. Huckemann, S.: Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Ann. Stat. **39**(2), 1098–1124 (2011)
25. Huckemann, S.: On the meaning of mean shape: manifold stability, locus and the two sample test. Ann. Inst. Stat. Math. **64**(6), 1227–1259 (2012)
26. Huckemann, S., Hotz, T.: On means and their asymptotics: circles and shape spaces. J. Math. Imaging Vis. (2013). doi:10.1007/s10851–013–0462–3

27. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). Statistica Sinica **20**(1), 1–100 (2010)
28. Jupp, P.E.: Residuals for directional data. J. Appl. Stat. **15**(2), 137–147 (1988)
29. Kendall, D.: Foundations of a theory of random sets. In: Stochastic Geometry, Tribute Memory Rollo Davidson, pp. 322–376. Wiley, New York (1974)
30. Kendall, D.G.: The diffusion of shape. Adv. Appl. Prob. **9**, 428–430 (1977)
31. Kendall, D.G.: Comment on "size and shape spaces for landmark data in two dimensions" by Fred l. Bookstein. Stat. Sci. **1**(2), 222–226 (1986)
32. Kendall, W.S., Le, H.: Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. Braz. J. Probab. Stat. **25**(3), 323–352 (2011)
33. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley, Chichester (1999)
34. Klassen, E., Srivastava, A., Mio, W., Joshi, S.: Analysis on planar shapes using geodesic paths on shape spaces. IEEE Trans. Pattern Anal. Mach. Intell. **26**(3), 372–383 (2004)
35. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry, vol. II. Wiley, Chichester (1969)
36. Le, H., Barden, D.: On the measure of the cut locus of a Fréchet mean (2013, preprint)
37. Mardia, K., Patrangenaru, V.: Directions and projective shapes. Ann. Stat. **33**, 1666–1699 (2005)
38. Marron, J.S., Alonso, A.M.: An overview of object oriented data analysis. Biomet. J. (2014, to appear)
39. Matheron, G.: Random Sets and Integral Geometry. Series in Probability and Mathematical Statistics. Wiley, New York (1975)
40. Molchanov, I.: Theory of Random Sets. Probability and Its Applications, vol. xvi. Springer, London (2005)
41. Munk, A., Paige, R., Pang, J., Patrangenaru, V., Ruymgaart, F.: The one- and multi-sample problem for functional data with application to projective shape analysis. J. Multivar. Anal. **99**, 815–833 (2008)
42. Pizer, S.M., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J.N., Huckemann, S., Marron, J.: Nested sphere statistics of skeletal models. In: Proceedings of Dagstuhl Workshop on Innovations for Shape Analysis: Models and Algorithms (2013, to appear)
43. Siddiqi, K., Pizer, S.: Medial Representations: Mathematics, Algorithms and Applications. Springer, Dordrecht (2008)
44. Skwerer, S., Bullitt, E., Huckemann, S., Miller, E., Oguz, I., Owen, M., Patrangenaru, V., Provan, S., Marron, J.: Tree-oriented analysis of brain artery structure. J. Math. Imaging Vis. (2014, accepted)
45. Telschow, F.J., Huckemann, S.F., Pierrynowski, M.: Discussion: asymptotics for object descriptors. Biometrical J. (2014, to appear)
46. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)
47. Zahn, C., Roskies, R.: Fourier descriptors for plane closed curves. IEEE Trans. Comput. **C-21**, 269–281 (1972)
48. Ziezold, H.: Expected figures and a strong law of large numbers for random elements in quasi-metric spaces. Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes A, pp. 591–602 (1977)
49. Ziezold, H.: Mean figures and mean shapes applied to biological figure and shape distributions in the plane. Biometrical J. **36**, 491–510 (1994)

An Investigation of Projective Shape Space

John T. Kent

1 Introduction

Consider a configuration $X_0(k \times m)$ of k points or landmarks in m-dimensional space. By identifying configurations which are related to one another by a certain group action, we obtain the concept of a "shape" as an equivalence class of configurations. The collection of equivalence classes forms a "shape space". Here are several important examples.

1. *Similarity shape space.* The similarity shape of X_0 can be described as the equivalence class of configurations

$$[X_0]_{\mathrm{SS}} = \{\beta X_0 R + 1_k \gamma^T : \beta > 0,\ R \in SO(m),\ \gamma \in R^m\},$$

   under similarity transformations; β is a scaling parameter, R represents an $m \times m$ rotation matrix, and γ represents a translation parameter. See, e.g., [2];
2. *Reflection similarity shape space.* As above, but now suppose R is an orthogonal matrix (so reflections are allowed). The reflection similarity shape of X_0 can be described as the equivalence class of configurations

$$[X_0]_{\mathrm{RSS}} = \{\beta X_0 R + 1_k \gamma^T : \beta > 0,\ R \in O(m),\ \gamma \in R^m\}$$

   (e.g. [3]);

J.T. Kent (✉)
Department of Statistics, University of Leeds, Leeds LS2 9JT, UK
e-mail: j.t.kent@leeds.ac.uk

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_8

3. *Affine shape space.* Replacing βR by a general nonsingular matrix yields an affine shape as the equivalence class of configurations

$$[X_0]_{\mathrm{AS}} = \{X_0 A + 1_k \gamma^T : A(m \times m) \text{ nonsingular}, \gamma \in R^m\};$$

4. *Projective shape space.* Let $P\mathscr{S}(k, m)$ denote the projective shape space of k landmarks in m dimensions. To describe this space, it is helpful to switch to homogeneous coordinates. Introduce an augmented matrix

$$X = \left[X_0 \; 1_k \right],$$

where 1_k is a k-vector of ones. Then X is a $k \times p$ matrix, where throughout the paper we set $p = m + 1$. If X is written in terms of its rows as

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_k^T \end{bmatrix},$$

then from the point of view of homogeneous coordinates each row x_i^T of X is well-defined only up to a scalar multiple. Then the projective shape of X is defined as the equivalence class of matrices (in homogeneous coordinates)

$$[X]_{\mathrm{PS}} = \{DXB^T : D(k \times k) \text{ diagonal nonsingular}, \; B(p \times p) \text{ nonsingular}\}.$$

Projective geometry is important in computer vision for identifying features in images which are invariant under the choice of camera view ([4, 6]). The matrix B holds information about the location of the focal point of the camera and its orientation. The matrix D is present because in a camera image of a point, it is not possible to determine how far away that point is in the real world.

Of course, as the group of transformations get larger, the number of distinct equivalence classes gets smaller. Unfortunately, working with equivalence classes is rather awkward from statistical point of view. Therefore various superimposition methods have been developed to facilitate quantitative comparisons between shapes. The most successful class of superimposition methods goes under the name of Procrustes analysis.

2 Review of Procrustes Methods

Consider a transformation group $\mathscr{G}$, with elements denoted by g, acting on configurations X_0 by taking X_0 to $g(X_0)$. Suppose that $\mathscr{G}$ can be split into a product of three subgroups, $\mathscr{G}_1, \mathscr{G}_2, \mathscr{G}_3$ in such a way that each $g \in \mathscr{G}$ can be decomposed

(in at least one way) as $g(X_0) = g_3(g_2(g_1(X_0)))$. In any particular application, one or more of the subgroups might be trivial. Write $g = (g_1, g_2, g_3)$ to represent this decomposition. Then, as discussed in [8], the Procrustes approach to shape analysis involves several steps.

1. *Standardization.* Remove the transformation parameters in $\mathcal{G}_1$ by standardization. For example, if $\mathcal{G}_1$ denotes the location-scale group, so $g_1(X_0) = \beta X_0 + 1_k \gamma^T$ for some $\beta > 0$ and $\gamma \in R^m$, it is common to choose X_0 from the equivalence class so that it is centered and scaled,

$$X_0^T 1_k = 0_m, \quad \text{tr}(X_0^T X_0) = 1; \tag{1}$$

2. *Embedding.* Embed the standardized shape into some Euclidean space in such a way as to remove the parameters in $\mathcal{G}_2$. That is, consider a mapping $\phi(X_0) = T$, say, where ϕ has the property that $\phi(X_0) = \phi(g_2(X_0))$ for all g_2 and all standardized configurations X_0;
3. *Optimization.* Define a (partial) Procrustes distance between the shapes of the configurations $X_0^{(1)}$ and $X_0^{(2)}$ by minimizing the Euclidean distance between the embedded objects $T^{(1)} = \phi(X_0^{(1)})$ and $T^{(2)} = \phi(g_3(X_0^{(2)}))$ over the remaining transformation parameters in $\mathcal{G}_3$,

$$d_{\text{PP}}^2(T^{(1)}, T^{(2)}) = \min_{g_3} \text{tr}\left\{\left(T^{(1)} - T^{(2)}\right)^T \left(T^{(1)} - T^{(2)}\right)\right\}; \tag{2}$$

4. *Metric comparisons.* The Procrustes distance (2) can be used directly to compare different shapes. Alternatively, as a slight variant, its infinitesimal version can be used to define a Riemannian metric on shape space and Riemannian distance can be used. In particular, each of the shape spaces under consideration can be viewed as a Riemannian manifold, other than perhaps at some singular points.

For this construction to be useful, two properties must be checked:

- *Symmetry:* $d_{\text{PP}}^2(T^{(1)}, T^{(2)}) = d_{\text{PP}}^2(T^{(2)}, T^{(1)})$;
- *Identifiability:* $d_{\text{PP}}^2(T^{(1)}, T^{(2)}) > 0$ for distinct shapes; that is, d_{PP} is a distance and not just a semi-distance.

Here are details for the three shape spaces described above.

For similarity shape, the standardization step involves centering ($X_0^T 1_k = 0_m$) and scaling ($\text{tr}(X_0^T X_0) = 1$) as in (1). The embedding step is trivial, $\phi(X_0) = X_0$. Only the rotation parameter remains at the optimization stage. For the special case $k = 3$, $m = 2$, it is well-known that the resulting shape space (shapes of triangles) can be identified with the usual sphere of radius 1/2 in R^3. A variant of partial Procrustes distance, known as full Procrustes distance, corresponds to chordal distance on the sphere, and Riemannian distance corresponds to great circle distance. For the purposes of this paper, these are called "Level 1" metrics.

For reflection similarity shape, there are two possible Procrustes approaches. In the first, essentially the same steps as in the previous paragraph can be applied,

with g_3 now denoting an orthogonal transformation (including reflection as well as rotation) and again leading to a Level 1 metric. However, as an alternative approach, it is more elegant at the embedding step to set $T = \phi(X_0) = X_0 X_0^T$ in terms of a configuration X_0 standardized as in (1), so that T is a $k \times k$ positive semi-definite symmetric matrix from which X_0 can be recovered up to an orthogonal transformation on the right. Hence the optimization step is not needed here. Euclidean distance between the embedded configurations will be called a "Level 2" metric in this paper and has been studied by Dryden et al. [3]. The Level 1 and Level 2 metrics are different from one another, even infinitesimally, especially when X_0 is singular or nearly singular.

For affine shape, the standardization step involves centering ($X_0^T 1_k = 0_m$) and orthonormalization ($X_0^T X_0 = I_m$), e.g. by Gram-Schmidt. The simplest embedding is given by $T = X_0 X_0^T$, which again removes the orthogonal transformations, so that no transformation parameters remain at the optimization step. Affine shape space can be identified with the Grassmann manifold of m-dimensional subspaces of R^{k-1} ($k-1$ rather than k to allow for the centering). Euclidean distance between the embedded configurations will be called "Grassmann Euclidean" distance and is another example of a Level 2 metric.

Affine shape space is already very familiar to statisticians from multiple linear regression analysis, where X_0, taken to be centered for simplicity, plus a column of ones represents the design matrix. If y is a centered k-vector of responses, then the ordinary least squares fit of y on X_0 is given by $\hat{y} = X_0(X_0^T X_0)^{-1} X_0^T$, which is unchanged if X_0 is replaces by $X_0 A$ for any nonsingular $p \times p$ matrix A. Note that X_0 and $X_0 A$ have the same column space, so that $\hat{y}$ depends only on the span of the columns of X_0, not on the individual columns themselves.

For projective shape, the standardization is more delicate. As shown in [8], it is possible to find a diagonal matrix D and a nonsingular matrix B such that after standardization the rows of the augmented configuration $X = [x_1, \ldots, x_k]^T$ are unit vectors ($x_i^T x_i = 1$) and the columns are orthonormal up to a scale factor ($X^T X = (k/p) I_p$). This standardization is known as "Tyler standardization" after [11, 12]. After this standardization, X is unique up to multiplication on the left by a diagonal matrix of plus and minus ones, and on the right by an orthogonal matrix. A nice way to remove these remaining indeterminacies at the embedding stage is to define $k \times k$ matrices M and N with entries

$$m_{ij} = |x_i^T x_j|, \quad n_{ij} = (x_i^T x_j)^2 = m_{ij}^2. \tag{3}$$

Then there are two versions of Procrustes distance between the projective shapes: the Euclidean distance between the M matrices (a Level 2 metric), or between the N matrices (a Level 4 metric), respectively. At least for $m = 1$, that is, $p = 2$, it can be shown that these constructions are identifiable.

One way to think about Tyler standardization for general k and m is in terms of a camera image of a scene of k points in m dimensions affinely situated in an "ambient space" R^p, $p = m + 1$. Tyler standardization is equivalent to using a camera with spherical film (rather than the more conventional flat film) and choosing the focal

point chosen so that the moment of inertia of the film image is proportional to the identity matrix I_p.

It is worth commenting on the naming conventions for the different levels of metric. A Level 1 metric involves direct comparisons between standardized configurations (after optimizing over the remaining transformation parameters). A Level 2 metric involves comparisons between second order moments of a standardized configuration, such as $X_0 X_0^T$, (after optimizing over any remaining transformation parameters). Finally the Level 4 metric involves fourth order moments of the original configurations (with no remaining transformation parameters in our projective shape example in $m = 1$ dimension). There is no Level 3 metric in this naming system.

3 Singularities

Here is a brief summary of the singularities that arise in the different shape spaces. Here X_0 denotes a standardized (i.e. centered and scaled) configuration, and X denotes a Tyler standardized augmented configuration. The nature of any singularities depends on the particular shape space and on the metric used.

1. *Similarity shape space, partial or full Procrustes distance.* When $m = 1, 2$, $k \geq 3$, there are no singularities. All that is required is that X_0 be at least a rank 1 matrix. In particular, if $m = 2$ there is no singularity when X_0 has rank $r = 1$. Indeed for $m = 2$, similarity shape space is homogeneous in the language of differential geometry, meaning that every point in shape space looks like every other point.

   However, singularities do arise in similarity shape space in dimensions $m \geq 3$ at configurations X_0 of rank $r \leq m - 2$. The simplest example is given by a set of collinear points in R^3; the singularity arises because the configuration is unchanged under a rotation in R^3 that leaves the axis of the line fixed. The high curvature near such points in shape space has been studied in [7];
2. *Reflection similarity shape space, Level 2 metric*. This space has more singularities. For all dimensions $m \geq 2$, there is a singularity whenever X_0 has rank $r < m$. So a set of collinear points in the plane is viewed as singular under the Level 2 metric for reflection similarity shape space, but not under the Level 1 metric for similarity shape space;
3. *Affine shape space, Grassmann Euclidean metric*. Here by definition the standardized configurations X_0 are assumed to have rank m. There are no singularities;
4. *Projective shape space, Level 1, 2 and 4 metrics*. Here the situation is more complicated. It is easiest to study for $m = 1$, where the Level 1, 2 and 4 metrics all have singularities at the same points in shape space [8].

4 Projective Shape Space $P\mathscr{S}(4,1)$

Projective shape spaces are considerably more complicated than similarity or affine shape spaces. In this section we summarize some of the challenges which appear even in the simplest case $P\mathscr{S}(4,1)$ of $k = 4$ collinear points in $m = 1$ dimension. Let the positions of the landmarks be given by 4 numbers, $u_j,\ j = 1,\ldots,4$. Then the projective shape can be described in terms of a single number known as the cross ratio, one version of which is defined by

$$\tau = \frac{(u_1 - u_2)(u_3 - u_4)}{(u_1 - u_3)(u_2 - u_4)}.$$

Each value of τ represents a different projective shape as τ ranges through the extended real line (with the limits $\pm\infty$ identified with one another).

In [8], it is emphasized that the cross ratio representation of projective shape is not suitable for metric comparisons. Instead various Procrustes distances are considered. Here is a sketch of the main results.

The M and N versions of Procrustes distance in (3) give rise to two simple geometric representations of $P\mathscr{S}(4,1)$. Under the M representation, $P\mathscr{S}(4,1)$ becomes a spherical equilateral triangle, most easily visualized as two great circle arcs along lines of longitude from the north pole to the equator separated by 90^o, together with an arc along the equator connecting them. Each edge of this spherical triangle is an arc of length 90^o.

For the N representation, $P\mathscr{S}(4,1)$ becomes a planar equilateral triangle, with each edge of length 1, say. A plot is given by the outline of the central triangle given in Fig. 4. One edge, AB, say, corresponds to the interval $\tau \in [0,1]$. The next edge, BC, corresponds to the interval $\tau \in [1,\infty]$, or equivalently, $(\tau - 1)/\tau) \in [0,1]$. The final edge, CA, corresponds to the interval $\tau \in [-\infty, 0]$, or equivalently, $1/(1-\tau) \in [0,1]$. The relevance of these nonlinear mappings of τ can be explained in terms of the effect on τ of various permutations of the labels of the landmarks.

However, the neatness of these geometrical representations hides some of the more subtle aspects of projective shape:

1. visualizing when different configurations have the same projective shape is intuitively difficult. Figure 1 illustrates several configurations for which the outer landmarks u_1 and u_4 are held fixed, but the inner two landmarks vary in such a way that the cross ratio remains fixed at $\tau = 0.3$. The human eye is not very good at recognizing that these configurations have the same cross ratio. (On the other hand the human eye is excellent at deducing depth information from stereo images!);
2. Tyler standardization provides a mathematically elegant way to standardize a configuration. However, in real world applications a film image of an underlying configuration will be observed subject to errors (either in the location of the landmarks in the ambient space R^p or in the location of the landmarks on the film image). Unfortunately the way these errors influence the distribution of projective

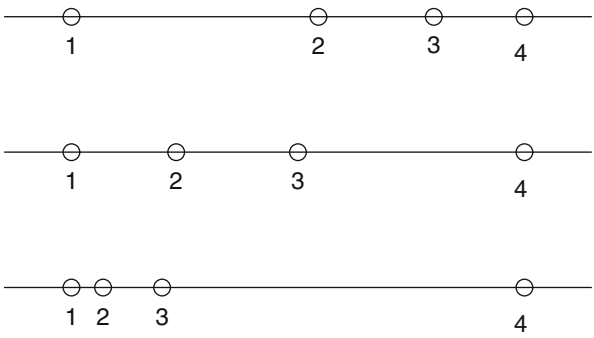

Fig. 1 Various configurations of four collinear points with the same cross ratio

shape depends on the "pose" of the original configuration in the ambient space in a considerably more complicated manner than is the case for similarity and affine shapes. A partial analysis is given in [8];

3. the singularities in projective shape space are well-defined mathematically, but are somewhat unexpected intuitively. In the current setting, $k = 4, m = 1$, they correspond to either a single pair coincidence (e.g. $u_1 = u_2 < u_3 < u_4$) or a double pair coincidence (e.g. $u_1 = u_2 < u_3 = u_4$). On the other hand, the singular points do not look very special when looked at in terms of the cross ratio;
4. the implications on statistical understanding of different metrics, e.g. those based on either Level 2 or Level 4 Procrustes analysis, is still not entirely clear;
5. a related issue is the difficulty in constructing useful and tractable models on projective shape space. The next sections give some initial suggestions.

5 Uniform Distributions on $P\mathscr{S}(4, 1)$

In this section we explore possible uniform distributions on $P\mathscr{S}(4, 1)$. In terms of the planar triangle representation, there are at least three general approaches. For each approach the density is the same on each side of the triangle and is symmetric about the midpoint of each side. Hence we limit attention to one side, parameterized by $0 < \tau < 1$. The form of the three densities is given as follows.

1. (Independent sampling) Take four independent points from a specified distribution on the line and compute the resulting distribution of the cross ratio. The distribution under normality was worked out by Maybank [9, 10] and the distribution under uniformity was worked out by [1]. The latter distribution is more tractable to write down here and can be expressed as

$$f_{\text{Ind}}(\tau) = f_0(\tau) + f_0(1 - \tau), \quad 0 < \tau < 1,$$

where

$$f_0(\tau) = \{(\tau + 1)\log\tau + 2(1-\tau)\}/(\tau - 1)^3;$$

2. (Level 2 metric) Consider a uniform distribution on the spherical triangle representation of $P\mathscr{S}(4,1)$. Thus on each edge, the angle $\delta \in (0, \pi/2)$ is uniformly distributed. After changing the variable from δ to $\tau = \sin^2\delta$ on the edge for $0 < \tau < 1$, the density becomes

$$f_{L2}(\tau) = 1/\{\pi\sqrt{\tau - \tau^2}\}, \quad 0 < \tau < 1;$$

3. (Level 4 metric) Consider a uniform distribution on the planar triangle representation of $P\mathscr{S}(4,1)$, so

$$f_{L4}(\tau) = 1, \quad 0 < \tau < 1.$$

In each case the density has been scaled to integrate to 1 over the interval $0 < \tau < 1$ and should be divided by 3 and repeated on the other two edges of the triangle to give the corresponding density over all of $P\mathscr{S}(4,1)$. A plot of these three densities on the τ scale is given in Fig. 2. Note that f_{Ind} and f_{L2} are difficult to tell apart and have poles at the endpoints $\tau = 0, 1$. In contrast f_{L4} is flat.

It is also of interest to plot these densities on the δ scale, where we rewrite each density as a function of δ and introduce the Jacobian factor $\mathrm{d}\tau/\mathrm{d}\delta = 2\sin\delta\cos\delta = \sin 2\delta$ for the change of variables. The resulting densities are in Fig. 3. Now all three densities are quite distinct. Both f_{Ind} and f_{L4} vanish at the endpoints, though f_{L4}

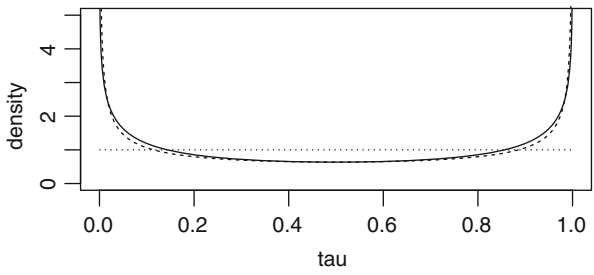

Fig. 2 Densities for three possible "uniform" distributions in τ coordinates, plotted for $0 < \tau < 1$: independence model (*solid*), level 2 uniform (*dashed*) and level 4 uniform (*dotted*)

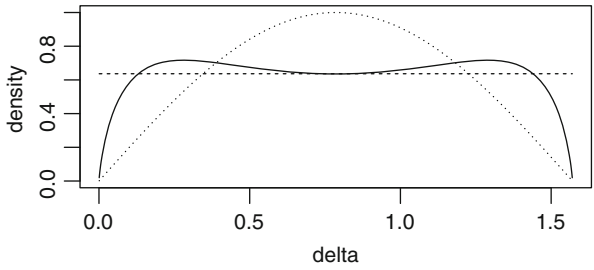

Fig. 3 Densities for three possible "uniform" distributions in δ coordinates, plotted for $0 < \delta < \pi/2$: independence model (*solid*), level 2 uniform (*dashed*) and level 4 uniform (*dotted*)

converges to 0 more quickly. In the middle of the interval f_{Ind} is bimodal whereas f_{L4} is unimodal. In contrast f_{L2} is flat throughout.

6 Constructing Distributions on $P\mathscr{S}(4,1)$ About a Preferred Shape

Both the spherical and planar representations of $P\mathscr{S}(4,1)$ are topological circles. Thus one modelling strategy is to ignore the corners and treat them as actual circles. Then fit a standard circular model such as a von Mises distribution, which can allows concentration about any specified projective shape. This strategy was explored by Goodall and Mardia [5].

However, in this paper we look at a different strategy, which is valid for any compact manifold which can be embedded in a Euclidean space. Namely, we construct an exponential family based on first, and possibly second, moments in the Euclidean coordinates. This strategy has been very common and successful in directional data analysis, yielding the Fisher and Bingham distributions and various generalizations.

For the spherical triangle representation of $P\mathscr{S}(4,1)$, the simplest strategy is to condition the linear-exponential function $\exp(a_1w_1 + a_2w_2 + a_3w_3)$ to lie on the spherical triangle, where w_1, w_2, w_3 are the Euclidean coordinates and a_1, a_2, a_3 are parameters. The resulting distribution will be truncated von Mises on each arc. However, since the cumulative distribution function of the von Mises distribution is a bit awkward to work with, we will not consider this density further here.

For the planar triangle representation of $P\mathscr{S}(4,1)$, the same strategy of using just first order Euclidean terms in the exponent of the density does not work very well. The resulting densities are not very flexible because the mode of the density must lie either at a vertex or uniformly along an edge.

Hence we consider an exponential family in the plane based on linear and quadratic terms in the Euclidean coordinates. The simplest version of this strategy is to condition an isotropic bivariate normal distribution to lie on the planar triangle. This distribution is explored in the next section.

7 Conditioned Normal Distribution for the Planar Representation of $P\mathscr{S}(4,1)$

In this section we look at a bivariate normal distribution $N_2(\mu, \sigma^2 I)$, conditioned to lie on the planar equilateral triangle representation of projective shape space.

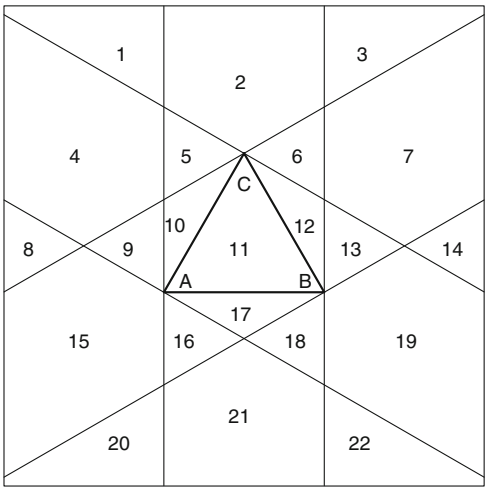

Fig. 4 The outline of the heavy central triangle is a Level 4 representation of projective shape space for four collinear points. The labels 1–22 demarcate 22 possible regions for the mean parameter μ in the conditioned isotropic bivariate normal distribution

An explicit coordinate representation of this equilateral triangle is given by the choice of vertices

$$v_A = \frac{1}{\sqrt{3}} \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ -\frac{1}{2} \end{bmatrix}, \quad v_B = \frac{1}{\sqrt{3}} \begin{bmatrix} \frac{\sqrt{3}}{2} \\ -\frac{1}{2} \end{bmatrix}, \quad v_C = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

so that the edges have length 1, edge AB is horizontal, and the center of the triangle is at the origin.

In order to study the shape of the bivariate normal distribution conditioned to lie on this triangle, it is helpful to divide the parameter space for $\mu \in R^2$ into 22 regions, as shown in Fig. 4. The shape space $P\mathscr{S}(4,1)$ is denoted by the central triangle with thick edges and vertices marked A,B,C. At each vertex, two lines have been plotted, orthogonal to each of the two edges at the vertex. These lines, plus the original triangle partition R^2 into 22 regions, as marked.

The behavior of the density on a particular edge, e.g. AB, depends on whether μ lies inside or outside the corresponding parallel lines perpendicular to that edge. If μ lies inside the parallel lines, then the quadratic form has a minimum, and hence the density has a local maximum, on the edge. If μ lies outside the parallel lines, then the quadratic form is monotone increasing (and the density is monotone decreasing) along the edge from the nearer parallel line to the further parallel line.

It is convenient to classify the behavior of the conditioned normal density into five types:

Type I. The density is unimodal with the mode within one edge (regions 4,7,21).
Type II. The density is unimodal with the mode at a vertex (regions 1,3,8,14, 20,22).

Type III. The density is bimodal with the modes within two edges (regions 5,6,9, 13,16,18).

Type IV. The density is bimodal with the modes at one vertex and within the opposite edge (regions 2,15,19).

Type V. The density is trimodal with a mode within each edge (regions 10,11, 12,17).

Figure 5 gives plots of each of these different types of density, with μ taking the values $-2v_A$, $-2v_A - 4v_B, v_A + .2(v_B - v_C), -2v_A - 2v_B$, 0, respectively. Note that Types I and II are likely to be the most relevant in practice.

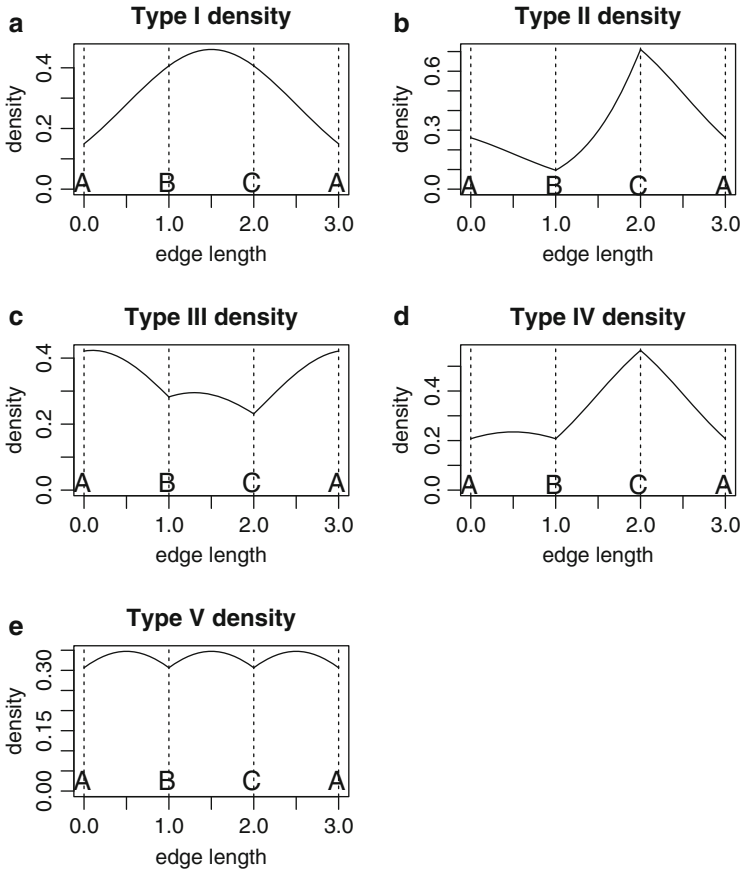

Fig. 5 Plots of the conditioned normal density for the planar triangle representation of $P\mathscr{S}(4,1)$ along the edges AB, BC, CA. Parts (**a**)–(**e**) illustrate densities of Types I–V, respectively

This section has focused on the case of an isotropic covariance matrix $\Sigma = \sigma^2 I$ for the underlying bivariate normal distribution. It is interesting to consider what happens if Σ is unrestricted. In this case it is possible to divide the parameter space for μ into different regions similarly to Fig. 4, except the thin lines are now orthogonal to the edges of the triangle in the Σ metric rather than the Euclidean metric. Provided the level of anisotropy is not too severe, there is still a split into 22 regions. However, if the level of anisotropy is extreme, the thin lines will cross over the edges of the triangle and the partition into regions will be more complicated.

8 Discussion

The Procrustes approach to projective shape offers an elegant mathematical framework to study projective shape. However, the construction of useful densities in this setting is considerably more complicated than in the more traditional setting of similarity shape analysis. In particular, there are at least three plausible candidates for the label of "uniform" distribution on $P\mathscr{S}(4,1)$ and a variety of approaches to construct more concentrated distributions. More work is needed to fully appreciate the implications of this framework for computer vision and statistical inference.

Acknowledgements This work benefited from a visit to the Statistical and Applied Mathematics Sciences Institute (SAMSI), North Carolina and a workshop at the Mathematical Biosciences Institute (MBI), Ohio, especially through helpful discussions with Ian Dryden, Thomas Hotz, Huiling Le, Stephan Huckemann, Kanti Mardia, Ezra Miller, and Vic Patrangenaru.

References

1. Åström, K., Morin, L.: Random cross ratios. In: Borgefors, G. (ed.) Proceedings of the 9th Scandinavian Conference on Image Analysis, pp. 1053–1061. Swedish Society for Automated Image Analysis, Göteborg (1995)
2. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, Chichester (1998)
3. Dryden, I.L., Kume, A., Le, H., Wood, A.T.A.: A multi-dimensional scaling approach to shape analysis. Biometrika **95**, 779–798 (2008)
4. Faugeras, O., Luong, Q.-T.: The Geometry of Multiple Images. MIT Press, Cambridge (2001)
5. Goodall, C.R., Mardia, K.V.: Projective shape analysis. J. Comput. Graph. Stat. **8**, 143–168 (1999)
6. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000)
7. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: geodesic principal component analysis for Riemannian manifolds modulo isometric Lie group actions (with discussion). Stat. Sinica **20**, 1–100 (2010)
8. Kent, J.T., Mardia, K.V.: A geometric approach to projective shape and the cross ratio. Biometrika **99**, 833–849 (2012)
9. Maybank, S.J.: Classification based on the cross ratio. In: Mundy, J.L. Zisserman, A., Forsyth, D. (eds.) Applications of Invariance in Computer Vision. Lecture Notes in Computer Science, vol. 825, pp. 453–472. Springer, Berlin (1994)

10. Maybank, S.J.: Probabilistic analysis of the application of the cross ratio to model based vision. Int. J. Comput. Vis. **16**, 5–33 (1995)
11. Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**, 234–251 (1987)
12. Tyler, D.E.: Statistical analysis for the angular central Gaussian distribution on the sphere. Biometrika **74**, 579–589 (1987)

Treelet Decomposition of Mobile Phone Data for Deriving City Usage and Mobility Pattern in the Milan Urban Region

Fabio Manfredini, Paola Pucci, Piercesare Secchi, Paolo Tagliolato, Simone Vantini, and Valeria Vitelli

1 Introduction

Interpretative tools for the identification of mobility practices in the contemporary cities are needed, not only for some known limitations of traditional data sources, but also because new forms of mobility are emerging, describing new city dynamics and time-variations in the use of urban spaces by temporary populations [3, 5, 15]. These practices challenge the analytical tools and the conventional data sources used for urban and mobility investigations (i.e. surveys, census), unable to describe adequately the space-time variability in the use of the city as well as the combined movements of people, objects and information in their complex relational dynamics [12, 15].

The operative challenges opened up by the new forms of mobility emerging in the contemporary city are measured in terms of their capacity to integrate different approaches. One approach employs the aggregate method (Origin/Destination flows) to study mobility as geographic displacement, recognizing a proportional

F. Manfredini • P. Pucci • P. Tagliolato
Dipartimento di Architettura e Studi Urbani, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: fabio.manfredini@polimi.it; paola.pucci@polimi.it; paolo.tagliolato@polimi.it

P. Secchi • S. Vantini (✉)
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: piercesare.secchi@polimi.it; simone.vantini@polimi.it

V. Vitelli
Department of Biostatistics, University of Oslo, Domus Medica, Sognsvannsveien 9, 0372 Oslo, Norway
e-mail: valeria.vitelli@medisin.uio.no

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_9

relationship between the utility and the cost/time of movement. Another approach explains mobility as a spatialized form of social interaction and considers mobility as a social capital and the territory as a space of social interactions facilitated by mobility. If both approaches are relevant to describe different patterns of mobility in their social and spatial differentiation, it becomes important to accompany the traditional quantitative approaches referred to a geographic displacement that tends to focus on movement in space and time, in an aggregate way and for limited time periods, with data sources able to describe fine grain over-time variation in urban movements.

In this perspective, an interesting contribution may be provided by mobile phone network data as a potential tool for the real-time monitoring of urban dynamics and mobile practices, as tested in several experimental studies [1, 4, 8]. The application researches focused on two different products. Some studies deal with aspects of representation of the data, emphasizing the aspects most directly evocative, to highlight how these data may represent the "Mobile landscapes" [8]. Other studies focus on data-mining analysis to building methods for managing large amounts of data, and on the construction of instruments capable of deriving summary information and relevant data on cell-phone [1]. As opposed to the more traditional methods of urban surveys, the use of aggregated and anonymous mobile phone network data has shown promise for large-scale surveys with notably smaller efforts and costs [9].

If we consider the observed and aggregated telephone traffic as the result of individual behaviors and habits, we can treat mobile phone data as a useful source on the real use of the cities, capturing for example traces of temporary populations, which are difficult to intercept by traditional data sources, but which, at the same time, increasingly affect urban practices both quantitatively and qualitatively. An increasing number of studies concerns the exploitation of mobile phone data in urban analysis and planning[2]. In particular an interesting issue regards the classification of urban spaces according to their users' practices and behaviors [9, 13]. In [14] the authors outline the fact that city areas are generally not characterized by just one specific use, and for this reason they introduce the use of fuzzy c-means, a fuzzy unsupervised clustering technique for land use classification, which returns for each area a certain grade of membership to each class. In the same paper fuzziness is then abandoned to favor the identification of areas with a clearly defined use. We want to drive the reader's intuition on the interesting point that different "basic" profiles of city usages can concur in the same place and that the overall observed usage of a certain place is the superimposition of layers of these basic profiles.

According to this synthetic framework on the challenges that mobility practices pose to traditional sources and approaches, in this article we experiment a novel geo-statistical unsupervised learning technique finalized to identify useful information on hidden patterns of mobile phone use regarding different usages of the city in time and in space which are related to individual mobility, outlining the potential of this technology in the urban planning community. The analysis return new maps of the region, each describing the intensity of one of the identified mobility pattern

on the territory. In detail, the territorial distribution of the intensity of these patterns allows us to reconstruct the density of use of urban spaces in different temporal, and territorial scales as a precondition:

- to identify temporary populations and different forms of mobility that structure the relationships in the contemporary city;
- to propose diversified management policies and mobility services that city users require, increasing the efficiency of the supply of public services.

2 Data

For the present research we had the opportunity to use the same data that feeds the CityLive platform developed by Telecom Italia for the real time evaluation of urban dynamics based on the anonymous monitoring of mobile phone networks. Telephone traffic is anonymously recorded by each cell of the network as the average number of concurrent contacts in a time unit. Telecom Italia elaborate then these measurements obtaining their distribution by means of weighted interpolations, throughout a tessellation of the territory in squared areas (pixels).

In the Telecom Italia database, the metropolitan area of Milan is divided into a uniform grid (lattice) S_0 of 97×109 pixels. For each pixel, Telecom Italia made available the average number of mobile phones simultaneously using the network for calling, for every 15-min time interval along a period of 14 days. This quantity is called Erlang and, to a first approximation, can be considered proportional to the number of active people in that pixel at that time interval, hence providing information about people density and mobility. Technically the Erlang $E_{\mathbf{x}j}$ relevant to the pixel $\mathbf{x} \in S_0$ and to the jth quarter of an hour is computed as:

$$E_{\mathbf{x}j} = \frac{1}{15} \Sigma_{q=1}^{Q} |T_{\mathbf{x}j}^{q}| \,, \tag{1}$$

where $T_{\mathbf{x}j}^{q}$ indicates the time interval (or union of intervals) in which the qth mobile phone is using the network for calling within pixel $\mathbf{x}$ and during the jth quarter of an hour. $|T_{\mathbf{x}j}^{q}|$ indicates the length of $T_{\mathbf{x}j}^{q}$ expressed in minutes. The number of potential phones using the network is indicated with Q. Even though the phone company uses Eq. (1) to compute $E_{\mathbf{x}j}$, the meaning of this quantity is better understood from its equivalent representation:

$$E_{\mathbf{x}j} = \frac{1}{15} \int_{15(j-1)}^{15j} N_{\mathbf{x}}(t) dt, \tag{2}$$

where $N_{\mathbf{x}}(t)$ indicates the number of mobile phones using the network within the pixel $\mathbf{x}$ at time t. Equation (2) shows that $E_{\mathbf{x}j}$ is the temporal mean over the jth quarter of an hour of the number of mobile phones using the network within pixel $\mathbf{x}$.

The Erlang data we deal with are recorded, with missing values, from March 18, 2009, 00:15 a.m., till March 31, 2009, 23:45 p.m., providing $p = 1,308$ records per pixel. The lattice of pixels S_0 covers an area of 757 km^2 included between latitudes 45.37 and 45.57 and longitude 9.05 and 9.35 which corresponds to the Milan core city and to the first ring of municipalities surrounding Milan, located along the ring road. It is divided in $N = 97 \times 109 = 10,573$ approximately rectangular pixels. On the whole, 13, 829, 484 records are available. To have a first idea of these data, in the top panel of Fig. 1 the aggregated Erlang, i.e. the sum of the Erlang measures for each pixel in the investigated area, $\sum_{\mathbf{x} \in S_0} E_{\mathbf{x}j}$, is represented as a function of the corresponding quarter of an hour $j = 1, \ldots, p$. Some peculiar features are already

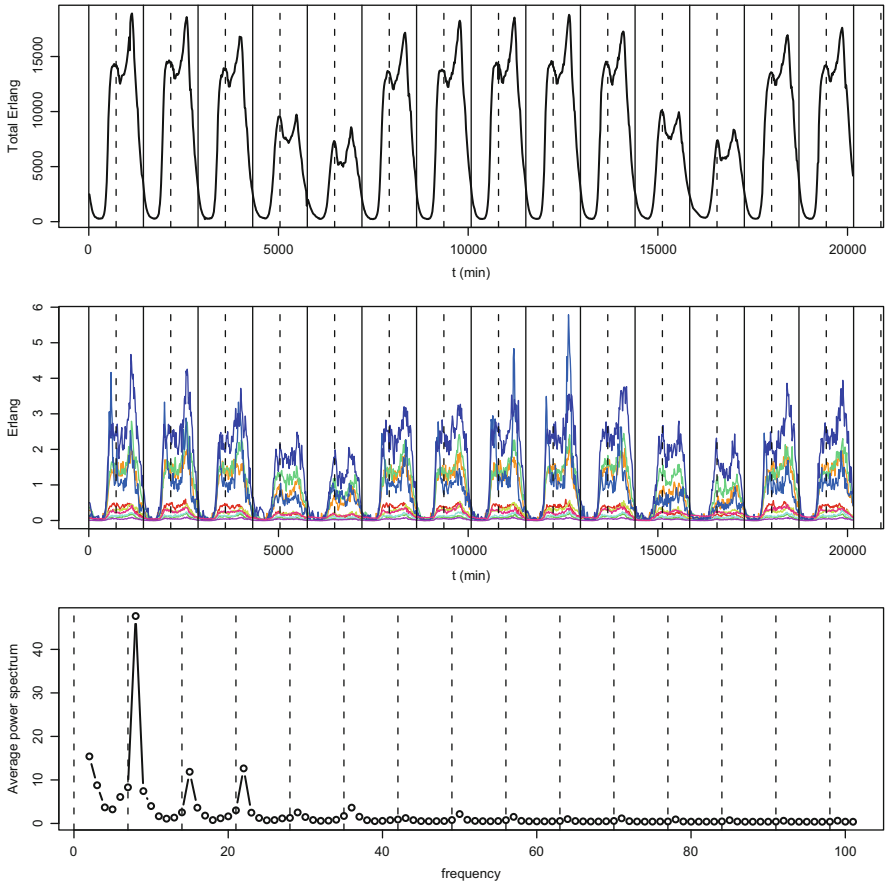

Fig. 1 Erlang data: (*top*) the aggregated Erlang of the investigated area as a function of time; (*middle*) a random selection of ten Erlang data, drawn at random among the sites of the lattice, as a function of time; the *solid vertical lines* are drawn at midnight of each day, and the *dotted vertical lines* at midday. The first day is Wednesday March 18, 2009. (*Bottom*) average power spectrum; *Dotted vertical lines* are drawn for multiples of 7

noticeable, such as the day/night effect and working/weekend day effect. The aim of the analysis is indeed to identify these global features together with those that are more local, both in terms of time and space.

3 Methodology

3.1 Data Preprocessing: Fourier Expansion

In each pixel, we may consider the process of the Erlang measures over time, which can be thought as a continuous process in time describing the average number of mobile phones using the network in that site (see Eq. 2). An example of the observed Erlang profiles along time is shown in Fig. 1 (middle): ten sites have been randomly selected in the lattice, and the Erlang measures recorded in each selected site have been plotted as a function of time. It is clear from the picture that, beside macro periodic behaviors due to week and daily-seasonality in the average use of mobile phone, Erlang data present strongly localized features.

Indeed, in each site of the lattice we observe a discrete version of the Erlang continuous process, recorded every quarter of an hour: due to discontinuities in information provided by the network antennas, the Erlang measure is missing at some time intervals, and hence the time grid of the Erlang measurements is non-uniform. We thus need to choose a proper basis expansion to reconstruct the functional form of the time-varying Erlang data on a common grid of time values.

To this purpose, we perform a pixel-wise smoothing of the Erlang data via a Fourier basis expansion of period one-week. The idea is to estimate each Erlang profile as a function in time obtained as a weighted sum of sinusoids of increasing frequency. Formally, the reconstructed Erlang profile relative to the pixel $\mathbf{x} \in S_0$ is a function $E_{\mathbf{x}}(t) = \frac{c_0^{\mathbf{x}}}{2} + \sum_{h=1}^{H} \left[a_h^{\mathbf{x}} \cos(h\omega t) + b_h^{\mathbf{x}} \sin(h\omega t)\right]$, where $t \in [0; T]$, $\omega = 2\pi/T$ and $T = 60{\cdot}24{\cdot}7$ is the period expressed in minutes with the coefficients $c_0^{\mathbf{x}}$, $a_h^{\mathbf{x}}$ and $b_h^{\mathbf{x}}$ estimated through ordinary least squares.

In Fig. 1 (bottom) the average power spectrum of the Telecom Italia database is reported. This plot shows, for each frequency, the relevant average contribution of the corresponding sinusoid to the Erlang profiles observed within the investigated area. From a graphical inspection of the plot, it is clear that the frequencies significantly contributing to the Erlang time variation are the smaller ones (all less than 7), capturing the difference among days or blocks of days (e.g., the working and weekend days variation), and the frequencies multiple of 7, capturing the recurring daily dynamics. Note that if only the frequencies multiple of 7 were present, the Erlang profiles would be daily-periodic. For an extensive description of smoothing procedures for functional data we refer to [7].

3.2 Dimensional Reduction: Treelet Decomposition

After data preprocessing, we now aim at the identification of a set of "reference signals" able to synthetically describe the different temporal patterns of utilization of the mobile phone network across the region; and of a set of "influence maps" pointing out site-by-site the contribution of each reference signal to the site Erlang profile.

Coherently, in this work, we assume that a limited number of time-varying basis functions, common to the entire area under investigation, are sufficient to describe pixel-by-pixel all the corresponding Erlang profiles. Indeed we will interpret the basis functions as describing the Erlang profile associated to a specific temporal dynamic related to a human activity. More formally, we assume the following model for the generation of Erlang data

$$E_{\mathbf{x}}(t) = \sum_{k=1}^{K} d_{\mathbf{x}k}\,\psi_k(t) + \epsilon_{\mathbf{x}}(t), \tag{3}$$

where $\{\psi_1(t), \ldots, \psi_K(t)\}$ is the set of time-varying basis functions and $d_{\mathbf{x}1}, \ldots, d_{\mathbf{x}K}$ describe their contribution to the Erlang profile relative to the pixel $\mathbf{x}$. The quantity $\epsilon_{\mathbf{x}}(t)$ represents an error term describing unstructured variability of the Erlang data.

In the statistical literature, the process leading to the identification of the finite dimensional basis $\{\psi_1(t), \ldots, \psi_K(t)\}$ and of the coefficients $d_{\mathbf{x}1}, \ldots, d_{\mathbf{x}K}$ is known as *dimensionality reduction*. A very common procedure for dimension reduction is Principal Component Analysis [7]. In this work we use a method known as *treelet analysis* introduced in [6].

Treelets (i.e., the estimates of the basis functions $\psi_1(t), \ldots, \psi_K(t)$) have been originally proposed as a surrogate of wavelets for dealing with unordered variables. Nevertheless, we found them to be an effective dimension reduction technique for Erlang profiles and, more generally, for data with peculiar functional features, like spikes, periodicity, outliers.

Similarly to wavelets, the treelet decomposition has the property of following a hierarchical structure interpretable in a multiscale framework. Differently from wavelets, treelets are data-driven. More specifically, the treelet analysis generates a sparse multiscale orthonormal set on functions iteratively detected through nested pairwise Principal Component Analysis. See [6] for further details.

Once the treelets $\psi_1(t), \ldots, \psi_K(t)$ have been identified, for each pixel $\mathbf{x} \in S_0$ their respective contributions $d_{\mathbf{x}1}, \ldots, d_{\mathbf{x}K}$ to the Erlang profile $E_{\mathbf{x}}(t)$ are obtained by orthogonal projection.

3.3 Spatial Smoothing: Bagging Voronoi Tessellations

The contribution $d_{\mathbf{x}r}$ of the rth treelet $\psi_r(t)$ to the local Erlang profile $E_{\mathbf{x}}(t)$ is expected to vary smoothly in space because of the spatial dependence between Erlang profiles recorded in close sites which is induced by the arbitrary segmentation of the area and by the mobility of phone users. Thus an improved estimate of $d_{\mathbf{x}r}$ for pixel $\mathbf{x}$ can be obtained by borrowing information from neighboring pixels. In an urban setting, the identification of an optimal neighborhood system is not a trivial issue because of the unisotropic and dishomogeneous nature of the urban matrix. Indeed, detecting close sites is more an aim of the analysis than a starting point.

For this reason we decided to exploit spatial dependence in a fully non parametric setting using a Bagging strategy based on Voronoi tessellations proposed in [11]. We refer to this paper for a deeper understanding of the rationale behind the Bagging Voronoi Tessellation strategy, which is however easily described:

(*i*) we build a neighborhood system by randomly generating a Voronoi tessellation covering the entire area under study;
(*ii*) for each neighborhood (i.e. for each element of the tessellation) we exploit spatial dependence by computing the median of the values $d_{\mathbf{x}r}$ relative to the pixels $\mathbf{x}$ within the neighborhood. We attribute the value of that median to each pixel $\mathbf{x}$ within the neighborhood;
(*iii*) we then repeat steps (*i*) and (*ii*) B times (known as bootstrap replicates).

At the end of the B bootstrap iterations, to each pixel $\mathbf{x}$ corresponds a sample of B medians; this sample is summarized by its median $\hat{d}_{\mathbf{x}r}$ which provides an improved estimate of $d_{\mathbf{x}r}$, taking into account spatial dependence. By plotting, for each pixel in the lattice S_0, the value $\hat{d}_{\mathbf{x}r}$ we obtain a smooth surface describing the variation in space of the contribution of the treelet $\psi_r(t)$ to the local Erlang profiles and thus identifying regions within the urban matrix that are similar with respect to the human activity characterized by the treelet $\psi_r(t)$.

On the whole, the entire methodology allows us to identify:

- a reference basis $\{\psi_1(t), \ldots, \psi_K(t)\}$, i.e the set of basis functions describing the specific effects on the Erlang data of some human activities recorded in the area. In the top panels of Figs. 2, 3, 4, 5, 6 and 7 a selection of the most easily interpretable treelets is reported;
- a set of maps, i.e. the set of spatially-varying functions $\{\hat{d}_1(\mathbf{x}), \ldots, \hat{d}_K(\mathbf{x})\}$) showing pixel-by-pixel the contribution of each treelet to the local Erlang profile. In the top panels of Figs. 2, 3, 4, 5, 6 and 7 the maps corresponding to the treelets illustrated on top are reported.

From a computational point of view, this procedure is of course more time consuming than a simple treelet analysis since it further requires: the generation of B random Voronoi maps; the computation of the local medians for each element of each Voronoi map and for each treelet; and then the computation of the bootstrap

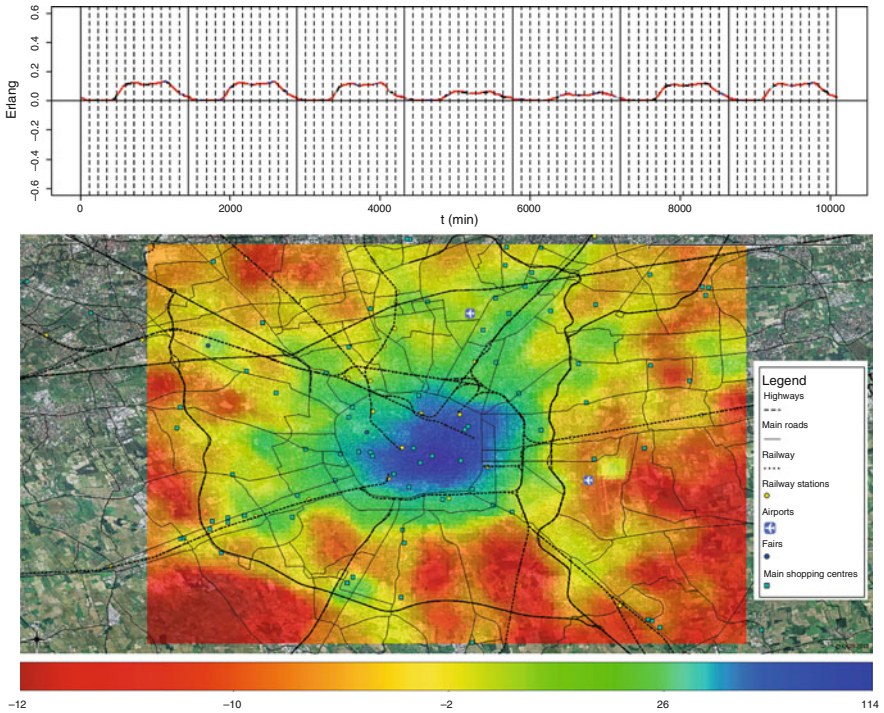

Fig. 2 Treelet 1—The "average use" treelet map. The treelet contains different temporal patterns of mobile phone activity (i.e. daily, working day versus week end) that fit with actual city usage

medians of the local medians for each site and for each treelet. Despite of these, if the procedure is suitably implemented, the increment in the computational time can be dramatically reduced. Indeed, the generation of the random Voronoi maps can be performed off-line since it only requires to know the location of sites and not the actual data; the computation of the local medians can be fully parallelized over maps, elements, and treelets; and finally the computation of the bootstrap medians can be fully parallelized over sites and treelets.

4 Case Study: Experimenting Milan Mobility Patterns

4.1 Case Study: Milan

In our work, urban planning expert knowledge showed the potential of this methodology. We discuss in the present paragraph the specific case of the Milan urban region. We analyzed and mapped "hidden mobile phone use patterns" derived from

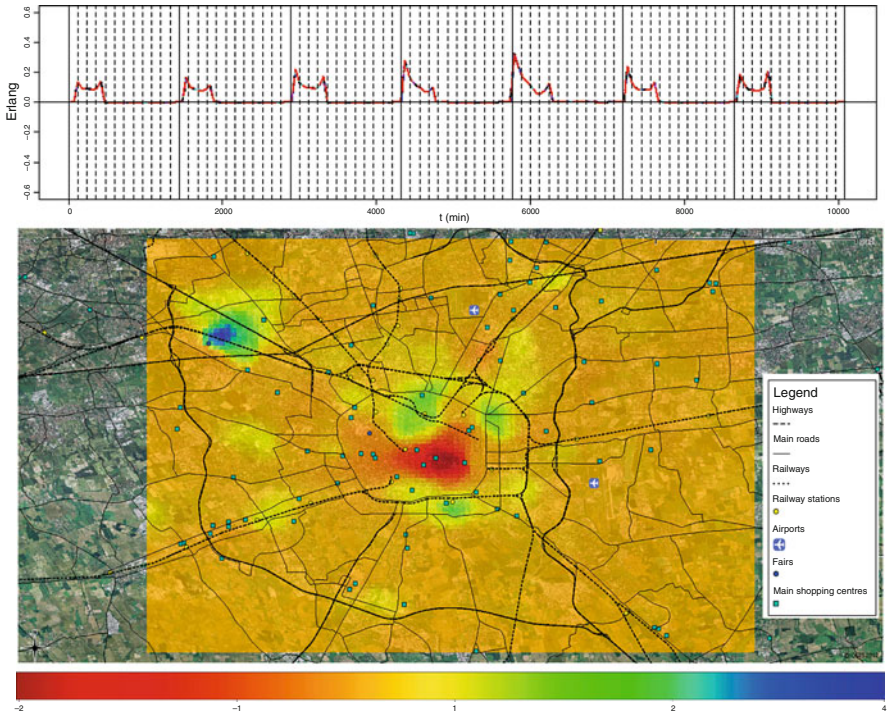

Fig. 3 Treelet 2—Nightly activity. *Hot spots* highlight the presence of night work: organization of International Fair at the Rho-Fiera exhibition site (North-West); delivering and distributing products in the Fruit and Vegetable Wholesale market (second circular ring of the city—South East); scarce nightly activity inside the city centre

the treelets analysis in order to verify the potential of this method for explaining spatial urban usage and mobility patterns.

Milan is an urban region which goes far beyond its administrative boundaries. The core city and the whole urban area have been affected in the last 20 years by relevant changes in their spatial structures and have generated new relationships between centre and suburbs. Daily mobility patterns are now even more complex than in the past when a hierarchical structures of cities was present and the physical relationship between jobs and homes was the main reason of mobility. Now the commuter flows describe only a minor part of the overall urban movements (about the 29 %, excluding returning home). Daily mobility is generated by many other reasons which are becoming increasingly relevant. These non-systematic flows are related to individual habits and are the effects of diversified and complex uses of the Milan urban region. For the intrinsic characteristics of this kind of mobility, it is difficult to measure its dimension and its intensity, in space and in time and systematic studies or sources which provide this information in Italy do not exist. Within the Milan urban region services and activities are distributed in a

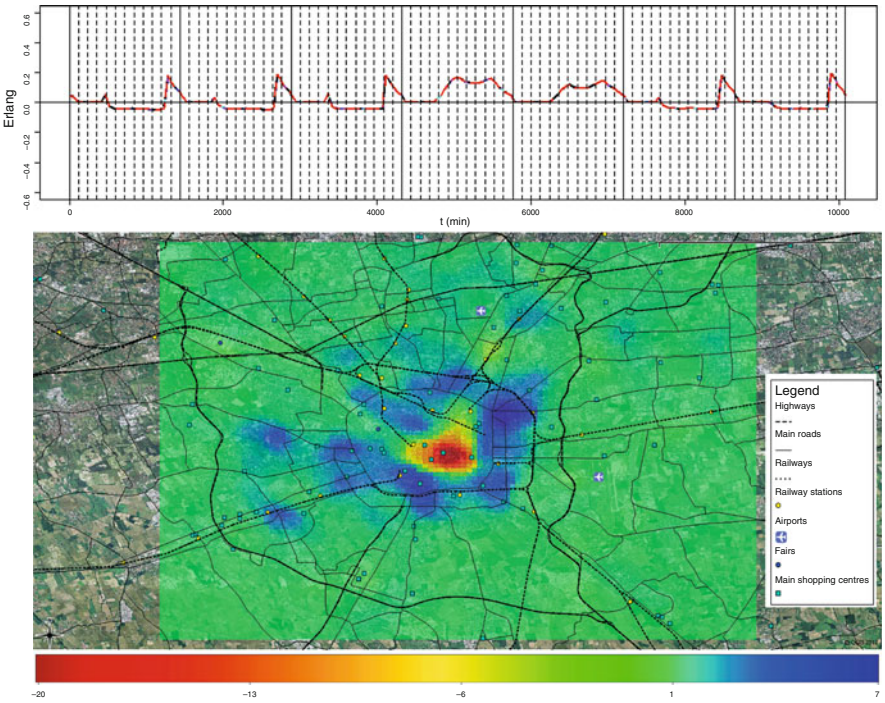

Fig. 4 Treelet 33—Concentration of activities during the evenings of working days and during daytime (from 8 a.m. until 8 p.m.) of the week end: residential districts of the Milan urban region

wide territory and there is a plurality of places with specific meanings for mobile populations. At the moment, the urban region of Milan is a densely populated, integrated area where 4,000,000 inhabitants live, where there are 370,000 firms, covered by huge flows of people moving daily in this wide area. Mapping overall mobility in space and in time therefore requires new data sources, able to adequately describe mobility patterns.

4.2 Testing the Treelets

Among the dozens of treelets produced applying the methodology explained in the previous section, we selected some results as significant for explaining specific mobility and city usages patterns and tested their significance and their interpretation from an urban analysis and planning perspective at the Milan scale.

On each image we added infrastructures (railways and main roads), main shopping centres, railway stations, localization of the city airport and of the fair trade centre in order to facilitate the interpretation of the map.

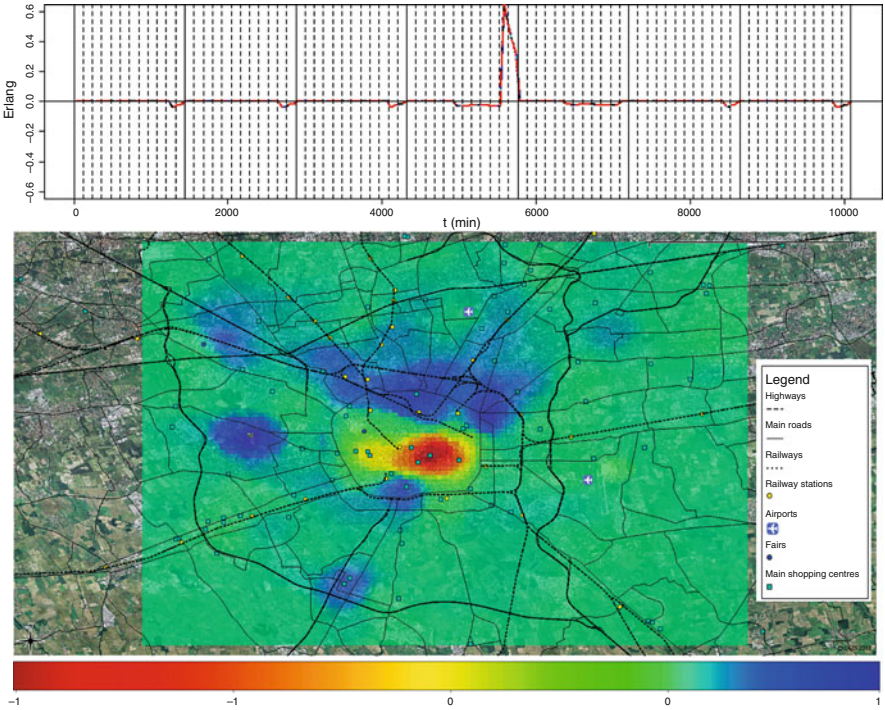

Fig. 5 Treelet 78—Density of activity during Saturday evening (8 p.m.-midnight). Saturday night population: leisure and hospitals

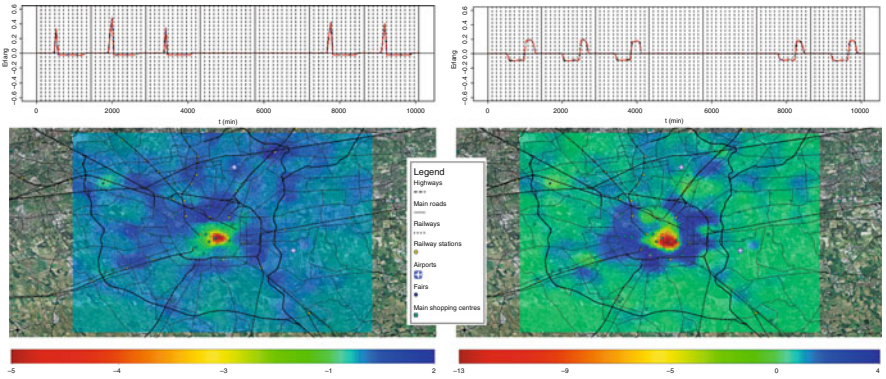

Fig. 6 Treelet 82 (*left*) and 83 (*right*)—Mobility practices. Weekdays commuting flows at the Milan urban region scale: morning rush hours (*left*) vs to evening rush hours (*right*)

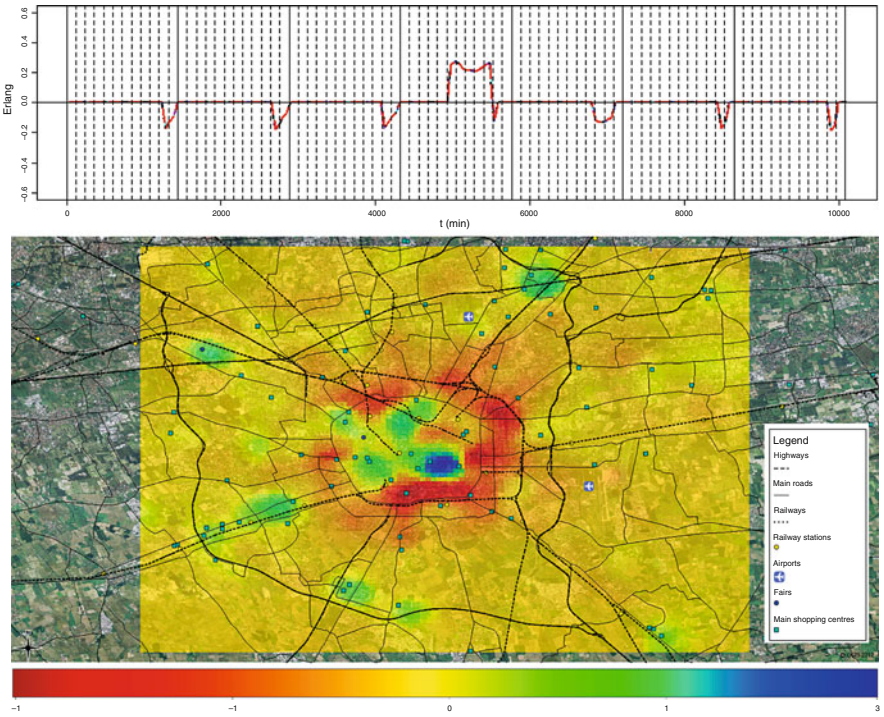

Fig. 7 Treelet 93—Mobility practices. Saturday (10 a.m.–8 p.m.), shopping and leisure activity

The "average use" treelet map (Fig. 2) highlights some urban districts characterized by specific telephonic patterns that are compatible with the real urban structure of the region. The treelet contains different temporal patterns of mobile phone activity (i.e. daily, working day versus week end) that fit with actual city usage. In particular we can observe the highest values in the Milan city centre and in others neighbourhoods where there is a strong attraction of urban populations during working day and, with minor intensity, during week end.

In other suburban districts, the intensity is lower, due to the presence of a less relevant mobile phone activity. In general we can conclude that the emerging spatial patterns represent well the highly populated areas versus the poorly populated areas. The mobile phone activity and the urbanized area produce in fact a similar image of the region.

The proposed methodology shows its advantages when we try to face with other, less evident spatial patterns which are difficult to intercept through traditional data sources.

Figure 3 (Treelet 2) is about the density of mobile phone activity late at night (in particular from midnight until 8 a.m.). We can observe here some interesting hot spots where the values are very high. For example, the exhibition district in the Northern Western side of the map. In the considered period an important Fair was

held and the peak fits well with the nightly activities necessary for the mounting and the organization of the site. Another point of interest is the Fruit and Vegetable Wholesale market in the South Eastern part of the region where consistent night work happens for delivering and distributing products that come from whole Italy and abroad. The city centre is characterized by a relative low value, according to the absence of relevant nightly activity inside it.

Figure 4 (Treelet 33) puts in evidence some locations with high concentration of mobile phone activity during the evening of the working days and during daytime (from 8 am until 8 pm) of the week end. It shows a significant correspondence with main residential districts of the Milan urban region. It highlights a relevant concentration of homes along the second circular ring of the city, where the density of resident population reaches the highest value of Milan, but also in some municipalities with a residential profile and social housing in the south, south-west and in the north of the metropolitan area (Corsico, Rozzano, Sesto S.G). The Milan city centre appears as a void and this is consistent with the changes that occurred in the last decades, namely a gradual replacement of the residents with activities mainly related to the service and the commercial sectors.

Figure 5 (Treelet 78) shows places with high density of activity during Saturday evening, from 8 p.m. until midnight. Focusing on the core city area, we notice several interesting patterns: a high activity in some places where there are many pubs and restaurants near the Milan Central Station, in the Navigli District, in the Isola Quarter and in other ambits characterized by the presence of leisure spaces (Filaforum Assago in the south of Milan) but also of activities in a continuous cycle as the hospitals. This treelet has proven to be effective in describing the temporal profile of the city lived by night populations during Saturday.

Figure 6 (Treelet 82 and Treelet 83) concerns more directly the representation of specific mobility patterns evident at the Milan urban region scale during working days, i.e. commuting flows. In fact, the Treelet 82 map is about the concentration of mobile phone activity during the morning rush hours and the Treelet 83 regards the activity during the evening rush hours. The emerging spatial patterns are quite different and show several interesting mobility practices within and outside the city. In fact, during the morning, we observe a concentration of traffic along the main roads, in proximity of relevant high roads junctions and in the surrounding areas. We can put this trend in relation with the daily commuting flows of people moving from homes, located in a wide area around Milan to job places. From an urban analysis point of view, it can be seen as a representation of the overall traffic generated mainly by cars, from 8 a.m. to 10 a.m. from Monday to Friday. Treelet 83 measures the mobile phone activity during the evening rush hours, which are longer than the morning ones, since they last more than 4 h, from 4 p.m. to 8 p.m. In this case, the hot spots are mainly located within the road ring. The map well represents the complex mobility pattern related to the exit from workplaces, when, before going home, chains of daily shifts take place, linked to a number of social practices (shopping, going to the gym, go get a family member or friend). The chain of daily moves becomes more articulated, and the daily rush hours are dilated. As it emerges from traditional sources [10]: the individual daily displacements in the Province of

Milan are 2,55 moves/person, with an average of two moves in sequence. The spatial pattern puts in evidence a relative low intensity in the city centre and an increasing density of activity in proximity of the main shopping centres, commercial streets (Vigevanese), and radial connections moving outward.

Figure 7 (Treelet 93) highlights another relevant mobility pattern, which is difficult to intercept through database traditionally used in urban studies: the shopping activity and in general the leisure activity. The map represents the density of mobile phone use during Saturday, from 10 a.m. to 8 p.m. Shopping and leisure are two of the main reasons of mobility in contemporary cities: they belong to the category of non systematic mobility, and they significantly contribute to the even more complex mobility patterns that can be observed in the Milan urban region due to the distribution of commercial centres, commercial streets and, in general, of activities (museums, touristic sites, cinemas, just to cite some) inside and outside the city. These places attract, especially in certain days of the week, a huge amount of population coming from a vast territory that goes far beyond the administrative boundaries of the city. The map is the result of this spatial pattern and shows an important concentration of mobile phone traffic in the city centre and in other several places outside the city (most of them corresponding to the presence of commercial centres). The mainly residential areas, recognized in the previous Fig. 6, are consequently characterized by the lowest value.

5 Future Works

The research allowed us to test the potential of the treelet decomposition analysis in explaining relevant urban usage and mobility patterns at the Milan urban scale. We plan to improve the integration of traditional database (i.e. land cover maps, distribution of activities, infrastructures and transport junctions) with mobile phone data pattern in order to reach a less descriptive and a more synthetic classification of the urban space according to its temporal and spatial usages, that could be useful for understanding the dynamic of temporary populations and of mobility patterns and for promoting more specific urban policies. This task may be possible from the recognition of mobile populations which are given the opportunity to choose among alternative forms of available mobility which can offer the greatest flexibility, range of connections, reversibility and the best means of accessing the various resources and destinations possible, but also which offer more oriented services.

Acknowledgements The authors would like to acknowledge Piero Lovisolo, Dario Parata and Massimo Colonna, Tilab—Telecom Italia for their collaboration during the research project. This work was supported by Telecom Italia. We also thank Paolo Dilda for helping us in the preparation of the figures.

References

1. Ahas, R., Mark, Ü.: Location based services–new challenges for planning and public administration? Futures **37**(6), 547–561 (2005)
2. Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Comput. **10**(4), 18–26 (2011)
3. Cresswell, T.: On the Move: Mobility in the Modern Western World. Routledge, London (2006)
4. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008). http://dx.doi.org/10.1038/nature06958. M3: 10.1038/nature06958; 10.1038/nature06958
5. Kaufmann, V.: Re-thinking Mobility Contemporary Sociology. Ashgate, Aldershot (2002)
6. Lee, A.B., Nadler, B., Wasserman, L.: Treelets – an adaptive multi-scale basis for sparse unordered data. Ann. Appl. Stat. **2**(2), 435–471 (2008)
7. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)
8. Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D.: Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plan. B Plan. Des. **33**(5), 727–748 (2006)
9. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Comput. **6**(3), 30–38 (2007). http://dl.acm.org/citation.cfm?id=1435641.1436493
10. Regione Lombardia, Direzione Generale Infrastrutture e mobilità: Indagine origine/destinazione regionale 2002 - sintesi. Tech. rep., Regione Lombardia (2002)
11. Secchi, P., Vantini, S., Vitelli, V.: Bagging voronoi classifiers for clustering spatial functional data. Int. J. Appl. Earth Obs. Geoinf. (2012). doi:10.1016/j.jag.2012.03.006
12. Sheller, M., Urry, J.: The new mobilities paradigm. Environ. Plan. A **38**(2), 207–226 (2006). http://www.envplan.com/abstract.cgi?id=a37268
13. Soto, V., Frías-Martínez, E.: Automated land use identification using cell-phone records. In: Proceedings of the 3rd ACM International Workshop on MobiArch, pp. 17–22. ACM (2011)
14. Soto, V., Frías-Martínez, E.: Robust land use characterization of urban landscapes using cell phone data. In: The First Workshop on Pervasive Urban Applications (PURBA) (2011)
15. Urry, J.: Sociology Beyond Societies: Mobilities for the Twenty-First Century. Routledge, London (2002)

Methodological Issues in the Use of Administrative Databases to Study Heart Failure

Cristina Mazzali, Mauro Maistriello, Francesca Ieva, and Pietro Barbieri

1 Introduction

The use of administrative data for clinical and epidemiological research is well-established; advantages in their use and criticisms are also well discussed (see, among others, [5, 7] and [8] and references therein). Among advantages we could remember large sample size, represented population, which may vary between different healthcare systems, absence of additional costs for gathering data, long observation periods, and sometimes possibility of linking different databases' information on the patient [11]. However, these databases were originally designed for administrative aims rather than for clinical research, therefore some drawbacks exist. Quality of collected data is better when it has financial or administrative implications, sometimes unique patient identification improves over years, some comorbidities are poorly recorded and their possible presence on admission is often unclear. At last misclassification of outcome or exposure is also possible.

C. Mazzali (✉)
Department of Biomedical and Clinical Sciences "L. Sacco", Università degli Studi di Milano, via G. B. Grassi 74, Milano, Italy

DIG (Department of Management, Economics and Industrial Engineering), Politecnico di Milano, via Lambruschini 4/b, Milano, Italy
e-mail: cristina.mazzali@unimi.it

M. Maistriello • P. Barbieri
A.O. Melegnano - Ospedale Uboldo, Via Uboldo 21, Cernusco sul Naviglio, Italy
e-mail: mauro.maistrello@aomelegnano.it; pietro.barbieri@fastwebnet.it

F. Ieva
Department of Mathematics "Federigo Enriques", Università degli Studi di Milano, via Saldini 50, Milano, Italy
e-mail: francesca.ieva@unimi.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_10

Moreover, statistical analysis of administrative data has particular characteristics; for example, it is important to distinguish between statistical significance, easily achieved, and clinical significance especially in model selection [17]. Dealing with administrative data instead of usual observational studied, a sort of change of perspective occurs. Actually, data are already available, while patients of interest are to be found. In order to avoid patient misclassification, or incorrect interpretation and use of clinical information, a preliminary work on data and on study design is needed. Several competences should be involved in an interdisciplinary team: clinicians, statisticians, experts of coding classification systems, experts of the databases used in the study, etc.

As a first step, data should be checked for accuracy, internal and external validity. Quality control allows to define reliable variables for the analysis, and also valid records. For example, cases missing patient identification may be excluded from the analysis because the linkage with vital statistics is not possible. Another important question concerns the choice of extraction criteria to select patients of interest. When dealing with hospital discharge records, selection is often made searching specific codes in diagnosis positions. For the most relevant diseases there are reviews of studies evaluating sensibility, positive predictive values and specificity of different selection criteria. These reviews could be useful in the choice of the best criteria for the purposes of the study.

Several choices are to be made before statistical analysis is performed; these choices could be source of misclassification and biases. It is important that the researchers using administrative databases make every decision explicit and clearly declared in the final report or article to let the reader know their possible effects.

The project "Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure" is funded by Ministero della Salute and is developed in Lombardia region (Italy). The purpose of the project is to study epidemiology, outcome and process of care of patients hospitalized for heart failure in Lombardia region on the basis of administrative data. At first, patients will be selected from the database of hospital discharge abstracts to study epidemiology of heart failure and its outcomes; subsequently, databases of outpatient care and drugs prescriptions will be the sources of information to study the process of care. The main aim of this article is to describe the necessary steps to make project data suitable for epidemiological and statistical analysis. We are also going to discuss methodological issues concerning data quality analysis, selection criteria, comorbidities detection and definition of observation units.

2 Dataset Description

The first goal of our work was the identification of patients hospitalized for heart failure among residents in Lombardia, which is an administrative region in the northern part of Italy. In order to do this, we required the regional database of

hospital discharge abstracts. In fact, to obtain reimbursement from National Health Service, almost all Italian hospitals, both private and public, are bounded to submit discharge summaries. Regional administrative databases of hospital discharges are characterized by universality, i.e. all the resident population is involved, and completeness, i.e., all the hospitalizations are reported.

We considered discharges from 2000 to 2012 of residents in Lombardia region. We also obtained hospitalizations of residents which occurred from 2000 to 2011 in Italian regions other than Lombardia. In particular, we considered discharges in Major Diagnostic Category MDC 01, 04, 05 and 11 from acute care facilities and rehabilitation services. Discharge abstract data contain information on sex, age and residence of patients. Among information of medical interest there are: date of admission and date of discharge, admission ward and internal transfers to other ward of the same hospital, principal diagnosis, up to five secondary diagnoses, up to six procedures. Diagnoses and procedures are reported using ICD-9-CM codes. Finally, source of admission and discharge status are available; the last one is used to identify in-hospital deaths.

A unique personal identifier, although encrypted, was supplied. It was used to detect subsequent admissions for the same patient. Possible date of death for each patient was linked from death registry by the institution that hosts the databases. Thanks to these data, it was possible to evaluate survival time for each patient. If a patient moved to another Italian region during the period between 2000 and 2012, the date of move was also provided.

3 Data Quality Control

Data collected with administrative purposes are usually evaluated for quality by the source agencies; however, their quality is not always suitable for research use. We performed an evaluation of data quality which was oriented to the use of data for the project aims. According to the framework proposed by the Manitoba Centre for Health Policy [10] we checked data for accuracy, internal and external validity.

Controls of accuracy included evaluation of variables completeness and correctness. We checked all the variables involved in the analysis for missing values, and for invalid values or out of range data.

Among the internal validity controls, we compared values of different variables in order to evaluate the coherence of information they provided; for example, date of death had to be less or equal to date of discharge. We also considered stability across time of a few number of variables, such as distribution of MDCs and discharge status. Thanks to this control, we found that, among patient hospitalized in other Italian regions, abnormal values in mortality rates were due to a change in coding discharge status occurred in this period. Imputation of values for incorrect data was bounded to few cases; for example, the use of different coding sets for the same variable through years was verified and corrected. Reliability of variables was

assessed, and sometimes they were excluded from further analyses. For example, date of internal transfers were excluded for lack of internal validity.

Close attention was paid to the analysis of missing data in patient's identifier, and to the characteristics of discharges in which it was missing.

4 Methodological Issues

4.1 Bibliographic Research

Administrative health care databases are a rich source of clinical and administrative information on broad populations and an increasingly volume of studies has been published using this source of data. A useful classification of the wide-ranging administrative database applications into a meaningful typology has been done by Schoenman et al. [15]. They identified eight type of applications:

- public safety and injury surveillance and prevention;
- public health, disease surveillance and disease registries;
- public health planning and community assessments;
- public reporting for informed purchasing and comparative reports;
- quality assessment and performance improvement;
- health services and health policy research applications;
- private sector and commercial applications;
- informing policy deliberations and legislation.

According to this report, there are various advantages on using administrative databases as a source: they are relatively inexpensive to obtain when compared to the cost of similar data collected through surveys or medical record abstraction; they are more reliable than other sources of data, such as patient self-reporting or physician reporting of specific conditions for disease surveillance; they are usually available for multiple years and they typically cover entire populations.

Researchers working with administrative databases should also be mindful of inherent limitations such as: lack of detailed clinical information, test results, functional status, severity of illness and behavioral risk factors. To conduct our study, using PubMed, we searched MEDLINE for articles published in the last 10 years combining text words and MeSH term in a quick search strategy: `(administrative data*[All Fields] OR claims data*[All Fields]) AND Heart Failure[Mesh]`. Subsequently, we refined our strategy taking advice from a recent project supported by the U.S. Food and Drug Administration (FDA). The FDA Mini-Sentinel initiative reviewed the literature to find validated methods for identifying health outcomes using administrative and claims data. A detailed review on congestive heart failure has been prepared by Saczynski et al. [13].

The specific search strategy for the HF review can be found in the full report available at http://www.mini-sentinel.org/workproducts/HealthOutcomes/MSHOICHFReport.pdf We identified additional studies from bibliographies of relevant articles and we also searched for reports in the web sites of major research institutions worldwide such as the Centers for Medicare & Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ).

4.2 Selection Criteria

The choice of criteria to select patient affected by a specific disease is one of the most relevant issues when dealing with administrative data. In order to avoid biases in patients selection, it is important to note that sensibility and specificity of codes may vary for different diseases and actually affect the possibility to study a disease using administrative data. Which codes are to be addressed to identify patients of interest should be decided involving several competences, such as clinicians or coding experts. Clinicians have the knowledge of the diseases and how they are treated; coding experts have a detailed knowledge of coding system and of coding habits, they are also aware of regional laws that could affect coding. Moreover, when dealing with data from several hospitals and years, the knowledge of variations over time in coding practice is also important.

Several studies (see [9, 13, 16] and [18], among others) address the issue of the best algorithms to identify hospitalizations for heart failure based on discharge abstracts. According to Saczynski et al. [13], in order to yield cases of heart failure, the use of codes 428.*x*, leads to high specificity and positive predictive value (PPV) but may have a low sensitivity. Some authors (for example, see [6]) look for the presence of other codes besides 428.*x*, resulting in lower specificity and PPVs but with an increase of sensibility values.

The Agency for Healthcare Research and Quality developed a set of Inpatient Quality Indicators that give an insight of quality of care using hospital administrative data. Mortality for heart failure is one of the indicators of mortality for conditions developed within the Inpatient Quality Indicators set. Discharges of patients affected by heart failure are identified searching for a list of ICD-9-CM codes (see Table 1) in principal diagnosis.

The Hierarchical Condition Categories by Centers for Medicare & Medicaid Services (CMS-HCC) is a risk adjustment model used to adjust capitation payments in agreement with health condition of patients. To construct the CMS-HCC the ICD-9-CM codes are classified into 805 groups of diagnostic codes, named DGXs. Subsequently, DGXs are aggregated into 189 Condition Categories (CC), which describe broader sets of diseases related both clinically and with respect to cost. Hierarchies are imposed among related CCs to create HCC. Thanks to hierarchy, a person with related conditions is coded for only the most severe one. Hierarchical Condition Category 80 (CMS-HCC version 12) is related to congestive heart failure and is defined by the codes reported in Table 1.

Table 1 Sets of ICD-9-CM codes used for the selection of patients with heart failure in AHRQ quality indicator and in CMS-HCC model

		HCC	AHRQ
398.91	RHEUMATIC HEART FAILURE		X
402.01	Mal hyperthrt dis w hf	X	X
402.11	Benign hypht dis w hf	X	X
402.91	Hypht dis NOS w ht fail	X	X
404.01	Mal hypht/kd I-IV w hf	X	X
404.03	MAL HYP HRT/REN W CHF&RF		X
404.11	Ben hypht/kd I-IV w hf	X	X
404.13	BEN HYP HRT/REN W CHF&RF		X
404.91	Hypht/kd NOS I-IV w hf	X	X
404.93	HYP HT/REN NOS W CHF&RF		X
415.0	Acute corpulmonale	X	
416.0	Primpulmhypertension	X	
416.1	Kyphoscolioticheartdis	X	
416.8	Chrpulmon heart dis NEC	X	
416.9	Chrpulmon heart dis NOS	X	
417.0	Arteriovenfistupulves	X	
417.1	Pulmonarteryaneurysm	X	
417.8	Pulmoncirculatdis NEC	X	
417.9	Pulmoncirculatdis NOS	X	
425.0	Endomyocardialfibrosis	X	
425.1	Hypertrobstrcardiomyop	X	
425.2	Obscafriccardiomyopath	X	
425.3	Endocardfibroelastosis	X	
425.4	Primcardiomyopathy NEC	X	
425.5	Alcoholiccardiomyopathy	X	
425.7	Metaboliccardiomyopathy	X	
425.8	Cardiomyopath in othdis	X	
425.9	Second cardiomyopath NOS	X	
428.0	CHF NOS	X	X
428.1	Left heartfailure	X	X
428.20	Systolichrtfailure NOS	X	X
428.21	Acsystolichrtfailure	X	X
428.22	Chrsystolichrtfailure	X	X
428.23	Ac on chrsysthrt fail	X	X
428.30	Diastolchrtfailure NOS	X	X
428.31	Acdiastolichrtfailure	X	X

Table 1 (continued)

428.32	Chrdiastolichrtfail	X	X
428.33	Ac on chrdiasthrt fail	X	X
428.40	Syst/diasthrt fail NOS	X	X
428.41	Ac syst/diastolhrt fail	X	X
428.42	Chrsyst/diastlhrt fail	X	X
428.43	Ac/chrsyst/diahrt fail	X	X
428.9	Heartfailure NOS	X	X
429.0	Myocarditis NOS	X	
429.1	Myocardialdegeneration	X	

Symbol "X" means that the code is used in the definition of the CMS-HCC model or AHRQ quality indicator

The two sets of codes by AHRQ and CMS-HCC partially overlap; common codes, i.e., 402.*xx*, 404.*x*1, and 428.*xx*, identify paradigmatic cases of heart failure. AHRQ codes, not included in CMS-HCC set, were 398.91 (Rheumatic heart failure - congestive) and 404.*x*3 (Hypertensive heart and chronic kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease). Codes 415.*x*, 416.*x*, 417.*x*, 425.*x*, and 429.*x* where included in CMS-HCC set but not in AHRQ one; they are used in coding: acute pulmonary heart disease, chronic pulmonary heart disease, Other diseases of pulmonary circulation, Cardiomyopathy, myocarditis (unspecified) and myocardial degeneration.

In the first step of data extraction the whole set of diagnostic codes was used in order not to miss any case of heart failure in the subsequent analysis. In the second step incidence and prevalence estimates were calculated on the whole set of heart failure patients population and in some subsets based on specific subgroups of codes; etiology was defined extracting information by secondary diagnoses and/or previous admissions. Selected codes were searched in each diagnosis position, that is the principal one and up to five secondary diagnoses.

4.3 Definition of the Observation Unit

The regional discharge database contains all hospitalizations occurred in Lombardia region or occurred to residents outside this region. Therefore, when a patient is transferred from a hospital to another, or is discharged by a hospital and admitted in another one, two discharge summaries are recorded in the regional database.

In order to correctly count heart failure events and appropriately estimate time elapsed between two heart failure events, we needed to identify hospitalizations occurred to a given patient for the same medical event. In absence of general clinical criteria that could be applied to discharge summaries, we decided, at least, on considering two subsequent hospitalizations of the same patient as a single event.

By subsequent hospitalizations we meant that the latter began the same day or the day after the former ended.

As a first step, we organized all the hospitalizations into groups of events, and then we considered those events with at least one hospitalization selected for heart failure. Hereafter, we call these cases heart failure events.

4.4 Comorbidity Detection

Indices of patient's comorbidities are needed for risk prediction and risk adjustment modelling. Several measures had been proposed in literature [14]. The best known are Charlson comorbidity index [1] and Elixhauser comorbidity measure [3].

Charlson index was developed for predicting 1 year mortality in elder patients admitted in acute care, and is based on clinical data. However, translations using ICD codes had been proposed by different authors (see [2] and [12], among others). On the other hand, Elixhauser comorbidity classification system has better predictive performance for length or stay, hospital charges and in-hospital mortality. In [4] a comorbidity score predicting mortality in elder patients which was developed combining conditions included in Romano–Charlson and Elixhauser comorbidities measures is proposed. According to the authors, this score performs better than its component scores in predicting short- and long-term mortality. We decided to adopt the combined score for risk prediction and comorbidity adjustment of statistical models.

In [14] the authors state that another factor that affects the predicting capability of comorbidity measures is the, so called, "look-back period", i.e., the period before hospitalization in which comorbidities could be searched. According to the authors, an adequate look-back period for improving prediction of mortality should be of 1 year, while a longer period would be better for readmission. In order to study time between subsequent events and its association with patient's worsening condition, the whole burden of comorbidities was calculated extracting information by all the previous available admissions.

5 Application of the Methodological Issues

The total number of discharges in MDC 01, 04, 05 and 11 from 2000 to 2012 occurred in Lombardia region or in other regions but related to patients resident in Lombardia was 6,636,611.

Out of 6,636,611 hospital discharges, in 889,060 (13.40 %) patient ID was missing. Among these, 579,423 (65.17 %) were related to patients not resident in Lombardia. The overall number of discharges related to non-resident patients was 643,932 (9.70 %). Excluding data with missing patient ID and data related to non-resident patients, the total number of discharges decreased to 5,683,042 (85.63 %

Table 2 Hospital discharge abstracts selected with AHRQ and CMS-HCC codes

Code set and position	Frequency	Percent
HCC80 ∩ AHRQ - principal diagnosis	318,373	42.86
HCC80 ∩ AHRQ - secondary diagnoses	213,321	28.72
HCC80 (not AHRQ)- secondary diagnoses	768,08	10.34
HCC80 ∩ AHRQ - principal diagnosis AND HCC80 (not AHRQ) - secondary diagnoses	58,426	7.87
HCC80 (not AHRQ) - principal diagnosis	37,725	5.08
HCC80 ∩ AHRQ - principal AND secondary diagnoses	23,338	3.14
HCC80 (not AHRQ) - principal diagnosis AND HCC80 ∩ AHRQ - secondary diagnoses	11,455	1.54
Other	3,319	0.45
Total	742,765	100.00

Table 3 Number of events per patient

Number of events per patient	Freq	%	Cumulative %
1	229,341	61.69	61.69
2	71,011	19.10	80.79
3	31,071	8.36	89.15
4	15,801	4.25	93.40
5	8,836	2.38	95.78
6	5,437	1.46	97.24
7	3,246	0.87	98.11
8	2,202	0.59	98.70
9	1,444	0.39	99.09
10	960	0.26	99.35
11+	2,417	0.65	100.00

of 6,636,611). Discharges with an ICD-9-CM code from AHRQ or CMS-HCC sets in any diagnosis position were 742,765. As shown in Table 2, more than 71 % of discharges were selected by codes common to AHRQ and CMS-HCC which were found in principal diagnosis (42.86 %) or in secondary diagnosis (28.72 %). About 15 % of discharges were selected as heart failure cases on the basis of a CMS-HCC code, not present in AHRQ set, in principal diagnosis (5.08 %) or secondary diagnoses (10.34 %).

Excluding records with missing personal identifier, the total number of hospital discharge records was grouped in 5,255,479 events. Events with at least one hospitalization for heart failure, i.e. heart failure events, occurred to residents were 701,701, concerning 371,766 patients. About 88.4 % of the events was made by a single hospitalization, and about 99 % of the events was composed by 3 or less events. The distribution of the number of events per patient is shown in Table 3; about 62 % of patients underwent a single event in the observation period; approximately 27 % of patients underwent 2 or 3 events.

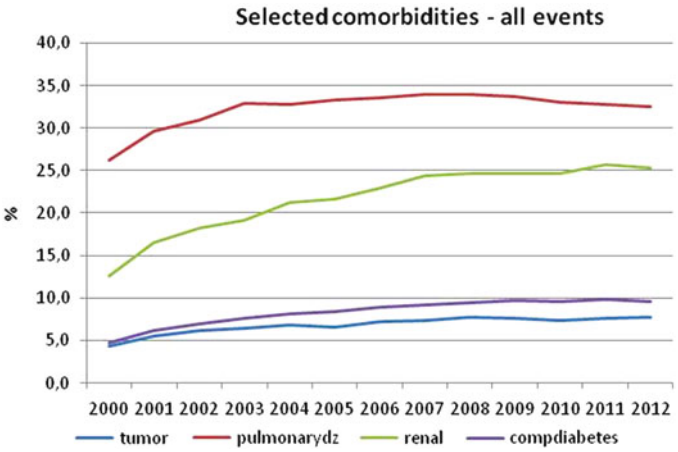

Fig. 1 Percentages of presence of selected comorbidities in patients' events over years

In Fig. 1 the distribution over years of selected comorbidities is showed. These comorbidities are commonly associated with heart failure and aged patients. The frequency of comorbidities is expected to be quite constant over time; an apparent increase may be due to the effect of the so-called "look-back period". Tumour and pulmonary comorbidities seem to reach their plateau within the first three or 4 years; while renal comorbidities and diabetes are increasing over a longer period and become quite constant in the last 4 or 5 years.

Conclusions

Dealing with administrative databases implies a strong interdisciplinary work. Shared decisions and definition of a common methodology are key points of this work. In our project, thanks to the process of quality control, the choice of selection criteria and observation units, and the method selected for comorbidity detection, administrative data were made suitable for clinical and epidemiological research. Every choice was discussed within the work groups, and is made clear for readers.

We choose broad selection criteria in order to gather information on a large number of patients affected by heart failure. However, the epidemiological and clinical analyses will be better conducted on subgroups of patients with possible different outcomes. At first, these subgroups of patients can be defined on the basis of sets of ICD codes; for example codes in the AHRQ set only, or in the CMS-HCC set only, or in the intersection of the two sets. Subsequently, more refined subgroups can be defined using etiological criteria on the basis of ICD codes. Further analysis and discussion within

(continued)

the group will be useful to determine the best possible look-back period for comorbidities detection, in order to make it more homogeneous for all the patients regardless when their incident event occurred.

At last, a separate analysis will be necessary for cases excluded from the study due to missing patient identification code. In particular, it's important to know their demographic characteristics, e.g. their sex, age or nationality, and their outcomes. Thanks to this work, we are confident that discharge admission abstracts of the entire regional population over a period of 13 years were made suitable for clinical research, through a shared and clearly declared process.

Acknowledgements This work was developed within the "Utilization of regional health service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure" project founded by Ministero della Salute and supervised by Dott. M. Frigerio

References

1. Charlson, M.E., Pompei, P., Ales, K.L., et al.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic. Dis. **40**, 373–383 (1987)
2. Deyo, R.A., Cherkin, D.C., Ciol, M.A.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J. Clin. Epidemiol. **45**, 613–619 (1992)
3. Elixhauser, A., Steiner, C., Harris, D.R., et al.: Comorbidity measures for use with administrative data. Med. Care. **36**, 8–12 (1998)
4. Gagne, J.J., Glynn, R.J., Avorn, J., Levin, R., et al.: A combined comorbidity score predicted mortality in elderly patients better than existing scores. J. Clin. Epidemiol. **64**, 749–759 (2011)
5. Gavrielov-Yusim, N., Friger, M.: Use of administrative medical databases in population-based research. J. Epidemiol. Community Health **68**(3), 283–287 (2014)
6. Goff, D.C., Pandey, D.K., Chan, F.A., et al.: Congestive heart failure in the United States. Is there more than meets the I(CD codes)? The corpus christi heart project. Arch. Intern. Med. **160**, 197–202 (2000)
7. Grimes, D.A.: Epidemiologic research using administrative databases—garbage in, garbage out. Obstet. Gynecol. **116** (5), 1018–1019 (2010)
8. Hoover, K.W., Tao, G., Kent, C.K., Aral, S.O.: Epidemiologic research using administrative databases: garbage in, garbage out. Letter to the editor. Obstet. Gynecol. **117**(3), 729–730 (2011)
9. Lee, D.S., Donovan, L., Austin, P.C., et al.: Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med. Care **43**, 182–188 (2005)
10. Manitoba Centre for Health policy: "MCHP Data Quality Framework", internal document version—July 2013. http://umanitoba.ca/faculties/medicine/units/communityhealthsciences/departmentalunits/mchp/protocol/media/DataQualityFramework.pdf (2013). Accessed 11 Dec 2013
11. Nguyen. L.L., Barshes N.R.: Analysis of large databases in vascular surgery. J. Vasc. Surg. **52**(3), 768–774 (2010)

12. Romano, P.S., Roos, L.L., Jollis, J.G.: Adapting a clinical comorbidityindex for use with ICD-9-CM administrative data: differing perspectives. J. Clin. Epidemiol. **46**, 1075–1079 (1993)
13. Saczynski, J.S., Andrade, S.E., Harrold, L.R., et al.: A systematic review of validated methods for identifying heart failure using administrative data. Pharmacoepidemiol. Drug. Saf. **21**(S1), 129–140 (2012)
14. Sharabiani, M.T.A., Aylin, P., Bottle, A.: Systematic review of comorbidity indices for administrative data. Med. Care **50**, 1109–1118 (2012)
15. Schoenman, J.A., Sutton, J.P., Kintala, S., Love, D., Maw, R.: The Value of Hospital Discharge Databases. Agency for Healthcare Research and Quality, Rockville (2005)
16. Schultz, S.E., Rothwell, D.M., Chen, Z., et al.: Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. Chronic Dis. Inj. Can. **33**(3), 160–166
17. Van Walraven, C., Austin, P.: Administrative database research has unique characteristics that can risk biased results. J. Clin. Epidemiol. **65**, 126–131 (2012)
18. Zarrinkoub, R., Wettermark, B., Wandell, P., et al.: The epidemiology of heart failure, based on data for 2.1 million inhabitants in Sweden. Eur. J. Heart Fail. **15**, 995–1002 (2013)

Bayesian Inference for Randomized Experiments with Noncompliance and Nonignorable Missing Data

Andrea Mercatanti

1 Introduction

The theoretical model of a randomized experiment with noncompliance is widely adopted in biomedical and social sciences for inferring causal effects. The original application stems from experimental settings, where the treatment received by an individual (or a generic statistical unit) sometimes differs from the treatment randomly assigned to him by the experimenter. The concept was later generalized to observational settings where the treatment is not randomized. The nonrandom nature of the treatment usually implies different distributions of the potential outcomes[1] between treatment groups, a situation known as self-selection, which makes the direct comparison of the outcome distributions between treatment groups a biased estimate for the treatment effect. In the presence of self-selection, one approach to obtain unbiased estimates for causal effects is to find a variable that can be viewed as a nonmanipulated, natural or quasi random assignment to the treatment to which the units not necessarily comply, and then to apply inferential methods for randomized experiment with noncompliance. In the following discussions, we use the familiar

The opinions expressed are those of author and do not commit the Bank of Italy.

[1] Under the potential outcome approach to causal inference [5, 11], potential outcomes are the set of the outcomes potentially observed for the same individual under the range of possible values for the assignment to treatment.

A. Mercatanti (✉)
Statistics Department, Bank of Italy, via Nazionale 91, 00184 Rome, Italy
e-mail: mercatan@libero.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_11

nomenclature of noncompliance irrespective of the experimental or observational nature of the study.

Missing data is a common problem in both experimental and observational settings. For example, Jo [8] considered a randomized experiment aimed to evaluate the effect of a U.S. public school intervention program in reducing early behavioral problems among school children. Apart from the random assignment to the intervention condition, each variable has missing for some of the sampled units, and the analysis there was conducted after listwise discarding of the units with missingness. Alternatively, when the model of an experiment with noncompliance is adopted in nonexperimental studies, information is generally collected from surveys where missing data due to item nonresponses to questionnaires is an even more diffuse problem. One type of information that has traditionally been difficult to obtain in survey is income data; for example: the Current Population Surveys, that includes approximately 50,000 U.S. households monthly, suffers from 20 to 25 % of nonresponse rate on many income items. Other type of information with high rate on nonresponse are those related to stigmatized activies like drinking behavior or use of drug [10]. More generally, nonresponses are common in surveys whenever the population consists of units such as individual people, households or businesses.

In general, standard methods for complete data cannot be immediately used to analyze the dataset with missing data. Moreover, possible biases can arise because the respondents are often systematically different from the nonrespondents, and these biases are difficult to eliminate since the reasons for nonresponse are usually not known. The existing literature on missing data in randomized experiment with noncompliance has mainly focused on the scenarios with missingness only in the outcome. However, missing data in the treatment, and/or in the assignment to treatment are common in practice, where usual ignorability conditions for the missing data mechanism are considered too restrictive. The paper proposes a range of models under weaker nonignorable conditions for missing data. Section 2 presents the basic setup. Section 3 proposes identified models for three special cases: missingness in each of the three basic variables with and without a binary pretreatment variable separately, plus missingness in the outcome and in the treatment received without pretreatment variables. Section 4 develops a Bayesian approach for inference, and the methods are illustrated by a simulated comparative analysis in Sect. 5. Section "Conclusions" concludes.

2 Basic Setup

Adopting the standard notation in the literature, let Z be the assignment to a binary treatment, D be the treatment received, and Y be the binary outcome. The population can be classified into four sub-groups, compliance types, based on their potential treatment status to both levels of assignment, denoted by $U = (D(0), D(1))$. Units for which $Z = 1$ implies $D = 1$ and $Z = 0$ implies $D = 0$ (*compliers*, $U = c$) are induced to take the treatment by the assignment. Units for

which $Z = 1$ implies $D = 0$ and $Z = 0$ implies $D = 0$ are called *never-takers,* $U = n$ because they never take the treatment, while units for which $Z = 1$ implies $D = 1$ and $Z = 0$ implies $D = 1$ are called *always-takers,* $U = a$ because they always take the treatment. Finally the units for which $Z = 1$ implies $D = 0$ and $Z = 0$ implies $D = 1$ do exactly the opposite of the assignment and are called *defiers* ($U = d$).

We maintain hereafter the standard assumptions to identify causal effects in randomized experiments with noncompliance [7].

A.1 *Stable unit treatment value assumption* (SUTVA). There are no different versions of any single treatment arm and no interference between units;

A.2 *Randomization of treatment assignment.* $P(Z|Y(0), Y(1), D(0), D(1)) = P(Z)$;

A.3 *Monotonicity.* $D_i(1) \geq D_i(0)$ for all i, ruling out the group of defiers;

A.4 *Exclusion restriction* (ER). $Y_i(1) = Y_i(0)$ for all noncompliers, implying that the assignment to treatment has no direct effect on the outcome for noncompliers.

When there is no missing, two models $P(Y, D, Z)$ and $P(Y, U, Z)$ are sufficient to describe the data.[2] To account for nonresponses we need to introduce the additional model for the missing data generation usually called missing data mechanism. Let $\mathbf{R} = (R_y, R_d, R_z)$ be the three-dimensional vector of missing data indicators, where R_v is the missing data indicator for $v = y, d, z$: $R_v = 1$ if v is observed, 0 otherwise; and Z^*, D^*, and Y^* be the observable quantities defined as $Z^* = Z$ if $R_z = 1$, $Z^* = *$ if $R_z = 0$, and analogously for D^* and Y^*. Also let $\mathbf{X}$ be the set (Y, D, Z), $\mathbf{X}_{obs}$ the observed part of $\mathbf{X}$ (namely the elements of $\mathbf{X}$ for which the missing data indicators result to be equal to one), and $\mathbf{X}_{mis}$ the unobserved part of $\mathbf{X}$ (the elements of $\mathbf{X}$ for which the missing data indicators result to be equal to zero); and $P(\mathbf{R} \mid \mathbf{X})$ be the missing data mechanism, that is the probabilities to observe the quantities (Y, D, Z).

The practice to apply models for complete data after listwise deletion of units with missingness is justified only under the Missing Completely at Random (MCAR) assumption for the missing data mechanism, $P(\mathbf{R} \mid \mathbf{X}) = P(\mathbf{R})$, that is the probability of missingness is the same for each unit. A weaker assumption is Missing at Random (MAR), $P(\mathbf{R} \mid \mathbf{X}) = P(\mathbf{R} \mid \mathbf{X}_{obs})$, that is, the probabilities of missingness depends only on the observed part of $\mathbf{X}$. In these cases the missing data mechanism is said to be ignorable after conditioning on the observable quantities, or

[2]The domain of the distribution $P(Y, U, Z)$ is only partially observed, even in case of no missingness, because the compliance status U is defined on counterfactual quantities that cannot be simultaneously observed for the same unit. Only under suitable conditions, like the exclusion restriction, U can be observed for a subset of the population.

more simply ignorable. The paper will focus on identification and estimation issues for some special cases under nonignorable missing data mechanism, namely when:

$$P(\mathbf{R} \mid \mathbf{X}) \neq P(\mathbf{R} \mid \mathbf{X}_{obs}) \neq P(\mathbf{R} \mid \mathbf{X}_{mis}) \neq P(\mathbf{R}).$$

3 Identification

Under assumptions A.1–A.4 and maintaining the parameter space for (Y, U, Z) separated from that of the missing data mechanism, the model for the observable data $(\mathbf{R}, Y^*, D^*, Z^*)$ can be written as:

$$\begin{aligned}
&P(\mathbf{R}, Y^*, D^*, Z^*) \\
&= \sum_{Y,D,Z} P(\mathbf{R}, Y, D, Z) \cdot I[P(Y, D, Z \mid Y^*, D^*, Z^*) > 0] \\
&= \sum_{Y,U,Z} P(\mathbf{R}, Y, U, Z; \pi, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \cdot I[P(Y, U, Z \mid Y^*, D^*, Z^*) > 0] \\
&= \sum_{Y,U,Z} P(\mathbf{R} \mid Y, U, Z; \boldsymbol{\alpha}) \cdot P(Y, U, Z; \pi, \boldsymbol{\omega}, \boldsymbol{\theta}) \cdot \\
&I[P(Y, U, Z \mid Y^*, D^*, Z^*) > 0]. \qquad (1)
\end{aligned}$$

The specification for the missing data mechanism as a function of (Y, U, Z) allows a direct quantification of the relationships between the probability of nonresponse and the compliance status. Moreover, it allows to relate the models hereafter proposed to others models that recently have appeared in the literature, such as those by Frangakis and Rubin [2], Mealli et al. [9], Chen et al. [1], Imai [6]. At the same time the interpretation of the missing data mechanism as a function of the treatment is not excluded because the latter is a function of both the compliance status and the assignment to treatment.

Dealing with a set of binary variables allows to relate the parameter identification for (1) to the analysis of the contingency table for the observable data (Y^*, D^*, Z^*), like proposed by Small and Cheng [12] for the special case of nonresponses only in the outcome. The identification rule to apply is the general one for contingency tables that prescribed to set the number of parameters for regular models at most equal to the number of the entries of the contingency table minus one. Model regularity implies that the probability of each entry has to be defined as a sequence of conditional probabilities; this condition is satisfied by all of the models proposed hereafter.

Three cases will be examined in the following: missingness in all of the three variables, Y, D, and Z, without and with a binary pretreatment variable separately, plus missingness in Y and D without pretreatment variables.

Table 1 Contingency table for missingness in Y, D and Z

	D^*, Z^*								
	$1,0$	$0,1$	$1,1$	$0,0$	$1,*$	$0,*$	$*,1$	$*,0$	$*,*$
$Y^* = 1$									
$Y^* = 0$									
$Y^* = *$									

3.1 Missingness in Y, D, and Z Without Pretreatment Variables

When the possibility of nonresponses exists for each of the three binary variables, Y, D, and Z, then the contingency table for the observables Y^*, D^*, and Z^*, will show $3 \times 9 = 27$ cells in total, Table 1, and consequently at most $27 - 1 = 26$ parameters will be allowed for model (1).

The specification for $P(Y, U, Z; \pi, \boldsymbol{\omega}, \boldsymbol{\theta})$ can be formalized following [7] as:

$$\begin{aligned} & P(Y, U, Z; \pi, \boldsymbol{\omega}, \boldsymbol{\theta}) \\ = & \pi^Z (1-\pi)^{1-Z} \omega_a^{I(U=a)} \omega_n^{I(U=n)} (1-\omega_a-\omega_n)^{I(U=c)} \\ & \times \theta_a^{Y\, I(U=a)} (1-\theta_a)^{(1-Y)\, I(U=a)} \theta_n^{Y\, I(U=n)} (1-\theta_n)^{(1-Y)\, I(U=n)} \\ & \times \theta_{c1}^{Y\, I(U=c)\, Z} (1-\theta_{c1})^{(1-Y)\, I(U=c)\, Z} \theta_{c0}^{Y\, I(U=c)\, (1-Z)} (1-\theta_{c0})^{(1-Y)\, I(U=c)\, (1-Z)}. \end{aligned} \tag{2}$$

Model (2), shows 7 parameters so that at most $26 - 7 = 19$ are allowed for the missing data mechanism $P(\mathbf{R}|\, Y, U, Z; \boldsymbol{\alpha})$ whose domain is composed by the 8 possible combinations $(R_Y, R_D, R_Z) : \{(1,1,1), (1,1,0), \ldots\}$. Any attempt to comply with the limit of 19 parameters via a multinomial logit model for the 8 categories of the missing data mechanism leads to very few parameters for each logit equation and consequently to unplausible models. To reduce the number of categories for $P(\mathbf{R}|\, Y, U, Z; \boldsymbol{\alpha})$, we introduce the restriction, whose plausibility in real applications should be evaluated on a case-by-case basis, that the three marginal missing data mechanisms $P(R_v|\, Y, U, Z)$, $v = Y, D, Z$, are mutually independent:

$$P(\mathbf{R}|\, Y, U, Z) = P(R_Y|\, Y, U, Z) \cdot P(R_D|\, Y, U, Z) \cdot P(R_Z|\, Y, U, Z).$$

The restriction allows to specify only three independent logit models $P(R_v|\, Y, U, Z)$, $v = Y, D, Z$. The following model specification for the missing

data mechanism $P(\mathbf{R}|\, Y,\, U,\, Z;\, \boldsymbol{\alpha})$ implies identifiability of $(\mathbf{R},\, Y^*,\, D^*,\, Z^*)$ in that it complies with the restriction on the maximum number of parameters:

$$\begin{aligned}\text{logit}[P(R_v = 1|\, Y,\, U,\, Z)] = \alpha_{0v} + \alpha_{1v}\, I(U = a) \\ +\alpha_{2v}\, I(U = n) + \alpha_{3v}\, I(U = a,\, Z = 1) \\ +\alpha_{4v}\, I(U = n,\, Z = 1) + \alpha_{5v}\, I(U = c,\, Z = 1),\end{aligned} \tag{3}$$

for $v = D, Z$, and

$$\begin{aligned}\text{logit}[P(R_Y = 1|\, Y,\, U,\, Z)] = \alpha_{0Y} + \alpha_{1Y}\, Y + \alpha_{2Y}\, I(U = a) + \alpha_{3Y}\, I(U = n) \\ +\alpha_{4Y}\, I(U = a,\, Z = 1) + \alpha_{5Y}\, I(U = n,\, Z = 1).\end{aligned} \tag{4}$$

Models (3) and (4) reflect nonignorable conditions for the missing data mechanism because the probability of missingness for v is affected by the value of v itself. In comparison to the model proposed by Frangakis and Rubin [2], Model (3) is based on weaker conditions because no kind of response exclusion restrictions[3] is imposed. Model (4) is based on weaker conditions than other proposals recently appeared in the literature for dealing with missingness only on the outcome. Compared to [2] and [9], model (4) does not imposes ignorability of Y conditionally on the compliance status, at the same time the response exclusion restriction is introduced only for one compliance status[4] (for the compliers group in the particular specification (4), even if the restriction can be analogously imposed only for one of the other two compliance statuses). Compared to Imai [6], model (4) is based on weaker conditions because the way that treatment affects missingness depends on the assignment to treatment. Analogously, given that the dependency between missingness and the assignment to treatment depends on the treatment received, Model (4) is weaker than the model proposed by Small and Cheng [12].

Specifications (3) and (4) do not allow interactions between the outcome and the compliance status. However, the interactions can be introduced if imposing response exclusion restrictions for all the compliance statuses:

$$\begin{aligned}\text{logit}[P(R_Y = 1|\, Y,\, U,\, Z)] = \alpha_{0Y} + \alpha_{1Y}\, Y + \alpha_{2Y}\, I(U = a) + \alpha_{3Y}\, I(U = n) \\ +\alpha_{4Y}\, Y \cdot I(U = a) + \alpha_{5Y}\, Y \cdot I(U = n).\end{aligned}$$

[3]Response exclusion restriction for a certain compliance status [2] imposes that missingness does not depend on the assignment to treatment conditionally for that compliance status.

[4]Frangakis and Rubin [2] and Mealli et al. [9] impose the response exclusion restriction for two compliance statuses.

3.2 Missingness in Y, D, and Z with a Binary Pretreatment Variable

In the previous Subsection, the double need to comply with the maximum number of parameters to ensure model (1) is identifiable and, at the same time, to avoid the proposal of unplausible models directed us to restrict the three marginal missing data mechanisms $P(R_v|\, Y,\, U,\, Z)$, $v = Y,\, D,\, Z$, to be mutually independent. The restriction can be relaxed by introducing a binary and always-observed pretreatment variable X. This implies a larger contingency table for the observables and a large numbers of parameters allowed for model (1) to be identified. The contingency table for the observable data (not shown here) has 3 rows and 18 columns that permit at most $(3 \times 18) - 1 = 53$ parameters for $(\mathbf{R},\, Y^*,\, D^*,\, Z^*,\, X)$.

Following [1] we assume that X does not enter in the missing data mechanism so that: $P(\mathbf{R},\, Y^*,\, D^*,\, Z^*,\, X) = P(\mathbf{R}\,|\, Y,\, U,\, Z) \cdot P(Y, U, Z, X)$, where

$$
\begin{aligned}
&P(Y,\, U,\, Z,\, X;\, \pi, \boldsymbol{\omega}, \boldsymbol{\theta}) \\
&= \pi^{Z}\,(1-\pi)^{1-Z}\,\omega_{1a}^{I(X=1,a)}\,\omega_{0a}^{I(X=0,a)}\,\omega_{1n}^{I(X=1,n)}\,\omega_{0n}^{I(X=0,n)} \\
&\quad \times \omega_{1c}^{I(X=1,c)}\,(1-\omega_{1a}-\omega_{0a}-\omega_{1n}-\omega_{0n}-\omega_{1c})^{I(X=0,c)}\,\theta_{1a}^{Y\,I(X=1,a)}\,\theta_{0a}^{Y\,I(X=0,a)} \\
&\quad \times (1-\theta_{1a})^{(1-Y)\,I(X=1,a)}\,(1-\theta_{0a})^{(1-Y)\,I(X=0,a)}\,\theta_{1n}^{Y\,I(X=1,n)}\,\theta_{0n}^{Y\,I(X=0,n)} \\
&\quad \times (1-\theta_{1n})^{(1-Y)\,I(X=1,n)}\,(1-\theta_{0n})^{(1-Y)\,I(X=0,n)}\,\theta_{1c1}^{Y\,I(X=1,c)\,Z}\,\theta_{0c1}^{Y\,I(X=0,c)\,Z} \\
&\quad \times (1-\theta_{1c1})^{(1-Y)\,I(X=1,c)\,Z}\,(1-\theta_{0c1})^{(1-Y)\,I(X=0,c)\,Z}\,\theta_{1c0}^{Y\,I(X=1,c)\,(1-Z)} \\
&\quad \times \theta_{0c0}^{Y\,I(X=0,c)\,(1-Z)}(1-\theta_{1c0})^{(1-Y)\,I(X=1,c)\,(1-Z)}\,(1-\theta_{0c0})^{(1-Y)\,I(X=0,c)\,(1-Z)}.
\end{aligned}
$$

Given the larger number of free parameters, the missing data mechanism can be now modeled as a multinomial logit for the 8 combinations $\mathbf{R} = (R_Y,\, R_D,\, R_Z)$: $\{(1,1,1),\,(1,1,0),\ldots\}$. Taking into account that 14 parameters are involved in specification (2) of $P(Y, U, Z, X)$, the following two plausible multinomial models for $P(\mathbf{R}\,|\, Y,\, U,\, Z)$ can be delineated.

The first one complies with the natural choice for a nonignorable missing data mechanism, that is the simple dependency of $P(\mathbf{R}\,|\, Y,\, U,\, Z)$ on the values of the variables subjected to missingness, Y, D, Z. The proposed linear specification for the logit of $P(\mathbf{R}\,|\, Y,\, U,\, Z)$ is:

$$
\begin{aligned}
&\text{logit}[P(R_Y,\, R_D,\, R_Z)] \\
&= \alpha_{0R_y R_d R_z} + \alpha_{1R_y R_d R_z}\, Y + \alpha_{2R_y R_d R_z}\, I(U = a,\, Z = 0) \\
&\quad + \alpha_{3R_y R_d R_z}\, I(U = n,\, Z = 1) + \alpha_{4R_y R_d R_z}\,[I(U = a,\, Z = 1) \\
&\quad + I(U = c,\, Z = 1)].
\end{aligned}
$$

To note that the treatment D indirectly enters into $\text{logit}[P(R_Y,\ R_D,\ R_Z)]$ by the compliance status that is a function of the couple of potential quantities $D(Z = z)$. The logit can be coherently rewritten as:

$$\begin{aligned}\text{logit}[P(R_Y,\ R_D,\ R_Z)] = \alpha_{0R_yR_dR_z} + \alpha_{1R_yR_dR_z}\, Y + \alpha_{2R_yR_dR_z}\, I(D = 1,\ Z = 0)\\ +\alpha_{3R_yR_dR_z}\, I(D = 0,\ Z = 1)\\ +\alpha_{4R_yR_dR_z}\, I(D = 1,\ Z = 1).\end{aligned}$$

The second specification imposes response exclusion restrictions for two compliance statuses; for example, response exclusion restriction for noncompliers:

$$\begin{aligned}\text{logit}[P(R_Y,\ R_D,\ R_Z)] = \alpha_{0R_yR_dR_z} + \alpha_{1R_yR_dR_z}\, Y + \alpha_{2R_yR_dR_z}\, I(U = a)\\ +\alpha_{3R_yR_dR_z}\, I(U = n) + \alpha_{4R_yR_dR_z}\, I(U = c,\ Z = 1).\end{aligned}$$

Each of the two proposed specifications imposes $7 \times 5 = 35$ parameters for the missing data mechanism.

3.3 Missingness in Y and D Without Pretreatment Variables

In economic and social sciences the model of a randomized experiment with noncompliance is often used in observational studies, where most common choices for the natural assignment to treatment Z are registry information (such as date of birth, location, etc.), usually unaffected by the problem of nonresponses. A plausible assumption is that there is no missingness in the treatment assignment: $P(R_z = 1) = 1$. Table 2 shows the contingency table for the observable data that now allows $(3 \times 6) - 1 = 17$ free parameters for $(\mathbf{R},\ Y^*,\ D^*,\ Z^*,\ X)$.

The model for $P(Y, U, Z)$ remains as specified in (2), while an identified missing data mechanism can be specified by a multinomial logit model where each logit depends on the values of the two variables subjected to missingness Y and D:

$$\begin{aligned}\text{logit}[P(R_Y,\ R_D)] = \alpha_{0R_yR_d}\\ +\alpha_{1R_yR_d}\, Y + \alpha_{2R_yR_d}\, [I(U = a) + I(U = c,\ Z = 1)].\end{aligned}$$

Table 2 Contingency table for missingness in Y and D without pretreatment variables

	$D^*,\ Z$					
	1, 0	0, 1	1, 1	0, 0	∗, 1	∗, 0
$Y^* = 1$						
$Y^* = 0$						
$Y^* = *$						

Again, the treatment indirectly enters into logit$[P(\mathbf{R}|\, Y,\, U,\, Z)]$ by the compliance status so that the logits can be rewritten as:

$$\text{logit}[P(R_Y,\, R_D)] = \alpha_{0R_yR_d} + \alpha_{1R_yR_d}\, Y + \alpha_{2R_yR_d}\, D.$$

4 Bayesian Inference

There is mainly a pragmatic justification in adopting a Bayesian instead of a maximum likelihood approach as frequently done in order to exploit the mixture structure of the likelihood. In principle, a ML estimation via the EM algorithm could be attractive also in case of nonignorable condition for the missing data mechanism because if Y, U and Z were known for all units, $P(Y, U, Z)$ would not involve mixture components. However, and contrary to the cases of ignorable missing data mechanisms, the augmented log-likelihood function would not be linear in the missing information. Consequently the EM algorithm could not work by simply filling-in missing data and then updating the parameter estimates, so that the expected augmented log-likelihood function should be computed at each iteration of the algorithm. Then, it is computationally more convenient to conduct Bayesian inference.

Because of the mixture structure of the observed likelihood, the posterior distributions can be sensitive to the choice of the prior distributions. As shown by Hirano et al. [4] in a similar context, standard diffuse improper prior can lead to improper posteriors. We adopt here proper prior distributions that correspond to add a number of observations to the likelihood. Under nonignorable conditions for the missing data mechanism we add 48 observations to the likelihood: one for each of the 48 combinations of the variables (Z, U, Y, R_Y, R_D). For example: for $(Z = 1, U = n, Y = 0, R_Y = 1, R_D = 1)$ the following term is added to the likelihood:

$$\pi \cdot \omega_n \cdot (1 - \theta_n) \cdot \exp(\alpha_{011}) \cdot (1 + \exp(\alpha_{011}) + \exp(\alpha_{010}) + \exp(\alpha_{001}))^{-1}.$$

The same arguments applies to the priors for the analysis under MAR and MCAR leading to 12 and 30 added observations to the likelihood respectively.

For what concern the choice of the algorithm on which to base the MCMC analysis, the Gibbs is not particularly attractive given there are no conjugate forms for the prior distribution and no standard forms for the conditional posterior distributions under nonignorable missing data. Consequently we base the analysis on the Metropolis-Hastings algorithm: a general term for a family of Markov Chain method useful for drawing samples from posterior distributions. It is an adaptation of a random walk that uses an acceptance/rejection rule to converge to the specified posterior.

5 Simulation

We now present a comparison of results obtained by different models applied to an artificial sample from a hypothetical population affected by nonignorable missing data in the binary outcome and in the binary treatment received. The simulation analysis shows the biases we incurred when simpler but wrong models such as those based on ignorable conditions for the missing data mechanism, MAR, or on listwise deletion of units with missing data, MCAR, are applied.

Table 3 reports the parameter values for the proposed hypothetical population. Table 4 reports the marginal and conditional probabilities of missingness implied by hypothetical missing data mechanism. This is clearly nonignorable because the probabilities of missingness for Y and D depend on their values.

A sample of size 5,000, was drawn from the hypothetical population, and four Markov Chains each of 200,000 iterations were run. The starting values were drawn from an overdisperded normal distribution, while the parameters were updated in batch with an Acceptance Rate of $\approx 20\,\%$. The convergence, for each parameter, was assessed by the potential scale reduction indicator, $\hat{R} = \sqrt{(W + (B/n))/W} < 1.1$ [3]. Finally, the posterior inferences were from the second half of the simulated sequences of parameters.

Table 5 reports the posterior means and standard deviations for a subset of the parameter vector. Of particular concern is the bias arising from the inference under the usual MAR and MCAR models. In particular the compliers average causal effect, $\mu_{c1} - \mu_{c0}$, that is the main quantity under study in the context of a randomized experiment with noncompliance, results clearly biased alongside the two compliance-status probabilities ω_a and ω_n under MAR and MCAR models.

Table 3 Hypothetical population: parameters values

$\pi = P(Z = 1)$	0.50	α_{011}	−2
$\omega_a = P(U = a)$	0.30	α_{111}	2
$\omega_n = P(U = n)$	0.45	α_{211}	0.5
$1 - \omega_a - \omega_n = P(U = c)$	0.25	α_{010}	0.5
μ_a	0.90	α_{110}	1.5
μ_n	0.20	α_{210}	1.5
μ_{c1}	0.70	α_{001}	−2
μ_{c0}	0.40	α_{101}	1.5
$\mu_{c1} - \mu_{c0}$	0.30	α_{201}	2

Table 4 Hypothetical population: marginal and conditional probabilities of missingness

$P(R_Y = 0)$	0.2544
$P(R_D = 0)$	0.2292
$P(R_D = 0 \mid D = 0)$	0.3429
$P(R_D = 0 \mid D = 1)$	0.0755
$P(R_Y = 0 \mid Y = 0)$	0.3646
$P(R_Y = 0 \mid Y = 1)$	0.1431

Table 5 Posterior means and standard deviations for some parameters from the application of the correct model alongside those calculated under the MAR and MCAR assumptions

	Real value	Nonignorable		MAR		MCAR	
π	0.50	0.506	(0.007)	0.506	(0.020)	0.521	(0.009)
ω_a	0.30	0.328	(0.015)	0.379	(0.025)	0.369	(0.012)
ω_n	0.45	0.435	(0.017)	0.387	(0.032)	0.396	(0.012)
ω_c	0.25	0.237	(0.014)	0.232	(0.038)	0.234	(0.016)
$\mu_{c1} - \mu_{c0}$	0.30	0.287	(0.071)	0.239	(0.075)	0.225	(0.063)

Table 6 Conditional probabilities of the missing data indicators: real values alongside the posterior means and standard deviations from the application of the correct model

	Real value	Posterior value	
$P(Ry = 1, Rd = 1 \mid Y = 1, D = 1)$	0.823	0.794	(0.028)
$P(Ry = 1, Rd = 1 \mid Y = 1, D = 0)$	0.739	0.706	(0.069)
$P(Ry = 1, Rd = 1 \mid Y = 0, D = 1)$	0.769	0.730	(0.057)
$P(Ry = 1, Rd = 1 \mid Y = 0, D = 0)$	0.565	0.588	(0.044)
$P(Ry = 1, Rd = 0 \mid Y = 1, D = 1)$	0.041	0.047	(0.009)
$P(Ry = 1, Rd = 0 \mid Y = 1, D = 0)$	0.100	0.077	(0.019)
$P(Ry = 1, Rd = 0 \mid Y = 0, D = 1)$	0.023	0.033	(0.013)
$P(Ry = 1, Rd = 0 \mid Y = 0, D = 0)$	0.046	0.052	(0.006)
$P(Ry = 0, Rd = 1 \mid Y = 1, D = 1)$	0.111	0.095	(0.012)
$P(Ry = 0, Rd = 1 \mid Y = 1, D = 0)$	0.061	0.044	(0.017)
$P(Ry = 0, Rd = 1 \mid Y = 0, D = 1)$	0.104	0.130	(0.046)
$P(Ry = 0, Rd = 1 \mid Y = 0, D = 0)$	0.046	0.056	(0.010)

Table 6 shows close to unbiased posterior means with low values of the standard deviations for the missing data mechanism evaluated under the correct model.

Conclusions

One major problem in dealing with nonignorable missing data in randomized experiments with noncompliance is parameter identifiability. Based on the analysis of contingency tables, this article investigates the identification issue in the models for studies with nonignorable missing data in the response, as well as in the assignment to treatment and/or in the treatment received. Our simulation results suggest that the causal estimates were sensitive to the assumption of the missing data mechanism, which merits special attention in practice. The identification conditions considered here can be extended to incorporate continuous outcomes. Moreover, within the Bayesian approach proposed here, one can further conduct sensitivity analysis regarding possible deviations from the identifiability conditions and varying proportions of missing data.

References

1. Chen, H., Geng, Z., Zhou, X.H.: Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. Biometrics **65**, 675–682 (2008)
2. Frangakis, C.E., Rubin, D.B.: Addressing complications of intention to treatment analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika **86**, 365–379 (1999)
3. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis. Chapman and Hall/CRC, Boca Raton (2004)
4. Hirano, K., Imbens, G.W., Rubin, D.B., Zhou, X.: Estimating the effect of an influenza vaccine in an encouragement design. Biostatistics **1**, 69–88 (2000)
5. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–970 (1986)
6. Imai, K.: Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. J. R. Stat. Assoc. Series C **58**, 83–104 (2009)
7. Imbens, G.W., Rubin, D.B.: Bayesian inference for causal effects in randomized experiments with noncompliance. Ann. Stat. **25**, 305–327 (1997)
8. Jo, B.: Estimation of intervention effects with non compliance: alternative model specifications. J. Educ. Behav. Stat. **27**, 385–409 (2002)
9. Mealli, F., Imbens, G.W., Ferro, S., Biggeri, A.: Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. Biostatistics **5**, 207–222 (2004)
10. Molinari, F.: Missing treatments. J. Bus. Econ. Stat. **28**, 82–95 (2010)
11. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
12. Small, D.S., Cheng, J.: Discussion of "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data" by Chen H., Geng, Z., Zhou, X.H. Biometrics **65** (2008)

Approximate Bayesian Quantile Regression for Panel Data

Antonio Pulcini and Brunero Liseo

1 Introduction

Quantile Regression (QR hereafter) and the use of panel—or longitudinal—data are two popular extension of the standard linear regression model and they represent two very active fields of research, both in statistical and econometric literature.

In particular, QR is a very useful tool when the response variable is known to have a non Gaussian distribution and/or we are particularly interested on the way in which the explanatory variables are associated with the occurrence of extreme values in the response variable.

On the other hand, the use of panel data, and the consequent adoption of mixed models, is nowadays current practice, due to the frequent availability of data observed on the same units in several different occasions. In a panel data set up, the main goal is to identify the factors that contribute to explain changes in the response variable over time. In general these factors express their influence not only in terms of location (i.e. mean or median) of the response; more generally their influence can modify the entire distribution and this can happen both at a fixed observation time and dynamically as well. As stressed in [13], covariates may influence the conditional distribution of the response variable in many ways: expanding its dispersion, stretching one tail of the distribution and/or compressing the other tail, and so on. In these situation the standard linear mixed-effect regression model might be inadequate and a QR approach may be more helpful.

A. Pulcini
ACEA ENERGIA SPA, Rome, Italy
e-mail: antoniopulcini@gmail.com

B. Liseo (✉)
MEMOETEF, Sapienza Università di Roma, Viale del Castro Laurenziano 9, Rome 00161, Italy
e-mail: brunero.liseo@uniroma1.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_12

The combined use of quantile regression methods and longitudinal data is not yet very popular, notwithstanding the excellent work of Koenker [12] and Geraci and Bottai [6], which consider the problem from a classical perspective. In particular Koenker [12] shows how to estimate quantile functions in a mixed model with fixed effects using the penalized interpretation of the classical random-effects estimator. Geraci and Bottai [6] have proposed a method to provide an estimator of the conditional quantile functions with quantile-dependent individual effects.

Bayesian analysis of quantile regression models have been mainly considered in the parametric settings [22]: in fact it is well known that a quantile estimation problem can be rephrased in terms of a statistical model based on the Asymmetric Laplace distribution (ALD, hereafter), and standard MCMC algorithms can be easily implemented.

However, in the recent years, some Bayesian semi and nonparametric approaches to quantile estimation have been proposed. Good references in this context are [8,9, 15].

In this paper we propose a novel approximate Bayesian approach for making inference in quantile regression models in the presence of repeated observations on the same units. The method is *approximated* since it does not use a "true" likelihood function. In fact we will consider the so-called substitution likelihood, a concept introduced by Jeffreys [11] and more recently analysed by Lavine [17]. The substitution likelihood is a sort of multinomial likelihood function which is able to estimate the quantiles of the unknown cdf F which truly generated the data, without making any distributional assumption on the variable of interest. Dunson and Taylor [4] have already considered a Bayesian use of the substitution likelihood in a linear model with cross-sectional data.

The paper is organized as follows: in Sect. 2, we briefly recall the standard quantile regression methods. In Sect. 3, we illustrate the concept of substitution likelihood in detail. In Sect. 4 we describe our novel method and prove the main theorem. Section 5 is devoted to numerical simulation and comparisons with existing methods.

2 Quantile Regression

In this section we briefly review the standard literature in QR, based on the notion of the asymmetric absolute loss function and on the ALD, with particular emphasis on the panel data setting.

2.1 Standard Model and Methods

Suppose data $(y_i, \mathbf{x}_i')$, $i = 1, \ldots, n$, are available, where the y_i's are independent and continuous random variables distributed according to a cumulative distribution

function (cdf) F_Y, and the $\mathbf{x}_i'$'s are K-dimensional vectors of covariates. The linear conditional τ-th quantile function is defined as

$$F_{Y_i}^{-1}(\tau|\mathbf{x}_i') = Q_{y_i}(\tau|\mathbf{x}_i') = \mathbf{x}_i'\beta(\tau) \qquad i = 1, \ldots, n$$

where $0 < \tau < 1$ is a fixed probability and $\beta \in \mathbb{R}^K$ is a (possibly) τ-dependent vector of unknown parameters, which will represent our main object of interest. Without any distributional assumption on F_Y, $\beta \in \mathbb{R}^k$ can be estimated through the following minimization problem:

$$\min_{\beta \in \mathbb{R}^K} V_n(\beta) = \min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\beta), \tag{1}$$

where $\rho_\tau(\cdot)$ is the asymmetric absolute loss function

$$\rho_\tau(\nu) = \nu(\tau - I(\nu \leq 0)), \quad \nu \in \mathbb{R}$$

and $I(\cdot)$ is the usual set indicator function. The above minimization problem is the "natural" quantile regression counterpart of the standard OLS [13]. The role which is played by the Gaussian distribution in standard linear regression model is now taken by the ALD, which represents a distributional bridge between the previous minimization problem and maximum likelihood theory for quantile regression problems. The ALD [14, 22] is a skewed distribution with three parameters: a skewness parameter $0 < \tau < 1$, a scale parameter $\sigma > 0$ and a location parameter $\mu \in \mathbb{R}$. A random variable $Y \in \mathbb{R}$, is distributed according to an ALD with parameters μ, σ and τ, if its density is given by:

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}.$$

Yu et al. [23] gave properties and generalizations of this distribution. Note that, for any fixed value of the parameters $F_Y(\mu) = \tau$. Then, setting $\mu_i = \mathbf{x}_i'\beta$, and if we assume to observe $Y_1, \ldots, Y_n$ independently from an ALD(μ_i, τ, σ) density, the corresponding likelihood function is

$$L(\beta, \sigma; (y), \tau) \propto \sigma^{-n} \exp\left\{-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \mu_i}{\sigma}\right)\right\}. \tag{2}$$

By taking σ as a nuisance parameter, the maximization of the likelihood in (2) with respect to β is equivalent to the minimization of the function $V_n(\beta)$ in (1). This explains why the ALD distribution is often used in the QR estimation instead of the asymmetric absolute loss function, in order to apply likelihood and Bayesian approaches for point estimation and hypotheses testing [14, 16, 22].

The above description is suitable for the estimation of a single quantile. In real applications, researchers are often interested to the simultaneous estimation of a set of quantiles. This case brings the additional difficulty that some monotonicity constraints should be considered since different quantile function (of the covariates) cannot cross.

2.2 QR for Panel Data

Suppose we have panel data in the form $(\mathbf{x}'_{it}, y_{it})$, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\mathbf{x}'_{it}$ are K-dimensional vectors of predictors of a design matrix X_i and y_{it} is the t-th measurement of a continuous response variable on the i-th statistical unit. The linear conditional τ-th quantile function is defined as

$$Q_{y_{it}}(\tau|\mathbf{x}'_{it}, \beta, \alpha_i) = \mathbf{x}'_{it}\beta(\tau) + \alpha_i \qquad i = 1, \ldots, n \quad t = 1, \ldots, T$$

where α_i represents the individual random effect, $i = 1, \ldots, n$. In order to estimate the parameters β and α_i's, two main approaches are described in [6, 12]. In the former, the Author, avoiding any distributional assumption for the shape for the response variable, and adopting the asymmetric absolute loss function as in (1), proposed a penalized approach, based on an L_1 penalty term, to simultaneously estimate m different quantiles.

$$\min_{\alpha,\beta} \sum_{l=1}^{m} \sum_{i=1}^{n} \sum_{t=1}^{T} \omega_l \rho_{\tau_l}(y_{it} - \mathbf{x}'_{it}\beta(\tau_l) - \alpha_i) + \lambda \sum_{i=1}^{n} |\alpha_i|. \tag{3}$$

In the above formula, each single weight ω_l controls the influence of the τ_l-th quantile in the estimation of the individual effect α_i, while λ may be considered as a penalization parameter. This approach shares some similarities with a LASSO procedure [21] on the random coefficients. This procedure have same potential drawbacks: the parameter λ must be arbitrarily set and this choice could play a role in the estimation of the $\beta(\tau_l)$'s. Also, the individual effects are not related to the τ_l's; last, but not least, as it is usual in mixed models, the number of parameters increases with the sample size.

To circumvent these drawbacks [6] based their model on the ALD loss and they assume that the individual effects are independent and identically distributed random variables with a distribution which possibly depends on the quantiles τ's; they also propose a Monte Carlo EM algorithm to estimate the parameters. In detail, Geraci and Bottai [6] proposed the following model:

$$Q_{Y_{it}|\alpha_i}(\tau|\mathbf{x}'_{it}, \alpha_i) = \mathbf{x}'_{it}\beta(\tau) + \alpha_i \qquad i = 1, \ldots, n \quad t = 1, \ldots, T \tag{4}$$

with

$$Y_{it} \overset{\text{ind}}{\sim} \text{ALD}(\mathbf{x}'_{it}\beta(\tau)+\alpha_i, \sigma, \tau) \quad \forall i, t \quad \text{and} \quad \alpha_i \overset{\text{iid}}{\sim} f_\alpha(\cdot|\varphi(\tau)) \qquad i = 1, \ldots, n.$$

The main concern with this model is related to the multiple quantile issue. When inference is needed for more than one quantile, the model needs to be separately calibrated for each corresponding τ. This approach need not automatically satisfy the non-crossing condition among the quantile curves. As a possible way to circumvent the above mentioned problems, in this paper we propose a Bayesian approach based on the idea of Substitution Likelihood [11, 17]. In particular we generalize the approach illustrated in [4], in the case of cross-sectional data. This approach naturally allows one to simultaneously estimate m different regression quantiles without assuming any parametric distributional form for the $Y_{it}|\mathbf{x}'_{it}, \alpha_i$. Our aim is to propose a simple and easy-to-use nonparametric method for estimating the regression quantile curves for panel data, when the distribution of the outcome variable cannot be safely assumed to be Asymmetric Laplace.

3 An Approximate Likelihood for Quantiles

In many empirical application, the distribution of the outcome variable is clearly non normal and standard regression methods might be unreliable. On the other hand, the use of asymmetric Laplace densities for quantile regression can only be justified in terms of computational convenience—unless one agrees to assume—incidentally—a specific form of the loss function. Under these conditions, Bayesian inference is non trivial and this general problem has motivated a vast literature in Bayesian nonparametric methods. For an excellent review on this approach see [10] and, for recent, specific applications in quantile regression see [15, 20].

Although Bayesian nonparametric methods may be useful in this context, their implementation is often quite difficult, due to computational complexity. An alternative and feasible solution is represented by the use of a simple and natural approximation of the likelihood function, which can then be complemented by some prior information into a standard Bayesian framework.

Suppose that inference concerns $\boldsymbol{\theta}$, a vector of some fixed quantiles of the true distribution F, and some prior knowledge is available only for those quantiles: Jeffreys [11] and later Lavine [17] have proposed the use of the so called "substitution likelihood", which we will describe in the next section. This idea has been then implemented in a quantile regression for cross-sectional data framework by Dunson and Taylor [4]. Similar, although different, approaches are based on the Bayesian version of the Empirical Likelihood [18] and on the Generalized Moment Method in [2].

3.1 The Substitution Likelihood

Suppose $y_1, \ldots, y_n$ is a random sample from un unknown distribution F_Y. Assume that some prior information is available only for $\boldsymbol{\theta} = (\theta(\tau_1), \ldots, \theta(\tau_m))'$, a vector of quantiles of F_Y, corresponding to a known vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)'$ such that $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = 1$. When inference concerns $\boldsymbol{\theta}$, one can replace the true likelihood function, based on $F_Y(\cdot)$, with a multinomial approximation, namely the "substitution likelihood" $s(\boldsymbol{\theta}; \mathbf{y})$, defined as

$$s(\boldsymbol{\theta}; \mathbf{y}) = \binom{n}{Z_1(\boldsymbol{\theta}), \ldots, Z_{m+1}(\boldsymbol{\theta})} \prod_{l=1}^{m+1} \Delta\tau_l^{Z_l(\boldsymbol{\theta})} \tag{5}$$

with $\theta_l = \theta(\tau_l)$ for $l = 1, \ldots, m$, $\theta_0 = -\infty$, and $\theta_{m+1} = +\infty$. In (5), $Z_l(\boldsymbol{\theta}) = \sum_{i=1}^{n} 1_{(\theta_{l-1}, \theta_l]}(y_i)$ is the number of observations falling in the interval $(\theta_{l-1}, \theta_l]$ and $\Delta\tau_l = \tau_l - \tau_{l-1}$, for $l = 1, \ldots, m+1$. In words, $s(\boldsymbol{\theta}; \mathbf{y})$ is a multinomial model which represents, conditionally on $\boldsymbol{\theta}$, a discrete version of the true model.

Here we briefly recall, following [4, 17], the main properties and characteristics of $s(\boldsymbol{\theta}; \mathbf{y})$:

1. the Substitution Likelihood is asymptotically conservative at the true F. This means that, for large samples, $s(\boldsymbol{\theta}; \mathbf{y})$ has less discriminating power than the "true" likelihood [17];
2. the Substitution Likelihood surface looks like a step function with data-dependent jumps and takes its maximum value within a region containing the empirical estimates of the quantiles;
3. the pseudo-posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto s(\boldsymbol{\theta}; \mathbf{y})\pi(\theta)$ is improper when the usual noninformative prior for location parameter is used, namely $\pi(\boldsymbol{\theta}) \propto 1_{(\boldsymbol{\theta} \in \Theta)}$, $\Theta = \{\boldsymbol{\theta} : \theta_1 < \ldots < \theta_m\}$. This is due to the fact that, beyond the most extreme quantiles, the substitution likelihood $s(\boldsymbol{\theta}; \mathbf{y})$ does not vanish to 0.

The last remark suggests that the loss of information resulting from the use of a sort of non parametric likelihood, must be paid in terms of partial elicitation and a proper prior density must be adopted for θ. The simplest solution to this problem, when genuine prior information is lacking, is based on a truncation of the parameter space. This can be done if a lower limit c_L can be assumed for the lowest quantile θ_1 and an upper limit c_U can be assumed for the largest quantile θ_m, with $c_L < c_U$. In this case the prior can be set as:

$$\pi(\boldsymbol{\theta}) \propto 1(\boldsymbol{\theta} \in \Theta) 1(\theta_1 \geq c_L) 1(\theta_m \leq c_U). \tag{6}$$

The use of the prior (6) is reasonable when prior information is available, at least, for c_L and c_U. Alternatively, one can follow the approach described in [4] and use the following prior:

$$\pi(\boldsymbol{\theta}) \propto 1(\boldsymbol{\theta} \in \Theta) g(\theta_1, \theta_m) \tag{7}$$

with

$$g(\theta_1, \theta_m) = \left\{e^{-\phi_L(c_L-\theta_1)}\right\}^{1(\theta_1<c_L)} \left\{e^{-\phi_U(\theta_1-c_U)}\right\}^{1(\theta_1>c_U)} \left\{e^{-\nu(\theta_m-\theta_1-d)}\right\}^{1(\theta_m-\theta_1>d)}. \tag{8}$$

In the above expression d is a known fixed constant which represents our prior beliefs about the range $(\theta_1; \theta_m)$; in addition, ϕ_L, ϕ_U and ν represent positive rate parameters that measure—respectively—how much θ_1 can be far from c_L and c_U, and how much the difference $\theta_m - \theta_1$ can be far from d. The above prior, conditionally on θ_1 and θ_m, is uniform for the rest of the components of θ, while it is exponentially decreasing for θ_1 outside the interval $[c_L, c_U]$ and if $\theta_m - \theta_1 > d$.

Prior assumptions on the extreme quantiles are of course crucial in the use of the substitution likelihood, since the posterior tail behaviour will be completely driven by the prior tails. However, this should not be seen as a drawback: substitution likelihood explicitly states that information for extreme quantiles cannot be easily obtained from the data, unless strong prior assumptions, either in the prior or in the model, are accepted.

4 Substitution Likelihood for Quantile Regression with Panel Data

Dunson and Taylor [4] have proposed the use of the substitutional likelihood for quantile regression in the presence of cross sectional data. Here we extend their approach to the case of panel data, where individual random effects need to be introduced.

Before dealing with the panel data setting, we briefly recall the cross sectional quantile regression model: let $\boldsymbol{\theta}_i$ the vector of the conditional quantiles for the i-th individual, corresponding to the probabilities $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_m)$ with $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = 1$, For each individual we assume to observe covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})' \in \mathscr{X}$. The quantile regression model is defined as:

$$\begin{bmatrix} Q_{y_i}(\tau_1|\mathbf{x}_i') \\ \vdots \\ Q_{y_i}(\tau_m|\mathbf{x}_i') \end{bmatrix} = \boldsymbol{\theta}_i = \begin{bmatrix} \theta_i(\tau_1) \\ \vdots \\ \theta_i(\tau_m) \end{bmatrix} = \begin{bmatrix} \beta(\tau_1)' \\ \vdots \\ \beta(\tau_m)' \end{bmatrix} \mathbf{x}_i = \begin{bmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_m' \end{bmatrix} \mathbf{x}_i = \mathfrak{B}\mathbf{x}_i \tag{9}$$

where $\mathfrak{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)'$ is a $m \times K$ matrix of unknown regression coefficients; each row of $\mathfrak{B}$ provides the regression coefficients for the corresponding conditional quantile. It is apparent that the conditional quantiles should satisfy the so-called no crossing condition, that is, in order to produce proper quantiles, it must be true that

$$\boldsymbol{\beta}_1'\mathbf{x}_i < \ldots < \boldsymbol{\beta}_l'\mathbf{x}_i < \ldots < \boldsymbol{\beta}_m'\mathbf{x}_i \quad \forall i \in 1, \ldots, m. \tag{10}$$

However, it is not sufficient that the above inequality is satisfied for all observed covariates pattern: in principle, it must be satisfied for all $\mathbf{x} \in \mathscr{X}$. Dunson and Taylor [4] provides the sufficient conditions in order to guarantee (10). Let $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$, where $\mathscr{X}_1 = \{(x_1, \ldots, x_{K_1}) : x_r \in \{a_r, a_r + 1, \ldots, b_r\}, r = 1, \ldots, K_1\}$ is a discrete space, and $\mathscr{X}_2$ is a compact convex subset of $\mathbb{R}^{K_2}$, where $K_1 + K_2 = K$. Then

Theorem 1 *Let $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$, where $\mathscr{X}_1 = \{(x_1, \ldots, x_{K_1}) : x_r \in \{a_r, a_r + 1, \ldots, b_r\}\}$, $\forall r$, is a discrete space and $\mathscr{X}_2$ is a compact convex subset of $\mathbb{R}^{K_2}$, with $K_1 + K_2 = K$. Let $\{\mathbf{a}, \tilde{\mathbf{x}}_{11}, \ldots, \tilde{\mathbf{x}}_{1,K_1-1}\}$ be a linearly independent set of vectors in $\mathscr{X}_1$ with $\mathbf{a} = (a_1, \ldots, a_{K_1})'$, and $\{\tilde{\mathbf{x}}_{2,1}, \ldots, \tilde{\mathbf{x}}_{2,K_2+1}\}$ is a set of extreme points of $\mathscr{X}_2$. Set $\tilde{\mathbf{X}}_1 = [\tilde{\mathbf{x}}_{11}, \ldots, \tilde{\mathbf{x}}_{1,K_1-1}]'$, $\tilde{\mathbf{X}}_2 = [\tilde{\mathbf{x}}_{2,1}, \ldots, \tilde{\mathbf{x}}_{2,K_2+1}]'$*

$$\tilde{\mathfrak{X}} = \begin{bmatrix} \tilde{\mathbf{X}}_1 & \boldsymbol{0}_{(K_1-1)\times K_2} \\ \mathbf{a}' \otimes \boldsymbol{I}_{(K_2+1)\times 1} & \tilde{\mathbf{X}}_2 \end{bmatrix},$$

and $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]' = \tilde{\mathfrak{X}}\mathfrak{B}'$. If

$$\boldsymbol{\gamma}_j \in \Theta \quad \textit{for} \quad j = 1, \ldots, K$$

then this implies that

$$\boldsymbol{\theta} = \mathfrak{B}\mathbf{x} \in \Theta$$

for all $\mathbf{x} \in \mathscr{X}$.

The above theorem gives an indirect way of constructing priors on the beta coefficients which automatically satisfies restriction (10).

We now describe in detail how to extend the use of the substitution likelihood to the case of quantile regression with panel data. The main change is that in this new framework, the individual random effects, say α_i's $i = 1, \ldots n$, should be introduced in such a way that the restriction (10) continues to be valid.

Suppose one observes data $(\mathbf{x}'_{it}, y_{it})$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Here the y_{it}'s are independent scalar observations from a continuous random variable Y whose shape is unknown and $\mathbf{x}'_{it}$ are K-dimensional row-vectors of predictors. We define the following linear mixed quantile function for the response y_{it} at a given probability level τ [6],

$$Q_{y_{it}|\alpha_i}(\tau|\mathbf{x}_{it}, \alpha_i) = \theta_{it|\alpha_i}(\tau) = \mathbf{x}'_{it}\beta(\tau) + \alpha_i$$

where $0 < \tau < 1$, and α_i represents the individual effect. Furthermore, we will assume that

1. conditionally on α_i, y_{it}'s are independent random variable with unspecified distribution $f_{Y|\alpha_i}$, $i = 1, \ldots n$;
2. α_i's are i.i.d. from a specified density f_α;

3. prior beliefs (in the form of bounds and/or distributional information) on the quantiles of the response variable are available for some specific points of predictor's space, i.e. for some $\mathbf{x} \in \mathscr{X}$.

Given the vector of probabilities:

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$$

we define the following conditional m-quantiles regression model:

$$\boldsymbol{\theta}_{it|\alpha_i} = \begin{bmatrix} \theta_{it|\alpha_i}(\tau_1) \\ \vdots \\ \theta_{it|\alpha_i}(\tau_m) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_m' \end{bmatrix} \mathbf{x}_{it} + \boldsymbol{\iota}_m \alpha_i = \mathfrak{B}\mathbf{x}_{it} + \boldsymbol{\alpha}_i$$

where $\mathfrak{B}$ is a $(m \times K)$ matrix of unknown regression coefficients and ι_m is a m-vector of 1's. The associated substitution likelihood is then:

$$s(\mathfrak{B}|\alpha_1, \ldots, \alpha_n, \mathbf{y}) = \binom{nT}{Z_1(\mathfrak{B}_{|\boldsymbol{\alpha}}) \cdots Z_{m+1}(\mathfrak{B}_{|\boldsymbol{\alpha}})} \prod_{l=1}^{m+1} \Delta\tau_l^{Z_l(\mathfrak{B}_{|\boldsymbol{\alpha}})}$$

with

$$Z_l(\mathfrak{B}_{|\boldsymbol{\alpha}}) = \sum_{i=1}^{n} \sum_{t=1}^{T} 1_{(\boldsymbol{\beta}_{l-1}'\mathbf{x}_{it} + \alpha_i, \boldsymbol{\beta}_l'\mathbf{x}_{it} + \alpha_i]}(y_{it}) \qquad l = 1, \ldots, m+1.$$

As a boundary condition we set with $\boldsymbol{\beta}_0'\mathbf{x}_{it} + \alpha_i = -\infty$, $\boldsymbol{\beta}_{m+1}'\mathbf{x}_{it} + \alpha_i = +\infty$, and $\Delta\boldsymbol{\tau} = (\tau_1, \tau_2 - \tau_1, \ldots, 1 - \tau_m)'$. The subscript "$|\boldsymbol{\alpha}$" reflects the previous first assumption that, conditionally on the individuals effects, the response variables are i.i.d. observations from an unknown distribution. Also, it is worth to keep in mind that when dealing with quantile regression, one is bound to include the intercept parameter, because otherwise the model would suggest that any quantiles must be zero at the origin, which is quite unreasonable for most of the empirical studies. Henceforth, we will then assume that $x_{i_1} = 1, \quad i = 1, \ldots, n$.

We now discuss prior specification for the parameters $\mathfrak{B}$ and α_i's. The quantile interpretation of our $\boldsymbol{\theta}_{it|\alpha_i}$'s and the no crossing condition require then that each single $\boldsymbol{\theta}_{it|\alpha_i}$ belongs to the restricted space $\Theta = \{\boldsymbol{\theta} : \theta_1 < \ldots < \theta_m\}$.

This restriction automatically implies that:

$$\boldsymbol{\beta}_1'\mathbf{x}_{it} + \alpha_i < \ldots < \boldsymbol{\beta}_l'\mathbf{x}_{it} + \alpha_i < \ldots < \boldsymbol{\beta}_m'\mathbf{x}_{it} + \alpha_i \qquad \forall i, t \tag{11}$$

for all $\mathbf{x} \in \mathscr{X}$.

Conditionally on the α_i's—and assuming that remain constant for all τ_l's, we can use again Theorem 1 adapted to a panel data settings. Let $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$, where $\mathscr{X}_1 = \{(x_1, \ldots, x_{q_1}) : x_r \in \{a_r, a_r + 1, \ldots, b_r\}, \quad r = 1, \ldots q_1\}$ is a discrete

space and $\mathscr{X}_2$ is a compact convex subset of $\mathbb{R}^{q_2}$, with $q_1 + q_2 = K$. In the light of Theorem 1, if we choose a set $\{\mathbf{a}, \tilde{\mathbf{x}}_{11}, \ldots, \tilde{\mathbf{x}}_{1,q_1-1}\}$, with $\mathbf{a} = (a_1, \ldots, a_{q_1})'$, of linearly independent vectors in $\mathscr{X}_1$, and another set $\{\tilde{\mathbf{x}}_{2,1}, \ldots, \tilde{\mathbf{x}}_{2,q_2+1}\}$ of extreme points of $\mathscr{X}_2$, we can define a prior distribution on the set

$$\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K]' = \tilde{\mathfrak{X}}\mathfrak{B}'$$

where

$$\tilde{\mathfrak{X}} = \begin{bmatrix} \tilde{\mathbf{X}}_1 & \mathbf{0}_{(q_1-1)\times q_2} \\ \mathbf{a}' \otimes \mathbf{1}_{(q_2+1)\times 1} & \tilde{\mathbf{X}}_2 \end{bmatrix},$$

and, in turn,

$$\tilde{\mathbf{X}}_1 = [\tilde{\mathbf{x}}_{11}, \ldots, \tilde{\mathbf{x}}_{1,q_1-1}]', \quad \tilde{\mathbf{X}}_2 = [\tilde{\mathbf{x}}_{2,1}, \ldots, \tilde{\mathbf{x}}_{2,q_2+1}]'$$

are such that the rows $\boldsymbol{\gamma}_j \in \Theta$ for $j = 1, \ldots, K$. This prior will induce a prior on $\boldsymbol{\theta}_{|\alpha} = \mathfrak{B}\mathbf{x} + \alpha \in \Theta$ for all possible $\mathbf{x} \in \mathscr{X}$. Then the problem of selecting a prior on $\mathfrak{B}$ essentially becomes a problem of choosing a suitable matrix $\tilde{\mathfrak{X}}$; this in turn implies the choice of a prior distribution on the coordinates of vectors $\boldsymbol{\gamma}'_j$ of Γ. The matrix Γ plays a crucial role for several reasons:

1. it defines linear constraints on the element of the matrix $\mathfrak{B}$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\gamma}'_1 \\ \vdots \\ \boldsymbol{\gamma}'_K \end{bmatrix} = \tilde{\mathfrak{X}}\mathfrak{B}' \quad \text{s.t.} \quad \boldsymbol{\gamma}_j \in \Theta, j = 1, \ldots K;$$

2. it defines the conditional—on $\tilde{\mathfrak{X}}$—quantiles.

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{K1} & \cdots & \gamma_{Km} \end{bmatrix} = \tilde{\mathfrak{X}}\mathfrak{B}'$$

$$= \begin{bmatrix} \tilde{\mathbf{X}}_1 & \mathbf{0}_{(q_1-1)\times q_2} \\ \mathbf{a}' \otimes \mathbf{1}_{(q_2+1)\times 1} & \tilde{\mathbf{X}}_2 \end{bmatrix} \begin{bmatrix} \beta_1(\tau_1) & \ldots & \beta_1(\tau_m) \\ \vdots & \ddots & \vdots \\ \beta_K(\tau_1) & \ldots & \beta_K(\tau_m) \end{bmatrix}.$$

In other words, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jm})$ represents the quantile vector of the response variable Y at specific values of the covariates space $\mathscr{X}$, which are given by the rows $\tilde{x}_j = (\tilde{x}_{j1}, \ldots, \tilde{x}_{jK})$ of $\tilde{\mathscr{X}}$.

From the above discussion, it is apparent that the matrix $\tilde{\mathfrak{X}}$ should be chosen, in such a way that its rows define points of the covariate space for which prior beliefs,

intrinsic bounds or any other prior informations, on the corresponding quantiles of Y are available. In regression models, when choosing the matrix $\tilde{\mathfrak{X}}$, one should take the value of $\tilde{x}_{11}$ not equal to zero, in order to avoid the collapse of all quantile curves at zero. In summary, prior elicitation should be considered for the parameter matrix Γ and the random effects α_i's, $i = 1, \ldots, n$. According to (7), we will adopt a uniform prior for each row of the Γ matrix, except for the penalizing factor on the tails of the lower and upper quantile, that is

$$\pi(\gamma_1, \gamma_2, \ldots, \gamma_K) = \pi(\gamma_1)\pi(\gamma_2|\gamma_1)\cdots\pi(\gamma_K|\gamma_1, \ldots, \gamma_{K-1}), \tag{12}$$

where, for each $j = 1, \ldots, K$,

$$\pi(\gamma_j|\gamma_1, \ldots, \gamma_{K-1}) = 1(\gamma_j \in \Theta)g(\gamma_{j1}, \gamma_{jm}),$$

as in (7).

No particular problems arise in the choice of the prior distribution for the α_i's. One should only notice that, since the model already contains an intercept, they should have a prior mean equal to zero In the following, we shall assume that the α_i's are conditionally on an hyper-parameter σ, independent and identically normally distributed, that is

$$\alpha_i|\sigma^2 \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$$

In this set-up σ represents a measure of dispersion of the random effects.

4.1 Posterior Computation

Posterior inference with the substitution likelihood must be based on the pseudo-posterior

$$\pi(\mathfrak{B}, \sigma^2, \alpha_1, \ldots, \alpha_n|\mathbf{y}) \propto s(\mathfrak{B}|\alpha_1, \ldots, \alpha_n, \mathbf{y})\pi(\mathfrak{B})\pi(\sigma^2)\prod_{i=1}^{n}\pi(\alpha_i|\sigma^2). \tag{13}$$

The non conjugate form of (13) prevents one from using both closed form calculations and Markov Chain Monte Carlo methods based on the direct sampling from the full conditional distributions as, for example, the Gibbs sampling algorithm. Also, the simple off-the-shelf- Metropolis algorithm is very difficult to tune up, given the potentially large number of parameters involved in the model. More precisely, our model contains $m \times K + 1$ parameters, corresponding to the elements of the matrix $\mathfrak{B}$ and to the hyperparameter σ. Our simulation experience with the simple Normal Symmetric Random Walk Metropolis (N-SRWM), were discouraging, since it was

quite difficult to be close to the optimal acceptance rate of about 30 % for all the parameters [19].

To circumvent this drawback we have implemented an Adaptive Markov Chain Monte Carlo (AMCMC). AMCMC algorithm are special MCMC methods which allows to get the required acceptance probability without specifying—a priori—the variance of the proposal density. The idea is that the proposal distribution evolves at each iteration via a self-learning structure. In particular, we have implemented the Algorithm 4 in [1].

Another issue in our MCMC algorithm was the fact that, by directly proposing values in the $\mathfrak{B}$ space, there is no guarantee that the proposed value will satisfy the linear constraints established by the matrix Γ.

We have not found yet a satisfactory and feasible solution to this problem and we are currently working in this direction; in this paper we have adopted the simple strategy of rejecting the proposed values of $\boldsymbol{\beta}$'s which did not satisfy the linear constraints [5].

We now describe in detail the proposed AMCMC algorithm, a sort of adaptive N-SRWM within Gibbs on $\boldsymbol{\beta}_l$, α_i and σ. In the algorithm $\lambda^{(h)}$ represents the tuning parameter of the proposal at time h.

1: At time 0, choose an internal point to start from.
2: At time $h+1$, sample a candidate for $(\alpha_1, \ldots, \alpha_n)$ using a Gaussian proposal with zero mean and fixed standard deviation.
3: At time $h+1$, draw a value for σ^2 form its full conditional distribution, which is an Inverse Gamma with shape $a + n/2$ and scale $b + \sum(\alpha_i^2)^{(h+1)}$
4: At time $h+1$, sample a candidate for the l-th row of $\mathfrak{B}^*$, $l = 1, \ldots, m$— from the proposal density:

$$\boldsymbol{\beta}_l^* \sim N_K(\boldsymbol{\beta}_l^{(h)} \lambda_l^{(h)} \Sigma_l^{(h)}).$$

5: Calculate $\boldsymbol{\Gamma}^* = \tilde{\mathfrak{X}}\mathfrak{B}^{*'}$. If $\boldsymbol{\gamma}_j^* \notin \Theta$, for some j, then reject the candidate and return to step 2. Otherwise set $\boldsymbol{\beta}_l^{(h+1)} = \boldsymbol{\beta}_l^*$ with probability

$$p(\boldsymbol{\beta}_l^{(h)}, \boldsymbol{\beta}_l^{(h+1)}) = \min \left\{ 1, \frac{s(\boldsymbol{\beta}_l^* | \boldsymbol{\beta}_{-l}, \alpha_1^{(h+1)}, \ldots, \alpha_n^{(h+1)}, \mathbf{y}) \pi(\boldsymbol{\beta}_l^*)}{s(\boldsymbol{\beta}_l^{(h)} | \boldsymbol{\beta}_{-l}, \alpha_1^{(h+1)}, \ldots, \alpha_n^{(h+1)}, \mathbf{y}) \pi(\boldsymbol{\beta}_l^{(h)})} \right\},$$

or $\boldsymbol{\beta}_l^{(h+1)} = \boldsymbol{\beta}_l^{(h)}$, with probability $1 - p(\boldsymbol{\beta}_l^{(h)}, \boldsymbol{\beta}_l^{(h+1)})$.

(continued)

6: Update the variance of the proposal density $N(\boldsymbol{\beta}_l^{(h)}, \lambda_l^{(h)} \Sigma_l^{(h)})$ through the following relations

$$\log(\lambda_l^{(h+1)}) = \log(\lambda_l^{(h)}) + \delta^{(h+1)}[p(\boldsymbol{\beta}_l^{(h)}, \boldsymbol{\beta}_l^{(h+1)}) - \bar{p}^*]$$
$$\overline{\boldsymbol{\beta}}_l^{(h+1)} = \overline{\boldsymbol{\beta}}_l^{(h)} + \delta^{(h+1)}(\boldsymbol{\beta}_l^{(h+1)} - \overline{\boldsymbol{\beta}}_l^{(h)})$$
$$\Sigma_l^{(h+1)} = \Sigma_l^{(h)} + \delta^{(h+1)}[(\boldsymbol{\beta}_l^{(h+1)} - \overline{\boldsymbol{\beta}}_l^{(h)})(\boldsymbol{\beta}_l^{(h+1)} - \overline{\boldsymbol{\beta}}_l^{(h)})' - \Sigma_l^{(h)}]$$

where $\bar{p}^*$ is the target acceptance probability and δ is $O(1/h)$.

7: Repeat steps 2–4 for $l = 1, \ldots, m$.

5 Simulation Study

A simulation study has been conducted to evaluate the performance of the proposed method under different scenarios. For the quartiles $\boldsymbol{\tau} = (0.25, 0.5, 0.75)$, we have considered a sample size $n = 100$ and $T = 5$ replications for each subject. Data have been generated using the location-shift model

$$y_{it} = 100x_{1,it} + 10x_{2,it} - 100x_{3,it} + \alpha_i + \varepsilon_{it} \quad i = 1, \ldots, 100 \quad t = 1, \ldots, 5$$

with $x_{1,it} = 1$, $x_{2,it} = t$ and $x_{3,it}$ randomly generated from a $U(0, 1)$ distribution. Random effects were generated from a standard normal distribution, i.e. α_i's $\overset{\text{iid}}{\sim} N(0, 1)$. Finally the error terms ε_i's have been independently generated from a χ_3^2 distribution re-centred at zero,

$$\varepsilon_{it} \overset{\text{iid}}{\sim} \chi_3^2 - 3.$$

In our simulation we have set $K = 3, m = 3, T = 5$ and $n = 100$. We have then defined prior distributions on each row of Γ, setting $\boldsymbol{c_L} = (c_{L_1}, \ldots, c_{L_K})$ and $\boldsymbol{c_U} = (c_{U_1}, \ldots, c_{U_K})$ to the true values, while $\tilde{\mathfrak{X}}$ has been set to be equal to

$$\tilde{\mathfrak{X}} = \begin{bmatrix} 1 & 5 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

The first row of $\tilde{\mathfrak{X}}$ has been chosen in order to obtain the maximum distance in $\mathscr{X}_1$.

Table 1 Posterior estimates of the parameters: mean and 95 % *HPD* (highest posterior density credibility intervals)

	Mean	95 % HPD	True value
$\beta_1(0.25)$	99.46	[97.91, 101.36]	98.21
$\beta_2(0.25)$	9.76	[9.33, 10.09]	10
$\beta_3(0.25)$	−100.89	[−102.55, −99.46]	−100
$\beta_1(0.50)$	100.5	[97.92, 103.03]	99.37
$\beta_2(0.50)$	9.83	[9.18, 10.50]	10
$\beta_3(0.50)$	−100.59	[−103, 42, −98.04]	−100
$\beta_1(0.75)$	101.53	[99.19, 103.82]	101.11
$\beta_2(0.75)$	9.96	[9.47, 10.49]	10
$\beta_3(0.75)$	−100.06	[−102.06, −97.92]	−100
σ^2	0.61	[0.12, 1.41]	1

Then, following (12), our final prior was:

$$\pi(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3) = \pi(\boldsymbol{\gamma}_3|\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2|\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_1)$$

$$\alpha_i|\sigma^2 \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\sigma^2 \sim \text{Inverse Gamma}(2, 1)$$

where the density of an Inverse Gamma(ψ, ρ) evaluated at x is proportional to $x^{-(\psi+1)} \exp(-\rho/x)$, and

$$\pi(\boldsymbol{\gamma}_j) \propto 1(\boldsymbol{\gamma}_j \in \Theta) g(\gamma_{j1}, \gamma_{jm})$$

with $\phi_L = 6$, $\phi_U = 100$. The parameters ν and d appearing in (8) have been set equal to 4 and $c_U - c_L$, respectively.

Chains of 200,000 iterations have been run with $\bar{p}^* = 0.3$, To reduce the autocorrelation effect, chains have been thinned by taking one draw every 40. Convergence for all parameters has been checked via the [7] diagnostic tools. Results are displayed in Table 1, and Figs. 1, 2, and 3. All parameters are accurately estimated and the MCMC chains have quickly reached convergence.

The results of simulations are encouraging, although some sort of bias is still present in the estimation of the regression coefficients. This is likely due to the presence of autocorrelation in the chains, and we are currently work on a more efficient version of the algorithm based on a Sequential Monte Carlo strategy [3].

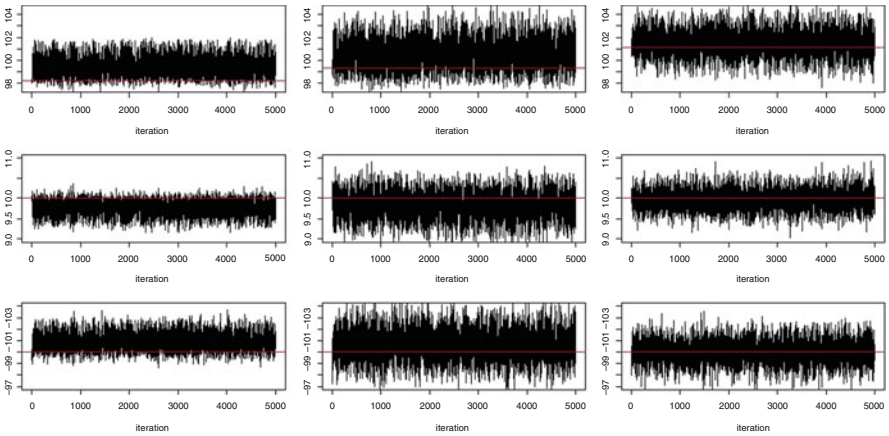

Fig. 1 Traceplots of the $\beta(\tau)$'s parameters in the MCMC run. Row i refers to the i-th β coefficient, $i = 1, 2, 3$. Column j refers to the j-th quartile. *Red lines* represent the true values of the parameters

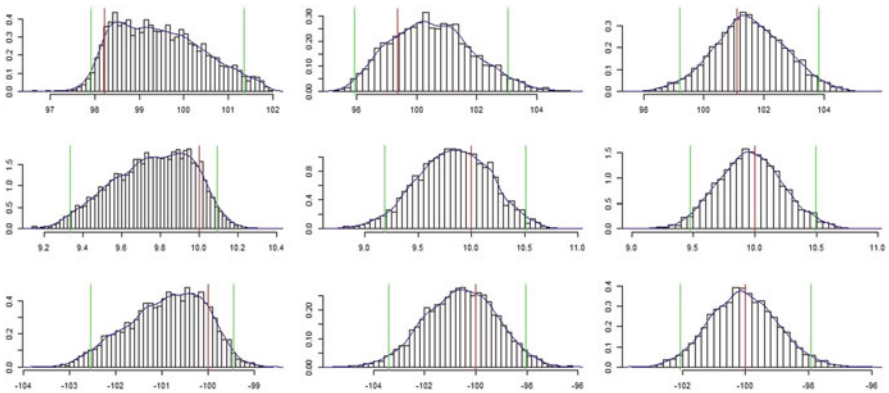

Fig. 2 Posterior densities of the $\beta(\tau)$'s parameters. Row i refers to the i-th β coefficient, $i = 1, 2, 3$. Column j refers to the j-th quartile. *Red lines* represent the true values of the parameters. *Green lines* represent the 95 % HPD intervals

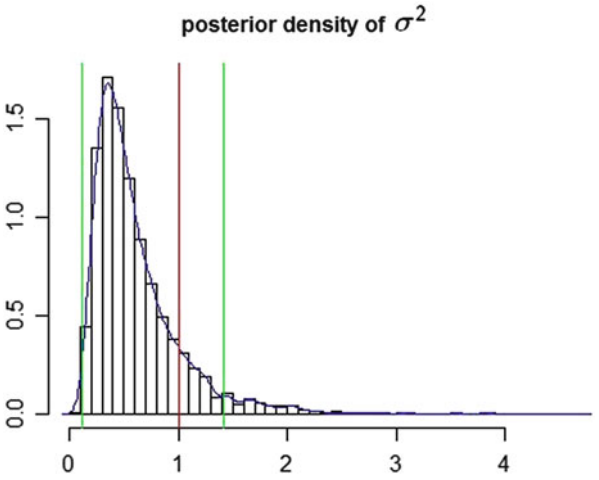

Fig. 3 Posterior histogram and approximated density from the MCMC run for σ^2; the *red line* represents the true value of the parameter; *green lines* indicate the 95 % HPD interval

References

1. Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**, 343–373 (2008)
2. Chernozhukov, V., Hong, H.: An MCMC approach to classical estimation. J. Econom. **115**, 293–346 (2003)
3. Doucet, A., Johansen, A.: Particle filtering and smoothing: fifteen years later. In: Crisan, D., Rozovsky, B. (eds.) Handbook of Nonlinear Filtering. Oxford University Press, Oxford (2011)
4. Dunson, D.B., Taylor, J.A.: Approximate Bayesian inference for quantiles. J. Nonparametr. Stat. **17**, 385–400 (2005)
5. Gelfand, A.E., Smith, A.F.M., Lee, T.M.: Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J. Am. Stat. Assoc. **87**, 523–532 (1992)
6. Geraci, M., Bottai, M.: Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics **8**, 140–154 (2007)
7. Geweke, J.: Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 4, pp. 169–193. Clarendon Press, Oxford (1992)
8. Hjort, N.L., Petrone, S.: Nonparametric quantile inference using Dirichlet Processes. In: Nair, V. (ed.) Advances in Statistical Modeling and Inference. Festschrift for Kjell Doksum, pp. 463–492. World Scientific, Singapore (2007)
9. Hjort, N.L., Walker, S.G.: Quantile pyramids for Bayesian nonparametrics. Ann. Statist. **37**, 105–131 (2009)
10. Hjort, N.L., Holmes, C., Muller, P., Walker, S.G. (eds.): Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 28. Cambridge University Press, Cambridge (2010)
11. Jeffreys, H.: Theory of Probability, 3rd edn. Clarendon Press, Oxford (1961)
12. Koenker, R.: Quantile regression for longitudinal data. J. Multivariate Anal. **91**, 74–89 (2004)
13. Koenker, R., Hallock, K.F.: Quantile regression. J. Econ. Perspect. **15**(4), 143–156 (2001)
14. Koenker, R., Machado, J.A.F.: Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. **94**, 1296–1310 (1999)
15. Kottas, A., Krnjajic, M.: Bayesian semiparametric modelling in quantile regression. Scand. J. Stat. **36**, 297–319 (2009)
16. Lancaster, T., Jun, S.J.: Bayesian quantile regression methods. J. Appl. Econom. **25**, 287–307 (2010)

17. Lavine, M.: On an approximate likelihood for quantiles. Biometrika **82**, 220–222 (1995)
18. Lazar, N.A.: Bayesian empirical likelihood. Biometrika **90**, 319–326 (2003)
19. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**, 110–120 (1997)
20. Taddy, M. A., Kottas, A.: A Bayesian nonparametric approach to inference for quantile regression. J. Bus. Econom. Stat. **28**, 357–369 (2010)
21. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
22. Yu, K., Moyeed, R.A.: Bayesian quantile regression. Stat. Probab. Lett. **54**, 437–447 (2001)
23. Yu, K., Van Kerm, P., Zhang, J.: Bayesian quantile regression: an application to the wage distribution in 1990's Britain. Sankhya **67**, 359–377 (2005)

Estimating Surfaces and Spatial Fields via Regression Models with Differential Regularization

Laura M. Sangalli

1 Introduction

This work briefly reviews progress to date on spatial regression with differential regularization. These are penalized regression models for the accurate estimation of surfaces and spatial fields, that have regularizing terms involving partial differential operators. Partial Differential Equations (PDEs) are commonly used to describe complex phenomena behavior in many fields of engineering and sciences, including Biosciences, Geosciences and Physical sciences. Here they are used to model the shape of the surface and the spatial variation of the problem under study.

In the simpler context of curve estimation and univariate smoothing problems, the idea of regularization with ordinary differential operators has already proved to be very effective and it is in general playing a central role in the functional data analysis literature. See, e.g., [23]. In the more complex case of surface estimation and spatial smoothing, a few methods have been introduced that use roughness penalties involving simple partial differential operators. A classical example is given by thin-plate-splines, while more recent proposals are offered for instance by Ramsay [22], Wood et al. [27], and Guillas and Lai [15]; see also the applications in [1, 11, 19]. Finally, although in a different framework, the use of simple forms of (stochastic) PDEs is also at the core of the Bayesian spatial models introduced by Lindgren et al. [18] and more generally of the larger literature on Bayesian inverse problems [25] and data assimilation in inverse problems [9].

Regression models with partial differential regularizations [3, 4, 8, 12, 24] merge advanced statistical methodology with numerical analysis techniques. Thanks to the interactions of these two areas, the proposed class of models have important

L.M. Sangalli (✉)
MOX - Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
e-mail: laura.sangalli@polimi.it

A.M. Paganoni, P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Contributions to Statistics,
DOI 10.1007/978-3-319-11149-0_13

advantages with respect to classical techniques used in spatial data analysis. Spatial regression with differential regularization is able to efficiently deal with data distributed over irregularly shaped domains with complicated geometries [24]. Moreover, it can comply with specific conditions at the boundaries of the problem domain [3, 24], which is fundamental in many applications to obtain meaningful estimates. The proposed models can also deal with data scattered over general bidimensional Riemannian manifold domains [8, 12]. Moreover, spatial regression with differential regularization has the capacity to incorporate problem-specific prior information about the spatial structure of the phenomenon under study [2–4], formalized in terms of a governing PDE. This also allows for a very flexible modeling of space variation, that accounts naturally for anisotropy and non-stationarity. Space-varying covariate information is included in the models via a semiparametric framework. The estimators have a penalized regression form, they are linear in the observed data values, and have good inferential properties. The use of advanced numerical analysis techniques, and specifically of finite elements, makes the models computationally very efficient. The method is implemented in both R [21] and Matlab.

This work gives a unified summary review of [3, 4, 8, 12, 24] and is organized as follows. Section 2 introduces the model in the simplest setting, characterizes the estimation problem (Sect. 2.2), illustrates how to compute the estimators using finite elements (Sect. 2.3) and reports some distributional properties of the estimators (Sect. 2.4). Section 3 shows how to include in the model prior information about the space variation of the phenomenon under study. Section 4 reviews the works on spatial regression over manifolds. The method is illustrated in various applied contexts in Sects. 2–4, including demographic data and medical imaging data. Finally, Sect. 5 outlines current research directions.

2 Regression Models with Partial Differential Regularization

Let $\{\mathbf{p}_i = (p_{i1}, p_{i2}); i = 1 \ldots, n\}$ be a set of n points in a bounded domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega \in C^2$. Let z_i be the value of a real-valued variable observed at location $\mathbf{p}_i$, and let $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})^t$ be a q-vector of covariates associated to observation z_i at $\mathbf{p}_i$.

Consider the semi-parametric generalized additive model

$$z_i = \mathbf{w}_i^t \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

where ϵ_i, $i = 1, \ldots, n$, are residuals or errors distributed independently of each other, with zero mean and variance σ^2. Vector $\boldsymbol{\beta} \in \mathbb{R}^q$ contains regression coefficients and the function $f : \Omega \to \mathbb{R}$ captures the spatial structure of the phenomenon.

In [24], generalizing the bivariate smoothing method introduced by [22], we porpose to estimate the vector of regression coefficients $\boldsymbol{\beta}$ and the surface or spatial

field f by minimizing the penalized sum-of-square-error functional

$$J_\lambda(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i^t \boldsymbol{\beta} - f(\mathbf{p}_i)\right)^2 + \lambda \int_\Omega \left(\Delta f(\mathbf{p})\right)^2 d\mathbf{p} \tag{2}$$

where λ is a positive smoothing parameter. Here Δf denotes the Laplacian of the function f. To define this partial differential operator let us consider a generic surface $f = f(\mathbf{p})$ defined on Ω. Denote the gradient of f by

$$\nabla f(\mathbf{p}) := \left(\frac{\partial f}{\partial p_1}(\mathbf{p}), \frac{\partial f}{\partial p_2}(\mathbf{p})\right)^t$$

where ^{t} is the transpose operator. Moreover, given a vector field $\mathbf{f} = \left(f_1(\mathbf{p}), f_2(\mathbf{p})\right)^t$ on Ω, where f_1 and f_2 are two surfaces on Ω, define the divergence of the vector field as

$$\operatorname{div} \mathbf{f}(\mathbf{p}) := \frac{\partial f_1}{\partial p_1}(\mathbf{p}) + \frac{\partial f_2}{\partial p_2}(\mathbf{p}).$$

Then, the Laplacian of the surface f is defined as

$$\Delta f(\mathbf{p}) := \operatorname{div} \nabla f(\mathbf{p}) = \frac{\partial^2 f}{\partial p_1^2}(\mathbf{p}) + \frac{\partial^2 f}{\partial p_2^2}(\mathbf{p}).$$

The Laplacian Δf provides a simple measure of the local curvature of the surface f defined on a planar domain Ω, and is invariant with respect to rigid transformations (rotations, translations and reflections) of the spatial coordinates of the domain. The use of the Laplace operator in the roughness penalty in (2) therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. This roughness penalty can be seen as a generalization of the penalty considered for one-dimensional smoothing splines, normally consisting in the L^2-norm of the second order derivative of the curve to be estimated. Likewise for one-dimensional splines, the higher the smoothing parameter λ, the more we are controlling the roughness of the field f, the smaller the smoothing parameter, the more we are allowing for local curvature of f.

The proposed model is able to efficiently handle data distributed over domains Ω having shapes with complicated geometries. This constitutes an important advantage with respect to classical methods for surface estimation, especially in all the applicative contexts where the shape of the problem domain is important for the behavior of the phenomenon under study. Classical methods for surface estimation, such as tensor product of unidimensional splines, thin-plate splines, bidimensional kernel smoothing, bidimensional wavelet-based smoothing and kriging, are in fact naturally defined on tensorized domains and cannot efficiently deal with more complex domains. Moreover, the method proposed can also comply with general

conditions at the boundary $\partial\Omega$ of the domain [3, 24], which in many applied problems is a crucial feature to obtain meaningful estimates. These boundary conditions, homogeneous or not, may involve the evaluation of the function and/or its normal derivative at the boundary, allowing for different behaviors of the surface at the boundary of the domain of interest.

As will be summarized in Sect. 3, instead of the Laplacian and other simple partial differential operators, the roughness penalty may also involve more complex partial differential operators. This model extension, developed in [2–4], is particularly interesting whenever prior knowledge is available on the phenomenon under study, coming for instance from the Physics, Chemistry, Biology or Mechanics of the problem at hand, that can be formalized in terms of a partial differential equation modeling the phenomenon under study. Moreover, it allows for a very flexible modeling of the space variation.

Another important result is that the models can be extended to handle data distributed over general bidimensional Riemannian manifold domains. In such a case the differential operator considered in the roughness penalty is computed over the manifold domain. This generalization, developed in [8, 12], is reviewed in Sect. 4.

2.1 Modeling Data Distributed over Irregular Domains and Complying with General Conditions at the Domain Boundaries

To illustrate the issue of spatial smoothing over irregularly shaped domains and with boundary conditions, consider the problem of estimating population density over the Island of Montréal (QC, Canada), starting from census data (1996 Canadian census). Figure 1 displays census tract locations over the Island of Montréal; population density and other census information are available at each census tract, together with a binary covariate indicating whether a tract is predominantly residential or industrial/commercial. The figure highlights two parts of the island without data: the airport and rail yards in the south and an industrial park with an oil refinery tank farm in the north-east; these two areas are not part of the domain of interest when studying population density, since people cannot live there. Notice that census quantities can vary sharply across these uninhabited parts of the city; for instance, in the south of the industrial park there is a densely populated area with medium-low income, but north-east and west of it are wealthy neighborhoods, with low population density to the north-east, and high population density to the west. Hence, whilst it seems reasonable to assume that population density and other census quantities feature a smooth spatial variation over the inhabited parts of the island, there is no reason to assume smooth spatial variation across uninhabited areas. The figure also shows the island coasts as boundaries of the domain of interest, as it does not make sense to smooth population density into the rivers. Those parts of the boundary that are

Fig. 1 Island of Montréal census data. *Dots* indicates the centroids of census enumeration areas, for which population density and other census information are available. The two parts of the island where there are no data, encircled by *yellow lines*, are areas where people cannot live (the airport and rail yards in the south and an industrial park with an oil refinery tank farm in the north-east). The island boundary is also outlined in *yellow* and *red*, with *red* sections indicating the harbor and two public parks. Adapted from [24]

highlighted in red correspond respectively to the harbor, in the east shore, and to two public parks, in the south-west and north-west shore; no people live by the river banks in these boundary intervals. We thus want to study population density, taking into account covariate information, being careful not to artificially link data across areas where people cannot live, and also efficiently including prior information concerning those stretches of coast where the population density should drop to zero.

2.2 Characterization of the Estimators

Denote by W the $n \times q$ matrix whose ith row is given by $\mathbf{w}_i^t$, the vector of q covariates associated with observation z_i at $\mathbf{p}_i$, and assume that W has full rank. Let P be the matrix that projects orthogonally on the subspace of $\mathbb{R}^n$ generated by the columns of W, i.e., $P := W(W^tW)^{-1}W^t$, and let $Q = I - P$, where I is the identity matrix. Furthermore, set $\mathbf{z} := (z_1, \ldots, z_n)^t$ and, for a given function f on Ω denote by $\mathbf{f}_n$ the vector of evaluations of f at the n data locations, i.e., $\mathbf{f}_n := (f(\mathbf{p}_1), \ldots, f(\mathbf{p}_n))^t$.

Let $H^m(\Omega)$ be the Hilbert space of all functions which belong to $L^2(\Omega)$ along with all their distributional derivatives up to the order m. The penalized sum-of-square-error functional (2) is well defined for $\boldsymbol{\beta} \in \mathbb{R}^q$ and $f \in H^2(\Omega)$. Furthermore, imposing appropriate boundary conditions on f ensures that the estimation problem has a unique solution. Three classic boundary conditions are Dirichlet, Neumann and Robin conditions. The Dirichlet condition controls the value of f at the boundary, i.e., $f|_{\partial\Omega} = \gamma_D$, the Neumann condition controls instead the flow across the boundary, concerning the value of the normal derivative

of f at the boundary, i.e., $\partial_{\boldsymbol{\nu}} f|_{\partial\Omega} = \gamma_N$, where $\boldsymbol{\nu}$ is the outward unit normal vector to $\partial\Omega$, while the Robin condition involves linear combinations of the above conditions, i.e., $\left(\partial_{\boldsymbol{\nu}} f + \alpha f\right)\big|_{\partial\Omega} = \gamma_R$. The functions γ_D, γ_N and γ_R have to satisfy some regularity conditions in order to obtain a well defined functional $J(f)$; when these functions coincide with null functions, the condition is said homogeneous. Moreover, it is possible to impose different boundary conditions on different portions of the boundary, forming a partition of $\partial\Omega$, as will be illustrated for instance in the application to Montréal census data. For simplicity of exposition, in the following we consider homogeneous Neumann (or homogeneous Dirichlet) boundary conditions, and $V(\Omega)$ will denote the subspace of $H^2(\Omega)$ characterized by the chosen boundary conditions. The interested reader is referred to [3, 4, 24] for the case of general boundary conditions.

The following Proposition characterizes the estimators (see [24]).

Proposition 1 *There exists a unique pair of estimators* $(\hat{\boldsymbol{\beta}} \in \mathbb{R}^q, \hat{f} \in V(\Omega))$ *which minimize (1). Moreover,*

- $\hat{\boldsymbol{\beta}} = (W^t W)^{-1} W^t (\mathbf{z} - \hat{\mathbf{f}}_n)$;
- $\hat{f}$ *satisfies*

$$\mathbf{h}_n^t \, Q \, \hat{\mathbf{f}}_n + \lambda \int_\Omega (\Delta h)(\Delta \hat{f}) = \mathbf{h}_n^t \, Q \, \mathbf{z} \tag{3}$$

for every $h \in V(\Omega)$.

Problem (3) is an infinite dimensional problem and cannot be solved analytically. We thus solve the problem numerically, reducing it to a finite dimensional one. Since (3) is a fourth order problem, a convenient way to tackle it is to first rewrite it as a coupled system of second order problems, by introducing an auxiliary variable. The latter problem is hence reformulated in order to involve only first order derivatives, i.e., in a suitable $H^1(\Omega)$ subspace. This weak or variational formulation can be obtained by integrating the differential equations against a test function and integrating by parts the second order terms. This problem reformulation is particularly well suited to be solved numerically, as H^1 spaces can be approximated by convenient finite dimensional spaces, and specifically by standard finite element spaces. Finite element analysis has been mainly developed and used in engineering applications, to solve partial differential equations; see, e.g., [20]. In the finite element space, solving the estimation problem reduces to solving a linear system. Next section describes the finite element spaces we shall be using.

2.3 Finite Element Solution to the Estimation Problem

To construct a finite element space, we start by partitioning the domain Ω of interest into small subdomains. Convenient domain partitions are given for instance by triangular meshes; Fig. 3, left panel, shows for example a triangulation of the domain

of interest for the Island of Montréal data. In particular, we consider a regular triangulation $\mathscr{T}$ of Ω, where adjacent triangles share either a vertex or a complete edge. Domain Ω is hence approximated by domain $\Omega_{\mathscr{T}}$ consisting of the union of all triangles, so that the boundary $\partial\Omega$ of Ω is approximated by a polygon (or more polygons, in the case for instance of domains with interior holes). It is assumed, therefore, that the number and density of triangles in $\mathscr{T}$, with the associated finite element basis, are sufficient to adequately describe the data. The triangulation points in $\mathscr{T}$ may or may not coincide with the data locations $\mathbf{p}_i$. In any case, for the inferential properties of the estimators, it is convenient to consider triangulations that are finer where there are more data points, and coarser where there are fewer data points.

Starting from the triangulation $\mathscr{T}$, we can introduce a locally supported basis that spans the space of continuous surfaces on $\Omega_{\mathscr{T}}$, coinciding with polynomials of a given order over each triangle of $\mathscr{T}$. The resulting finite element space, denoted by $H^1_{\mathscr{T}}(\Omega)$, provides an approximation of the infinite dimensional space $H^1(\Omega)$. Linear finite elements are for instance obtained considering a basis system where each basis function ψ_j is associated with a triangle vertex ξ_j, $j = 1, \ldots, N$, in the triangulation $\mathscr{T}$. This basis function ψ_j is a piecewise linear polynomial which takes the value one at the vertex ξ_j and the value zero on all the other vertices of the mesh, i.e., $\psi_j(\xi_l) = \delta_{jl}$, where δ_{jl} denotes the Kronecker delta symbol. Figure 2 shows an example of such linear finite element basis function on a planar mesh, highlighting the locally supported nature of the basis.

Now, let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_N)^t$ be the column vector collecting the N basis functions associated with the N vertices ξ_j, $j = 1, \ldots, N$. Then, each function h in the finite element space $H^1_{\mathscr{T}}$ can be represented as an expansion in terms of the basis function $\psi_1, \ldots, \psi_N$. In particular,

$$h(\cdot) = \sum_{j=1}^{N} h(\xi_j)\psi_j(\cdot) = \mathbf{h}^t \boldsymbol{\psi}(\cdot) \tag{4}$$

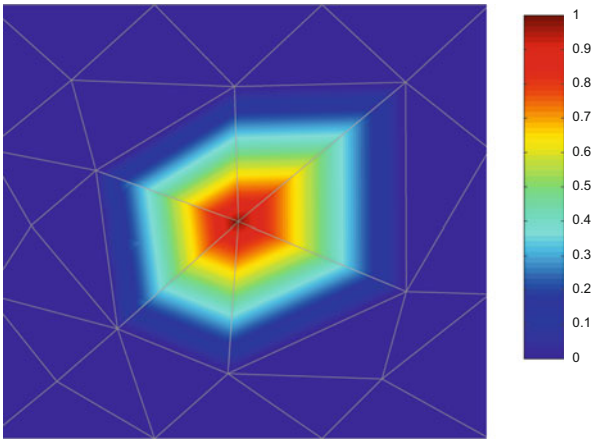

Fig. 2 Example of a linear finite element basis on a planar triangulation

where

$$\mathbf{h} = (h(\xi_1), \dots, h(\xi_N))^t \tag{5}$$

is the column vector of the evaluations of h at the N mesh nodes. Each function $h \in H^1_{\mathscr{T}}$ is thus uniquely identified by its evaluations $\mathbf{h}$ on the mesh nodes.

Let Ψ be the $n \times N$ matrix of the evaluations of the N basis at the n data locations $\mathbf{p}_1, \cdots, \mathbf{p}_n$,

$$\Psi = \begin{bmatrix} \boldsymbol{\psi}^t(\mathbf{p}_1) \\ \vdots \\ \boldsymbol{\psi}^t(\mathbf{p}_n) \end{bmatrix}$$

and consider the $N \times N$ matrices

$$R_0 := \int_{\Omega_{\mathscr{T}}} (\boldsymbol{\psi}\,\boldsymbol{\psi}^t) \qquad R_1 := \int_{\Omega_{\mathscr{T}}} \nabla\boldsymbol{\psi}^t \nabla\boldsymbol{\psi}.$$

Corollary 1 shows that, once recast in the finite element space, solving the estimation problem reduces to solving a linear system.

Corollary 1 *There exists a unique pair of estimators* $(\hat{\boldsymbol{\beta}} \in \mathbb{R}^q, \hat{f} \in H^1_{\mathscr{T}}(\Omega))$ *which solve the discrete counterpart of the estimation problem. Moreover,*

- $\hat{\boldsymbol{\beta}} = (W^t W)^{-1} W^t (\mathbf{z} - \hat{\mathbf{f}}_n)$;
- $\hat{f} = \hat{\mathbf{f}}^t \boldsymbol{\psi}$, *with* $\hat{\mathbf{f}}$ *satisfying*

$$\begin{bmatrix} -\Psi^t Q \Psi & \lambda R_1 \\ \lambda R_1 & \lambda R_0 \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\Psi^t Q\,\mathbf{z} \\ \mathbf{0} \end{bmatrix}. \tag{6}$$

Solving the linear system (6) is fast. In fact, although the system is typically large, being of order $2N$, it is highly sparse because the matrices R_0 and R_1 are highly sparse, since the cross-products of nodal basis functions and of their partial derivatives are mostly zero. As an example, for the Island of Montréal census data, we used 626 nodes and only about 1 % of the entries of R_0 and 0.2 % of the entries of R_1 were non-zero.

2.4 Properties of the Estimators

Corollary 1 highlights that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{f}$ are linear in the observed data values $\mathbf{z}$. Moreover $\hat{f}$ has a penalized regression form, being identified by the vector

$$\hat{\mathbf{f}} = (\Psi^t Q \Psi + \lambda R_1 R_0^{-1} R_1)^{-1} \Psi^t Q\,\mathbf{z}$$

where the positive definite matrix $R_1 R_0^{-1} R_1$ represents the discretization of the penalty term in (2). Notice that, thanks to the variational formulation of the estimation problem, this penalization matrix only involves the computation of first order derivatives.

Denote by S_f the $n \times n$ matrix

$$S_f = \Psi(\Psi^t Q \Psi + \lambda R_1 R_0^{-1} R_1)^{-1} \Psi^t Q.$$

Using this notation,

$$\hat{\mathbf{f}}_n = S_f \, \mathbf{z}$$
$$\hat{\boldsymbol{\beta}} = (W^t W)^{-1} W^t \{I - S_f\} \, \mathbf{z}.$$

Some distributional properties of the estimators are straightforward to derive and classic inferential tools can be obtained. Recalling that $E[\mathbf{z}] = W\boldsymbol{\beta} + \mathbf{f}_n$ and $Var(\mathbf{z}) = \sigma^2 I$, and exploiting the properties of the matrices involved (e.g., Q is symmetric and idempotent, $Q\,W = \mathbf{O}_{n\times n}$, $Q\,W(W^tW)^{-1} = (W^tW)^{-1}W^t\,Q = \mathbf{O}_{n\times n}$), with a few simplifications we obtain the means and variances of the estimators $\hat{\mathbf{f}}_n$ and $\hat{\boldsymbol{\beta}}$:

$$E[\hat{\mathbf{f}}_n] = S_f \, \mathbf{f}_n \tag{7}$$
$$Var(\hat{\mathbf{f}}_n) = \sigma^2 \, S_f \, S_f^t \tag{8}$$

and

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (W^t W)^{-1} W^t \big(I - S_f\big)\mathbf{f}_n \tag{9}$$
$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (W^t W)^{-1} + \sigma^2 (W^t W)^{-1} W^t \{S_f \, S_f^t\} W (W^t W)^{-1}. \tag{10}$$

Denote now by S the smoothing matrix $S := P + Q\,S_f$ and consider the vector $\hat{\mathbf{z}}$ of fitted values at the n data locations

$$\hat{\mathbf{z}} = W\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}}_n = S\,\mathbf{z}.$$

The fitted values $\hat{\mathbf{z}}$ are thus obtained from observations $\mathbf{z}$ via application of the linear operator S, independent of $\mathbf{z}$. A commonly used measure of the equivalent degrees of freedom for linear estimators is given by the trace of the smoothing matrix; see, e.g., [6], who first introduced this notion. The equivalent degrees of freedom of the estimator $\hat{\mathbf{z}}$ are thus given by

$$tr(S) = q + tr(S_f),$$

and coincide with the sum of the q degrees of freedom of the parametric part of the model (q being the number of covariates considered) and of the $tr(S_f)$ equivalent

degrees of freedom of the non-parametric part of the model. We can now estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n - tr(S)} \left(\mathbf{z} - \hat{\mathbf{z}}\right)^t \left(\mathbf{z} - \hat{\mathbf{z}}\right).$$

This estimate, together with expressions (10) and (8), may be used to obtain approximate confidence intervals for $\boldsymbol{\beta}$ and approximate confidence bands for f. Furthermore, the value of the smoothing parameter λ may be selected by Generalized-Cross-Validation, see, e.g., [23] and references therein:

$$GCV(\lambda) = \frac{1}{n\left(1 - tr(S)/n\right)^2} \left(\mathbf{z} - \hat{\mathbf{z}}\right)^t \left(\mathbf{z} - \hat{\mathbf{z}}\right).$$

Finally, the value predicted for a new observation, at point $\mathbf{p}_{n+1}$ and with covariates $\mathbf{w}_{n+1}$, is given by

$$\hat{z}_{n+1} = \mathbf{w}_{n+1}^t \hat{\boldsymbol{\beta}} + \hat{f}(\mathbf{p}_{n+1}) = \mathbf{w}_{n+1}^t \hat{\boldsymbol{\beta}} + \hat{\mathbf{f}}^t \boldsymbol{\psi}(\mathbf{p}_{n+1}),$$

whose mean and variance can be obtained from expressions above; correspondingly, approximate prediction intervals may be also derived.

The expressions (7) and (9) highlight that the estimators are biased. In particular, there are two sources of bias in the proposed model. The first source is the discretization and it is common to any model employing a basis expansion. This source of bias disappears as the number n of observations increases (in the sense of infill asymptotic) if meanwhile the mesh is correspondingly refined. The second source is the penalty term, and this is typical of regression models involving a roughness penalty: unless the true function f is such that it annihilates the penalty term, this term will of course induce a bias in the estimate. As shown in [3], this source of bias disappears as n increases, if the smoothing parameter λ decreases with n. This appears to be a natural request since having more observations decreases the need to impose a regularization. In all the simulations we carried out, the bias always appeared negligible.

2.5 *Applied Illustrative Problem: Island of Montréal Census Data*

Figure 3, left panel, displays a triangulation of the domain of interest for Montréal census data application. We shall estimate population density, measured as 1,000 inhabitants per km^2, using as covariate the binary variable that indicates whether a tract is predominantly residential (1) or commercial/industrial (0). We use here mixed boundary conditions: homogeneous Dirichlet along the stretches of coast corresponding to the harbor and the public parks, implying that the estimate of

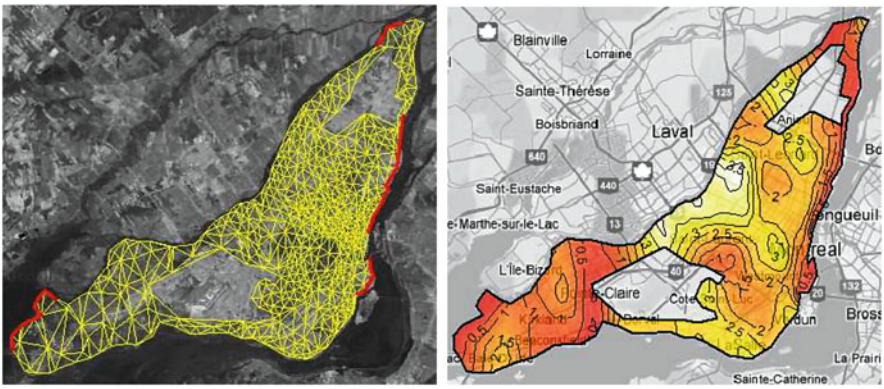

Fig. 3 *Left*: constrained Delaunay triangulation of the Island of Montréal. *Right*: estimate of spatial structure for population density over the Island of Montréal. Adapted from [24]

population density should drop to zero along these stretches, and homogeneous Neumann along the remaining shores, meaning there is no immigration-emigration across shores. Figure 3, right panel, shows the estimated spatial structure of population density. The estimate complies with the specified boundary conditions and does not artificially link data points on either side of the uninhabited parts; see for instance the densely populated areas just in the south and in the west of the industrial park, with respect to the low population density neighborhood north-east of it. The β coefficient that corresponds to the binary covariate is estimated to be 1,300; this means that census tracts that are predominantly residential are in average expected to have 1,300 more inhabitants per km^2, with respect to those predominantly commercial; the approximate 95 % confidence interval is given by [0.755; 1.845].

3 Incorporating Prior Knowledge About the Phenomenon Under Study

In this section we briefly summarize the generalization of the models developed in [2–4]. Suppose that prior knowledge is available on the phenomenon under study, that can be formalized through a partial differential equation $Lf = u$ modeling the phenomenon (where $u \in L^2(\Omega)$ is some forcing term). Partial differential models are indeed commonly used to describe complex phenomena behaviors in many fields of engineering and sciences. The prior knowledge, formalized in the PDE, is thus incorporated into the statistical model, looking for estimates of $\boldsymbol{\beta}$ and f that minimize the functional

$$J_\lambda(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i^t \boldsymbol{\beta} - f(\mathbf{p}_i)\right)^2 + \lambda \int_\Omega \left(Lf(\mathbf{p}) - u(\mathbf{p})\right)^2 d\mathbf{p} \qquad (11)$$

with respect to $f \in V$. The penalized error functional hence trades off a data fitting criterion, the sum-of-square-error, and a model fitting criterion that penalizes departures from a PDE problem-specific description of the phenomenon. The proposed method can be seen as a regularized least square analogous to the Bayesian inverse problems presented, e.g., in [25]. In particular, the least square term in $J(f)$ corresponds to a log-likelihood for Gaussian errors, while the regularizing term effectively translates the prior knowledge on the surface. With respect to [25], besides the different model framework and estimation approaches, we also deal with a larger class of operators, including also non-stationary (i.e., spatially inhomogeneous) and anisotropic diffusion, transport and reaction coefficients. This also allows for a very flexible modeling of space variation, that accounts naturally for anisotropy and non-stationarity (space inhomogeneity).

In particular, in [3, 4] we consider phenomena that are well described in terms of linear second order elliptic operators L and forcing term $u \in L^2(\Omega)$ that can be either the null function $u = 0$, homogeneous case, or $u \neq 0$, non-homogeneous case. The operator L is a general differential operator that can, for instance, include second, first and zero order differential operators. Consider a symmetric and positive definite matrix $\mathbf{K} = \{\mathbf{K}_{ij}\} \in \mathbb{R}^{2\times 2}$, named diffusion tensor, a vector $\mathbf{b} = \{\mathbf{b}_j\} \in \mathbb{R}^2$, named transport vector, and a positive scalar $c \in \mathbb{R}^+$, named reaction term. Then, the operator can include: second order differential operators as the divergence of the gradient, i.e.,

$$\operatorname{div}(\mathbf{K}\nabla f) = \frac{\partial}{\partial p_1}\left(\mathbf{K}_{11}\frac{\partial f}{\partial p_1} + \mathbf{K}_{12}\frac{\partial f}{\partial p_2}\right) + \frac{\partial}{\partial p_2}\left(\mathbf{K}_{21}\frac{\partial f}{\partial p_1} + \mathbf{K}_{22}\frac{\partial f}{\partial p_2}\right),$$

first order differential operators as the gradient, i.e.,

$$\mathbf{b}\cdot\nabla f = \mathbf{b}_1\frac{\partial f}{\partial p_1} + \mathbf{b}_2\frac{\partial f}{\partial p_2},$$

and also zero order operators, i.e., cf. The general form that we consider is

$$Lf = -\operatorname{div}(\mathbf{K}\nabla f) + \mathbf{b}\cdot\nabla f + cf. \tag{12}$$

Moreover, the parameters of the differential operator L can be space-varying on Ω; i.e., $\mathbf{K} = \mathbf{K}(p_1, p_2)$, $\mathbf{b} = \mathbf{b}(p_1, p_2)$ and $c = c(p_1, p_2)$. The three terms that compose the general second order operator (12) induce an anisotropic and non-stationary smoothing, providing different regularizing effects. The diffusion term $-\operatorname{div}(\mathbf{K}\nabla f)$ induces anisotropic and non-stationary smoothing in all directions; the transport term $\mathbf{b}\cdot\nabla f$ induces a non-stationary smoothing only in the direction specified by the transport vector $\mathbf{b}$. Finally, the reaction term cf has instead a non-stationary shrinkage effect, since penalization of the L^2 norm of f induces a shrinkage of the surface to zero. Setting $\mathbf{K} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$, $c = 0$ and u equal to the null function we obtain the special case described in Sect. 2, where the penalization of the Laplacian Δf induces an isotropic and stationary smoothing.

In [2–4] the estimation problem is shown to be well defined. Its discretization with the finite element space follows the same lines described in Sect. 2.3, with the important difference that the matrix R_1 is now defined as

$$R_1 = \int_{\Omega_{\mathscr{T}}} \left(\nabla \boldsymbol{\psi}^t \mathbf{K} \nabla \boldsymbol{\psi} + \nabla \boldsymbol{\psi}^t \mathbf{b} \boldsymbol{\psi} + c\, \boldsymbol{\psi} \boldsymbol{\psi}^t \right).$$

This change is due to the different penalty in (11) with respect to (2). The matrix R_1 is in fact used in the discretization of the penalty term. In the case where the considered forcing term $u \in L^2(\Omega)$ is not homogeneous ($u \neq 0$), the vector $\mathbf{0}$ in the right hand side of (6) is replaced by the discretization $\mathbf{u} = (u(\xi_1), \ldots, u(\xi_N))^t$ of the forcing term. Moreover, when the forcing term is homogeneous the estimator properties are as in Sect. 2.4; otherwise, additional terms need instead to be considered, but the estimators remain linear in the observed data values, and their properties follows along the lines described in Sect. 2.4.

In [3] we study in detail the numerical convergence properties of the Finite Element approximation to the estimation problem. In addition, the case of areal data is considered in [3, 4].

3.1 Applied Illustrative Problem: Blood-Flow Velocity Field Estimation

The motivating applied problem driving the generalization to more complex penalty terms concerns the estimation of the blood-flow velocity field on a cross-section of the common carotid artery, using data provided by echo-color doppler acquisitions. This applied problem arises within the research project MAthematichs for CARotid ENdarterectomy@MOX (MACAREN@MOX), which aims at studying atherosclerosis pathogenesis. Carotid Echo-Color Doppler (ECD) is a medical imaging procedure that uses reflected ultrasound waves to create images of an artery and to measure the velocity of blood cells in some locations within the artery. Specifically, the ECD data measure the velocity of blood within beams located on the considered artery cross-section; see Fig. 4, top left panel. For this study, during the ECD scan, 7 beams are considered, located in the cross-shaped pattern shown in Fig. 4, bottom left panel.

In this applied problem we have prior knowledge on the phenomenon under study that could be exploited to derive accurate physiological estimates. There is in fact a vast literature devoted to the study of fluid dynamics and hemodynamics; see for example [13] and references therein. This prior information concerns both the shape of the field, which can be conveniently described via a linear second order elliptic PDE, and the conditions at the boundary of the problem domain, i.e., specifically, at the wall of the carotid cross-section. The proposed method efficiently uses the prior information on the phenomenon under study and gives a realistic and physiological estimate of the blood flow velocity field, which is not affected by the pattern of

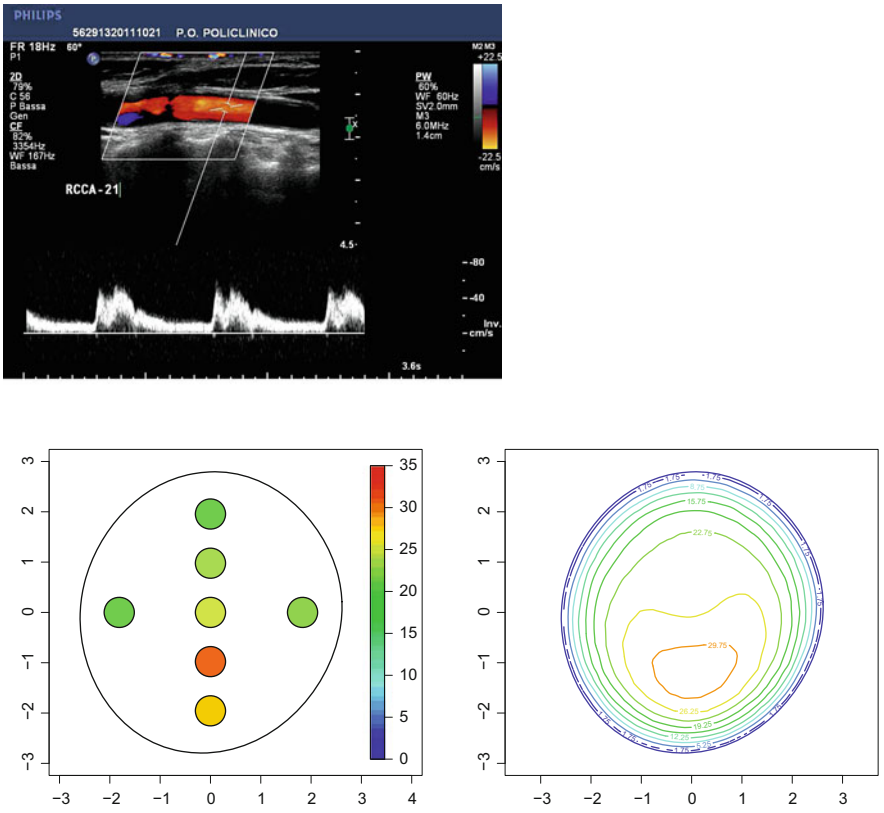

Fig. 4 *Top left*: ECD image corresponding to the central point of the carotid section located 2 cm before the carotid bifurcation. *Bottom left*: MRI reconstruction of the cross-section of the carotid artery located 2 cm before the bifurcation; cross-shaped pattern of observations with each beam colored according to the mean blood-velocity measured in the beam at systolic peak time. *Bottom right*: estimate of the blood-flow velocity field in the carotid section. Adapted from [4]

the observations; see Fig. 4, bottom right panel. Moreover the estimates accurately highlight important features of the blood flow, such as eccentricity, asymmetry and reversion of the fluxes, that are of interest to physicians, in order to understand how the local hemodynamics influences atherosclerosis pathogenesis. See [4].

4 Modeling Data over Manifold Domains

In [12], the models are extended to non-planar domains. Few methods are available in literature to deal with data on non-planar domains (see, e.g., [5, 7, 14, 16–18, 26]); to the best of our knowledge, none of these methods is currently devised to

handle the data structures we are considering, with variables of interest together with space-varying covariates, both distributed over general bi-dimensional Riemannian manifolds.

Consider n fixed data locations $\{\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i}) : i = 1, \ldots, n\}$ lying on a general bi-dimensional Riemannian manifold Γ, and let the variable of interest z_i and covariates $\mathbf{w}_i$ be observed at location $\mathbf{x}_i$, for $i = 1, \ldots, n$. Likewise in (1), assume the model

$$z_i = \mathbf{w}_i^t \boldsymbol{\beta} + f(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{13}$$

where the spatial field f is now defined on the manifold Γ, $f : \Gamma \to \mathbb{R}$. In analogy to (2), the vector of regression coefficients $\boldsymbol{\beta}$ and the surface or spatial field f are estimated by minimizing the penalized sum-of-square-error functional

$$J_{\Gamma,\lambda}(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i^t \boldsymbol{\beta} - f(\mathbf{x}_i)\right)^2 + \lambda \int_{\Gamma} \left(\Delta_{\Gamma} f(\mathbf{x})\right)^2 \, d\mathbf{x} \tag{14}$$

where $\Delta_{\Gamma} f$ denotes the Laplace-Beltrami operator, a generalization of the standard Laplacian to functions $f = f(x_1, x_2, x_3)$, defined on a non-planar domain Γ, with $(x_1, x_2, x_3) \in \Gamma$. The definition of the Laplace-Beltrami operator requires the computation of the gradient operator ∇_{Γ} and of the divergence operator div_{Γ} over the non planar domain; see, e.g., [10]. The Laplace-Beltrami operator of f is then defined as

$$\Delta_{\Gamma} f(\mathbf{x}) = \mathrm{div}_{\Gamma} \nabla_{\Gamma} f(\mathbf{x}).$$

Similar to the standard Laplacian, the Laplace-Beltrami operator provides a simple measure of the local curvature of the function f, as defined on the curved domain Γ. Likewise to the standard Laplacian, the Laplace-Beltrami is invariant with respect to rigid transformations (rotations, translations and reflections) of the spatial coordinates of the non-planar domain. Hence, the employment of the Laplace-Beltrami operator as a roughness penalty ensures that the concept of smoothness does not depend on the orientation of the coordinate system or on the orientation of the domain Γ itself.

In [12] the estimation problem (14) is recast over a planar domain, via a conformal reparametrization of the non-planar domain Γ. The reparametrization is obtained by a continuously differentiable map

$$\begin{aligned} X : \Omega &\to \Gamma \\ \mathbf{p} = (p_1, p_2) &\mapsto \mathbf{x} = (x_1, x_2, x_3) \end{aligned} \tag{15}$$

where Ω is an open, convex and bounded set in $\mathbb{R}^2$ and the boundary of Ω, denoted $\partial\Omega$, is piecewise C^{∞}. The map X essentially provides a change of variable between the planar coordinates $\mathbf{p} = (p_1, p_2)$ and the non-planar coordinates $\mathbf{x} = (x_1, x_2, x_3)$,

and is unique up to dilations, rotations and translations. Consider the first order partial derivatives of X with respect to the planar coordinates p_1 and p_2, $\frac{\partial X}{\partial p_1}(\mathbf{p})$ and $\frac{\partial X}{\partial p_2}(\mathbf{p})$, that are column vectors in $\mathbb{R}^3$. We denote by $\langle \cdot, \cdot \rangle$ the Euclidean scalar product of two vectors and by $\| \cdot \|$ the corresponding norm. Consider the (space-dependent) metric tensor

$$G(\mathbf{p}) := \nabla X(\mathbf{p})' \nabla X(\mathbf{p}) = \begin{pmatrix} \left\| \frac{\partial X}{\partial p_1}(\mathbf{p}) \right\|^2 & \langle \frac{\partial X}{\partial p_1}(\mathbf{p}), \frac{\partial X}{\partial p_2}(\mathbf{p}) \rangle \\ \langle \frac{\partial X}{\partial p_1}(\mathbf{p}), \frac{\partial X}{\partial p_2}(\mathbf{p}) \rangle & \left\| \frac{\partial X}{\partial p_2}(\mathbf{p}) \right\|^2 \end{pmatrix}.$$

Let $\mathscr{W}(\mathbf{p}) := \sqrt{det(G(\mathbf{p}))}$, and define the matrix $\mathbf{K}(\mathbf{p}) = \mathscr{W}(\mathbf{p})\, G^{-1}(\mathbf{p})$, where $G^{-1}(\mathbf{p})$ denotes the inverse of $G(\mathbf{p})$. Then for $f \circ X \in \mathscr{C}^2(\Omega)$, the Laplace-Beltrami operator be re-expressed in terms of the planar coordinates $\mathbf{p}$ as

$$\Delta_\Gamma f(\mathbf{x}) = \frac{1}{\mathscr{W}(\mathbf{p})} \mathrm{div}(\mathbf{K}(\mathbf{p}) \nabla f(X(\mathbf{p}))) \tag{16}$$

where div and ∇ denote the standard divergence and gradient operators over planar domains, defined in Sect. 2. Thus, by considering the map X and the corresponding planar parametrization (16) for the Laplace-Beltrami operator, after setting $\mathbf{u}_i = X^{-1}(\mathbf{x}_i)$, we can reformulate the estimation problem (14) over the manifold Γ as an equivalent problem over the planar domain Ω as follows: find $\boldsymbol{\beta}$ and the function $f \circ X$, defined on Ω, that minimizes

$$\begin{aligned} J_{\Omega,\lambda}(f \circ X) = & \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i' \boldsymbol{\beta} - f(X(\mathbf{p}_i))\right)^2 \\ & + \lambda \int_\Omega \frac{1}{\mathscr{W}(\mathbf{p})} \Big(\mathrm{div}\big(\mathbf{K}(\mathbf{p}) \nabla f(X(\mathbf{p}))\big) \Big)^2 d\Omega. \end{aligned} \tag{17}$$

As previously remarked, the map X and its corresponding planar domain Ω are unique only up to dilations, rotations and translations. However, the terms $\mathbf{K}(\mathbf{u})$ and $\mathscr{W}(\mathbf{u})$, that account for the change of variable from the original non-planar coordinates to the planar ones, adjust for each considered map X and for the corresponding planar domain Ω. Consequently this leads to different planar parameterizations of the function f as well as of the estimation problem (17), all equivalent to problem (14) on the original manifold Γ.

The conformal reparametrization is computed resorting to non-planar finite elements. These are defined likewise planar finite elements, but over non-planar meshes. See [12] for details. The discretization of the conformal map via non-planar finite elements provides a planar mesh for the corresponding planar domain Ω, and the discretization of the terms $\mathscr{W}$ and $\mathbf{K}$ in (17). These terms, together with the planar mesh, are then used for the solution of the equivalent estimation

problem (17) over the planar domain. The discretization of problem (17) with planar finite elements follows the same lines described in Sect. 2.3, with the important difference that the matrices R_0 and R_1 are now defined as

$$R_0 := \int_{\Omega_{\mathscr{T}}} \mathscr{W}(\boldsymbol{\psi}\boldsymbol{\psi}^t) \qquad R_1 := \int_{\Omega_{\mathscr{T}}} \nabla\boldsymbol{\psi}^t \mathbf{K} \nabla\boldsymbol{\psi}.$$

This change is due to the different penalty in (17) with respect to (2), with the terms $\mathscr{W}(\mathbf{p})$ and $\mathbf{K}(\mathbf{p})$ accounting for the change of variable from the original non-planar coordinates to the planar ones. The estimator properties are as in Sect. 2.4.

4.1 *Applied Illustrative Problems: Modeling Hemodynamical Stresses on Cerebral Arteries and Studying Cortical Surface Data*

The extension to manifold domains has fascinating fields of applications. In [12] the models are applied to the study of hemodynamical forces exerted by blood flow on the wall of cerebral arteries affected by aneurysms; see Fig. 5. These data come from three-dimensional angiographies and computational fluid dynamics simulations, and belong to the AneuRisk project, a scientific endeavor that aimed at investigating the role of vessel morphology, blood fluid dynamics and biomechanical properties of the vascular wall, on the pathogenesis of cerebral aneurysms; see http://mox.polimi.it/it/progetti/aneurisk/. In [8] we discuss instead an application in the neurosciences by analyzing cerebral cortex thickness data; see Fig. 6.

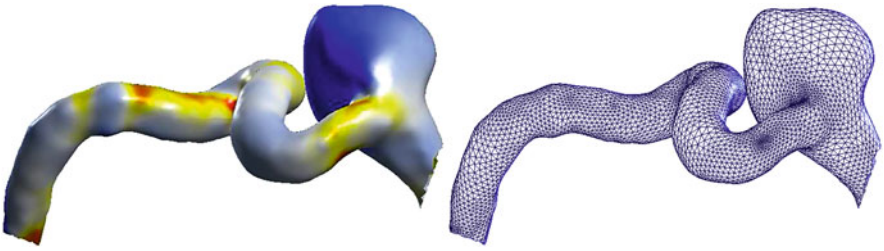

Fig. 5 *Left*: estimate of shear stress (modulus of shear stress at systolic peak) exerted by blood flow on the wall of an internal carotid artery affected by an aneurysm. *Right*: triangular mesh reconstruction of the wall of the internal carotid artery in *left panel*. Adapted from [12]

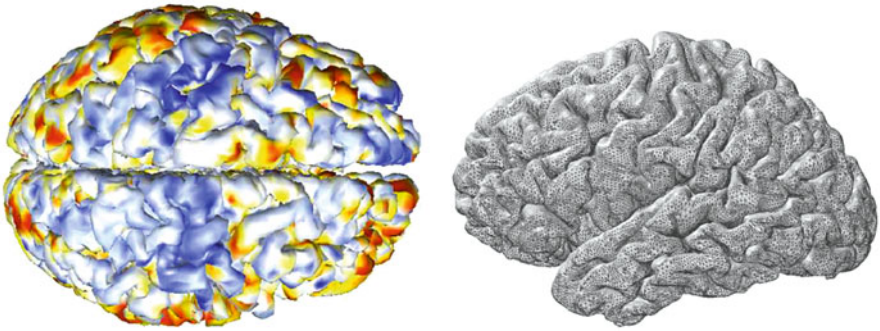

Fig. 6 *Left*: estimate of cortical thickness. *Right*: triangular mesh reconstruction of cortical surface

5 Discussion

We are currently extending this approach in various directions. One important generalization concerns for instance the modeling of space-time data, both over planar and over non-planar domains, which is of interest in several applications, including those briefly mentioned in Sects. 2.1, 3.1 and 4.1. Moreover, via a generalized linear framework, we are extending the models in order to handle outcomes having general distributions within the exponential family, including binomial, Poisson and gamma outcomes, further broadening the applicability of the proposed models. Combining all these features may in fact create a class of models that aims at handling data structures for which no statistical modeling currently exists.

Acknowledgements This paper reviews joint works with Laura Azzimonti, Bree Ettinger, Fabio Nobile, Simona Perotto, Jim Ramsay and Piercesare Secchi. This research has been funded by the research program Dote Ricercatore Politecnico di Milano—Regione Lombardia, project "Functional data analysis for life sciences", and by the starting grant *FIRB Futuro in Ricerca*, MIUR Ministero dell'Istruzione dell'Università e della Ricerca, research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" (http://mox.polimi.it/users/sangalli/firbSNAPLE.html).

References

1. Augustin, N.H., Trenkel, V.M., Wood, S.N., Lorance, P.: Space-time modelling of blue ling for fisheries stock management issue. Environmetrics **24**(Part 2), 109–119 (2013)
2. Azzimonti, L.: Blood flow velocity field estimation via spatial regression with PDE penalization. Ph.D. Thesis, Politecnico di Milano (2013)
3. Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P.: Mixed finite elements for spatial regression with PDE penalization. SIAM/ASA J. Uncertain. Quantif. **2**(Part 1), 305–335 (2013)

4. Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F.: Blood flow velocity field estimation via spatial regression with PDE penalization. J. Am. Stat. Assoc. Theory Methods. (2014) doi:10.1080/01621459.2014.946036
5. Baramidze, V., Lai, M.-J., Shum, C.K.: Spherical splines for data interpolation and fitting. SIAM J. Sci. Comput. **28**, 241–259 (2006)
6. Buja, A., Hastie, T., Tibshirani, R.: Linear smoothers and additive models. Ann. Stat. **17**, 453–555 (1989)
7. Chung, M.K., Robbins, S.M., Dalton, K.M., Davidson, R.J., Alexander, A.L., Evans, A.C.: Cortical thickness analysis in autism with heat kernel smoothing. NeuroImage **25**, 1256–1265 (2005)
8. Dassi, F., Ettinger, B., Perotto, S., Sangalli, L.M.: A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain. Technical Report MOX 31/2013, Dipartimento di Matematica, Politecnico di Milano (2013)
9. D'Elia, M., Perego, M., Veneziani, A.: A variational data assimilation procedure for the incompressible Navier-Stokes equations in hemodynamics. SIAM J. Sci. Comput. **52**(Part 4), 340–359 (2012)
10. Dierkes, U., Hildebrandt, S., Sauvigny, F.: Minimal Surfaces, vol. 1, 2nd edn. Springer, Heidelberg (2010)
11. Ettinger, B., Guillas, S., Lai, M.-J.: Bivariate splines for ozone concentration forecasting. Environmetrics **23**(Part 4), 317–328 (2012)
12. Ettinger, B., Perotto, S., Sangalli, L.M.: Spatial regression models over two-dimensional manifolds. Technical Report MOX 54/2012, Dipartimento di Matematica, Politecnico di Milano (2012)
13. Formaggia, L., Quarteroni, A., Veneziani, A.: Cardiovascular mathematics: modeling and simulation of the circulatory system. Springer, Berlin (2009)
14. Gneiting, T.: Strictly and non-strictly positive definite functions on spheres. Bernoulli **19**, 1087–1500 (2013)
15. Guillas, S., Lai, M.: Bivariate splines for spatial functional regression models. J. Nonparametr. Stat. **22**(Part 4), 477–497 (2010)
16. Hagler, D.J., Saygin, A.P., Sereno, M.I.: Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. NeuroImage **33**, 1093–1103 (2006)
17. Jun, M.: Non-stationary cross-covariance models for multivariate processes on a globe. Scand. J. Stat. **38**, 726–747 (2011)
18. Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. **73**, 423–498, with discussions and a reply by the authors (2011)
19. Marra, G., Miller, D., Zanin, L.: Modelling the spatiotemporal distribution of the incidence of resident foreign population. Stat. Neerl. **66**, 133–160 (2012)
20. Quarteroni, A.: Numerical Models for Differential Problems. Springer, Milan (2013)
21. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2011)
22. Ramsay, T.: Spline smoothing over difficult regions. J. R. Stat. Soc. Ser. B Stat. Methodol. **64**, 307–319 (2002)
23. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, Berlin (2005)
24. Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. J. R. Stat. Soc. Ser. B Stat. Methodol. **75**(Part 4), 681–703 (2013)
25. Stuart, A.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)
26. Wahba, G.: Spline interpolation and smoothing on the sphere. SIAM J. Sci. Stat. Comput. **2**, 5–16 (1981)
27. Wood, S.N., Bravington, M.V., Hedley, S.L.: Soap film smoothing. J. R. Stat. Soc. Ser. B Stat. Methodol. **70**, 931–955 (2008)