Springer Proceedings in Mathematics & Statistics

Asis Kumar Chattopadhyay
Gaurangadeb Chattopadhyay *Editors*

Statistics and its Applications

Platinum Jubilee Conference, Kolkata, India, December 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 244

Springer Proceedings in Mathematics & Statistics

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

More information about this series at http://www.springer.com/series/10533

Asis Kumar Chattopadhyay
Gaurangadeb Chattopadhyay
Editors

Statistics and its Applications

Platinum Jubilee Conference, Kolkata, India, December 2016

Springer

Editors
Asis Kumar Chattopadhyay
Department of Statistics
University of Calcutta
Kolkata, West Bengal, India

Gaurangadeb Chattopadhyay
Department of Statistics
University of Calcutta
Kolkata, West Bengal, India

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-981-13-1222-9 ISBN 978-981-13-1223-6 (eBook)
https://doi.org/10.1007/978-981-13-1223-6

Library of Congress Control Number: 2018945880

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Fragmentation of Young Massive Clusters: A Hybrid Monte Carlo Simulation Study

Abisa Sinha

Abstract To study the hierarchical fragmentation procedure in Young Massive Clusters, a stochastic model has been developed. Binary fragments along with individual stars are primarily studied in this work. Stellar masses for individual stars have been generated from the univariate truncated Pareto distribution and the stellar masses for binary stars have been generated from the truncated bi-variate Skew Normal Distribution using the Hamiltonian Monte Carlo method. The above distribution is used by observing the fitted bi-variate distribution of masses of all type of binary stars viz. visual binaries, spectroscopic binaries and eclipsing binaries. The resulting mass spectrum computed at different projected distances are observed under opacity limited fragmentation procedure and they display signature of mass segregation along the core to radius, whereas degree of segregation becomes reduced due to inclusion of all type of binary fragments in comparison to inclusion of eclipsing binaries only.

Keywords Initial mass function · Binary stars · Bivariate skew normal
Hybrid Monte Carlo

1 Introduction

The Initial Mass Function (IMF) of fragmented masses of molecular clouds undergoing gravitational collapse is of fundamental interest in many fields of astronomy and astrophysics. First observed by Salpeter (1955), the IMF is a power-law of the form $\xi = \frac{dN}{d \log m} \propto m^{\Gamma}$, where m is the mass of a star and N is the number of stars in the mass range $\log m$ and $(\log m + d \log m)$. His work favored an exponent of $\Gamma \sim -1.35$ for $0.4 M_{\odot} \leq m \leq 10 M_{\odot}$. Kroupa et al. (1993) found $\Gamma \sim -1.3$ (i.e. $\alpha \sim 2.3$) above half a solar mass, but introduced $\alpha \sim 1.3$ between $0.08 M_{\odot} - 0.5 M_{\odot}$ and $\alpha \sim 0.3$ below $0.08 M_{\odot}$ by proposing the IMF to be of segmented power-law form, where $\alpha = 1 - \Gamma$ in linear mass units of the form $\frac{dN}{dm} \propto m^{-\alpha}$. More modern

A. Sinha (✉)
Department of Statistics, Bethune College, Kolkata, India
e-mail: abisa.sinha@gmail.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_1

IMFs (segmented power-laws) appeared in the literature in recent years (e.g. Chabrier (2003)), but here our main intention was to find how the observed mass distribution depends on the slope of the fundamental IMF. Also, we have studied the effect of mass segregation in Young Massive Clusters (YMC) as a result of inclusion of binary fragments. Milone et al. (2012) have investigated the behavior of binary fraction for the Globular Clusters (GC) of Milky Way and have found that mass segregation is smaller than the field binaries. We have modeled the random fragmentation scenario together with line of sight effect as proposed by Chattopadhyay et al. (2011, 2016). Stars are generally born as binary or multiple systems and their binary nature is visible only for a small percentage. Chattopadhyay et al. (2016) have studied the binary fragments for eclipsing binary stars only. Here, we have tried to include all types of binary stars viz. visual binaries (consists of two stars, usually of different brightness and observed visually), spectroscopic binaries (consisting of a pair of stars where the spectral lines in the light emitted from each star shifts first towards the blue, then towards the red, as each moves first towards the observer, and then away from the observer, during its motion about their common center of mass, with the period of their common orbit and observed by periodic changes in spectral lines) and eclipsing binaries (consisting of stars in which the orbit plane of the two stars lies so nearly in the line of sight of the observer that the components undergo mutual eclipses and observed only by studying their respective Light Curves and Velocity Curves). Several authors have studied the masses, mass ratios of these binary stars (Tokovinin 2014; Kouwenhoven et al. 2007).

The percentage of binary contribution to the final form of fragments is of considerable debate. Over the past few years, several authors have considered the binary fragments in a coventional way (Abt 1983 for B stars; Duquennoy and Mayor (1991) for G dwarfs; Fischer and Marcy (1992) for M dwarfs; Kouwenhoven et al. 2007 for A and B stars; Goodwin et al. (2007) and references therein). Similar observations were noted for nearby and associated clusters (Duchéne 1999; Duchéne et al. 2007). Binary fractions of distant clusters were not observed previously because of observational limitations of measuring devices. Fortunately, by some alternative technique, some authors have become able to calculate the binary fraction. For example, by studying the morphology of colour-magnitude diagram, Romani and Weinberg (1991) determined the observed binary fractions in $M92$ and $M30$ at $\leq 9\%$ and 4% respectively. Rubenstein and Bailyn (1997) investigated binary fraction of stars in the range 15.8 mag $<$ V $<$ 28.4 mag in 13.5 Gyr old Galactic GC, NGC6752 as 15–38% inside the inner core, falling to 16% at larger radii with a power-law mass ratio distribution. Ballazzin et al. (2002) estimated the binary fraction in NGC288 for stars 20 mag $<$ V $<$ 23 mag ($\sim$0.54–0.77 $M_\odot$) as 8–38% inside cluster half mass radius. Zaho and Bailyn (2005) found 6–22% of main sequence binaries for $M3$ within core radius whereas Cool and Bolten (2002) derived a binary fraction of 3% for Galactic GC, NGC6397. Romani and Weinberg (1991) and Hurley et al. (2007) estimated a binary fraction $5.1 \pm 1.0\%$ within the inner region of NGC6397. All the clusters are dynamically evolved systems and are expected to significantly alter the initial binary population. Hu et al. (2010) have studied the young star cluster NGC1818 in LMC (age $\sim$15–25 Myr) and derived a binary fraction as high as 55%. Chattopadhyay et al.

(2016) has considered the contribution of binary fragments as close as 50%, while considering the rest as single stars. Malkov and Zinnecker (2001) have even claimed that the contribution of binary fragments is as close to 100%.

In the present work, we have considered random fragmentation of YMCs and have taken the binary contribution to be of 80% of the total fragments whereas 20% constitutes single stars. We have simulated 80% of binary stars from the truncated Bi-variate Skew Normal Distribution by Hybrid Monte Carlo method and the rest 20% of single stars from truncated Pareto distribution, truncated at minimum and maximum masses. The pattern of the bi-variate distribution is investigated and fitted to an appropriate form. In Sect. 2 we have discussed the data set, Sect. 3 gives the form of bivariate distribution, Sect. 4 gives the simulation procedures and Sect. 5 gives the results and discussions.

2 Data Set of Binary Stars

We have used a data-set of 2096 binary stars comprising of visual binaries, spectroscopic binaries and eclipsing binaries, among which 1875 sets of masses are taken from Tokovinin (2014) constituting only those binary stars which may be observed through telescope (viz. visual binaries, spectroscopic binaries), 78 sets of masses taken from Kouwenhoven et al. (2007) (also constituting visually observable stars) and the rest 143 sets of masses of eclipsing binaries (observed from their light-curves and velocity-curves) from Chattopadhyay et al. (2016). The method used to calculate binary masses from their observed mass-ratios as in Tokovinin (2014) and Kouwenhoven et al. (2007) has been found compatible for use (Fig. 3).

3 Bi-variate Distribution

3.1 Distribution Fit

Initially, we displayed the data of binary masses, in a bi-variate plot (Bivariate histogram (Fig. 1)) which displays a positive skewed pattern. The above data set is then fitted to bi-variate skew normal distribution of the form:

$$f_{Z_1,Z_2}(z_1, z_2) = 2\phi_2(\mathbf{z} - \xi; \Omega)\Phi(\alpha'\omega^{-1}(\mathbf{z} - \xi)) \qquad (1)$$

where z_1, z_2 are the random variables representing the masses of binary stars m_1, m_2 respectively, $f_{Z_1,Z_2}(z_1, z_2)$ is the probability density function and z_1, z_2 are particular values of Z_1, Z_2 respectively. Here $\xi = \left[\xi_1\ \xi_2\right]'$ is the location parameter, Ω is the correlation matrix (technically known as scale parameter) and $\alpha = \left[\alpha_1\ \alpha_2\right]'$ is the shape parameter which needs to be estimated.

We estimate the parameters by the help of Maximum Likelihood Estimation (MLE) by using the E-M Algorithm (Lachos et al. 2014) as follows: Let $\mathbf{y}$ denote the observed data and $\mathbf{s}$ denote the missing data. Let $\mathbf{y_c}$= $(\mathbf{y}, \mathbf{s})$ be $\mathbf{y}$ augmented with $\mathbf{s}$. We denote as $l_c(\theta \mid \mathbf{y_c})$, $\forall\theta \in \Theta$, (here $\theta = (\xi, \Omega, \alpha)$) the complete data log likelihood function and as $Q(\theta \mid \hat{\theta}) = E[l_c(\theta \mid \mathbf{y_c}) \mid y, \hat{\theta}]$, the expected data log-likelihood function. The Expectation and Maximization steps are respectively:

- E-step: $Q(\theta \mid \theta^{(r)})$ is computed as a function of θ.
- M-step: $Q^{(r+1)}$ is obtained confirming that $Q(\theta^{(r+1)} \mid \theta^{(r)}) = \max_{\theta\in\Theta} Q(\theta \mid \theta^{(r)})$.

By applying the above algorithm, the MLEs of the parameters came out to be:

$$\begin{aligned} \xi &= \begin{bmatrix}1.463 \ 0.851\end{bmatrix}' \\ \Omega &= \begin{bmatrix}2.6628 \ 0.5909 \\ 0.5909 \ 1.9085\end{bmatrix} \\ \alpha &= \begin{bmatrix}0.9076 \ 0.9922\end{bmatrix}' \end{aligned} \tag{2}$$

Based on the above parameters, we proceed henceforth in fitting the data to our proposed distribution.

3.2 Goodness of Fit Test

The goodness of fit test is based on the Moment Generating Function (MGF) of the bi-variate skew normal distribution as proposed by Meintanis et al. (2010), given as:

$$M(t) = 2exp[\frac{1}{2}(t_1^2 + 2\omega t_1 t_2^2)]\Phi(\alpha_1 t_1 + \alpha_2 t_2) \tag{3}$$

when ω is the co-rrelation parameter and $\alpha = \begin{bmatrix}\alpha_1 \ \alpha_2\end{bmatrix}'$ is the shape parameter. We know the MGF of any bivariate random vector $\mathbf{X} = \begin{bmatrix}X_1 \ X_2\end{bmatrix}'$ with $\mathbf{t} = \begin{bmatrix}t_1 \ t_2\end{bmatrix}' \in \mathbb{R}^2$ is defined by:

$$M(t_1, t_2) = E(e^{t'\mathbf{X}})$$

Putting $\vartheta = (\alpha_1, \alpha_2, \omega)$ in Eq. (3) and making the transformation $\mathbf{X} = \hat{\Omega^{-1}}(\mathbf{Z} - \hat{\xi})$ in our original variable (Eq. 1), the test statistic is (Meintanis et al. 2010):

$$T_{n,W}(\hat{\vartheta}) = n\int_{\mathbb{R}^2} D_n^2(t_1, t_2; \hat{\vartheta})W(t_1, t_2)dt_1dt_2 \tag{4}$$

We reject the null hypothesis for large values of $T_{n,W}(\hat{\vartheta})$, where

$$D_n^2(t_1, t_2; \hat{\vartheta}) = \alpha_2\frac{\partial M_n(t_1, t_2)}{\partial t_1} - \alpha_1\frac{\partial M_n(t_1, t_2)}{\partial t_2} - [(\alpha_2 - \omega\alpha_1)t_1 - (\alpha_1 - \omega\alpha_2)t_2]M_n(t_1, t_2)$$

with W being the weight function, $W(t_1, t_2) = W(\mathbf{t})$ satisfying $0 < \int_{\mathbb{R}^2} W(\mathbf{t})d\mathbf{t} < \infty$ and

$$M_n(t_1, t_2) = \frac{1}{n}\sum_{j=1}^{n} e^{t_1 X_{1j} + t_2 X_{2j}}$$

with $X_j = \left[X_{1j}\ X_{2j}\right]'$, $j = 1, 2, \ldots n$ are independent copies of $\mathbf{X}$. Our null hypothesis is

$$H_\circ : X \sim BVSN(\vartheta)$$

against general alternatives. Using the weight function $W(t) = e^{-at^2}$, $a > 0$ and satisfying $a \geq 2$ (Gradshteyn and Ryzhik 1994), $T_{n,W}(\hat{\vartheta})$ can be remodeled as:

$$T_{n,W}(\hat{\vartheta}) = \frac{\pi}{a} n(\hat{\alpha}_{2n}\bar{X}_1 - \hat{\alpha}_{1n}\bar{X}_2)^2 + O(a^{-1}),$$

$O(a^{-1}) \to \infty$ as $a \to 0$. We will declare the test statistic to be significant at nominal level α if $T_{n,W}(\hat{\vartheta})$ exceeds the $(1-\alpha)100\%$ level of significance. The p-value may be obtained from $P^*(T^*_{n,W}(\hat{\vartheta}) > T_{n,W}(\hat{\vartheta}))$, [$*$ denotes that $T_{n,W}(\hat{\vartheta})$ is computed from bootstrap sample]. To obtain $T_{n,W}(\hat{\vartheta})$, we proceed as follows:

- First, we collect samples, say 5000 from a bivariate skew normal distribution by the method as stated in Sect. 3.1 and using the MLEs of the parameters, as given in Eq. 2 and then standardize the data set.
- We repeat the above procedure for $B = 100$ bootstrap samples from $BVSN(\hat{\vartheta})$, estimate the respective location, scale and shape parameters and then standardize the data.
- For each bootstrap sample, we compute $\hat{\vartheta}^*_n = (\alpha_1, \alpha_2, \omega)$ and the corresponding value of $T^*_{n,W}(\hat{\vartheta}^*)$.

The result of our test is displayed in Table 1. So, we accept the null hypothesis at both 1% and 5% level of significance for $a = 3$ and conclude that the simulated data fits well to $BVSN(\vartheta)$.

We henceforth proceed towards simulation from $BVSN(\xi, \Omega, \alpha)$ for binary stars. In our previous work (Chattopadhyay et al. 2016), we have considered the value of efficiency factor ϵ (the ratio of stellar masses to the total mass of the parent cloud)

Table 1 Test of bivariate skew normality for the observed data-set for $B = 100$ bootstrap samples and 5000 Monte Carlo samples from BVSN (ϑ)

$T_{n,W}(\hat{\vartheta})$	α	$(1-\alpha)100\%$	p-value	a
63.05	0.05	95	0.0129	2.5
	0.01	99		
91.27	0.05	95	0.3512	3
	0.01	99		

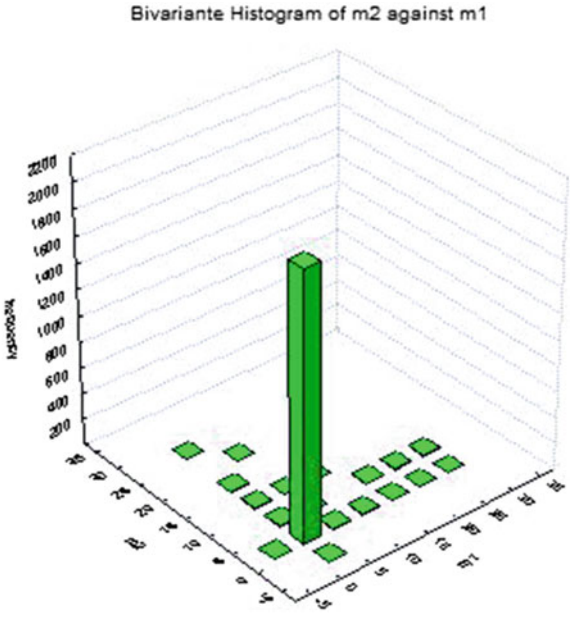


Fig. 1 Bivariate histogram of the Logarithm of Masses of Observed Binary stars. Masses m_1 and m_2 are in $M_\odot$

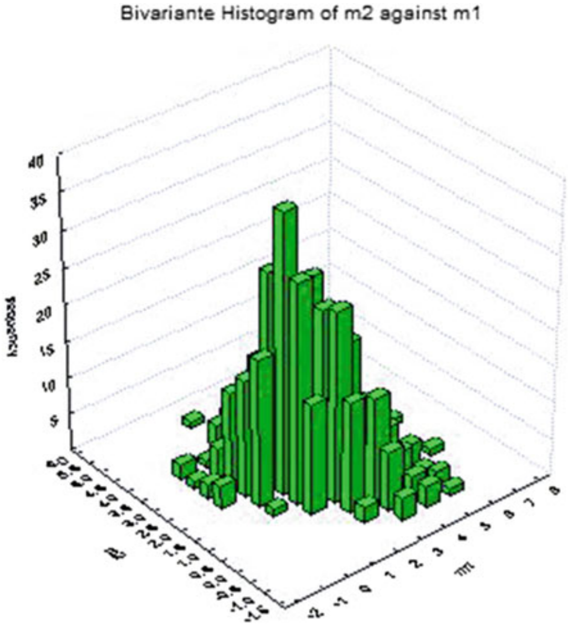


Fig. 2 Bivariate histogram of the logarithm of masses of simulated binary stars. Masses m_1 and m_2 are in $M_\odot$

as 0.5 (Lada et al. 1984; Elmegreen and Clemens 1985; Verschueren 1990). In the present work, we have retained the same choices of m_f, ϵ and have included our estimated parameters (vide Eq. 2). A bi-variate histogram of the simulated values along with that for the observed ones are displayed in Figs. 1 and 2 respectively.

4 Fragmentation and Mass Distribution

Hierarchical fragmentation of molecular cloud in YMCs and in other external galaxies and the final form of initial mass function have been a considerable matter of discussion during the last few decades. Random fragmentation of YMCs using a Monte Carlo simulation have been considered by Chattopadhyay et al. (2011). In their model, number of fragments, mass of the fragments and time between successive fragmentation are all considered as random variables. In our previous work, we have considered fragment masses simulated from bi-variate Gumbel Exponential distribution for binary stars and from Truncated Pareto Distribution for single stars. We simulated 50% of the total stellar mass of the parent cloud as binary stars and the rest half for single stars. In the present work, we simulate 80% of the total stellar masses of the parent cloud as binary stars and the rest as single stars. As pointed out earlier, here we are considering all binary types (i.e. visual, spectroscopic, resolvable and eclipsing binaries) and are simulating the binary pairs from our fitted distribution i.e. Skew Normal Distribution, whereas the single stars from Truncated Pareto distribution (Sect. 3.2).

4.1 Simulation of Binary Stars

Since our data directed us to bi-variate skew normal distribution, we draw random samples from the same distribution using Hamiltonian Dynamics (Neal 2012). This method, also known as Hamiltonian Monte Carlo method or the hybrid Monte Carlo simulation which uses the leap-frog scheme for simulating random numbers.

With the choice of parameters as in Eq. 2, we proceed as follows: We use Hamiltonian dynamics as a proposal function for a Markov chain in order to explore the target (canonical) density $p(x)$ defined by $U(x)$ more efficiently. Introducing the auxiliary variable p, more commonly known as kinetic energy function, defined as $K(\mathbf{P}) \sim \frac{\mathbf{p}'\mathbf{p}}{2}$, the Hamiltonian function $H(\mathbf{x}, \mathbf{p})$ is given as:

$$H(\mathbf{x}, \mathbf{p}) = -\log U(\mathbf{x}) + \frac{\mathbf{p}'\mathbf{p}}{2}$$

where $\frac{\partial U(x)}{\partial x_i}$ is the gradient function and $-\log U(\mathbf{x})$ is the log-likelihood to gradient. We take p to be a zero-mean bivariate Gaussian distribution with unit variance.

Starting with an initial state $[x_\circ, p_\circ]$, we simulate random numbers using Hamiltonian dynamics for a short time using the Leap-Frog method. If t be our t-th iteration step then to draw M random samples from the target density, the steps are:

- Set $t = 0$.
- Generate an initial position state $x^{(\circ)}$.
- Repeat until $t = M$.
  Set $t = t + 1$
  Sample a new initial momentum variable from the momentum canonical distribution $p_\circ \sim p(\mathbf{P})$. Set $x_\circ = x^{(t-1)}$. Run Leap-frog algorithm starting at $[x_\circ, p_\circ]$ for L steps and step-size δ to obtain the proposed states $[x_*, p_*]$.
- Calculate the Metropolis acceptance probability:

$$\alpha = \min[1, e^{(-U(x^*)+U(x_\circ)-K(p^*)+K(p_\circ))}] \tag{5}$$

- Draw a random number u from $U(0, 1)$. If $u \leq \alpha$, accept the proposed state position x^* and set the new state in the Markov chain $x^{(t)} = x^*$. Else set $x^{(t)} = x^{(t-1)}$. Return to Step 3.

We choose $p \sim BVN(\mathbf{0}, I)$, $M =$ number of samples required such that the total sum is $\leq 0.8m_f$, $L = 50$, $\delta = 0.25$.

4.2 Simulation of Single Stars

The method of generating random samples from the Truncated Power Law distribution is as follows (Chattopadhyay et al. 2015, 2016): The segmented power law is of the form:

$$\xi_{IMF}(m) = \frac{dN}{dm} = \begin{cases} Am^{-\alpha_1};\ m_{min} < m \leq m_c \\ Bm^{-\alpha_2};\ m_c < m \leq m_{max} \end{cases} \tag{6}$$

when A and B are solved to make:

$$\begin{gathered} \hat{A} = \hat{B} m_c^{\alpha_1-\alpha_2} \\ \hat{B} = [\tfrac{m_c^{\alpha_1-\alpha_2}}{1-\alpha_1}(m_c^{1-\alpha_1} - m_{min}^{1-\alpha_1}) + \tfrac{1}{1-\alpha_2}(m_{max}^{1-\alpha_2} - m_c^{1-\alpha_2})]^{-1} \end{gathered} \tag{7}$$

The values of m_{min}, m_{max}, m_c, α_1 and α_2 are taken from Chattopadhyay et al. (2011), and the random numbers are generated using the inverse transformation method for pseudo-random number sampling i.e. for generating random samples from the probability distribution given its cumulative distribution function (c.d.f), i.e. we get m-masses as

$$\begin{aligned} F_1(m) &= \tfrac{F(m)-F(m_{min})}{F(m_c)-F(m_min)} = u_1 \\ F_2(m) &= \tfrac{F(m)-F(m_c)}{F(m_{max})-F(m_c)} = u_2 \end{aligned}$$

Table 2 Initial values of parameters for simulation of binary fragments obtained through Hybrid Monte Carlo method with a $L = 50$ Leap-frog steps and stepsize $\delta = 0.25$

Name	m_f	ϵ	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\Omega}_{11}$	$\hat{\Omega}_{12}$	$\hat{\Omega}_{22}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
$NGC330$	$10^{5.8}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$M31Vdb0$	10^5	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$M31B2570$	10^5	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$LMCNGC2164$	$10^{5.2}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$LMCNGC2214$	$10^{5.4}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$NGC4038S2_3$	$10^{5.4}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$NGC4038S1_5$	$10^{5.6}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922
$NGC4038S2_1$	$10^{6.0}$	0.5	1.463	0.851	2.6628	0.5909	1.9085	0.9076	0.9922

Note—m_f is the mass of the parent cloud, ϵ is the efficiency

Solving for m, we get:

$$\begin{aligned} m &= [u_1(m_c^{1-\alpha_1} - m_{min}^{1-\alpha_1}) + m_{min}^{1-\alpha_1}]^{\frac{1}{1-\alpha_1}};\ m_{min} < m \le m_c \\ m &= [u_2(m_{max}^{1-\alpha_2} - m_c^{1-\alpha_2}) + m_c^{1-\alpha_2}]^{\frac{1}{1-\alpha_2}};\ m_c < m \le m_{max} \end{aligned} \tag{8}$$

when $u_1, u_2 \sim U(0, 1)$. Thus when $u_1 = 0$, $m = m_{min}$ and when $u_1 = 1$, $m = m_c$. Similarly, for $m_c < m \le m_{max}$. We simulate from $F_1(m)$ as long as the total stellar mass of the embedded cluster is equal to 0.02 times the efficient mass of the parent cloud and then simulate the rest 0.18 part from $F_2(m)$.

Combining the total number of stellar masses obtained as binary fragments and single stars, we form the segmented power law and obtain the critical masses, the slopes for different segments as well as errors (Fig. 3). The result is displayed in Table 4 and slopes for M31 Vdb0.

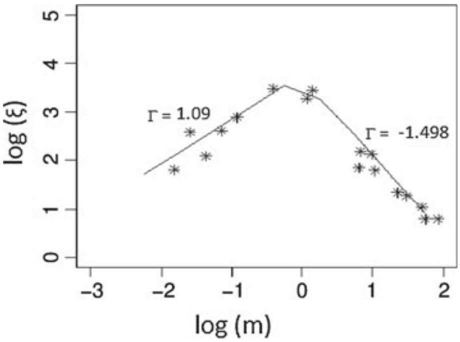


Fig. 3 Segmented power-law fit at $b = 1pc$ for $M31Vdb0$, with simulated values (asterisk). m is in $M_\odot$

Table 3 Initial values of parameters for simulation of individual fragments

Name	m_f	b	m_{max}	$logm_c$	$m_{min} \leq m \leq m_c$		$m_c \leq m \leq m_{max}$	
					Γ	α	Γ	α
	$(M_\odot)$	(pc)	$(M_\odot)$	$(M_\odot)$	$(M_\odot)$		$(M_\odot)$	
NGC 330	$10^{5.8}$	1	132.41 ± 8.87	−0.61 ± 0.02	1.10 ± 0.19	−0.10	−1.20 ± 0.19	2.20
		2		−0.60 ± 0.021	1.12 ± 0.05	−0.12	−1.49 ± 0.02	2.49
		12		−0.64 ± 0.082	1.32 ± 0.14	−0.32	−1.73 ± 0.28	2.73
M31 Vdb0	10^5	1	139.85 ± 12.23	−0.66 ± 0.05	1.28 ± 0.27	−0.28	−1.15 ± 0.02	2.15
		2		−0.61 ± 0.035	1.12 ± 0.021	−0.12	−1.33 ± 0.06	2.33
		12		−0.58 ± 0.17	1.33 ± 0.082	−0.33	−1.43 ± 0.39	2.43
M31 B2570	10^5	1	138.28 ± 15.6	−0.58 ± 0.028	1.07 ± 0.31	−0.07	−1.28 ± 0.58	2.28
		2		−0.57 ± 0.061	1.02 ± 0.06	−0.02	−1.53 ± 0.12	2.53
		12		−0.59 ± 0.091	1.28 ± 0.11	−0.28	−1.77 ± 0.092	2.77
LMC NGC2164	$10^{5.2}$	1	141.08 ± 8.02	−0.64 ± 0.015	1.04 ± 0.62	−0.04	−1.13 ± 0.34	2.13
		2		−0.58 ± 0.028	1.25 ± 0.032	−0.25	−1.25 ± 0.051	2.25
		12		−0.60 ± 0.053	1.05 ± 0.091	−0.05	−1.27 ± 0.348	2.27
LMC NGC2214	$10^{5.4}$	1	140.33 ± 3.44	−0.63 ± 0.018	1.30 ± 0.33	−0.30	−1.12 ± 0.33	2.12
		2		−0.60 ± 0.028	1.01 ± 0.05	−0.01	−1.28 ± 0.02	2.28
		12		−0.64 ± 0.082	1.18 ± 0.032	−0.18	−1.55 ± 0.27	2.55
NGC $4038S2_3$	$10^{5.4}$	1	123.01 ± 13.8	−0.59 ± 0.036	1.08 ± 0.25	−0.08	−1.12 ± 0.29	2.12
		2		−0.63 ± 0.011	1.25 ± 0.06	−0.25	−1.43 ± 0.05	2.43
		12		−0.57 ± 0.071	1.06 ± 0.091	−0.06	−1.62 ± 0.29	2.62
NGC $4038S1_5$	$10^{5.6}$	1	113.88 ± 21.38	−0.56 ± 0.18	1.12 ± 0.36	−0.12	−1.37 ± 0.19	2.37
		2		−0.58 ± 0.062	1.15 ± 0.017	−0.15	−1.57 ± 0.046	2.57
		12		−0.60 ± 0.087	1.18 ± 0.15	−0.18	−1.69 ± 0.47	2.69
NGC $4038S2_1$	$10^{6.0}$	1	133.37 ± 11.06	−0.64 ± 0.061	1.06 ± 0.51	−0.06	−1.04 ± 0.15	2.04
		2		−0.62 ± 0.028	1.24 ± 0.04	−0.24	−1.49 ± 0.15	2.49
		12		−0.61 ± 0.092	1.27 ± 0.17	−0.27	−1.49 ± 0.102	2.49

Note—Col. 1 represents the name of the galaxy, Col. 2 (m_f) gives the mass of YMC in that galaxy, Col. 3 (b) is the distance from the cloud center, Cols. 4, 5, 6, 7 show maximum mass(m_{max}), logarithm of the critical mass (m_c) and slopes (Γ and α) of the initial mass functions in low mass and high mass regimes as a result of random fragmentation

Table 4 Segmented Power-law fits to the simulated fragments as a result of random fragmentation of YMCs including binary fraction constituting 80% of the total active mass of the cloud (viz. ϵm_f)

Name	m_f	b	$m_{min} \leq m \leq m_c$		$m_c \leq m \leq m_{max}$	
			Γ	α	Γ	α
	$(M_\odot)$	(pc)	$(M_\odot)$		$(M_\odot)$	
NGC 330	$10^{5.8}$	1	1.13 ± 0.06	-0.13	-1.72 ± 0.13	2.72
		2	1.19 ± 0.07	-0.19	-1.91 ± 0.13	2.91
		12	1.22 ± 0.04	-0.22	-2.12 ± 0.08	3.22
M31 Vdb0	10^5	1	1.09 ± 0.24	-0.09	-1.498 ± 0.04	2.498
		2	1.17 ± 0.05	-0.17	-1.686 ± 0.16	2.686
		12	1.27 ± 0.14	-0.27	-1.801 ± 0.13	2.801
M31 B2570	10^5	1	1.13 ± 0.08	-0.13	-1.76 ± 0.18	2.76
		2	1.21 ± 0.92	-0.21	-1.83 ± 0.08	2.83
		12	1.29 ± 0.11	-0.29	-1.885 ± 0.29	2.885
LMC NGC2164	$10^{5.2}$	1	1.17 ± 0.14	-0.17	-1.79 ± 0.11	2.79
		2	1.24 ± 0.18	-0.24	-1.95 ± 0.168	2.95
		12	1.29 ± 0.25	-0.29	-2.01 ± 0.24	3.01
LMC NGC2214	$10^{5.4}$	1	1.212 ± 0.05	-0.212	-1.69 ± 0.36	2.69
		2	1.25 ± 0.21	-0.25	-1.85 ± 0.207	2.85
		12	1.27 ± 0.11	-0.27	-2.17 ± 0.44	3.17
NGC 4038$S2_3$	$10^{5.4}$	1	1.208 ± 0.22	-0.208	-1.85 ± 0.23	2.85
		2	1.27 ± 0.17	-0.27	-1.98 ± 0.39	2.98
		12	1.301 ± 0.28	-0.301	-2.11 ± 0.301	3.11
NGC 4038$S1_5$	$10^{5.6}$	1	1.19 ± 0.27	-0.19	-1.892 ± 0.32	2.892
		2	1.29 ± 0.14	-0.29	-2.12 ± 0.50	3.12
		12	1.311 ± 0.22	-0.311	-2.17 ± 0.35	3.17
NGC 4038$S2_1$	$10^{6.0}$	1	1.181 ± 0.11	-0.181	-2.08 ± 0.32	3.08
		2	1.31 ± 0.28	-0.28	-2.13 ± 0.35	3.13
		12	1.35 ± 0.205	-0.35	-2.27 ± 0.29	3.27

5 Result and Discussion

5.1 Initial Values of the Parameters

For fragmentation of YMCs we have used the YMCs in external galaxies (Zwart et al. 2010) whose masses varies from $10^5 M_\odot - 10^6 M_\odot$. In our present work, we have used efficiency value as $\epsilon = 0.5$. The minimum (m_{min}), maximum (m_{max}), critical masses (m_c) of the fragments and the indices of α (viz. α_1, α_2) are chosen from Table 2 of Chattopadhyay et al. (2011), in case of individual fragments. For bivariate simulation, the binary pairs have the computed mass regimes using Eqs. 1 and 2.

The projected distances from the cloud center are chosen as 1parsec, 2parsec and 12parsec respectively (Table 3).

5.2 Resulting Mass Spectrum

The resulting mass spectra generated using Eqs. 1, 2 and 8 are fitted by segmented power laws in different mass regimes and are shown in Table 4 along with the corresponding errors. The errors are computed by running each simulation for a number of times. It is clear from Table 4 that the mass spectrum are getting steeper in all zones compared to previous cases (Chattopadhyay et al. 2011, 2016). Also, it is observed that the number of high-mass stars gets reduced considerably, even less than what we got in our previous work (Chattopadhyay et al. 2016). It might be due to the cause that sufficient binaries were not observed previously, not withstanding the fact that if the low-mass limit is below the detectability limit, we have unseen components (brown dwarfs) in orbit around the field stars (Malkov and Zinnecker 2001). It is also observed that short periods with large mass ratios could preferably appear in multiple systems (Tokovinin 2014). As a result most binaries in multiple system were observed as single stars. The resultant mass spectrum, in our case, have more intermediate mass stars in comparison to low-mass and high-mass fragments, which is quite conformable to physical observation since when one single star is braked into a binary pair the presence of intermediate mass-stars and low-mass stars (brown dwarfs) is expected. Our result is confirmed by Halbwachs et al. (2003), Kobulnicky and Fryer (2007), Delfosse et al. (2004) and many other authors who have carried out similar studies in this area.

Table 4 also reflects the presence of mass segregation along the line of sight as we move from core to radius (i.e. as b changes from 1parsec to 12parsec). Mass segregation is generally associated with the gradual equipartition of energy via stellar encounters in old globular clusters as well as young massive clusters (Spitzer 1987; Hillenbrand and Hartmann 1998). The mass segregation reflects the following causes:

- Due to inclusion of binary fragments covering 80% of the total stellar mass, the innermost binaries segregate towards the core leaving a minimum in the radial binary frequency distribution that marches outward with time (Geller et al. 2008). Dynamical mass segregation is observed to proceed more rapidly at the core, on the order of a few crossing times (de Grijs et al. 2002a).
- Since stellar mass black holes eject one another from the system within a few relaxation times, so eventually these systems fully develop the amount of mass segregation observed in runs starting from Miller and Scalo IMF. As a result the low mass stars contribute very little to the overall luminosity in many cases, and thus the total luminosity will be very little similar to that of the brighter member (Gill et al. 2008). Hence, presence of low-luminosity stars in the envelope are suspected as a result of mass segregation.

Hence the steepness of the IMF is interpretated as resulting from the correction of unresolved binaries (Sagar and Richtler 1991) that were considered as single massive stars. Also, presence of brown dwarfs are suspected at the envelope.

Acknowledgements I deeply acknowledge the help, guidance and support received from my joint guides Prof. Asis Kumar Chattopadhyay and Prof. Tanuka Chattopadhyay. I also acknowledge the timely help and suggestions from Prof. R. M. Neal.

References

Bellazini, M., Fusi Pecci, F., Messineo, M., Monaco, L.,& Rood, R. T. (2002). *Astrophysical Journal*, *123*, 1509.

Chabrier, G. (2003). Star formation in molecular clouds. *Astronomical Society of the Pacific*, *115*, 763–795.

Chattopadhyay, T., Chattopadhyay, A. K., Sinha, A. (2011). Modelling of the initial mass function using the Metropolis-Hastings algorithm. *Astrophysical Journal*, *736*, 152.

Chattopadhyay, T., De, T., Warlu, B., & Chattopadhyay, A. K. (2015). Cosmic history of the integrated galactic stellar initial mass function: a simulation study. *Astrophysical Journal*, *808*, 24.

Chattopadhyay, T., Sinha, A., Chattopadhyay, A. K. (2016). Influence of binary fraction on the fragmentation of young massive Clustersa Monte Carlo simulation. *Astronomy & Space Science*, *361*, 120.

Cool, A. M., Bolton, A. S. (2002). Blue stars and binary stars in NGC 6397: Case study of a collapsed-core globular cluster. Astronomical Society of the Pacific, San Francisco, vol. 263, p. 163.

de Grijis, R., Bastian, N., Lamers, H. J. G. L. M, (2002a) Star cluster systems in interacting and starburst galaxies: A multicolour approach. Monthly Notices of the Royal Astronomical Society, 340, 197.

Delfosse, X., Beuzit, J. L., Marchal, L., Bonfils, X. C., Perrier, C. (2004). M dwarfs binaries: Results from accurate radial velocities and high angular resolution observations. In: R. W. Hilditch, H. Hensberge & K. Pavlovski (eds.) Spectroscopically and Spatially Resolving the Components of the Close Binary Stars, Proceedings of the Workshop Held 20–24 October 2003 in Dubrovnik, Croatia. ASP Conference Series, vol. 318, (pp. 166–174). San Francisco: Astronomical Society of the Pacific.

Duchéne, G., Beck, A. K., Twin, P. J., France, G., de Curien, D., Han, L., Beausang, C. W., Bentley, M. A., Nolan, P. J. & Simpson, J. (1999). The Clover: A new generation of composite Ge detectors. *Nuclear Instruments and Methods in Physics Research Section A*, *432*, 90–110.

Duchéne, G., Bontemps, S., Bouvier, J., Andre, P., Djupvik, A. A., & Ghez, A. M. (2007). Multiple protostellar systems. II. A high resolution near-infrared imaging survey in nearby star-forming regions. *Astronomy & Astrophysics*, *476*, 229–242.

Duquennoy, A., & Mayor, M. (1991). Multiplicity among solar-type stars in the solar neighbourhood. II - Distribution of the orbital elements in an unbiased sample. *Astronomy and Astrophysics, 248*(2), 485–524.

Elmegreen, B. G., & Clemens, C. (1985). On the formation rate of galactic clusters in clouds of various masses. *Astrophysical Journal*, *294*, 523–532.

Fischer, D. A., Marcy G. W. (1992). Multiplicity among M dwarfs. *The Astrophysical Journal, 396*, 178–194.

Geller, A. M., Mathieu, R. D., Harris, H. C., & McClure, R. D. (2008). WIYN open cluster study. XXXII. Stellar radial velocities in the OldOpen cluster NGC 188. *The Astrophysical Journal*, *135*, 2264–2278.

Gill, M., Trenti, M., & Miller, M. C. (2008). Intermediate-mass black hole induced quenching of mass segregation in star clusters. *The Astrophysical Journal*, *686*, 303–309.

Goodwin, S. P., Kroupa, P., Goodman, A., & Burkert, A. (2007). In: Reipurth, V. B., Jewitt, D., & Keil, K. (eds.) Protostars and Planets, p. 133. University of Arizona Press, Tuscon.

Gradshteyn, I. S. & Ryzhik, I. M. (1994). *Tables of integrals series and products*. 5th Edition, Academic Press, New York.

Halbwachs, J. L., Mayor, M., Udry, S., & Arenou, F. (2003). Multiplicity among solar-type stars. III. Statistical properties of the F7-K binaries with periods up to 10 years. *Astronomy & Astrophysics*, *397*, 159–175.

Hillenbrand, L., & Hartmann, L. (1998). A preliminary study of the orion nebula cluster structure and dynamics. *The Astrophysical Journal*, *492*(2), 540–553.

Hu, Y., Deng, L., de Grijs, R., Goodwin, S. P., & Liu, Q. (2010). The binary fraction of the young cluster NGC 1818 in the large magellanic cloud. *Astrophysical Journal*, *724*, 649.

Hurley, J. R., Aarseth, S. J., & Shara, M. M. (2007). The core binary fractions of star clusters from realistic simulations. *Astrophysical Journal*, *665*, 707–718.

Kobulnicky, H. A. & Fryer, C. L. (2007). A new look at the binary characteristics of massive stars. *Bulletin of the American Astronomical Society*, *39*, 726.

Kouwenhoven, M. B. N., Brown, A. G. A., Portegies Zwart S. F., & Kaper L. (2007). The primordial binary population. II. Recovering the binary population for intermediate mass stars in Scorpius OB2. *Astronomy & Astrophysics*, *474*, 77–104.

Kroupa, P., Tout, C. A., & Gilmore, G. (1993). The distribution of low-mass stars in the galactic disc. *Monthly Notices of the Royal Astronomical Society*, *262*, 545.

Lachos, V. H., Labra, F. V., Ghosh, P. (2014). Multivariate skew-normal/independent distributions: properties and inference, JUSTOR 2014, pp. 517–535.

Lada, C. J., Margulis, M., & Dearborn, D. (1984). The formation and early dynamical evolution of bound stellar systems. *Astrophysical Journal*, *285*, 141–152.

Malkov, O., & Zinnecker, H. (2001). Binary stars and the fundamental initial mass function. *Monthly Notices of the Royal Astronomical Society*, *321*, 149–134.

Meintanis, S. G., & Hlávka, Z. (2010). Goodness-of-fit tests for bivariate and multivariate skew-normal distributions, *37*, 701–714. (JUSTOR 2010)

Milone, A. P., Piotto, G., Bedin, L. R., et al. (2012). Multiple stellar populations in the globular clusters NGC1851 and NGC6656 (M22). *Astrophysical Journal*, *744*, 58.

Neal, R. M. (2012). *Handbook of Markov Chain Monte Carlo. MCMC using Hamiltonian dynamics*. CRC Press

Romani, R. W., & Weinberg, M. (1991). Limits on cluster binaries. *Astrophysical Journal*, *372*, 487.

Rubenstein, E. P., & Bailyn, C. D. (1997). Hubble space telescope observations of the post-core-collapse globular cluster NGC 6752. II. A large main-sequence binary population. *Astrophysical Journal*, *474*, 701–709.

Salpeter, E. E. (1955). The luminosity function and stellar evolution. *Astrophysical Journal*, *121*, 161.

Sagar, R., & Richtler, T. (1991). Astronomy and Astrophysics. *Mass Functions of 5 young Large Magellanic Cloud star clusters, 250*, 324–339.

Spitzer, L. (1987). *Dynamical evolution of globular clusters* (p. c1987). Princeton: Princeton University Press.

Tokovinin A. (2014). From binaries to multiples. I. Data on F and G Dwarfs Within 67 pc of the Sun. *The Astronomical Journal*, *147*, 86.

Verschueren, W. (1990). Collapse of young stellar clusters before gas removal. *Astronomy & Astrophysics*, *234*, 156–163.

Zhao, B., & Bailyn, C. D. (2005). *Astrophysical Journal*, *129*, 1934.

Zwart, S. F. P., McMillan, S. L. W., & Gieles, M. (2010). Young massive star clusters. *Annual Review of Astronomy and Astrophysics*, *48*, 431.

A Study on DNA Sequence of Rice Using Scoring Matrix Method and ANOVA Technique

Anamika Dutta and Kishore K. Das

Abstract In this paper, 12 accession numbers of rice has been used. The accession numbers have been taken from the article Cho et al. where it has already been used for other studies. The accession number for DNA, i.e., A, C, G and T along with the gap character (–) have been converted into alignment matrix with 5 rows and 7473 columns. The alignment has been done using ClustalX software. The 7473 columns have been alienated into 5 parts with different dimensions. Later for each part scoring has been done separately. Highest scores from all the 5 parts have been noted down. To minimize the data, the common regions between these 5 parts have been taken into consideration. Later one way ANOVA (Huck and McLean in Psychological Bulletin, 82(4), 511–518,1975; Mukhopadhyay in Applied statistics. Books and Allied (P) Ltd., Kolkata, 2011) has been constructed and conclusions are drawn accordingly.

Keywords Scoring matrix · Alignment matrix · Weight matrix · One way ANOVA

1 Introduction

The field of protein has a great name in the area of research. Rice is the main grain harvested in India; consequently, the FASTA formats of 12 accession number of rice have been selected for our study. The accession numbers comprise have been collected from an article Cho et al. where it has already been used for other purpose study. For scoring the matrices, alignment and weight matrix (Hertz and Stormo 1999; Shu et al. 2012) have been used. Alignment has been done using ClustalX software. The Accession Numbers for rice (Cho et al. 2000) are:

A. Dutta (✉) · K. K. Das
Department of Statistics, Gauhati University, Guwahati, India
e-mail: anamika.dut268@gmail.com

K. K. Das
e-mail: daskkishore@gmail.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_2

D17586, D16221, M36469, D78609, X58877, Z11920, D14000, L37528, X07515, U12171, U33175 and D30794

The main objectives of this paper are:

(a) To split the entire data into parts using alignment and weight matrix.
(b) Scoring the split parts and applying ANOVA one way technique.

2 Description of Data

2.1 Source of Data

The FASTA format of the accession numbers have been collected from NCBI using nucleotide as database. By means of ClustalX software and multiple sequence alignment the DNA sequence of the accession number have been precised. The arrangement and counting of the DNA sequences have been done using R software with the help of "sequinr" package.

At this moment, let us explain what is known as multiple sequence alignment. Multiple sequence alignment is defined as an alignment of similar sequences. The main criterion of multiple sequence alignment is that there should be more than two sequences or a minimum of three sequences however they may not be of same length (Wallace et al. 2005; Pei 2008).

2.2 Arrangement of Data

The 12 accession numbers have been transformed into alignment matrix with 5 rows and 7473 columns. Consequently, 7473 columns have been divided into 5 parts. Let us explain how the alignment has been done.

The sequences have been written one after another in a horizontal manner. Subsequently, we have placed 12 accession numbers in order. Consequently, a matrix with 12 rows and 7473 columns has been formed. Subsequently from each column we have written vertically how persistently A, C, G, T and gap character (–) have been repeated. Hence 5 rows have been formed with 7473 columns. Later than the entire 7473 columns have been separated into small matrices with dimensions $m = 5$ and $n = 8$, 16, 24, 32 and 40 respectively. Thereafter 5 matrices with dimensions 5×8, 5×16, 5×24, 5×32 and 5×40 have been constructed. Subsequently, we have found the scores for each dimension for the entire 7473 columns.

3 Methodology

Using the above method we have found the alignment for 12 sequences of rice, with 5 rows and 7473 columns. Let us explain how we can find the score of the above matrix.

The score of the DNA alignment matrix has been found using the formula (Hertz and Stormo 1999; Shu et al. 2012):

$$\ln \frac{\left(n_{i,j} + p_i\right)/(N+1)}{p_i}$$

Where

$n_{i,j}$ = The letter i is observed at position j of its alignment.
p_i = It is the priori probability of letter i.
N = It is the total number of sequences

Later, ANOVA one way technique has been applied to test the significance difference between the DNA varieties.

4 Results and Discussion

Using the formula of alignment matrix, we have found the weights of entire 7473 columns. Subsequently the entire region has been partitioned into small matrices with dimensions 5 × 8, 5 × 16, 5 × 24, 5 × 32 and 5 × 40. Partitioning the matrices we have found 943 matrices of dimensions 5 × 8, 467 matrices of dimensions 5 × 16, 311 matrices of dimensions 5 × 24, 233 matrices of dimensions 5 × 32 and 186 matrices of dimensions 5 × 40. Next the highest scores for matrices of each dimension have been found individually. Only the regions of highest scores for each dimension have been selected for our study. The method of intersection has been introduced to reduce the dimension of the data concerned.

Between the five highest scoring matrices, the common regions within the matrices have been pointed out. There exists a common region between the dimension of 5 × 8 and 5 × 16 matrices and the common region is of dimension 5 × 8. Also the common region between 5 × 32 and 5 × 40 matrices and the common region is of dimensions 5 × 24. There is no region common with 5 × 24 matrix, so it has been dropped from the study.

The significance of A, C, G, T and gap character (–) for the two common regions have been studied using one way ANOVA technique (Huck and McLean 1975; Mukhopadhyay 2011).

The first common region is:

$$\begin{array}{c} \text{DNA VARIETY} \quad \text{OBSERVATION} \\ \begin{bmatrix} - & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 1 & 4 & 2 & 2 & 5 & 0 & 3 & 4 \\ C & 6 & 1 & 1 & 3 & 1 & 10 & 5 & 1 \\ G & 3 & 7 & 6 & 6 & 5 & 0 & 2 & 7 \\ T & 1 & 0 & 3 & 1 & 1 & 2 & 2 & 0 \end{bmatrix} \end{array}$$

The null hypothesis to be tested is:

H_0 There is no significant difference between the DNA varieties

Against the alternative hypothesis,

H_1 There is significant difference between the DNA varieties

The ANOVA table is as follows (Table 1):

From the above ANOVA table, we have seen that the observed value of F is greater than the tabulated value of F at 5% level of significance. Hence the variance ratio is significant and we have reason to reject the null hypothesis which infers that possibly there is significant difference between the DNA varieties.

As the result is found to be significant, which means that at least one of the DNA Variety groups differ from the other DNA Variety group. Hence, we shall find the Post Hoc Tests of multiple comparisons using the method of Fisher's Least Significance Difference (LSD) test.

The formula for Least Significance Difference is (Williams and Abdi 2010):

$$\text{LSD} = t_{\nu,\alpha}\sqrt{\text{MSW}\left(\frac{1}{S_I} + \frac{1}{S_J}\right)}$$

Where t is the critical value from the t-distribution table; MSW is mean square within obtained from the ANOVA table; S is the number of scores used to calculate means and ν is the error df.

Let us construct the following table showing the multiple comparisons using LSD (Table 2):

Post Hoc Test has been analysed using PASW Statistics 18.

Table 1 ANOVA table for dimension 5 × 8

Source	df	SS	MS	F (obs)	F (tab) (5%)
DNA variety	4	97.350	24.338	5.669	2.641
Within Groups	35	150.250	4.293	–	–
Total	39	247.600	–	–	–

Table 2 Post hoc analysis test of multiple comparisons using LSD

(I) Data variation		(J) Data variation		Mean difference (I − J)	Sig.
	–		A	-2.500^{a}	0.021
			C	-3.375^{a}	0.002
			G	-4.375^{a}	0.000
			T	−1.125	0.285
	A		–	2.500^{a}	0.021
			C	−0.875	0.404
			G	−1.875	0.079
			T	1.375	0.193
	C		–	3.375^{a}	0.002
			A	0.875	0.404
			G	−1.000	0.341
			T	2.250^{a}	0.037
	G		–	4.375^{a}	0.000
			A	1.875	0.079
			C	1.000	0.341
			T	3.250^{a}	0.003
	T		–	1.125	0.285
			A	−1.375	0.193
			C	-2.250^{a}	0.037
			G	-3.250^{a}	0.003

[a] = The mean difference is significant at the 0.05 level

It can be seen from the above table that the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G); (C, T) and (G, T) differ significantly at 5% level of significance. We have the reason to reject the null hypothesis and infer that there exists significant difference between the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G); (C, T) and (G, T). And the other pairs are not significant.

Similarly, we have tried for the significance of A, C, G, T and gap character (–) for the second common region which has 24 columns.

The null hypothesis to be tested is:

H_0: There is no significant difference between the DNA varieties.

Against the alternative hypothesis:

H_1: There is significant difference between the DNA varieties.

The ANOVA table for the second common region is as follows (Table 3):

Similarly, from the above ANOVA table, we have seen that the observed value of F is greater than the tabulated value of F at 5% level of significance. Hence the

Table 3 ANOVA table for dimension 5 × 24

Source	df	SS	MS	F (obs)	F (tab) 5%
DNA variety	4	150.133	37.533	8.165	2.451
Error	115	528.667	4.597		
Total	119	678.800			

variance ratio is significant and we have reason to reject the null hypothesis which infers that possibly there is significant difference between DNA varieties.

Here also the result is found to be significant, which means that at least one of the DNA Variety groups differs from the other DNA Variety group. Again we shall construct the table for multiple comparisons using LSD.

As we know, the formula for Least Significance Difference is (Williams and Abdi 2010):

$$\mathrm{LSD} = t_{\nu,\alpha}\sqrt{\mathrm{MSW}\left(\frac{1}{S_I} + \frac{1}{S_J}\right)}$$

Where t is the critical value from the t-distribution table; MSW is mean square within obtained from the ANOVA table; S is the number of scores used to calculate means and ν is the error df (Table 4).

It can be seen from the above table that the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G) and (Gap Charcter, T) differ significantly at 5% level of significance. We have the reason to reject the null hypothesis and infer that there exists significant difference between the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G) and (Gap Character, T). And the other pairs are not significant.

5 Conclusion

Using alignment matrix and weight matrix the entire region have been divided small parts. In this paper, we have tried to study the significance of two common regions which are of 5 × 8 and 5 × 24 dimension using ANOVA one way method. And we got significant difference between the DNA varieties of rice for both the matrices. Hence there exist significant difference between A, C, G, T and Gap Character (–). After that we have performed pair wise comparison between the DNA varieties using the critical difference method for both the samples. From the first sample, the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G); (C, T) and (G, T) are significant. And from the second sample, the pairs (Gap Character, A); (Gap Character, C); (Gap Character, G) and (Gap Character, T) are significant and rest all pairs from both the samples are not significant.

Applying ANOVA technique and Post Hoc analysis test, it can be seen that in the first sample where there are 5 rows and 8 columns we got 5 pairs to be significant.

Table 4 Post hoc analysis test of multiple comparisons using LSD

(I) Data variation	(J) Data variation	Mean difference (I − J)	Sig.
–	A	−2.583[a]	0.000
	C	−3.167[a]	0.000
	G	−2.667[a]	0.000
	T	−2.542[a]	0.000
A	–	2.583[a]	0.000
	C	−0.583	0.348
	G	−0.083	0.893
	T	0.042	0.946
C	–	3.167[a]	0.000
	A	0.583	0.348
	G	0.500	0.421
	T	0.625	0.315
G	–	2.667[a]	0.000
	A	0.083	0.893
	C	−0.500	0.421
	T	0.125	0.840
T	–	2.542[a]	0.000
	A	−0.042	0.946
	C	−0.625	0.315
	G	−0.125	0.840

[a] = The mean difference is significant at the 0.05 level

In the second sample where there are 5 rows and 24 columns we got 4 pairs to be significant and all the 4 pairs are attached with the Gap Character (–) which doesn't has much value in our study. All other pairs of DNA's in the second sample are not significant. Here it concludes that the more we increase the dimension of our matrix, the pairs of DNA's will be not significant. Hence we may conclude that the pairs of DNA's with dimension 5 × 7473 will definitely be not significant.

Acknowledgements The author Miss. Anamika Dutta thank to Department of Science and Technology (DST), India for providing financial assistance for carrying out this work as an INSPIRE Fellow. Also we thank the reviewer for their thorough review and highly appreciate the comments and suggestions which substantially contributed to improving the class of the paper.

Appendix

The alignment of matrix (Hertz and Stormo 1999) has been shown with an example. Let us take some DNA sequences of different length say:

A – A C G T T C C
A C A C G T A C A
G C A A G A T – C
A C A C G T T C C

Gap character (–) come to view when ClustalX software is used. It happens due to multiple sequence alignment.

The above alignment has been created by ClustalX software. Now from the above DNA sequences, the alignment matrix can be formed which has been shown below:

$$\begin{bmatrix} \text{-} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ A & 3 & 0 & 4 & 1 & 0 & 1 & 1 & 0 & 1 \\ C & 0 & 3 & 0 & 3 & 0 & 0 & 0 & 3 & 3 \\ G & 1 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ T & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 \end{bmatrix}$$

Weight matrix using for the above example is given by:

$$\begin{bmatrix} \text{-} & -3.912 & -1.040 & -3.912 & -3.912 & -3.912 & -3.912 & -3.912 & -1.040 & -3.912 \\ A & 0.759 & -1.609 & 1.023 & -0.168 & -1.609 & -0.168 & -0.168 & -1.609 & -0.168 \\ C & -1.609 & 0.702 & -1.609 & 0.702 & -1.609 & -1.609 & -1.609 & 0.702 & 0.702 \\ G & 0.488 & -1.609 & -1.609 & -1.609 & 1.777 & -1.609 & -1.609 & -1.609 & -1.609 \\ T & -1.609 & -1.609 & -1.609 & -1.609 & -1.609 & 1.374 & 1.374 & -1.609 & -1.609 \end{bmatrix}$$

The highest weights of the above weight matrix are:

$$\begin{bmatrix} 0.759 & 0.702 & 1.023 & 0.702 & 1.777 & 1.374 & 1.374 & 0.702 & 0.702 \end{bmatrix}$$

Hence the score of the above matrix is:

$$0.759 + 0.702 + 1.023 + 0.702 + 1.777 + 1.374 + 1.374 + 0.702 + 0.702 = 9.115$$

This was a counter example of alignment and weight matrix.

References

Cho, Y. G., Ishii, T., Temnykh, S., Chen, X., Lipovich, L., McCouch, R. S., Park, D. W., Ayres, N., & Cartinhour, S. (2000). Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice. (*Oryza sativa L.*) *Theor Appl Genet*, *100*, 713–722. Springer-Verlag.

Hertz, Z. G., & Stormo, D. G. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics, 15*(7/8), 563–577.

Huck, W. S., & McLean, A. R. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*(4), 511–518.

Pei, J. (2008). Multiple protein sequence alignment. In *Current opinion in structural biology* (Vol. 18, pp. 382–386). Elsevier.

Shu, J. J., Yong, Y. K., & Chang, K. W. (2012). An improved scoring matrix for multiple sequence alignment. In *Mathematical problems in engineering* (Vol. 2012, no. 490649, pp. 1–9).

Mukhopadhyay, P. (2011). *Applied statistics*. Books and Allied (P) Ltd.

Wallace, M. I., Blackshields, G., & Higgins, G. D. (2005). Multiple sequence alignments. In *Current opinion in structural biology* (Vol. 15, p. 261–266). Elsevier.

Williams, J. L., & Abdi, H. (2010). Fisher's least significant difference (LSD) test. In N. Salkind (ed.), *Encyclopedia of research design* (pp. 1–6).

Regressions Involving Circular Variables: An Overview

Sungsu Kim and Ashis SenGupta

Abstract The last 40 years have seen a vigorous development of regression analysis involving circular data. A large body of results and techniques is now disseminated throughout the literature. In this paper, we provide a review of the literature on regressions involving circular variables that will be useful as a unified and up-to-date account of these methods for practical use. Examples and theoretical details are referred to corresponding papers herein, and omitted in our paper. Some of future topics of interest are also provided. Bayesian and non-parametric regression models involving a circular variable(s) are not included in this paper and will appear elsewhere.

Keywords Circular-circular regression · Circular-linear regression · Multiple circular regression · Linear-circular regression · Asymmetric generalized von Mises distribution

1 Introduction

Circular variables refer to random variables that are periodic in nature. Two main types of circular measurements are those of direction and time/day/ week/month. The early roots of circular data analysis reach back at least as far as the mid-18th Century. In 1767, the Reverend Jon Mitchell FRS analysed angular separations between stars. Circular variable analyses have appeared in numerous subject areas as diverse as Molecular/Structural Biology, Ecology, Econometrics, Environmental Sciences, Geography, Geology, Meteorology, Medicine, Oceanography, Physics, and Engineering. For example, geologists use circular variables to study paleocurrents in order to

S. Kim (✉)
Department of Mathematics, University of Louisiana at Lafayette, Lafayette, USA
e-mail: dr.sungsu@gmail.com

A. SenGupta
Applied Statistics Unit, Indian Statistical Institute, Kolkata, India
e-mail: amsseng@gmail.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_3

make inference about the direction of river flow in the past, and to study pole-reversal and continental drift (Fisher and Powell 1989). Typical of the problems of interest to biological scientists are those of bird navigation (Schmidt-Koenig 1963) and general orientations selected by particular creatures in response to experimental variation of their natural habitat (Stephens 1969; Wehner and Strasser 1985).

The problem of extracting practical information from circular data would appear to be tantalisingly close to the same problem for linear data, especially for concentrated data sets. Approximate linearity of a small arc would seem to justify application of linear methods and so make special treatment of circular data largely unnecessary. While linear approximations may solve ad hoc data analysis problems, they are not suitable for routine data processing. For example, it is easy to check that the usual arithmetic mean can not estimate the reference or circular mean direction for a sample of three points 359, 1° and 3°, which is called the crossover problem. Our readers can refer to various statistical methods developed for circular variables in Jammalamadaka and SenGupta (2001).

The last 40 years have seen a vigorous development of regression analysis involving a circular variable(s). A large body of results and techniques is now disseminated throughout the literature. This paper aims to present a unified and up-to-date account of theses methods for practical use. Applications in regression involving a circular variable(s) are ubiquitous in various disciplines. The Gould's (1969) paper is considered to be the earliest appearance of a regression with a circular response variable. He introduced a multiple regression model with a circular response variable and linear independent variables, which was fitted by the maximum likelihood (ML) estimation. Later, a number of papers argued that the model in Gould (1969) produces non-unique ML estimates (Johnson and Wehrly 1978; Fisher and Lee 1992). In Johnson and Wehrly (1978), a multiple regression model was presented as an improvement upon the Gould's model, using a completely specified cdf of a linear independent variable. Following this work, Lund (1999) presented a multiple regression model with a circular response and a combination of linear and circular predictors. Lund (1999) showed that the least circular distance estimators (LCDEs) are the same as ML estimates when the circular dependent variable is assumed to follow a von Mises distribution. Regression with a circular predictor and a linear response variables appeared in Mardia and Sutton (1978) and SenGupta and Kim (2016).

A regression involving both response and predictor as circular variables was discussed in Sarma and Jammalamadaka (1993), Downs and Mardia (2002), and SenGupta and Kim (2016). In Sarma and Jammalamadaka (1993), a circular regression model is translated into a linear regression model where sine and cosine functions of dependent circular variable were regressed on an analogous set of trigonometric functions of independent circular variable. Downs and Mardia (2002) used the tangent-tangent mean direction link, where the range of circular variables is shrunk from 2π to π by using the half tangent function in order to avoid issues resulting from many to one mappings. SenGupta and Kim (2016) extended the Downs and Mardia (2002) model by introducing an intercept parameter to add flexibility.

A multiple circular regression having more than one circular and/or linear independent variables appeared in Gould (1969), Johnson and Wehrly (1978), Fisher and Lee 1992, Sarma and Jammalamadaka (1993), Lund (1999), SenGupta and Ugwuowo (2006), and Kim and SenGupta (2016). Two multivariate multiple regressions involving circular response variables and circular independent variables were suggested in Kim and SenGupta (2016). Inverse regressions involving a circular variable(s) were discussed in SenGupta et al. (2013) and Kim and SenGupta (2016).

Circular distributions encountered in practice are usually asymmetric (SenGupta and Ugwuowo 2006). In fact, this case is also addressed in linear statistical analysis (Arnold and Beaver 2000). As alternatives to the circular normal distribution, asymmetric circular distributions applied in regression involving a circular response variable appeared in Pewsey (2000), Abe and Pewsey (2011), SenGupta et al. (2013), and Kim and SenGupta (2016).

In this paper, we will use the following notation for circular regression: we write a dependent variable first then followed by an independent variable(s). In the cases where a model was proposed to model more than one independent variable, they are listed without an order. For example, a circular-circular regression refers to having both circular dependent and circular independent variables; a circular-linear regression to having a circular response and linear predictor variables; and a circular-circular-circular regression to having a circular dependent and two circular independent variables. In our review, we used the same notations appeared in the original references. Throughout the paper, angles are measured in radians and assume values in the interval $(0, 2\pi]$.

2 Circular-Linear Regression Models

Let (x, θ) be a cylindrical random variable. Johnson and Wehrly (1978) introduced an alternative to the Gould's approach (1969), which avoids the drawbacks in his model. Their model, based on the conditional distribution of one of their proposed angular-linear distributions is given by

$$\theta = \mu + 2\pi F(x) + \epsilon,$$

where ϵ follows a centered von Mises distribution, and $f(x)$ is completed specified with cdf F.

Kim and SenGupta (2015) proposed the following link function for mean direction:

$$\mu_{\theta|x} = \mu_\theta + 2\arctan(\alpha + \beta x),$$

which was employed in the problem of inverse circular-linear regression in their paper.

3 Circular-Circular Regression Models

Let (ϕ, θ) be a bivariate circular random variable. As one of the earliest models appeared in the literature, Sarma and Jammalamadaka (1993) proposed the following link functions for mean direction, using trigonometric polynomials of degree m:

$$\cos(\mu_{\theta|\phi}) = \sum_{k=0}^{m}(A_k \cos(k\phi) + B_k \sin(k\phi))$$

$$\sin(\mu_{\theta|\phi}) = \sum_{k=0}^{m}(C_k \cos(k\phi) + D_k \sin(k\phi)),$$

for a suitable choice of m.

Downs and Mardia (2002) proposed the following link function for mean direcrion:

$$\mu_{\theta|\phi} = \mu_\theta + 2\arctan\left\{\beta \tan\left(\frac{\phi - \mu_\phi}{2}\right)\right\}. \tag{1}$$

Note that arctangent has double solutions in $[0, 2\pi)$, but by restricting to half-angles, an one-to-one mapping between θ and ϕ is found, provided that β is not equal to zero.

As an extension to Downs and Mardia (2002), SenGupta and Kim (2016) proposed the following link function for mean direction:

$$\mu_{\theta|\phi} = \mu_\theta + 2\arctan\left\{\alpha + \beta \tan\left(\frac{\phi - \mu_\phi}{2}\right)\right\}. \tag{2}$$

Using the link function in Downs and Mardia (2002), it is assumed that the conditional mean direction value of θ is its unconditional mean direction μ_θ, i.e. $\mu_{\theta|\phi} = \mu_\theta$, when the value of ϕ is its unconditional mean direction μ_ϕ. However, this is not always obviously appropriate and therefore, needs not to be generally assumed. Then α has a roll of adding the rotation from μ_θ, by the amount of $2\arctan(\alpha)$, when conditioning on $\phi = \mu_\phi$.

4 Linear-Circular Regression Models

Let (θ, x) be a cylindrical random variable. As one of the earliest models appeared in the literature, Mardia and Sutton (1978) proposed the following model:

$$E(x|\theta) = b_0 + b_1 \cos\theta + b_2 \sin\theta, \quad V(x|\theta) = \sigma^2(1 - \rho^2),$$

where $b_0 = \mu - b_1 \cos\mu_0 - b_2 \sin\mu_0, b_1 = \sigma\kappa^{\frac{1}{2}}\rho_1$ and $b_2 = \sigma\kappa^{\frac{1}{2}}\rho_2$. The parameters μ, μ_0, κ, ρ_1, ρ_2 and σ are from the cylindrical distribution proposed in their paper. The second order model which is a natural extension of the Mardia and Sutton (1978) first order model was presented in Anderson-Cook (2000).

Kim and SenGupta (2015) proposed the following link function for mean:

$$\mu_{x|\theta} = \alpha + \beta\cos(\theta' - \mu_{\theta'}),$$

where $\theta' = \frac{\theta+\pi}{2}$. The model was employed in the problem of inverse linear-circular regression in their paper.

5 Multivariate Multiple Circular Regression Models

The following regression model was proposed by Gould (1969):

$$\theta = \mu_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon,$$

where θ is a circular response variable, x_j's are linear predictor variables, and ϵ follows a circular distribution. His model was criticized for having infinitely many equally large peaks in the likelihood function.

Fisher and Lee (1992) also proposed the following link function for mean direction, as an alternative to the Gould's model (1969):

$$\mu = \mu_0 + g(\beta' x),$$

where $g(x) = 2\tan^{-1}(sgn(x)|x|^{\lambda})$ and λ can be estimated from the data, analogously to the estimation of Box-Cox transformation. When the covariate lie in a bounded region, the following link function was suggested:

$$\mu = \mu_0 + 2\pi g(x),$$

where $g(x)$ is a member of some flexible parametric family of $k-$dimensional distribution.

Combining two models proposed in Fisher and Lee (1992) and Sarma and Jammalamadaka (1993), Lund (1999) proposed the following link function for mean direction:

$$\mu = g_1(\phi, \beta_1) + g_2(\beta_2' x),$$

where

$$g_1(\phi, \beta_1) = \sum_{k=0}^{m} (A_k \cos(k\phi) + B_k \sin(k\phi))$$
$$g_2(\beta_2' x) = 2 \arctan(\beta_2' x).$$

However, it should be noted that $g_1(\phi, \beta_1)$ does not map the unit circle onto the unit circle.

In Johnson and Wehrly (1978), a multivariate linear-linear-circular regression model was proposed based on their joint distribution of $(\theta_1, \ldots, \theta_p, x_1, \ldots, x_q)$, where each linear response component $x_i, i = 1, \ldots, r(< q)$, has the following mean link:

$$x_i = \nu_0 + \sum_{i=r+1}^{q} \nu_i x_i + \sum_{i=r+1}^{q} \sum_{j=1}^{p} \sum_{n=1}^{k} \left(\gamma_{ijk} \cos(k\theta_j) + \delta_{ijk} \sin(k\theta_j) \right),$$

the best Fourier series of nth degree was used for the individual θ's. Their multivariate model is considered as an extension of the model for cylindrical variables proposed in Mardia and Sutton (1978).

SenGupta and Ugwuowo (2006) proposed the following linear-linear-circular models.

$$Y_i = M + \sum_{j=1}^{k} \beta_j x_{ji} + A \cos \omega(t_i - \phi) + \epsilon_i,$$

where Y is the linear response variable, M is the mean level, β_js are the regression coefficients, x_j's are the linear independent variables, A is the amplitude, ω is the angular frequency, t is the circular independent variable (usually time) subject to a certain period T, ϕ is the acrophase and ϵ_i is the random error component. When the peaks and troughs do not follow each other, it implies that there is a skew and they proposed the following non-linear model:

$$Y_i = M + \sum_{j=1}^{k} \beta_j x_{ji} + A \cos(\psi_i + \mu \sin \psi) + \epsilon_i,$$

where $\psi = \omega t_i - \phi$ and μ is the parameter of skewness. When the oscillations are sharply peaked or flat-topped, they proposed

$$Y_i = M + \sum_{j=1}^{k} \beta_j x_{ji} + A \cos(\psi_i + \nu \cos \psi) + \epsilon_i,$$

where ν is the parameter of kurtosis.

Kim and SenGupta (2016) proposed the following circular-circular-circular arctangent link functions:

$$\tan\left(\frac{\theta-\mu_\theta}{2}\right)=\alpha+\beta_1\tan\left(\frac{\phi_1-\mu_{\phi_1}}{2}\right)+\beta_2\tan\left(\frac{\phi_2-\mu_{\phi_2}}{2}\right)$$

$$\tan\left(\frac{\theta-\mu_\theta}{2}\right)=\alpha+\beta_1\sin\left(\frac{\phi_1-\mu_{\phi_1}}{2}\right)+\beta_2\sin\left(\frac{\phi_2-\mu_{\phi_2}}{2}\right).$$

A guideline for choosing between two models was given in their paper. Two multivariate multiple regressions involving circular variables were suggested in Kim and SenGupta (2016) as an extension to the multiple circular models in the same paper.

6 Asymmetric Circular Probability Distributions

Circular data encountered in practice are usually asymmetric (SenGupta and Ugwuowo 2006). In fact, this case is also addressed in linear statistical analysis (Arnold and Beaver 2000). As alternatives to the circular normal (CN) distribution, asymmetric circular distributions applied in regression involving a circular response variable appeared in Pewsey (2000), Abe and Pewsey (2011), SenGupta et al. (2013), and Kim and SenGupta (2016).

Umbach and Jammalamadaka (2009) discussed a method of introducing asymmetry into any symmetric circular model and developed general classes of nonsymmetric circular distributions, which included a resulting variation of the classical von Mises distribution. Pewsey (2000) presented the wrapped skew normal distribution by wrapping the skew normal distribution (Azzalini 1985) on the unit circle, where he proposed the method of moment for estimating the parameters. Abe and Pewsey (2011) proposed the sine-skewed family of circular distributions, which is a special case of the construction due to Umbach and Jammalamadaka (2009). Then, the sine-skewed Jones-Pewsey distribution is introduced as a particularly flexible model of this type. Kim and SenGupta (2012) proposed an asymmetric circular distribution called the asymmetric generalized von Mises (AGvM) distribution, which is flexible to model asymmetric or bimodal circular data. The density of AGvM is shown here:

$$f_\Theta(\theta)=\frac{\exp\left[\kappa_1\cos(\theta-\mu)+\kappa_2\sin 2\{\theta-(\mu-\delta)\}\right]}{\int_{-\pi}^{\pi}\exp\left[\kappa_1\cos(\theta-\mu)+\kappa_2\sin 2\{\theta-(\mu-\delta)\}\right]d\theta},$$

where $\mu\in(0.2\pi]$ is a location parameter, $\delta\in[0,2\pi)$, $\kappa_1\in R$ and $\kappa_2\in R$ are shape parameters, related to concentration and skewness, respectively.

Asymmetric distributions can occur in situations when the observed variables represent a sample that has been truncated with respect to some hidden or available covariable (Arnold and Beaver 2000), where the underlying joint distribution is symmetric. Kim (2009) applied the hidden truncation method to symmetric circular bivariate distributions, such as the generalized circular normal conditional density, the wrapped bivariate normal density and the wrapped bivariate cauchy density.

7 Future Interesting Topics in Directional Regression Models

Directional data refer to measurements in a smooth manifold. Although regression analysis and applications involving spherical variables have appeared in the literature, spherical-circular, spherical-linear and circular-spherical regressions are not found in the literature at our best knowledge. With numerous possible applications of these models in various disciplines, we think that developing statistical models of such is highly in demand.

References

Abe, T., & Pewsey, A. (2011). Sine-skewed circular distributions. *Statistical Papers*, *52*, 683–707.

Anderson-Cook, C. M. (2000). A second order model for cylindrical data. *Journal of Statistical Computation and Simulation*, *66*, 51–65.

Arnold, B. C., & Beaver, R. J. (2000). Hidden truncation models. *The Indian Journal of Statistics: Series A*, *62*, 23–35.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178.

Downs, T. D., & Mardia, K. V. (2002). Circular regression. *Biometrika*, *89*, 683–698.

Fisher, N. I., & Lee, A. J. (1992). Regression models for an angular response. *Biometrics*, *48*, 665–677.

Fisher, N. I., & Powell, C. McA. (1989). Statistical analysis of two-dimensional palaeocurrent data: Methods and examples. *Austrian Journal of Earth Sciences*, *36*, 91–107.

Gould, A. L. (1969). A regression technique for angular variates. *Biometrics*, *25*, 683–700.

Jammalamadaka, S., & SenGupta, A. (2001). *Topics in circular statistics*. New York: World Scientific.

Johnson, R. A., & Wehrly, T. E. (1978). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, *66*, 255–256.

Kim, S. (2009). Inverse circular regression with possibly asymmetric error distribution. Ph.D. Dissertation. University of California, Riverside.

Kim, S., & SenGupta, A. (2012). A three-parameter generalized von Mises distribution. *Statistical Papers*, *54*, 685–693.

Kim, S., & SenGupta, A. (2015). Inverse circular-linear/linear-circular regression. *Communications in Statistics: Theory and Methods*, *44*, 4772–4782.

Kim, S., & SenGupta, A. (2016). Multivariate and multiple circular regression. *Journal of Statistical Computation and Simulation*, *87*, https://doi.org/10.1080/00949655.2016.1261292.

Lund, U. (1999). Least circular distance regression for directional data. *Journal of Applied Statistics*, *26*, 723–733.

Mardia, K. V., & Sutton, T. W. (1978). A model for cylindrical variables with applications. *Journal of Royal Statistical Society: Series B*, *40*, 229–233.

Pewsey, A. (2000). The wrapped skew-normal distribution on the circle. *Communications in Statistics: Theory and Methods*, *29*, 2459–2472.

Sarma, Y. R., & Jammalamadaka, S. (1993). Circular regression. In *Proceedings of the Third Pacific Asia Statistical Conference* (pp. 109–128).

Schmidt-Koenig, K. (1963). On the role of the loft, the distance and the site of release in pigeon homming. *Biological Bulletin*, *125*, 154–164.

SenGupta, A., Kim, S., & Arnold, B. C. (2013). Inverse circular-circular regression. *Journal of Multivariate Analysis*, *119*, 200–208.

SenGupta, A., & Kim, S. (2016). Statistical inference for homologous gene pairs between two circular genomes: A new circular-circular regression model. *Statistical Methods and Applications*, *25*, 421–432.

SenGupta, A., & Ugwuowo, F. I. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics*, *13*, 299–309.

Stephens, M. A. (1969). Tests for the von Mises distribution. *Biometrika*, *56*, 149–160.

Umbach, D., & Jammalamadaka, S. R. (2009). Building asymmetry into circular distributions. *Statistics and Probability Letters*, *79*, 659–663.

Wehner, R., & Strasser, S. (1985). The POL area of the honey bee's eye: Behavioural evidence. *Physiological Entomology*, *10*, 337–349.

On Construction of Prediction Interval for Weibull Distribution

Ramesh M. Mirajkar and Bhausaheb G. Kore

Abstract In this article we have proposed the simple method of construction of an exact Prediction Interval (PI) for a single future observation from Weibull distribution. The method is based on a proposed simple pivotal statistic, assuming that the Weibull shape parameter is known. A simulation study is carried out by MATLAB, R 2012a. A simulation based comparison reveals that the proposed PI has smallest expected lengths for smaller shape parameter and percentage coverage is round about 95% than the existing ones for all sample sizes. Furthermore, it is computationally much simpler than most existing methods, which is an added advantage for non-statistical users. Application of the proposed PI to real data set is presented.

Keywords Coverage probability · Expected length · Prediction interval
Weibull distribution

1 Introduction

The Weibull distribution is widely used for modeling variability in almost all disciplines, and more effectively in reliability analysis, environmental sciences and water resource management among others. The probability density function of a Weibull random variable X is,

$$\begin{aligned} f(x;\ a, b) &= \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} \exp\left[-\left(\frac{x}{a}\right)^{b}\right], \quad x > 0, \quad a > 0, \quad b > 0, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

R. M. Mirajkar (✉)
Dr. Babasaheb Ambedkar College, Peth Vadgaon, 416112 Kolhapur, Maharashtra, India
e-mail: rmm2stats@gmail.com

B. G. Kore
Department of Statistics, Balwant College, Vita, 415311 Sangli, Maharashtra, India
e-mail: korebg2005@yahoo.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_4

Where ‘a’ and ‘b’ are scale and shape parameters of the Weibull distribution respectively. For $b < 1\,(> 1)$ the hazard function of the Weibull distribution is decreasing (increasing) and is constant for $b = 1$, which corresponds to the well-known exponential distribution. This being a scale-shape family, the distribution of $(X - a)^b$ is parameter free, namely, the standard exponential distribution.

In this work we attempt to develop a prediction interval (PI) for a single future observation from a Weibull distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Weibull (a, b) distribution, where initially we assume the shape parameter ‘b’ is to be known. Let $X_{(n)}$ be the nth order statistics, i.e. the largest among X_i’s and let X denotes a future independent observation from the same distribution. The proposed PI are based on the fact that the statistic $(X/X_{(n)})^b$ is distribution free, as a consequence of Weibull being a scale-shape family. This PI is compared on the basis of expected length and expected coverage probability of PI with other existing methods. The study reveals that the proposed method outperforms the existing methods.

Section 2 gives the review of existing methods in the literature for PI for a future single observation from a Weibull distribution. In Sect. 3 we develop the proposed PI for known shape parameter. Section 4 presents details of the comparison among these methods based on a simulation study. Section 5 illustrates the methods to real data sets. Concluding remarks are summarized in Sect. 6.

2 Existing Methods

In this section we take review of the existing methods in the literature of PI for a future observation from a Weibull distribution. There are three methods of construction of PI for a single future observation from Weibull distribution such as;

(1) Fertig et al. (FMM) (1980) have proposed the pivotal quantity (based on the fact that the extreme value distribution is a location-scale family) the statistic, $Q = (\bar{u} - w)/\bar{d}$ for constructing prediction limits for the smallest of a set of k future observations from a Weibull or extreme-value distribution, where $\bar{u}$ and $\bar{d}$ are best linear invariant estimators (BLIE). For $k = 1$, $(1 - \alpha)\,100\%$ PI given by [L, U] for a future observation z from a Weibull distribution as,

$$[L,\quad U] = \left[\exp\left(\bar{u} - Q_{1-\alpha/2}\bar{d}\right),\ \exp\left(\bar{u} - Q_{\alpha/2}\bar{d}\right)\right], \tag{1}$$

where Q_α is the $100\,\alpha$ th percentile of the distribution of the statistic Q.

(2) Engelhardt and Bain (EBM) (1982) have proposed a statistic $T = (\overline{u^*} - w)/\overline{d^*}$ for construction of prediction limits for the smallest of a set of k future observations from a Weibull or extreme-value distribution where $\overline{u^*}$ and $\overline{d^*}$ are best linear unbiased estimators (BLUE) or simplified BLUE’s of the location parameter u = log(a) and the scale parameter d = 1/b of the type-I extreme value distribution (for minimum) respectively, and w is a future observation from the

same extreme value distribution. For k = 1, $(1-\alpha)100\%$ PI given by [L, U] for a future observation z from a Weibull distribution as,

$$[L,\ U] = \left[\exp\left(\bar{u} - T_{1-\alpha/2}\overline{d^*}\right),\ \exp\left(\bar{u} - T_{\alpha/2}\overline{d^*}\right)\right] \tag{2}$$

(3) Yang et al. (YSXM) (2003) have suggested a PI for single future observation by using Box-Cox transformation. Using normal distribution based PI on C and transforming back to Weibull, the resultant PI for a Weibull future observation z is given by

$$\left\{1 + \lambda\left(\bar{c}\, t_{n-1,1-\alpha/2}\, s_c\sqrt{1 + 1/n}\right)\right\}^{1/\lambda} \tag{3}$$

where $\bar{c}$ and s_c are sample mean and standard deviation of the transformed vector C using Box-Cox transformation on x and $t_{m,\alpha}$ is the α^{th} quantile of students t distribution with m degrees of freedom. When the shape parameter 'b' is unknown, one can replace the parameter, in (3) by $\hat{\lambda} = 0.2654\,\hat{b}$, where $\hat{b}$ is m.l.e. of 'b'.

3 The Proposed Prediction Interval (PI)

3.1 PI for Known Shape Parameter

Let X_1, X_2, …,X_n be an observed random sample from a W(a, b) distribution. We assume that the shape parameter 'b' is known. Let $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$ and let Y denote an independent future observation from the same population. Then as mentioned above, as a consequence of shape-scale property of the Weibull distribution, the statistic $T = \left(\frac{Y}{X_{(n)}}\right)^{\hat{b}}$ is free from any parameters and hence is a pivot quantity. We further note that T has the same distribution as that of $Z/W_{(n)}$, where Z is an independent future observation and $W_{(n)}$ is the nth order statistics based on a standard exponential distribution. Noting that Z and $W_{(n)}$ are independently distributed with densities $f_Z(z) = e^{-z},\ z > 0$ and $f_{w(n)} = n\left[1 - e^{-w}\right]^{n-1} e^{-w},\ w > 0$.

Then the distribution function of $T = \frac{Z}{W_{(n)}}$ is given by

$$F_T(t) = P[T \le t] = P\left[\frac{Z}{W_{(n)}} \le t\right] = P\left[\frac{W_{(n)}}{Z} \ge \frac{1}{t}\right]$$

$$= \int_0^\infty P\left[W_{(n)} > \frac{Z}{t}/Z = z\right] e^{-z} dz$$

$$= \int_{0}^{\infty} \left\{1 - \left[1 - e^{-z/t}\right]^{n}\right\} e^{-z} dz$$

$$= 1 - t\,\beta(t,\ n + 1)$$

Equating this to α and solving for ‘t’ this leads to the expression for the αth quantile of the distribution of T given by

$$t_{\alpha}\,\beta(t_{\alpha},\ n + 1) = 1 - \alpha$$

It follows that $P\left[t_{\alpha} < \left(\frac{Y}{X_{(n)}}\right)^{b} < t_{1-\alpha}\right] = 1 - \alpha$
giving a $(1 - \alpha)$ 100% PI for a future observation Y to be

$$(L,\ U) = \left(t_{\alpha}^{1/b} X_{(n)},\ t_{1-\alpha}^{1/b} X_{(n)}\right) \tag{4}$$

3.2 PI for Unknown Shape Parameter

In practice the shape parameter ‘b’ is unknown and has to be replaced by an appropriate estimator, such that $\hat{b}$, the m.l.e. Note that the $(1 - \alpha)$100% PI for a single future observation from Weibull (a, b) is given by

$$\left(\widehat{L},\ \widehat{U}\right) = \left\{\left(t_{\alpha}^{1/\hat{b}} X_{(n)},\ t_{1-\alpha}^{1/\hat{b}} X_{(n)}\right)\right\} \tag{5}$$

4 Comparison Among the Methods

We have developed MATLAB for computation of all the above PI’s for a future Weibull observation. We compare the above mentioned four methods with respect to the expected lengths and coverage probabilities based on a simulation study.

4.1 Simulation Study

The comparison of the methods is attempted based on the simulated coverage probabilities and expected lengths of PI, for various fixed sample sizes n and α =0.05. For the values of the scale parameter a =1, 5, 10 and shape parameter b =0.5, 1, 3.5, 5 and for each of the fixed sample sizes n =6, 10, 15, 20, 30; 5000 simulated samples are generated from Weibull (a, b) distribution. For each of these 5000 samples, lower and upper prediction limits, namely, L_i and U_i, $i = 1, 2, \ldots, 5000$; for each of the four methods given in Eqs. 1–3 and 4 are obtained. Finally, the simulated expected lengths of PI obtained by averaging the 5000 quantities $(U_i - L_i)$; i = 1, 2,…, 5000 and the expected coverage which is proportion of the intervals that covered the simulated single future observation from the same Weibull (a, b) are computed.

4.2 Results of Simulation Study

Table below shows the percentage simulated coverage probabilities and expected lengths for 95% PI for all four above mentioned methods, for sample sizes $n = 6$, 10, 15, 20, 30.

The following prominent facts are clearly visible from the above Tables

1. When the shape parameter is unknown, for all sample sizes, proposed method has excellent coverage probabilities and uniformly outperform the FMM, EBM and YSX method.
2. The proposed method has uniformly smaller expected lengths than FMM, EBM and YSX method when shape parameter is small and hence exhibits the best performance.
3. Percentage coverage of proposed method is round about 95% as compared to other methods.
4. FMM and EBM have equivalent performance, for all sample sizes.
5. For all sample sizes, coverage of YSXM is more than 95%.

5 Application

In this section we analyze a real data set extracted from Yadav et al. (2010) representing inter-occurrence times between successive earthquakes (in days) that occurred in Northeast India and adjoining region of magnitude M > 7.0 which are listed below: 7872.0, 6029.0, 4545.45, 4200.54, 1258.10, 2237.34, 1889.50, 2697.91, 4652.16, 1055.64, 320.93, 1113.00, 460.81, 854.58, 2512.38, 5593.21, 4452.44, 882.60, and 1648.27.

Application of the Kolmogorov-Smirnov test to above data set for fitting Weibull distribution gave p-value 0.762 indicating that Weibull distribution is a good model for the above data set. Here the m.l.e.'s of scale and shape parameter based on first 18 observations are $\hat{a} = 3190.6$ and $\hat{b} = 1.3$ respectively.

The 95% PIs and their length using the above four methods based on the first 18 observations are as follows.

We conclude that the next earthquake is expected to occur in the duration of 07th Oct. 1991 to 19th Sep. 2015. Here we note that the actual earthquake (19th observation) has occurred on 11th July 1995, which falls within each interval.

Table 1 Simulated expected lengths along with percentage coverage for 95% PI for a future observations from Weibull (a, b) for $n = 6$; for scale parameter $a = 1, 5, 10$ and shape parameter $b = 0.5, 1, 3.5, 5$

n = 6					
a	b	Proposed method	FMM	EBM	YSXM
1	0.5	39.0312(94.84)	174.0279(95.22)	174.0271(94.84)	90.5181(96.83)
	1	5.1437(95.07)	8.6613(95.45)	8.6575(95.07)	7.186(96.98)
	3.5	1.5346(95.10)	1.4614(94.84)	1.444(94.42)	1.6681(97.07)
	5	1.345(94.92)	1.0879(95.15)	1.0699(94.74)	1.4251(96.84)
5	0.5	195.4513(94.99)	789.2778(94.98)	789.2738(94.53)	453.726(96.88)
	1	25.6719(94.93)	47.1458(95.29)	47.1247(94.91)	35.9462(96.68)
	3.5	7.6791(94.92)	7.1795(94.82)	7.0943(94.46)	8.0578(96.51)
	5	6.7198(94.91)	5.3309(94.86)	5.2414(94.40)	6.3709(96.35)
10	0.5	395.1686(94.94)	1462.4154(95.16)	1462.4079(94.74)	909.5245(96.84)
	1	51.8913(94.89)	80.4486(94.85)	80.4064(94.44)	72.48(96.71)
	3.5	15.3336(94.99)	14.7178(94.83)	14.5445(94.46)	15.7082(96.38)
	5	13.4515(95.05)	10.2748(94.00)	10.0964(93.58)	12.3143(96.28)

Table 2 Simulated expected lengths along with percentage coverage for 95% PI for a future observations from Weibull (a, b) for $n = 10$; for scale parameter $a = 1, 5, 10$ and shape parameter $b = 0.5, 1, 3.5, 5$

n = 10					
a	b	Proposed method	FMM	EBM	YSXM
1	0.5	19.978(95.08)	36.4241(94.53)	36.4236(94.12)	34.0322(97.41)
	1	4.0595(95.20)	5.3648(95.36)	5.3617(95.09)	5.2349(97.44)
	3.5	1.4609(95.04)	1.2685(95.02)	1.2549(94.67)	1.5687(97.40)
	5	1.3007(94.94)	0.9283(94.68)	0.9148(94.34)	1.3677(97.23)
5	0.5	101.3233(95.06)	174.1315(95.13)	174.129(94.86)	171.8366(97.18)
	1	20.3089(95.14)	28.0428(94.93)	28.0244(94.56)	26.1867(97.25)
	3.5	7.3062(95.16)	6.4436(95.58)	6.3769(95.30)	6.8637(96.07)
	5	6.5054(95.18)	4.6842(95.04)	4.616(94.71)	5.0329(96.02)
10	0.5	201.9258(95.01)	440.5799(95.19)	440.5755(94.87)	343.7446(97.33)
	1	40.5494(95.09)	54.2212(95.06)	54.1868(94.76)	52.2705(97.18)
	3.5	14.5977(95.07)	12.8111(95.30)	12.6755(94.94)	13.2387(95.84)
	5	13.0137(95.16)	9.3917(94.95)	9.2556(94.64)	9.9751(96.07)

Table 3 Simulated expected lengths along with percentage coverage for 95% PI for a future observations from Weibull (a, b) for $n = 15$; for scale parameter $a = 1, 5, 10$ and shape parameter $b = 0.5, 1, 3.5, 5$

n = 15					
a	b	Proposed method	FMM	EBM	YSXM
1	0.5	15.5877(95.01)	24.9322(95.07)	24.932(94.82)	24.5453(97.53)
	1	3.6755(95.09)	4.6209(95.66)	4.6187(95.42)	4.6575(97.54)
	3.5	1.4289(95.11)	1.2005(95.08)	1.1904(94.86)	1.5305(97.57)
	5	1.2821(95.15)	0.8872(94.82)	0.8773(94.59)	1.3453(97.59)
5	0.5	77.8314(95.05)	118.8113(94.86)	118.8102(94.67)	122.8556(97.58)
	1	18.3236(95.01)	23.065(94.95)	23.0518(94.71)	23.245(97.52)
	3.5	7.1428(94.83)	5.9534(94.88)	5.9029(94.62)	6.3383(95.51)
	5	6.4113(95.01)	4.4022(94.69)	4.3529(94.43)	4.6519(95.74)
10	0.5	155.665(94.99)	252.6155(95.61)	252.6136(95.40)	244.4971(97.58)
	1	36.7461(95.01)	46.4001(94.86)	46.373(94.62)	46.4019(97.39)
	3.5	14.28(95.02)	12.1431(95.13)	12.0425(94.43)	12.4205(95.77)
	5	12.823(94.84)	9.0532(95.34)	8.9533(95.40)	9.31(95.66)

Table 4 Simulated expected lengths along with percentage coverage for 95% PI for a future observations from Weibull (a, b) for $n = 20$; for scale parameter $a = 1, 5, 10$ and shape parameter $b = 0.5, 1, 3.5, 5$

n = 20					
a	b	Proposed method	FMM	EBM	YSXM
1	0.5	13.8978(94.94)	20.7733(95.25)	20.7732(95.04)	21.1694(97.56)
	1	3.5237(95.02)	4.2768(94.33)	4.2744(94.11)	4.3936(97.60)
	3.5	1.4154(94.73)	1.1683(94.85)	1.1603(94.65)	1.5124(97.47)
	5	1.2741(94.90)	0.8633(94.81)	0.8555(94.62)	1.3345(97.62)
5	0.5	69.4052(94.92)	103.3386(94.93)	103.3378(94.76)	105.7852(97.59)
	1	17.5808(95.15)	20.4782(95.33)	20.4696(95.17)	21.9725(97.67)
	3.5	7.0858(95.01)	5.7693(94.54)	5.7294(94.34)	6.1007(95.54)
	5	6.3729(94.96)	4.2903(94.56)	4.2513(94.38)	4.5031(95.45)
10	0.5	138.7374(94.94)	230.3449(95.10)	230.3431(94.86)	211.3183(97.60)
	1	35.1818(95.22)	42.8391(95.10)	42.8196(94.91)	43.9817(97.74)
	3.5	14.1725(95.07)	11.6954(94.93)	11.6157(94.76)	12.0752(95.53)
	5	12.7475(94.91)	8.6213(94.67)	8.544(94.47)	9.0097(95.60)

Table 5 Simulated expected lengths along with percentage coverage for 95% PI for a future observations from Weibull (a, b) for $n = 30$; for scale parameter $a = 1, 5, 10$ and shape parameter $b = 0.5, 1, 3.5, 5$

n = 30					
a	b	Proposed method	FMM	EBM	YSXM
1	0.5	12.8091(95.16)	17.272(94.96)	17.2719(94.83)	18.5104(97.66)
	1	3.4165(94.91)	4.1639(94.82)	4.1623(94.63)	4.1829(97.68)
	3.5	1.4082(94.90)	1.1524(95.07)	1.1467(94.90)	1.4969(97.67)
	5	1.27(94.99)	0.8557(95.13)	0.8502(94.95)	1.3252(97.64)
5	0.5	63.5879(95.12)	81.5368(94.39)	81.5362(94.26)	92.3134(97.64)
	1	17.1271(94.93)	19.8703(94.82)	19.8631(94.70)	20.9407(97.70)
	3.5	7.0378(94.86)	5.7011(94.84)	5.6724(94.66)	5.8812(95.36)
	5	6.3473(95.15)	4.3317(95.25)	4.3038(95.13)	4.37(95.63)
10	0.5	128.584(95.18)	154.9903(94.45)	154.9892(94.30)	185.4985(97.75)
	1	34.1772(94.86)	40.486(94.65)	40.4701(94.51)	41.8427(97.50)
	3.5	14.0815(94.83)	11.1093(94.08)	11.052(93.94)	11.7379(95.36)
	5	12.694(94.99)	8.4734(94.91)	8.418(94.78)	8.7347(95.50)

Table 6 Computation of PI

Method	95% PI	Length of PI
Proposed method	[264.6, 7861.4]	7596.7
FMM	[197.9, 8955.8]	8757.9
EBM	[208.3 8433.9]	8225.5
YSXM	[115.7, 9578.4]	9462.6

6 Conclusion

From Tables 1, 2, 3, 4 and 5 we come across that the proposed method gives smaller expected length when shape parameter is small, for all sample sizes. We conclude that the proposed PI method has uniformly better performance than the existing known methods such as FMM, EBM, and YSXM. Proposed method is based on pivotal approach which avoids complicated and lengthy calculations which finds in known methods. From real data set we observed that the earthquake event has occurred in the expected length (Table 6).

Acknowledgements Authors are highly grateful to the learned referee for his valuable suggestions to improve the quality of this paper.

References

Engelhardt, M., & Bain, L. J. (1982). On prediction limits for samples from a Weibull or extreme-value distribution. *Technometrics, 24,* 147–150.

Fertig, K. W., Meyer, M. E., & Mann, N. R. (1980). On constructing prediction intervals for samples from a Weibull or extreme value distribution. *Technometrics, 22,* 567–573.

Yadav, R. B., Tripati, J. N., Rastogi, B. K., Das, M. C., & Chopra, (2010). Probabilistic assessment of earthquake recurrence in northeast india and adjoining region. *Pure and Applied Geophysics, 167,* 1331–1342.

Yang, Z. L., See, S. P., & Xie, M. (2003). Transformation approaches for the construction of Weibull prediction interval. *Computational Statistics and Data Analysis, 43,* 357–368.

Combining High-Dimensional Classification and Multiple Hypotheses Testing For the Analysis of Big Data in Genetics

Thorsten Dickhaus

Abstract We present the so-called COMBI method for evaluating genome-wide association studies which we have developed in prior work. In contrast to traditional locus-by-locus analyses, COMBI is a multivariate procedure which takes dependencies between different genomic loci into account. This is done by combining methods from machine learning and multiple testing. In a first stage of data analysis, a support vector machine (which is an inherently multivariate classification method) is trained. In a second stage, only the genomic positions with the largest contributions to the resulting classification rule are explicitly tested for association with the phenotype of interest, yielding a drastic dimension reduction. The thresholding of the association p-values for the selected positions is performed by means of a resampling procedure. Some remarks on the performance and on software implementations of COMBI are made.

Keywords COMBI method · Genome-wide association studies
Machine learning · p-value · Statistical genetics · Support vector machine

1 Introduction and Motivation

In genetic association studies, associations between a (potentially very large) set of genetic markers and a phenotype of interest are analyzed. For concreteness, we consider here only binary phenotypes (e.g., disease indicators), although our approach can straightforwardly be extended to categorical phenotypes with more than two categories. This setup results in a particular multiple test problem which has several challenging aspects, for instance the high dimensionality of the statistical parameter and the discreteness of the statistical model; cf. Chap. 9 of Dickhaus (2014) and the references therein for further details. Assuming that m bi-allelic single nucleotide polymorphisms (SNPs, corresponding to genomic loci) are simultaneously under

T. Dickhaus (✉)
Institute for Statistics, University of Bremen, P.O. Box 330 440, 28344 Bremen, Germany
e-mail: dickhaus@uni-bremen.de

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_5

Table 1 Schematic representation of data for an association test problem at genetic locus j, where the two possible alleles are denoted by $A_1^{(j)}$ and $A_2^{(j)}$.

Genotype	$A_1^{(j)}A_1^{(j)}$	$A_1^{(j)}A_2^{(j)}$	$A_2^{(j)}A_2^{(j)}$	$\sum$
Cases	$x_{11}^{(j)}$	$x_{12}^{(j)}$	$x_{13}^{(j)}$	$n_{1.}$
Controls	$x_{21}^{(j)}$	$x_{22}^{(j)}$	$x_{23}^{(j)}$	$n_{2.}$
Absolute count	$n_{.1}{}^{(j)}$	$n_{.2}{}^{(j)}$	$n_{.3}{}^{(j)}$	N

consideration and a case-control study design with $n_{1.}$ cases and $n_{2.}$ controls is at hand, the statistical task is to analyze m contingency tables simultaneously, where the data for one particular locus $1 \leq j \leq m$ can be summarized as in Table 1, with $N = n_{1.} + n_{2.}$ denoting the total sample size.

One standard way to proceed with the data analysis is to compute for every locus $1 \leq j \leq m$ the chi-square statistic

$$Q_{\text{assoc}}^{(j)}(\mathbf{x}^{(j)}) = \sum_{r=1}^{2}\sum_{c=1}^{3} \frac{(x_{rc}^{(j)} - e_{rc}^{(j)})^2}{e_{rc}^{(j)}} \tag{1}$$

for association, where r runs over the rows and c over the columns of the contingency table $\mathbf{x}^{(j)}$ pertaining to locus j. In (1), the numbers $e_{rc}^{(j)} = n_{r.}n_{.c}^{(j)}/N$ denote the expected cell counts under the null hypothesis that the genotype at locus j is not associated with the binary phenotype of interest, conditionally to the marginal counts $n_{1.}$, $n_{2.}$, $n_{.1}{}^{(j)}$, $n_{.2}{}^{(j)}$, and $n_{.3}{}^{(j)}$. Large values of $Q_{\text{assoc}}^{(j)}$ indicate evidence against this null hypothesis of no association. A corresponding (asymptotic) p-value $p_{\text{assoc}}^{(j)}$ is given by $p_{\text{assoc}}^{(j)}(\mathbf{x}^{(j)}) = 1 - F_{\chi_2^2}(Q_{\text{assoc}}^{(j)}(\mathbf{x}^{(j)}))$, and the multiple test can be constructed by thresholding the p-values $(p_{\text{assoc}}^{(j)}(\mathbf{x}^{(j)}) : 1 \leq j \leq m)$, where the threshold is chosen according to an appropriate multiplicity correction, in the simplest case the Bonferroni correction.

The major drawback of this "locus-by-locus" analysis is that interactions (i.e., dependencies) between different genomic loci are not taken into account, although such dependencies are known to exist due to the biological mechanism of inheritance and other biological factors. For example, the concept of linkage disequilibrium (LD) quantifies the strength of the (bivariate) linear dependencies of the loci (cf., e.g., Dickhaus et al. (2015), Stange et al. (2016), and the references therein). Thus, it is near at hand to design multivariate statistical procedures which incorporate the dependencies. For example, methods based on the "effective number of tests" (cf. Dickhaus and Stange (2013) and the references therein) explicitly exploit LD to relax the multiplicity correction.

2 Proposed Methodology

In Mieth et al. (2016), we proposed a novel methodology, termed “COMBI”, which is an implicitly multivariate method for identifying significant SNP-phenotype associations in genome-wide association studies. The main idea behind COMBI is a two-step algorithm consisting of (i) a machine learning and SNP selection step that reduces the number of candidate SNPs by selecting only a small subset of the most predictive SNPs (the size of which can be controlled by the user), and (ii) a statistical testing step where only the SNPs selected in step (i) are tested for association with the phenotype.

In the first step, a support vector machine (SVM, which is an inherently multivariate classification method) with an appropriately designed kernel is trained. Per autosome, only the k genomic loci with largest (absolute) SVM weights are carried over to step (ii) of COMBI, and their p-values for association are computed in the usual manner, as described in Sect. 1. All other loci are not considered in step (ii) of COMBI, which technically means that their association p-values are set to one without explicit computation. The thresholding of the p-values computed in step (ii) is a delicate issue, because the same data are used in both steps of the COMBI method. Therefore, the chi-square distribution is prone not to be a valid (asymptotic) null distribution for the p-values anymore. In Mieth et al. (2016), we developed a fully data-driven resampling algorithm to determine appropriate thresholds. Furthermore, we discussed appropriate choices for the tuning parameter k.

3 Summary of Results

We have compared the COMBI method with several other state-of-the-art methods for evaluating genetic association studies. On simulated data, COMBI exhibited a very good performance in terms of type I and type II errors. Detailed results can be found in Mieth et al. (2016). Furthermore, we developed a validation pipeline for the analysis of real genome-wide association studies. Namely, we evaluated historical studies and checked whether SNPs which are declared to have a statistically significant association with the phenotype of interest only by the COMBI method have been detected in later, typically larger (with respect to the sample size) “validation” studies. We could demonstrate that this is the case for most of these SNPs, in particular when taking the study by The Wellcome Trust Case Control Consortium (2007) as the historical study and utilizing the GWAS catalog (an established reference database) for finding appropriate validation studies.

4 Implementation and Software

The COMBI method is implemented in Matlab/Octave, R and Java as a part of the GWASpi toolbox 2.0 (https://bitbucket.org/gwas_combi/gwaspi/).

Acknowledgements The article by Mieth et al. (2016) is part of a larger project on the genetics of social and economic behavior. The idea for the research described in this short note arose in the workshop that regularly took place in the context of this project at the University of Zurich, and which is based on the collaboration of teams at universities in Barcelona, Berlin, Bremen, Mainz, and Zurich. The author likes to acknowledge all cooperation partners.
Special thanks are due to the organizers of the Platinum Jubilee International Conference on Applications of Statistics (PJICAS) 2016 for the successful conference.

References

Dickhaus, T. (2014). *Simultaneous statistical inference with applications in the life sciences*. Berlin: Springer.

Dickhaus, T., & Stange, J. (2013). Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. *Calcutta Statistical Association Bulletin*, *65*(257–260), 123–144. https://doi.org/10.1177/0008068320130108.

Dickhaus, T., Stange, J., & Demirhan, H. (2015). On an extended interpretation of linkage disequilibrium in genetic case-control association studies. *Statistical Applications in Genetics and Molecular Biology*, *14*(5), 497–505. https://doi.org/10.1515/sagmb-2015-0024.

Mieth, B., Kloft, M., Rodriguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports*, *6*, 36671.

Stange, J., Dickhaus, T., Navarro, A., & Schunk, D. (2016). Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate. *Statistics and Probability Letters*, *111*, 32–40.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7), 661–678.

The Quantile-Based Skew Logistic Distribution with Applications

Tapan Kumar Chakrabarty and Dreamlee Sharma

Abstract In this paper, a modified form of the quantile based skew logistic distribution of Van Staden and King (Stat Probab Lett, 96:109–116, 2015) originally introduced by Gilchrist (Statistical modelling with quantile functions. CRC Press, 2000) has been studied. Some classical and quantile based properties of the distribution have been obtained. L-moments and L-ratios of the distribution have been obtained in closed form. The nature of L-Skewness and L-kurtosis of the distribution have been studied in detail. A brief study on the order statistics of the distribution has been done. The estimation of parameters of the proposed model is approached by the methods of matching L-moments estimation. Finally, we apply the proposed model to real datasets and compare the fit with the quantile based skew logistic distribution of Van Staden and King (Stat Probab Lett, 96:109–116, 2015).

Keywords Flattened skew logistic distribution · Quantile function · L-moment Asymptotic variance · Order statistics

1 Introduction

The logistic distribution finds an important place in the literature of continuous symmetric distributions due to its simplicity and wide applications. Various asymmetric generalizations of the logistic distribution have been proposed in the literature (Nadarajah 2009; Asgharzadeh et al. 2013; Wahed and Ali 2001)

T. K. Chakrabarty · D. Sharma (✉)
Department of Statistics, North-Eastern Hill University, Shillong 793022, Meghalaya, India
e-mail: dreamleesharma@yahoo.in

T. K. Chakrabarty
e-mail: tapankumarchakrabarty@gmail.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_6

Recently, Van Staden and King (2015) used a quantile based approach to define a skew logistic distribution, written hereafter as SLD_{QB}, which was originally introduced by Gilchrist (2000). The quantile function of SLD_{QB} is defined by

$$Q_{SLD}(p) = \alpha + \beta\,[(1-\delta)\ln(p) - \delta\ln(1-p)] \qquad (1)$$

where α is the location parameter, β is the scale parameter and $0 \le \delta \le 1$ is a skewness parameter that controls the tail shape of the distribution. The distribution has been fitted successfully to the concentration of polychlorinated biphenyl (PCB) in yolk lipids of pelican eggs (Thas 2010). In practical life, there may be skewed data that have a lower peak than the distribution in (1) in which case the SLD_{QB} may not provide a very acceptable fit. Hence we introduce another quantile-based distribution ideal for fitting skewed data with flatter peak than the distribution in (1). The quantile function of the proposed distribution is obtained by adding a multiple $\kappa \ge 0$, of uniform (0, 1) quantile function to Q_{SLD} of (1) as,

$$Q(p) = \alpha + \beta\,[(1-\delta)\ln(p) - \delta\ln(1-p)] + \kappa p \qquad (2)$$

where $\kappa \ge 0$ is a parameter regulating the flatness of the peak of the distribution. Hence we name it as the flattened skew logistic distribution (FSLD) which is presented in Definition 1.

In this paper, we shall study the properties and applications of this modified quantile distribution. The support of the distribution is $(Q(0), Q(1)) = (-\infty, \infty)$. Like many other quantile based distributions such as Tukey's lambda distribution (Tukey 1960) and its generalizations (Ramberg and Schmeiser 1972, 1974; Ramberg 1975; Freimer et al. 1988), the Davies distribution (Gilchrist 2000; Hankin and Lee 2006), the Govindarajulu Distribution (Govindarajulu 1977; Nair et al. 2012) and the quantile-based skew logistic distribution and generalized skew logistic distribution (Gilchrist 2000; Van Staden and King 2015; Balakrishnan and So 2015), closed form expressions for either the cumulative distribution function or the probability density function for the distribution do not exist except for the special case when $\kappa = 0$ and $\delta = \frac{1}{2}$. For recent studies on the use of quantile based functions, the readers are referred to Aljarrah et al. (2014); Balakrishnan and So (2015); Midhu et al. (2013, 2014); Thomas et al. (2015) and Sankaran et al. (2016).

Definition 1 A real-valued random variable X is said to have a quantile-based flattened skew logistic distribution, denoted by $X \sim FSLD_{QB}(\alpha, \beta, \delta, \kappa)$ if its quantile function is given by

$$Q(p) = \alpha + \beta\,[(1-\delta)\ln(p) - \delta\ln(1-p)] + \kappa p$$

where α is the location parameter, $\beta > 0$ is the scale parameter, $0 \le \delta \le 1$ is a skewness parameter, and $\kappa \ge 0$ is a parameter regulating the flatness of the peak of the distribution.

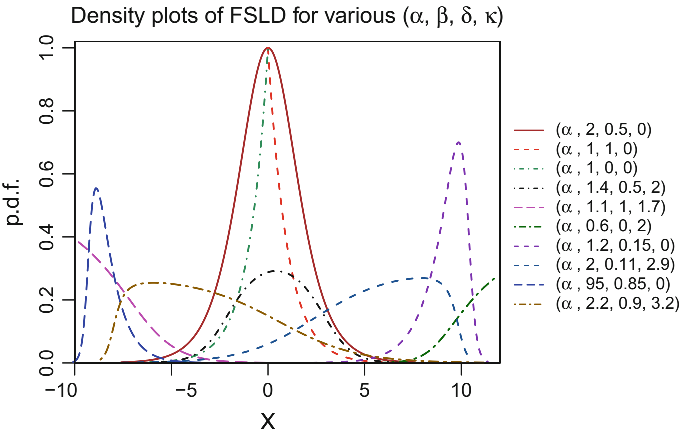


Fig. 1 Density plots of FSLD

2 Special Cases

1. For $\kappa = 0$, the FSLD is the quantile based skew logistic distribution, $Q(p) = \alpha + \beta\,[(1-\delta)\ln(p) - \delta\ln(1-p)]$
2. For $\alpha = 0$, $\beta = 2$, $\delta = \frac{1}{2}$ and $\kappa = 0$, the FSLD is a standard logistic distribution, $Q(p) = \ln\left(\frac{p}{1-p}\right)$
3. For $\delta = \frac{1}{2}$ and $\kappa = 0$, the FSLD is a location scale based logistic distribution, $Q(p) = \alpha + \frac{\beta}{2}\ln\left(\frac{p}{1-p}\right)$
4. For $\alpha = 0$, $\delta = 1$ and $\kappa = 0$ the FSLD is an exponential distribution, $Q(p) = -\beta\ln(1-p)$
5. For $\alpha = 0$, $\delta = 0$ and $\kappa = 0$ the FSLD is a reflected exponential distribution, $Q(p) = \beta\ln(p)$

3 Shape of the Distribution

The FSLD has a rich varieties of shapes. Figure 1 shows plots of possible shapes of FSLD.

The possible shapes of FSLD(α, β, δ, κ) are discussed below:-

(1) The FSLD(0, β, 1, 0) has an exponential shape.
(2) The FSLD(0, β, 0, 0) has a reflected exponential shape.
(3) The FSLD(0, 2, 0.5, 0) has a symmetric standard logistic shape.
(4) The FSLD(α, β, 0.5, 0) has a symmetric logistic shape.
(5) The FSLD(α, β, δ, $\kappa = 0$) has a skewed shape.

(6) The FSLD($\alpha, \beta, 0.5, \kappa > 0$) has a flattened logistic shape.
(7) The FSLD($\alpha, \beta, \delta < 0.5, \kappa > 0$) has a flattened negatively skewed shape.
(8) The FSLD($\alpha, \beta, \delta > 0.5, \kappa > 0$) has a flattened positively skewed shape.
(9) The FSLD($\alpha, \beta, \delta < 0.5, \kappa = 0$) has a negatively skewed j shaped curve.
(10) The FSLD($\alpha, \beta, \delta > 0.5, \kappa = 0$) has a positively skewed inverted j shaped curve.

4 Properties

4.1 Quantile Density Function

Theorem 1 *The quantile density function ($q(p)$) of the flattened skew logistic distribution is*

$$q(p) = \beta\left[\frac{1-\delta}{p} + \frac{\delta}{(1-p)}\right] + \kappa \tag{1}$$

Proof Since, $q(p) = \frac{d}{dp}Q(p)$, the proof follows. □

4.2 Density Quantile Function

Theorem 2 *The density quantile function or the p-p.d.f. of the flattened skew logistic distribution is*

$$f_p(p) = \frac{p(1-p)}{\beta(1-\delta+p(2\delta-1)) + \kappa p(1-p)} \tag{2}$$

Proof The proof follows from the definition of *p-p.d.f.*, $f_p(p) = \frac{1}{q(p)}$. □

4.3 Moments

Theorem 3 *If $X \sim FSLD(\alpha, \beta, \delta, \kappa)$, then the mean ($\mu$) and variance (Var(X)) of X are respectively given by*

$$\mu = \alpha + \beta(2\delta - 1) + \frac{\kappa}{2} \tag{3}$$

$$Var(X) = \beta^2(1 - 4\delta(1-\delta)) + \frac{\kappa}{2}\left(\beta + \frac{\kappa}{6}\right) + \beta^2\frac{\pi^2}{3}\delta(1-\delta) \tag{4}$$

where as the $3rd$ and $4th$ order central moments of FSLD are respectively

$$\begin{aligned}\mu_3 &= 2\beta^3[-1+2\delta\{3-2\delta(3-2\delta)\}] + \frac{\beta\kappa}{4}(2\delta-1)\left(\frac{\kappa}{3}+3\beta\right) - 6\beta^3\delta\zeta(3)[1-\delta(3-2\delta)] \\ \mu_4 &= 9\beta^4[1-8\delta\{1-\delta(3-2\delta(2-\delta))\}] + 9\beta^3\kappa\left[\frac{1}{2}-2\delta(1-\delta)\right] + \frac{2}{9}\beta^2\kappa^2[5-11\delta(1-\delta)]+ \\ &\quad \frac{\kappa^3}{2}\left(\frac{\beta}{3}+\frac{\kappa}{40}\right) + \frac{\beta^4\delta\pi^4}{5}\left[\frac{4}{3}-\delta\left\{\frac{13}{3}-3\delta(2-\delta)\right\}\right] + 2\beta^4\delta\pi^2[1-\delta\{5-4\delta(2-\delta)\}] \\ &\qquad + \beta^2\kappa\pi^2\left(\beta+\frac{\kappa}{6}\right)\delta(1-\delta)\end{aligned} \tag{5}$$

where $\zeta()$ is the Reimann's zeta function.

Proof The proof is given in the Appendix. □

4.4 Quartiles

Theorem 4 *The quartiles of the FSLD are respectively given by*

$$\begin{aligned}\textit{Lower Quartile}, LQ &= Q(0.25) = \alpha - \beta\, ln(3)\delta - \beta ln(4)(1-2\delta) + \frac{1}{4}\kappa \\ \textit{Median}, M &= Q(0.5) = \alpha - \beta ln(2)(1-2\delta) + \frac{\kappa}{2} \\ \textit{Upper Quartile}, UQ &= Q(0.75) = \alpha + \beta\, ln(3)(1-\delta) - \beta ln(4)(1-2\delta) + \frac{3}{4}\kappa\end{aligned}$$

Corollary 4.1 *The inter quartile range, IQR is given by*

$$IQR = UQ - LQ = \beta ln(3) + \frac{\kappa}{2}$$

Corollary 4.2 *The identification quantile function of the distribution is*

$$IQ(p) = \frac{\beta(1-\delta)ln(p) - \beta\delta ln(1-p) + \kappa\left(p-\frac{1}{2}\right) + \beta ln(2)(1-2\delta)}{2\beta ln(3) + \kappa}$$

4.5 Reflection of FSLD

Theorem 5 *If X is a FSLD with parameters α, β, δ and κ, then the reflection of X is a FSLD with parameters $-(\alpha+\kappa)$, β, $t = 1-\delta$ and κ*

4.6 L-Moments

Theorem 6 *If $X \sim FSLD(\alpha, \beta, \delta, \kappa)$, then the rth order L-moment is given by*

$$\lambda_r = \begin{cases} \alpha - \beta(1-2\delta) + \frac{\kappa}{2}, & \mathit{if}\, r = 1 \\ \frac{\beta}{2} + \frac{\kappa}{6}, & \mathit{if}\, r = 2 \\ \dfrac{\beta(2\delta-1)^{r \; mod \; 2}}{r(r-1)}, & \mathit{if}\, r = 3, 4, 5, ... \end{cases} \tag{6}$$

and the rth order L-moment ratio is given by

$$\tau_r = \frac{6\beta(2\delta-1)^{r \; mod \; 2}}{r(r-1)(3\beta+\kappa)}, \quad r = 3, 4, 5, ... \tag{7}$$

Proof The proof is given in the Appendix. □

Corollary 6.1 *The first 4 L-moments of FSLD are*

$$\begin{aligned} \lambda_1 &= \alpha - \beta(1-2\delta) + \frac{\kappa}{2} \\ \lambda_2 &= \frac{\beta}{2} + \frac{\kappa}{6} \\ \lambda_3 &= \frac{\beta}{6}(2\delta - 1) \\ \lambda_4 &= \frac{\beta}{12} \end{aligned} \tag{8}$$

Corollary 6.2 *The L-coefficient of variation, L-skewness and L-kurtosis of the FSLD are,*

$$\begin{aligned} \tau_2 &= \frac{3\beta+\kappa}{3(2\alpha - 2\beta(1-2\delta)+\kappa)} \\ \tau_3 &= \frac{\beta(2\delta-1)}{3\beta+\kappa} \\ \tau_4 &= \frac{\beta}{2(3\beta+\kappa)} \end{aligned}$$

Since $0 \leq \delta \leq 1$ and $\kappa \geq 0$, the L-skewness of the FSLD lies between $-\frac{1}{3}$ and $\frac{1}{3}$. Figure 2 gives plot of L-skewness against δ and κ. It can be seen that for $\delta < 0.5$, the FSLD has negative L-skewness, for $\delta > 0.5$, the FSLD has positive L-skewness and for $\delta = 0.5$, the FSLD has no L-skewness, this corresponds to the black line at $y = 0$ in Fig. 2.

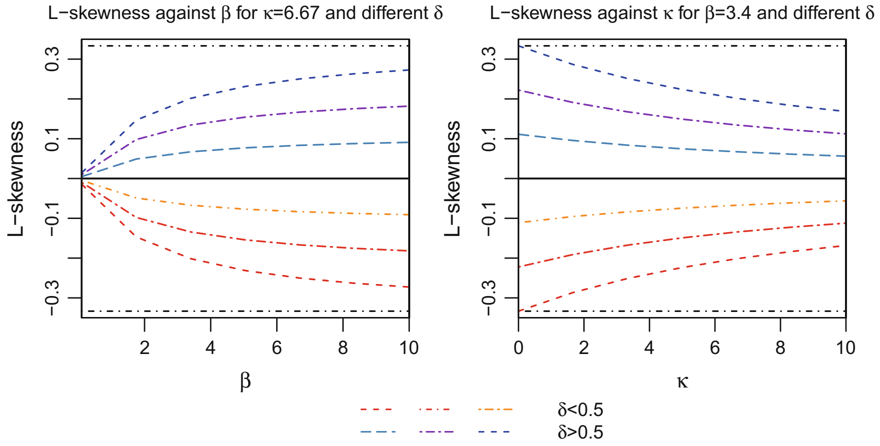


Fig. 2 Plot of L-skewness against β and κ for different values of δ

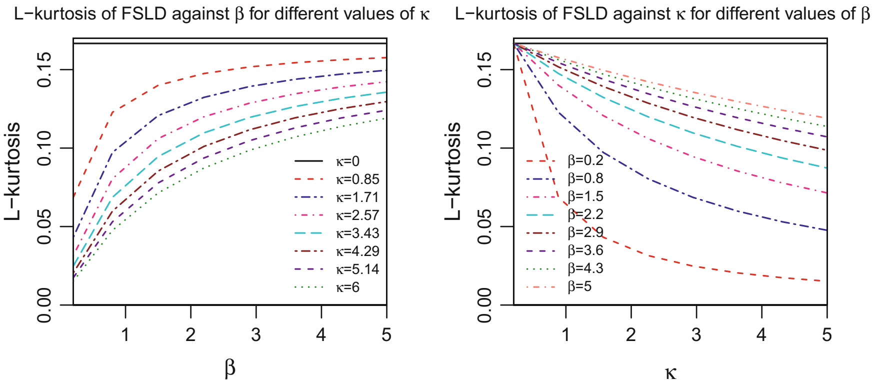


Fig. 3 Plot of L-kurtosis for different values of β and κ

Also, it is clear that with increase in β, the L-skewness of FSLD increases for $\delta > \frac{1}{2}$ where as it decreases for $\delta < \frac{1}{2}$. While for increase in κ, the L-skewness of FSLD decreases for $\delta > \frac{1}{2}$ where as it increases for $\delta < \frac{1}{2}$.

Since for $\kappa = 0$ the FSLD has a fixed L-kurtosis of $\frac{1}{6}$, the L-kurtosis of FSLD can never exceed $\frac{1}{6}$. Also, from Corollary 6.2 it is clear that $\tau_4 > 0$ since $\beta \neq 0$. Hence, $0 < \tau_4 \leq \frac{1}{6}$. Figure 3 gives plots of L-kurtosis of the flattened skew logistic distribution for increasing values of β and κ.

The black continuous line is the line where the y-axis is $\frac{1}{6}$. It can be seen that L-kurtosis of FSLD always lie below this line. The figure clearly indicates that L-kurtosis of FSLD increases with increase in β and decreases with increases in κ.

5 Random Number Generation

We can simulate a random sample of size n using the quantile function of the FSLD as defined in (2). Let U be a uniform ($U(0, 1)$) *r.v.* and let $Q(p)$, $0 \leq p \leq 1$ be the quantile function of the FSLD, then by uniform transformation rule, (Gilchrist 2000) the variable X, where $x = Q(u)$, has a distribution with quantile function $Q(p)$. Thus, by using the uniform transformation rule, a random sample of size n can be easily simulated from the FSLD by generating a random sample of the same size from a $U(0, 1)$ distribution.

6 Order Statistics

The order statistics are random variables that satisfy $X_{(1:n)} \leq X_{(2:n)} \leq \ldots \leq X_{(n:n)}$. For a detailed study on order statistics one can refer to Arnold et al. (1992), Balakrishnan and Rao (1998a, b) and Reiss (1989). Sample ordered values play a major role in modelling with quantile defined distributions.

Theorem 7 *If $X_{1:n}, X_{2:n}, \ldots, X_{r:n}$ denotes the order statistics in a random sample of size n from the FSLD, then the quantile function of the smallest, rth and largest order statistics are respectively given by (1).*

$$\begin{aligned}
Q_{(1)}(p) &= \alpha + \beta(1-\delta)ln\left[1-(1-p)^{\frac{1}{n}}\right] - \frac{\beta\delta}{n}ln(1-p) + \kappa\left[1-(1-p)^{\frac{1}{n}}\right] \\
Q_{(r)}(p) &= \alpha + \beta(1-\delta)ln\left[I_p^{-1}(r, n-r+1)\right] - \beta\delta\, ln\left[1 - I_p^{-1}(r, n-r+1)\right] \\
&\quad + \kappa I_p^{-1}(r, n-r+1) \\
Q_{(n)}(p) &= \alpha + \frac{\beta(1-\delta)}{n}ln(p) - \beta\delta ln\left[1-p^{\frac{1}{n}}\right] + \kappa p^{\frac{1}{n}}
\end{aligned} \tag{1}$$

where, $I_p^{-1}(r, n-r+1)$ is the inverse of the regularized incomplete beta function.

Proof The proof follows from the expressions of quantile functions of smallest, rth and largest order statistics (Gilchrist 2000) given in Eq. 2

$$\begin{aligned}
Q_{(1)}(p) &= Q\left(1-(1-p)^{\frac{1}{n}}\right) \\
Q_{(r)}(p) &= Q\left[I_p^{-1}(r, n-r+1)\right] \\
Q_{(n)}(p) &= Q\left(p^{\frac{1}{n}}\right)
\end{aligned} \tag{2}$$

□

Theorem 8 *Let $X_{r:n}$ denotes the rth order statistic in a random sample of size n from the FSLD, then the expectation of $X_{(r:n)}$ can be expressed in a closed form expression given in (3)*

$$E(X_{r:n}) = \alpha + \frac{\kappa r}{n+1} + \frac{\beta}{B(r,c)}\left[(1-\delta)\sum_{j=0}^{c-1}\frac{(-1)^{j+1}}{(r+j)^2}\binom{c-1}{j} - \delta\sum_{j=0}^{r-1}\frac{(-1)^{j+1}}{(c+j)^2}\binom{r-1}{j}\right] \tag{3}$$

where, $c = n - r + 1$.

Proof The proof is given in the Appendix. □

Corollary 8.1 *The expectation of the smallest and largest order statistics are respectively given by*

$$E(X_{1:n}) = \alpha + \frac{\kappa}{n+1} + \frac{\beta\delta}{n} + \beta n(1-\delta)\sum_{j=0}^{n-1}\frac{(-1)^{j+1}}{(j+1)^2}\binom{n-1}{j}$$

and

$$E(X_{n:n}) = \alpha + \frac{\kappa n}{n+1} - \frac{\beta(1-\delta)}{n} - \beta\delta n\sum_{j=0}^{n-1}\frac{(-1)^{j+1}}{(j+1)^2}\binom{n-1}{j}$$

7 Inference and Goodness of Fit

7.1 Inference

In the literature, there are various methods for the estimation of parameters for quantile based distributions, viz., method of minimum absolute deviation, method of least squares, method of maximum likelihood estimation (MLE) and method of matching L-moments estimation (Gilchrist 2000). The method of matching L-moments estimation (MLM) gives more robust estimates as compared to the traditional moments. With small and moderate samples the method of matching L-moments is more efficient than MLE. Moreover, closed form expression of the density function of the FSLD doesn't exist, hence the method of matching L-moments estimation is more appealing. The L-moments of the distribution have been obtained earlier so the parameters of the distribution can be estimated by the method of matching L-moments estimation. Hence to obtain the estimates of α, β, δ, and κ, the first four L-moments, λ_1, λ_2, λ_3 and λ_4 of the distribution have been matched with the corresponding sample L-moments, l_1, l_2, l_3 and l_4.

Theorem 9 *Let $X \sim FSLD(\alpha, \beta, \delta, \kappa)$ and let $l_1, l_2, ..., l_n$ be the sample L-moments of a sample of size n, then the matching L-moments estimates of α, β, δ and κ is given by,*

$$\begin{aligned}\widehat{\alpha} &= l_1 - 3(l_2 + 2l_3 - 6l_4)\\ \widehat{\beta} &= 12l_4\\ \widehat{\delta} &= \frac{l_3}{4l_4} + \frac{1}{2}\\ \widehat{\kappa} &= 6(l_2 - 6l_4)\end{aligned} \tag{1}$$

with asymptotic variance-covariance matrix given by

$$n\ var \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \\ \widehat{\delta} \\ \widehat{\kappa} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} \end{pmatrix} \tag{2}$$

where

$$\begin{aligned}
\theta_{11} &= \frac{\beta^2}{35}\,[8\delta(2626 - 2731\delta) + 1153] + \frac{2\kappa}{21}\left[\frac{1}{5}\,(299\beta + 67\kappa) - 29\beta\delta\right] - \frac{200}{3}\beta^2\pi^2\delta(1-\delta)\\
\theta_{22} &= \frac{4}{7}\left[2\beta^2\,\{398\delta(1-\delta) + 9\} + \frac{\kappa}{5}(9\beta + \kappa)\right] - 48\beta^2\pi^2\delta(1-\delta)\\
\theta_{33} &= \frac{1}{7}\left[\frac{1}{5}\,\{\delta(2927 - 18847\delta) + 27\} + 3184\delta^3(2-\delta)\right]\\
&\quad + \frac{\kappa}{\beta}\left[\frac{1}{2}\left(\frac{\kappa}{5\beta} + \frac{9}{14}\right) - \frac{8\delta}{35}(1-\delta)\left(3 + \frac{\kappa}{\beta}\right)\right] + 3\pi^2\delta[16\delta^2(\delta - 2) + 19\delta - 3]\\
\theta_{44} &= \frac{1}{21}\left[20\beta^2\{83\delta(1-\delta) + 2\} + \frac{\kappa}{3}\left(19\beta + \frac{23}{4}\kappa\right)\right] - \frac{25}{3}\beta^2\pi^2\delta(1-\delta)\\
\theta_{12} = \theta_{21} &= \frac{2\beta^2}{7}(61 + 1961\delta - 1996\delta^2) + \frac{\kappa}{7}\left[3\kappa + \frac{\beta}{5}(74 - 19\delta)\right] - 60\beta^2\pi^2\delta(1-\delta)\\
\theta_{13} = \theta_{31} &= \frac{\beta}{7}\left[\frac{1}{5}\left(\frac{267}{2} + 11246\delta - 31431\delta^2 + 3992\delta^3\right)\right]\\
\theta_{14} = \theta_{41} &= \frac{2}{7}\beta^2[2\delta(415\delta - 408) - 27] - \frac{1}{63}\beta\kappa\left(\frac{143}{2} - 29\delta\right) - \frac{23}{84}\kappa^2 + 25\beta^2\pi^2\delta(1-\delta)\\
\theta_{23} = \theta_{32} &= \frac{\beta}{7}[15 + 2\delta\{781 - 796\delta(3 - 2\delta)\}] + \frac{\kappa}{7}(1 - 2\delta)\left(3 + \frac{4}{5}\kappa\right) - 24\beta\delta\pi^2\{1 - \delta(3 - 2\delta)\}\\
\theta_{24} = \theta_{42} &= \frac{\beta^2}{7}\left[-29 - \frac{3992}{3}(1-\delta)\right] - \frac{\kappa}{7}\left(\frac{43}{10}\beta + \kappa\right) + 20\beta^2\pi^2\delta(1-\delta)\\
\theta_{34} = \theta_{43} &= \frac{\beta}{7}\left[\frac{1996}{3}\delta^2(3 - 2\delta) - \frac{1951}{3}\delta - \frac{15}{2}\right] - \frac{\kappa}{14}\left(\frac{229}{60} + \frac{\kappa}{\beta}\right)(1 - 2\delta) + 10\beta\delta\pi^2[1 - \delta(3 - 2\delta)]
\end{aligned} \tag{3}$$

Proof The proof is included in the Appendix. □

Corollary 9.1 *The asymptotic standard errors of the estimates are*

$$\begin{aligned}\sigma_\alpha &= \Bigg\{\frac{1}{n}\Bigg[\frac{\beta^2}{35}\,[8\delta(2626-2731\delta)+1153]+\frac{2\kappa}{21}\Bigg[\frac{1}{5}\,(299\beta+67\kappa)-29\beta\delta\Bigg]\\ &\qquad -\frac{200}{3}\beta^2\pi^2\delta(1-\delta)\Bigg]\Bigg\}^{\frac{1}{2}}\end{aligned}$$

$$\sigma_\beta = \Bigg\{\frac{1}{n}\Bigg[\frac{4}{7}\Bigg[2\beta^2\,\{398\delta(1-\delta)+9\}+\frac{\kappa}{5}(9\beta+\kappa)\Bigg]-48\beta^2\pi^2\delta(1-\delta)\Bigg]\Bigg\}^{\frac{1}{2}}$$

$$\begin{aligned}\sigma_\delta &= \Bigg\{\frac{1}{n}\Bigg[\frac{1}{7}\Bigg[\frac{1}{5}\,\{\delta(2927-18847\delta)+27\}+3184\delta^3(2-\delta)\Bigg]+3\pi^2\delta[16\delta^2(\delta-2)+19\delta-3]\\ &\qquad +\frac{\kappa}{\beta}\Bigg[\frac{1}{2}\left(\frac{\kappa}{5\beta}+\frac{9}{14}\right)-\frac{8\delta}{35}(1-\delta)\left(3+\frac{\kappa}{\beta}\right)\Bigg]\Bigg]\Bigg\}^{\frac{1}{2}}\end{aligned}$$

$$\sigma_\kappa = \Bigg\{\frac{1}{n}\Bigg[\frac{1}{21}\Bigg[20\beta^2\{83\delta(1-\delta)+2\}+\frac{\kappa}{3}\left(19\beta+\frac{23}{4}\kappa\right)\Bigg]-\frac{25}{3}\beta^2\pi^2\delta(1-\delta)\Bigg]\Bigg\}^{\frac{1}{2}}$$

7.2 *Goodness of Fit Test*

In the literature, there are several methods of validation for quantile based distributions. The well-known density probability plot and Q-Q plot of the validation data against the corresponding fitted values give an idea of the visual goodness of fit of the model. Few other methods includes residual plot, unit exponential spacing control chart, chi-square goodness of fit test and uniformity test (Gilchrist 2000). We shall be testing the model validation using uniformity test and also the density probability plot and Q-Q plot as a means of visual validation.

7.2.1 Uniformity test

Let $\widehat{Q}(p)$ be the quantile function of the FSLD fitted to a set of data $x_1, x_2, ...x_n$ by the method of matching L-moments. Using this information we can derive numerically a corresponding set of $p_{(r)}$; $r = 1, 2, 3, ..., n$ such that $\widehat{Q}(p_{(r)}) \approx x_{(r)}$. If the model is valid, these will be a set of ordered variables from a uniform distribution (Gilchrist 2000). We can therefore test the validity of the model by testing the uniformity of the distribution of p. There are a variety of such tests. In this article, we shall be using the Kolmogorov Smirnov (D_n), Anderson Darling (A_n) and Cramér von Mises (W_n) goodness of fit tests (Thas 2010) and obtain the asymptotic and bootstrap p-values of these tests. Thas (2010) has shown that since a distribution can be equally characterized by either its quantile function, distribution function or characteristic function, comparing them based on any of these functions are equivalent. We can thus use tests based on empirical distribution function, viz., (D_n), (A_n) and (W_n) to test the goodness of fit of the FSLD. Let $\widehat{q}(p_{(j)})$ be the fitted quantile density function. The algorithm to find $p_{(r)}$ (Gilchrist 2000) is given below.

Algorithm 1 Algorithm to determine $p_{(r)}$

```
Set initial $p_{(r)}$'s say $p_{0(1)}, p_{0(2)}, ..., p_{0(n)}$
Iterate for $j$,
    $p_{j+1} = p_j + \frac{x_{(i)} - \widehat{Q}(p_{(j)})}{\widehat{q}(p_{(j)})}$
    $p_{j+1} = 0.0000001 \quad \text{if } p_j < 0$
    $p_{j+1} = 0.9999999 \quad \text{if } p_j > 1$
    until $|\widehat{Q}(p_{(j)}) - x_{(i)}|$ is a very small quantity.
Repeat iterations for all $i = 1, 2, 3, ...., n$.
```

8 Data Fitting

8.1 *Simulation Study*

It has been discussed under Sect. 5 that a random sample of size n can be generated from a FSLD using its quantile function. Let $\theta = (\alpha, \beta, \delta, \kappa)'$ denote the vector of parameters of FSLD. Random samples of size $n = 100$, 1000, 3000, 5000, 8000 and 10000 have been generated from the FSLD with known values of θ. The generated datasets have been fitted by the method of matching L-moments as discussed in Sect. 7.1 and the asymptotic standard errors of the estimates for all the datasets have been obtained. The goodness of fit test is performed using the procedure discussed in Sect. 7.2. Table 1 shows the L-moment estimator of θ and the corresponding standard error of estimates and goodness of fit (GOF) results for a few simulated data.

Table 1 shows that as the sample size n increases, the estimates obtained by the method of L-moments are closer to the true value with smaller standard error. The standard errors of estimates obtained for each samples are plotted against sample size n. Figure 4 shows the plot of asymptotic standard errors for increasing sample size.

Figure 4 shows decreasing trends for the asymptotic standard errors with increasing sample size for all the drawn samples which is expected for the asymptotic behavior of the standard error. Also, the GOF tests performed gives a very acceptable p-value, so that for all the simulated samples the FSLD gives a very good fit by the method of matching L-moments and the GOF tests used. Hence the method of L-moment estimation and GOF tests described can be practically used to fit some real life data.

Table 1 Parameter estimates of FSLD, their Std. Err. and GOF test

Samples drawn from	Sample size	Estimators and Std. Err. in parenthesis				GOF statistic and the p-value in parenthesis		
	n	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\kappa}$	$\widehat{\delta}$	(D_n)	(A_n)	(W_n)
FSLD (2, 4, 0.2, 8)	n = 100	1.273043	4.566356	0.1945418	8.552020	0.017	0.2785	0.0462
		(2.7821943)	(1.4187999)	(0.10229787)	(0.69077842)			
	n = 5000	1.860817	3.886132	0.1834837	8.448516			
		(0.3510242)	(0.1758050)	(0.01550991)	(0.08711373)	(0.9354)	(0.9534)	(0.8989)
	n = 10000	2.041024	3.962356	0.1936960	8.067968			
		(0.2468837)	(0.1247017)	(0.01059837)	(0.06135400)			
FSLD (0, 1, 0.5, 0.6)	n = 100	−0.103678049	1.0859501	0.4717887	0.6577006	0.0172	0.2726	0.0466
		(0.43895430)	(0.27137122)	(0.08693686)	(0.12493950)			
	n = 5000	−0.028071515	0.9787647	0.4953523	0.6852927			
		(0.05583966)	(0.03485652)	(0.01247524)	(0.01614791)	(0.9293)	(0.9572)	(0.8966)
	n = 10000	−0.001257032	0.9877830	0.4967504	0.6312495			
		(0.03934916)	(0.02473196)	(0.00874943)	(0.01141319)			
FSLD(−3, 2.3, 0.7, 0.5)	n = 100	−3.595422	2.442099	0.7119531	5.190860	0.0164	0.2732	0.0449
		(1.1616700)	(0.73732562)	(0.10601849)	(0.37053297)			
	n = 5000	−3.116763	2.211760	0.7012052	5.337546			
		(0.1580952)	(0.09629170)	(0.01577213)	(0.04939556)	(0.9496)	(0.9568)	(0.9067
	n = 10000	−3.019786	2.248935	0.6981372	5.149672			
		(0.1110742)	(0.06843781)	(0.01091591)	(0.03483832)			
FSLD(0.4, 1, 0.9, 2)	n = 100	0.1303901	1.0286045	0.9377901	2.053320	0.0163	0.288	0.0452
		(0.45713635)	(0.35172713)	(0.11998304)	(0.16851111)			
	n = 5000	0.3454842	0.9598071	0.9107752	2.152905			
		(0.06351476)	(0.04629673)	(0.01720119)	(0.02263387)	(0.9541)	(0.9469)	(0.9049)
	n = 10000	0.3838221	0.9734695	0.9036203	2.081946			
		(0.04460694)	(0.03281565)	(0.01178148)	(0.01596199)			

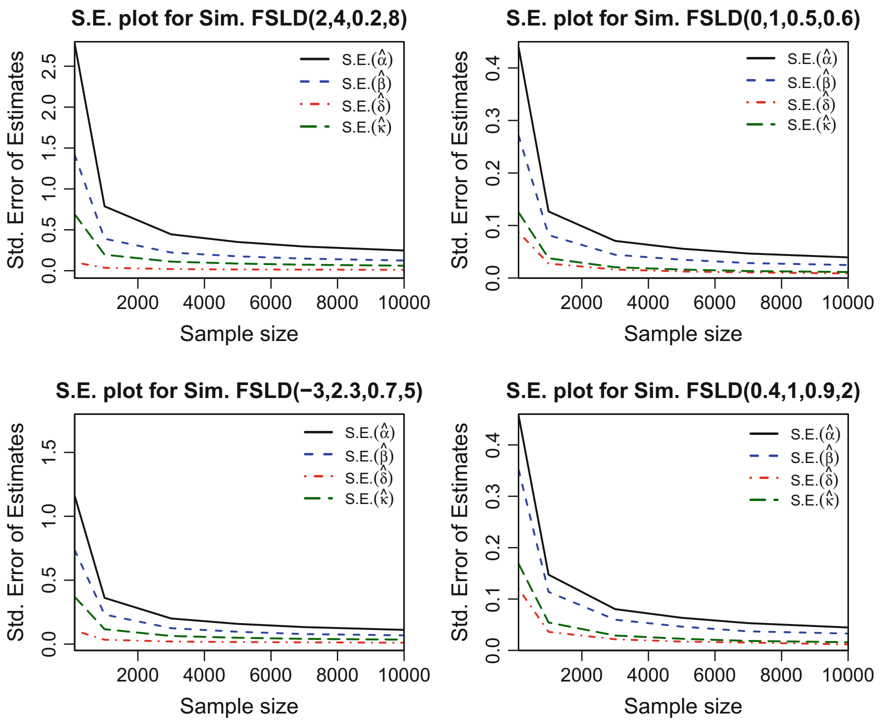


Fig. 4 Standard Error of $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\delta}$ and $\widehat{\kappa}$ for increasing sample size

8.2 Application to Real Life Data

We show the utility of the model in practical situations by applying it to real data sets. The first data represent the strength data measured in GPA, for single carbon fibers and impregnated 1,000-carbon fiber tows (CARBON data). The dataset has been taken from Gupta and Kundu (2010) originally considered by Bader and Priest (1982). The second dataset has been taken from the R-package 'sn' and it gives the percentage of alcohol in wine (ALCOHOL data). The datasets have been fitted to the FSLD and SLD_{QB} by the method of matching L-moments. The results obtained is summarized in Table 2.

It can be seen that L-kurtosis of both the datasets are same as the L-kurtosis of the FSLD where as the SLD_{QB} has a fixed L-kurtosis of 0.166667. The goodness of fit tests are performed for both FSLD and SLD_{QB} by the method discussed in Sect. 7.2. The bootstrap p-values have been determined from 10000 parametric bootstrap samples. Table 3 gives the results obtained from these tests.

Table 2 Parameter estimates of FSLD and their corresponding standard error

Data used	Estimates	Estimated value	Standard error of estimates	L-Kurtosis of data	L-Kurtosis of fitted FSLD	L-Kurtosis of SLD_{QB}
Strength of Carbon fibre	$\widehat{\alpha}$	2.398117	0.2394746	0.09969171	0.09969171	0.166667
	$\widehat{\beta}$	0.417986	0.1636575			
	$\widehat{\delta}$	0.7870514	0.1319874			
	$\widehat{\kappa}$	0.8424351	0.08067969			
Percentage of alcohol in wine	$\widehat{\alpha}$	11.99657	0.2202775	0.04310323	0.04310323	0.166667
	$\widehat{\beta}$	0.2418097	0.08685792			
	$\widehat{\delta}$	0.4261055	0.1711473			
	$\widehat{\kappa}$	2.079578	0.05863612			

Table 3 Goodness-of-fit statistics and their corresponding asymptotic and bootstrap p-values

Data used	Model used	Test used	Statistic	Asymptotic p-value	Bootstrap p-value
Strength of Carbon fibre	FSLD	D_n	0.0613	0.9717	0.9593
		A_n	0.2111	0.9871	0.9886
		W_n	0.0356	0.9555	0.9566
	SLD_{QB}	D_n	0.0913	0.6704	0.6384
		A_n	0.4524	0.795	0.7928
		W_n	0.0818	0.6834	0.6873
Percentage of alcohol in wine	FSLD	D_n	0.0349	0.9816	0.9768
		A_n	0.195	0.9917	0.9924
		W_n	0.0252	0.9893	0.9896
	SLD_{QB}	D_n	0.0829	0.1733	0.1667
		A_n	1.7062	0.1341	0.1367
		W_n	0.2779	0.1562	0.1524

The p-values of all the tests indicates that FSLD gives a much better fit to both the data as compared to the SLD_{QB}. This is further confirmed by the density plot and Q-Q plot for the fitted FSLD depicted in Figs. 5 and 6 which is indicative of a visually good fit of the FSLD to both the data. Hence it can be said that FSLD is a better candidate than the SLD_{QB} for skewed data with a flat peak.

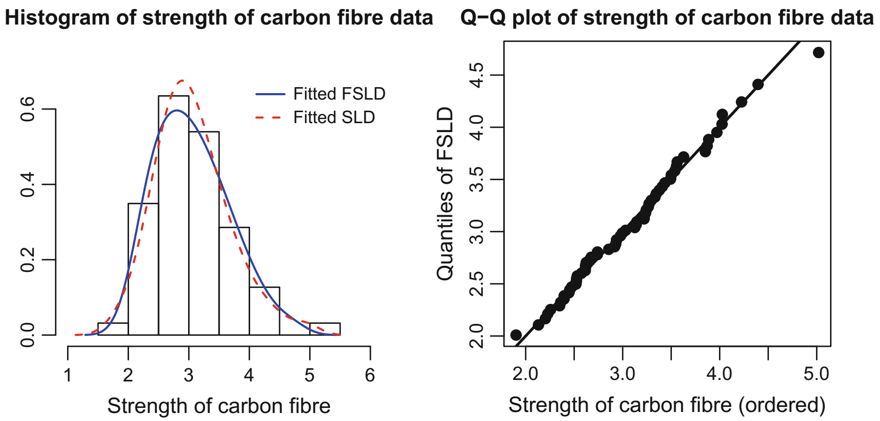


Fig. 5 Histogram of the CARBON data with the *p.d.f.* of fitted FSLD and the Q-Q plot of the fitted FSLD

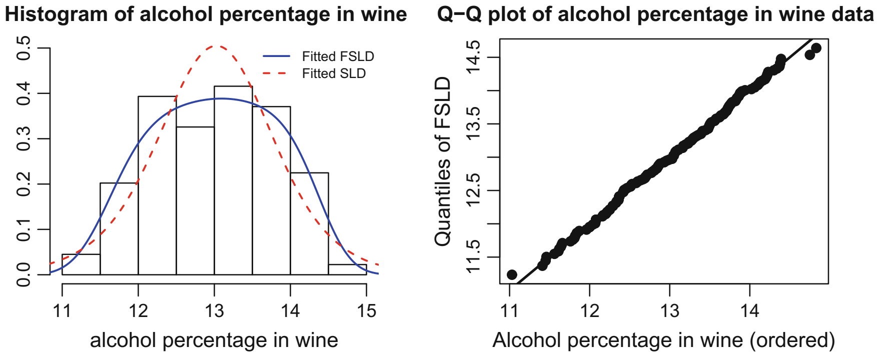


Fig. 6 Histogram of the ALCOHOL data with the *p.d.f.* of fitted FSLD and the Q-Q plot of the fitted FSLD

9 Conclusion

A new quantile based distribution ideal for fitting skewed data with a flat peak has been proposed and its various distributional properties (classical and quantile-based) have been derived. Closed form expression of the moments, L-moments and order statistics of the distribution and their corresponding expectations have been derived. The nature of L-kurtosis and L-skewness of the distribution have been studied in detail. The method for simulating random sample from FSLD has been discussed. The estimation of parameters is approached by the method of matching L-moments and the standard error of estimates have been derived. Using simulated data it has been shown that the method can provide reasonably good estimates of the parameters with smaller standard deviations of the estimates for increased sample size. Applications based on real life data shows a good fit based on some well known goodness of fit methods. Hence this distribution is ideal for any skewed data with a flat peak.

Acknowledgements The corresponding author acknowledges the Department of Science and Technology (DST), Government of India for her financial support through DST-INSPIRE fellowship with award no. IF130343. The authors are thankful to the anonymous reviewer for their various suggestions on the earlier version.

Appendix

1. **Proof of Theorem** 3:

Proof If X has a quantile distribution $Q(p)$, then the mean (μ), variance (Var(X)), third and fourth order central moments (μ_3 and μ_4 respectively) of X in terms of quantile function, $Q(p)$ are respectively defined as

$$\mu = \int_0^1 Q(p)dp \tag{1}$$

$$\text{Var}(X) = \int_0^1 [Q(p) - \mu]^2 dp \tag{2}$$

$$\mu_3 = \int_0^1 [Q(p) - \mu]^3 dp \tag{3}$$

$$\mu_4 = \int_0^1 [Q(p) - \mu]^4 dp \tag{4}$$

Simplifying these expressions using results ($\xi(i, j)$ and ($\nu(i, j)$) given in (5) and (6), the proof follows.
$\xi(i, j)$ is defined as

$$\begin{aligned}\xi(i, j) &= \int_0^1 ln^i(p) ln^j(1-p)dp \\ &= \frac{\partial^{i+j}}{\partial u^i \partial v^j} B(u+1, v+1)\bigg|_{u=v=0} \quad ; i, j = 1, 2, 3, 4, ...\end{aligned} \tag{5}$$

The expression is simplified using Leibnitz rule, and $\nu(i, j)$ is obtained from expression (2.6.3.2) in Prudnikov et al. (1986),

$$\begin{aligned}\nu(i,j) &= \int_0^1 p^i ln^j(p)dp \\ &= (-1)^j \frac{j!}{(i+1)^{j+1}} \quad ; i,j = 1,2,3,4,... \end{aligned} \tag{6}$$

□

2. **Proof of Theorem 6**:

Lemma 1 *The L-moments λ_r, $r = 1, 2, ...$ of a real-valued random variable X exist if and only if X has finite mean (Hosking 1990) and the rth order L-moment of X can be written in terms of quantile function as*

$$\lambda_r = \int_0^1 P^*_{r-1}(p)Q(p)dp$$

where,

$$P^*_r(p) = \sum_{l=0}^{r} (-1)^{r+l} \binom{r}{l}\binom{r+l}{l} p^l \tag{7}$$

is the rth order shifted Legendre polynomial. .

Proof Since the FSLD has a finite mean defined in (3), its L-moments exist. λ_1 is the mean as defined in (3) and needs no further proof.

Since, $\int_0^1 P^*_{r-1}(p)dp = 0$ for $r > 1$ & $P^*_{r-1}(p) = (-1)^{r-1}P^*_{r-1}(1-p)$

$$\begin{aligned}\therefore \lambda_r &= \int_0^1 (\alpha + \beta(1-\delta)ln(p) - \beta\delta ln(1-p) + \kappa p)\, P^*_{r-1}(p)dp \\ &= \beta(2\delta-1)^{r \ mod \ 2} \int_0^1 (-ln(p))P^*_{r-1}(p)dp + \kappa \int_0^1 pP^*_{r-1}(p)dp\end{aligned}$$

Now, using the representation of $(-ln(p))$ in terms of shifted Legendre polynomial and after substantial simplification using the orthogonality relation we get,

$$\int_0^1 (-ln(p))P^*_{r-1}(p)dp = \frac{(-1)^{r-1}}{r(r-1)}$$

and

$$\int_0^1 pP^*_{r-1}(p)dp = \begin{cases} \frac{1}{2} & , r = 1 \\ \frac{1}{6} & , r = 2 \\ 0 & , r > 2 \end{cases}$$

Using these results and simplifying we get the rth order L-moment of FSLD as given in (6).

The rth order L-moment ratio is defined as

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3, 4, 5, ...$$

Hence using the expression for rth order L-moment in the definition of L-moment ratio, we get the result given in (7). □

3. **Proof of Theorem 8**

Proof The expectation of rth order statistics in terms of the quantile function (Gilchrist 2000) is given by

$$E(X_{r:n}) = \frac{1}{B(r, n-r+1)} \int_0^1 Q(p) p^{r-1} (1-p)^{n-r} dp \tag{8}$$

Thus, for a ordered sample from FSLD,

$$\begin{aligned} E(X_{r:n}) &= \frac{1}{B(r, n-r+1)} \int_0^1 [\alpha + \beta(1-\delta)\ln(p) - \beta\delta\ln(1-p) + \kappa p]\, p^{r-1}(1-p)^{n-r} dp \\ &= \alpha + \frac{\kappa r}{n+1} + \frac{\beta}{B(r, n-r+1)} \left[(1-\delta) \int_0^1 \ln(p) p^{r-1} (1-p)^{n-r} dp \right. \\ &\qquad \left. -\delta \int_0^1 \ln(p)(1-p)^{r-1} p^{n-r} dp \right] \end{aligned} \tag{9}$$

Equation (9) is simplified using expression (2.6.5.3) in Prudnikov et al. (1986) and the final result in (3) is obtained. □

4. **Proof of Theorem 9**:

Proof The first part of the proof is straight forward and so can be easily proved by equalizing the sample L-moments l_1, l_2, l_3 and l_4 with the corresponding L-moments of the FSLD given in Theorem 6. □

Lemma 2 *Let X be a real-valued random variable with cumulative distribution function F, quantile density function $q(p)$, L-moments λ_r and finite variance. Let l_r, $r = 1, 2, ..., m$ be sample L-moments calculated from a random sample of size n drawn from the distribution of X. Let $\widehat{\theta}$ be the L-moment estimator of θ. Using the asymptotic sampling distribution for l_r (Hosking 1986, 1990), Van Staden (2013) and Van Staden and King (2015) have shown that as $n \to \infty$, $n^{1/2}(\widehat{\theta_r} - \theta_r)$ converge in distribution to the multivariate normal distribution $N(0, \Theta)$, where the elements $\Theta_{rs}(r, s = 1, 2, ..., m)$ of $\Theta = G \Lambda G^{'}$ are given by*

$$\Theta_{rs} = \sum_{u=1}^{m} \sum_{v=1}^{m} G_{ru} \Lambda_{uv} G_{sv} \tag{10}$$

with

$$G_{rs} = \frac{\partial \theta_r}{\partial \lambda_s} \tag{11}$$

So that using Lemma 2 the asymptotic variance covariance matrix of L-moments estimator $\widehat{\theta}$ of θ is given by

$$nVar(\widehat{\theta}) = n\Theta = nG\Lambda G' \tag{12}$$

where

$$G = \begin{pmatrix} \frac{\partial\alpha}{\partial\lambda_1} & \frac{\partial\alpha}{\partial\lambda_2} & \frac{\partial\alpha}{\partial\lambda_3} & \frac{\partial\alpha}{\partial\lambda_4} \\ \frac{\partial\beta}{\partial\lambda_1} & \frac{\partial\beta}{\partial\lambda_2} & \frac{\partial\beta}{\partial\lambda_3} & \frac{\partial\beta}{\partial\lambda_4} \\ \frac{\partial\delta}{\partial\lambda_1} & \frac{\partial\delta}{\partial\lambda_2} & \frac{\partial\delta}{\partial\lambda_3} & \frac{\partial\delta}{\partial\lambda_4} \\ \frac{\partial\kappa}{\partial\lambda_1} & \frac{\partial\kappa}{\partial\lambda_2} & \frac{\partial\kappa}{\partial\lambda_3} & \frac{\partial\kappa}{\partial\lambda_4} \end{pmatrix}$$

and Λ is a symmetric matrix given by

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & \lambda_{24} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \lambda_{34} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \lambda_{44} \end{pmatrix}$$

with the (r, s)th $(r, s = 1, 2, 3, 4)$ element as

$$\begin{aligned} \lambda_{rs} &= \lim_{n\to\infty} \\ &= \int_0^1 \int_0^v \left[P^*_{r-1}(u)P^*_{s-1}(v) + P^*_{s-1}(u)P^*_{r-1}(v)\right] u(1-v)q(u)q(v)dudv \end{aligned} \tag{13}$$

where $P^*_r(p)$ is the rth shifted Legendre polynomial as defined in (7).

After simplification we get,

$$G = \begin{pmatrix} 1 & -3 & -6 & 18 \\ 0 & 0 & 0 & 12 \\ 0 & 0 & \frac{3}{\beta} & \frac{6(1-2\delta)}{\beta} \\ 0 & 1 & 0 & -6 \end{pmatrix}$$

Now, Van Staden and King (2015) has obtained $\Xi(i, j)$ using expressions (4.291.4) and (4.293.8) in Gradshteyn and Ryzhik (2007) as,

$$\Xi(i, j) = \begin{cases} \dfrac{\pi^2}{6} - \sum_{m=1}^{i-1} \dfrac{1}{m^2}, & j = -1 \\ \frac{1}{j+1}(\psi(j+2) - \psi(1)) - \sum_{m=1}^{i-1} \frac{1}{m(m+j+1)}, & j > -1 \end{cases} \tag{14}$$

for $i = 2, 3, 4, ...$, where $\psi(i)$ is the digamma function. Hence, using the result from (13) and (14) and simplifying we obtain the elements of the matrix Λ as

$$\lambda_{11} = \beta^2[1 - 4\delta(1-\delta)] + \frac{\kappa}{2}\left(\beta + \frac{\kappa}{6}\right) + \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{22} = \frac{\beta^2}{3}[1 + 8\delta(1-\delta)] - \frac{\kappa}{18}\left(\beta - \frac{\kappa}{10}\right) - \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{33} = \frac{\beta^2}{15}[2\beta^2 - 53\delta(1-\delta)] + \frac{\kappa}{30}\left(\frac{\beta}{2} + \frac{\kappa}{7}\right) + \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{44} = \frac{\beta^2}{7}\left[\frac{1}{2} + \frac{199}{9}\delta(1-\delta)\right] + \frac{\kappa}{70}\left(\frac{\beta}{2} + \frac{\kappa}{9}\right) - \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{12} = \lambda_{21} = \beta\left(\frac{\beta}{2} + \frac{\kappa}{9}\right)(2\delta - 1)$$

$$\lambda_{13} = \lambda_{31} = \frac{\beta^2}{3}\left[\frac{1}{2} - 11\delta(1-\delta)\right] - \frac{\kappa}{12}\left(\frac{\beta}{2} + \frac{\kappa}{5}\right) + \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{14} = \lambda_{41} = \frac{\beta}{12}\left(\beta - \frac{\kappa}{10}\right)(2\delta - 1)$$

$$\lambda_{23} = \lambda_{32} = \frac{\beta}{6}\left(\beta + \frac{\kappa}{60}\right)(2\delta - 1)$$

$$\lambda_{24} = \lambda_{42} = \frac{\beta^2}{3}\left[\frac{1}{4} + \frac{28}{3}\delta(1-\delta)\right] - \frac{\kappa}{60}\left(\frac{\beta}{2} + \frac{\kappa}{7}\right) - \frac{1}{3}\beta^2\pi^2\delta(1-\delta)$$

$$\lambda_{34} = \lambda_{43} = \frac{\beta}{12}\left(\beta + \frac{\kappa}{35}\right)(2\delta - 1)$$

The final result in (2) is obtained by multiplying the matrix G with Λ and then with G'. □

References

Aljarrah, M. A., Lee, C., & Famoye, F. (2014). On generating tx family of distributions using quantile functions. *Journal of Statistical Distributions and Applications*, *1*(1), 1–17.

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A first course in order statistics* (Vol 54). Siam.

Asgharzadeh, A., Esmaeili, L., Nadarajah, S., & Shih, S. (2013). A generalized skew logistic distribution. *REVSTAT-Statistical Journal*, *11*(3), 317–338.

Bader, M., & Priest, A. (1982). Statistical aspects of fibre and bundle strength in hybrid composites. In *Progress in science and engineering of composites* (pp. 1129–1136).

Balakrishnan, N., & Rao, C. (1998a). Handbook of statistics. In *Order statistics: applications* (Vol. 17).

Balakrishnan, N., & Rao, C. R. (1998b). *Order statistics: theory and methods* (Vol. 16). Elsevier Amsterdam.

Balakrishnan, N., & So, H. (2015). A generalization of quantile-based skew logistic distribution of van staden and king. *Statistics & Probability Letters*, *107*, 44–51.

Freimer, M., Kollia, G., Mudholkar, G. S., & Lin, C. T. (1988). A study of the generalized tukey lambda family. *Communications in Statistics: Theory and Methods*, *17*(10), 3547–3567.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. CRC Press.

Govindarajulu, Z. (1977). A class of distributions useful in life testing and reliability with applications to non-parametric testing. In C. P. Tsokos, & I. Shimi (Eds.), *The theory and applications of reliability*. New York: Academic Press.

Gradshteyn, I., & Ryzhik, I. (2007). In Jeffrey, A. & Zwillinger, D. (Eds.), *Table of integrals, series and products* (7th ed.). Burlington, Massachusetts: Academic Press.

Gupta, R. D., & Kundu, D. (2010). Generalized logistic distributions. *Journal of Applied Statistical Science*, *18*(1), 51.

Hankin, R. K., & Lee, A. (2006). A new family of non-negative distributions. *Australian & New Zealand Journal of Statistics*, *48*(1), 67–78.

Hosking, J. (1986). *The theory of probability weighted moments*. TJ Watson Research Center: IBM Research Division.

Hosking, J. R. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* (pp. 105–124).

Midhu, N., Sankaran, P., & Nair, N. U. (2013). A class of distributions with the linear mean residual quantile function and it's generalizations. *Statistical Methodology*, *15*, 1–24.

Midhu, N., Sankaran, P., & Nair, N. U. (2014). A class of distributions with linear hazard quantile function. *Communications in Statistics-Theory and Methods*, *43*(17), 3674–3689.

Nadarajah, S. (2009). The skew logistic distribution. *AStA Advances in Statistical Analysis*, *93*(2), 187–203.

Nair, N. U., Sankaran, P., & Vineshkumar, B. (2012). The govindarajulu distribution: Some properties and applications. *Communications in Statistics-Theory and Methods*, *41*(24), 4391–4406.

Prudnikov, A., Brychkov, Y., & Marichev, O. (1986). *Integrals and series: volume 1, elementary functions (Gordon and Breach Science Publishers, New York, NY)* (Vol. 1). Gordon and Breach Science Publishers.

Ramberg, J. S. (1975). A probability distribution with applications to monte carlo simulation studies. In *A modern course on statistical distributions in scientific work* (pp. 51–64). Springer.

Ramberg, J. S., & Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM*, *15*(11), 987–990.

Ramberg, J. S., & Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, *17*(2), 78–82.

Reiss, R.-D. (1989). *Approximate distributions of order statistics*. Springer.

Sankaran, P., Nair, N. U., & Midhu, N. (2016). A new quantile function with applications to reliability analysis. *Communications in Statistics-Simulation and Computation*, *45*(2), 566–582.

Thas, O. (2010). *Comparing distributions*. Springer.

Thomas, B., Midhu, N., & Sankaran, P. (2015). A software reliability model using mean residual quantile function. *Journal of Applied Statistics*, *42*(7), 1442–1457.

Tukey, J. W. (1960). The practical relationship between the common transformations of percentages of counts and of amounts. *Statistical Techniques Research Group Technical Report*, 36.

Van Staden, P. J. (2013). *Modeling of generalized families of probability distribution in the quantile statistical universe*. Ph.D. thesis, Department of Statistics, University of Pretoria, Pretoria, South Africa.

Van Staden, P. J., & King, R. A. (2015). The quantile-based skew logistic distribution. *Statistics & Probability Letters*, *96*, 109–116.

Wahed, A., & Ali, M. M. (2001). The skew-logistic distribution. *Journal of Statistical Research*, *35*, 71–80.

A Note on Monte-Carlo Jackknife: McJack and Big Data

Jiming Jiang

Abstract This short note explains how Big Data computing issues can occur even when the data size is not large compared to what a computer scientist may call Big Data, in obtaining measures of uncertainty using a Monte-Carlo jackknife method.

Keywords Big data · Jackknife · Measure of uncertainty · Monte-Carlo

Resampling methods have played fundamental roles in statistical inference. As data are usually considered as samples from a population, rather than the population itself, any attempt in learning about the population through the data must take into account uncertainty in making the inference. In many complex problems the uncertainty measure is not easy to obtain. In such cases, resampling methods such as the jackknife and the bootstrap often have advantages. Our interest here is particularly in statistical inference post-model selection. Jiang et al. (2016) proposed a Monte-Carlo jackknife method, called McJack, for such settings.

Suppose that a quantity of interest, say, mean squared prediction errors (MSPE), is associated with a vector, ψ, of parameters. Let $b(\psi)$ denote the quantity of interest. Suppose that the observations are divided into m independent clusters. Jiang et al. (2002) showed that, under assumptions,

$$b(\widehat{\psi}) - \frac{m-1}{m}\sum_{j=1}^{m}\{b(\widehat{\psi}_{-j}) - b(\widehat{\psi})\} \tag{1}$$

is a second-order unbiased estimator of $b(\psi)$. Here $\widehat{\psi}_{-j}$ is the M-estimator of ψ (Jiang et al. 2002) obtained without cluster j.

Computationally, a key condition for the jackknife to work is that one knows how to compute $b(\psi)$ given ψ; in other words, the function $b(\cdot)$ is known. In light of this, Jiang et al. (2002) assumes that $b(\cdot)$ has an analytic expression. Jiang et al. (2016) extended (1) to cases where $b(\cdot)$ does not have an analytic expression. In the latter

J. Jiang (✉)
University of California, Davis, CA, USA
e-mail: jimjiang@ucdavis.edu

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_7

case, the computation is done via Monte-Carlo simulation, provided that one knows how to compute $b(\psi)$ if data can be generated repeated under ψ.

The above result on the second-order unbiasedness of McJack is a statistical theory, not a computational one. First, the condition is quite tough to meet, computationally. For example, in analysis of a big energy data (Kim et al. 2015), a main characteristic of interest is peak energy usage, which is expressed as a non-linear mixed effect. In the context of estimation, or prediction, of mixed effects, a standard measure of uncertainty is the MSPE. The well-known Prasad-Rao method of MSPE estimation (e.g., Jiang and Lahiri 2006) does not apply to the current non-linear mixed effects; thus, as an alternative, resampling methods are considered.

More specifically, McJack was proposed for estimating a smooth, monotone function of the MSPE. The McJack estimator can be expressed in the form of (1) but with assistance of Monte-Carlo simulation. Namely, a difficulty for (1) is that $b_i(\cdot)$ does not have an analytic form. Jiang et al. (2016) proposes to approximate the $b_i(\cdot)$ function via Monte-Carlo simulation. It is shown that the Monte-Carlo sample size, K, needs to be of higher order than m^2 in order to achieve the second-order unbiasedness. For the energy data, m is about 100,000, which is the number of different M-estimators, $\hat{\psi}_{-j}$, plus $\hat{\psi}$, that one needs to compute. Given the M-estimators, one needs to evaluate each of the 100,000 terms in the summation in $\sum_{j=1}^{m}\{b_i(\hat{\psi}_{-j}) - b_i(\hat{\psi})\}$, via the Monte-Carlo simulation of size K, which, by any conservative estimate, would require, in all, at least 10^{20} repeated computations of the non-linear mixed effects, which themselves do not have analytic expressions. With the size of the data, and our current computational resources, it would take months to complete the MSPE estimation. If, furthermore, one intends to evaluate the empirical performance of the above MSPE estimation procedure, the total computing time can easily be multiplied by a thousand! In fact, in the latter case, even if m is much smaller than for the energy data, say, a few thousands, the amount of computation is still quite formidable!

In a way, we have created a Big Data problem ourselves. Of course, with a much more advanced high-speed computer, such a problem would no longer be challenging, but one cannot just wait, and hope that one day the dream of the high-speed computer comes true (it will, on some day, of course). Plus, not everyone has access to such a high-speed computer, even if it exists. There is something we can do by making McJack more efficient. For example, if the original sample size is relatively small, there is no need to use a very large Monte-Carlo sample size, because the statistical (chance) error is going to dominate anyway. On the other hand, for large or huge sample size, such as in a Big Data situation, one likely has to entertain an even larger Monte-Carlo sample size. Is it still necessary to insist the second-order unbiasedness? These questions lead to some important and interesting research topics regarding computational efficiency of McJack. What needs to be solved is an optimization problem in terms of balances between statistical precision and computational cost. In the literature of statistics, computer science, and computational mathematics, such results are relatively rare. It would take a combined skills from these fields to solve the problem.

References

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *TEST*, *15*, 1–96.

Jiang, J., Lahiri, P., & Nguyen, T. (2016). A unified Monte-Carlo jackknife for small area estimation after model selection. *Annals of Mathematical Sciences and Applications* in press.

Jiang, J., Lahiri, P., & Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics*, *30*, 1782–1810.

Kim, T., Lee, D., Choi, J., Spurlock, A., Sim, A., Todd, A., et al. (2015). Extracting baseline electricity usage with gradient tree boosting. In *Proceedings of International Conference on Big Data Intelligence and Computing*, Chengdu, Sichuan, China.

A Review of Exoplanets Detection Methods

G. Jogesh Babu

Abstract A brief introduction to the discovery of planets outside the solar system is presented. Statistical challenges in the analysis of noisy Exoplanets data are indicated, with emphasis on NASA's Kepler mission.

Keywords ARMA · ARFIMA · Kepler mission · Pulsar timing method · Radial velocity method · Transits

1 Introduction

Major advances in our understanding of our planetary system used to occur on century-long timescales. The Copernican heliocentric model first proposed that Earth orbited around the Sun, and the Earth came to be considered a planet in 1543; Galileo first observed the moons of Jupiter in 1609, and Uranus was discovered by William Herschel on 13 March 1781. The planet Neptune was first predicted mathematically by Urbain Le Verrier, and the astronomer Johann Gottfried Galle confirmed it on the night of 23 September 1846 by using his observations at the Berlin Observatory.

But philosophers and scientists continued to ask the question: Do other stars have their own solar systems and, if so, do they resemble our system? In the mid–1800s, parallax measurement confirmed that the stars are like our Sun. But despite millennia of speculation, there was no evidence they had their own planetary systems until 1995 when a 'hot Jupiter' was unexpectedly found orbiting 51 Peg with a period of 3 days. This led to a revolution in astronomy: a rush to improve precision of instruments, races to discover more 'exoplanets'—the nickname for planets in other solar systems orbiting other stars—by different means, and the beginning of characterization of their atmospheres, interiors and interactions. Thousands of (mostly young) astronomers dove into the field, telescopes changed purposes, major

G. J. Babu (✉)
Department of Statistics and Department of Astronomy & Astrophysics,
The Pennsylvania State University, University Park, PA, USA
e-mail: babu@psu.edu

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_8

resources were diverted from other projects, and almost every prize was awarded to leading researchers. This is an exciting time when we are starting to be able to answer the longstanding questions about exoplanets.

The particular quest for other worlds like our Earth has been rejuvenated by the intense excitement and popular interest surrounding the discovery of hundreds of planets orbiting other stars. There is now a clear evidence for substantial numbers of three types of exoplanets: (a) gas giant Jupiters sometimes very close to the host star; (b) rocky Super-Earths in short period orbits (exoplanets with a mass higher than the Earth's, but substantially below the mass of the Solar System's smaller gas giants Uranus and Neptune); (c) ice giants in long period orbits.

The Extrasolar Planets Encyclopedia http://exoplanet.eu/, NASA Exoplanet Archive http://exoplanetarchive.ipac.caltech.edu/, and New Worlds Atlas https://exoplanets.nasa.gov/newworldsatlas/ track the day-by-day increase in new discoveries. These websites provide information on the characteristics of the planets as well as those of the stars they orbit. The challenge now is to find planets like Earth: 0.5–2 radius of the Earth, especially those in the *Habitable Zone* (the range of orbits around a star within which a planetary surface can support liquid water on the surface of the planet necessary for the support of life).

It is estimated that our galaxy has more than 100 billion stars, and many of them are likely to have planets orbiting them. However, extremely tiny number of these planets can be 'seen' by direct observation. As planets do not emit any electromagnetic radiation of their own, direct observation of them are beyond the reach of current technology. We can only infer their existence by indirect methods. Light emitted by stars mask the presence of planets orbiting them. Even the reflected light from their host star is only a tiny fraction and not detectable due to enormous distances of these stars from the Earth. For example, a planet orbiting the nearest star Proxima Centauri would be 7,000 times more distant than 'Pluto'.

2 Methods of Exoplanet Detection

Earth has 10^{-7} the mass and 10^{-4} the surface area of the Sun and the nearest stars are 10^{6} times more distant than the Sun. Detecting planets is thus exceedingly difficult by any method: gravitational pull changing the host star's radial velocity; periodic eclipses of a tiny portion of the starlight; direct imaging in telescopes; or gravitational lensing effects involving stars that fortuitously lie along the line-of-sight. But as the precision of instruments improved, it was recognized that a major impediment to discovery is that the stars themselves show temporal variability, both in brightness and in radial velocity. The causes are readily seen in the Sun—dark sunspots that come and go with the Sun's $\sim$30 day rotation; stochastic occurrences of short-lived prominences and flares; seething motions on the surface due to convection, but were found to have greater amplitude on a large fraction of stars.

The most important efforts are photometric surveys (where the brightness of 10^{5}–10^{8} stars are monitored with great precision to find periodic planetary transits) and spectroscopic surveys (where the spectra of 10^1–10^2 stars are monitored with great precision to find periodic radial velocity variations). Efforts use both ground-based telescopes (typical costs \$1–10 million) and space–based telescopes (typical costs \$0.1–1 billion). A particularly important project is NASA's Kepler mission that made ∼70,000 extremely precise evenly-spaced photometric observations of ∼200,000 stars every ∼29 min during 2009–2013, resulting in the discovery of several thousand probable planetary systems by the Kepler Team.

Some of the exoplanet detection methods are briefly described below.

2.1 Doppler Shift/Radial Velocity Method

As a planet orbits a star, the star also moves in a small orbit around their common center of mass. An exoplanet orbiting a star thus produces periodic changes in position and velocity of the star. The radial velocity (RV) method for planet detection is based on the detection of variations in the velocity of the central star, due to the changing direction and the gravitational pull from an (unseen) exoplanet as it orbits the star. When the star moves towards us, its spectrum is blue-shifted, while it is red-shifted when it moves away from us. By regularly looking at the spectrum of a star—and so, measuring its Doppler velocity—one can see if it moves periodically due to the influence of a companion. The RV method has successfully detected hundreds of mostly-Jupiter-mass exoplanets since 1995.

Empirical detection using a generalized Fourier transform, known as the Lomb-Scargle periodogram, can sometimes be effective for discovering planets from irregularly spaced RV time series. But for more complex multi-planet systems, high-dimensional Bayesian statistical modeling of the RV data is effective because the shape of the variations must satisfy deterministic Newtonian laws, and prior probability distributions over a space of orbital parameters are based on physical considerations. Markov chain Monte Carlo (MCMC) methods are used in its implementation. Until ∼2012, the radial-velocity method was by far the most productive technique used by planet hunters. The method is distance independent in theory, but requires high signal-to-noise ratios to achieve high precision, and so is generally only used for relatively nearby stars. RV method is also known as Doppler spectroscopy.

2.2 Pulsar Timing Method

A pulsar is a highly magnetized rapidly rotating neutron star that emits a beam of electromagnetic radiation, like beacon from a light house. Pulsars emit radio waves extremely regularly as they rotate. The radiation is detected on Earth as precisely timed pulses that are more accurate than an atomic clock so that slight anomalies in the timing of its observed radio pulses due to an orbiting planet can be tracked. Alexsander Wolszczan (Penn State) and Dale Frail (National Radio Astronomy

Observatory in Socorro, New Mexico) used this Pulsar Timing method to detect the first confirmed exoplanet in 1992. Pulsars are so rare that very few exoplanets have been detected by this method. This method was not originally designed for the detection of planets, but is so sensitive that it is capable of detecting planets far smaller than any other known method can, down to less than one tenth the mass of Earth.

The main drawback of the pulsar-timing method is that pulsars occur only after a very massive star explodes (supernova), so these post-supernova planets have little to do with ordinary planets that form simultaneously with the star formation.

2.3 Direct Imaging Method

Planets are extremely faint light sources compared to stars, emitting in the infrared band and reflecting starlight in the visible band. This faint light tends to be lost in the glare from their parent star. So in general, it is very difficult to detect and resolve them directly from their host star. Planets orbiting far enough from stars can be resolved in the best telescopes, even though they reflect very little starlight. The images are made at infrared where the planet is brighter than it is at visible wavelengths. Coronagraphs

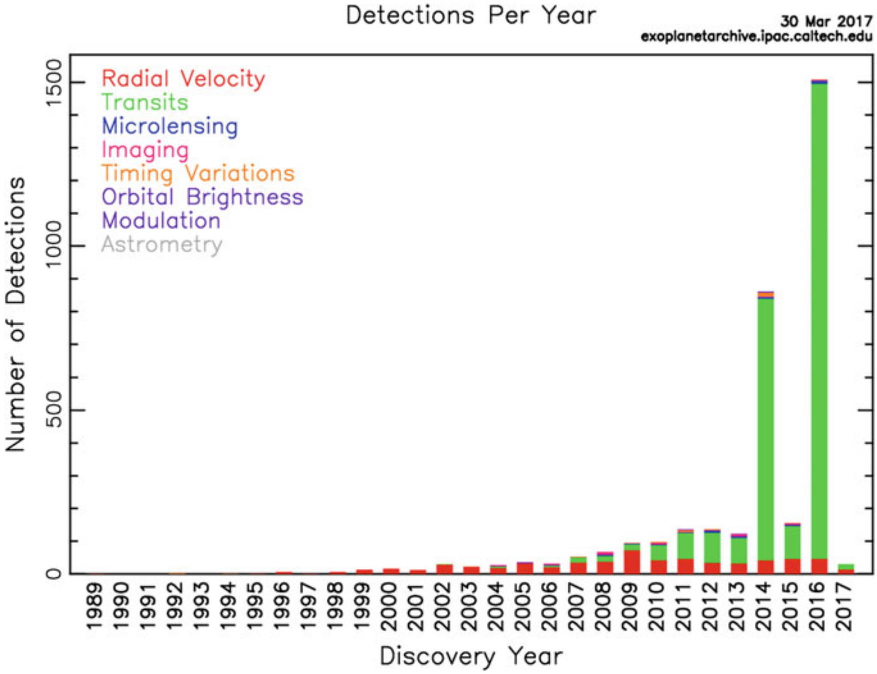


Fig. 1 Planet detection methods that have been proved successful. *NASA Exoplanet Archive*

are used to block light from the star while leaving the planet visible. Only a few cases are detectable. This method is most effective when (Fig. 1):

- the star system is relatively near to the Sun
- the planet is especially large (considerably larger than Jupiter)
- the planet is widely separated from its parent star
- the planet is young and massive so that it emits more intense infrared radiation.

In 2004 the European Southern Observatory's Very Large Telescope (VLT) array in Chile, which operates at visible and infrared wavebands detected a planet by this method. The planet is several times more massive than Jupiter, and have an orbital radius grater than 40 Astronomical Units (AU). One AU is the mean distance between the Earth and the Sun. The recently commissioned Gemini Planet Finder instrument is expected to significantly increase the number of directly imaged planets. However, currently, this method is limited to giant planets at large distances from their host stars. Direct imaging method can be used to measure the planet's orbit.

2.4 *Gravitational Microlensing Method*

This was the first method capable of detecting planets of Earth-like mass around ordinary main sequence stars and is most sensitive to detect planets around 1–10 AU away from Sun-like stars. This method derives from one of the insights of Albert Einstein's theory of general relativity: gravity bends space-time. The method was proposed in 1991 to look for binary companions to stars and used to detect exoplanets in 1992. Successes with the method dates back to 2002, when a group of Polish astronomers developed a workable technique. During one month, they found several possible planets, though limitations in the observations prevented clear confirmation. First microlensing and imaging planets were discovered in 2004. Since then, several confirmed extrasolar planets have been detected using microlensing.

This method has some disadvantages. Lensing cannot be repeated because the chance alignment never occurs again. Only a few cases are seen despite extensive monitoring of millions of stars. The detected planets will tend to be several kiloparsecs away, so follow-up observations with other methods are usually impossible. A parsec is the distance from the Sun to an astronomical object that has a parallax angle of one arc-second; the nearest stars are roughly 1 parsec (2.06×10^5 AU or 3.25 light years) away.

2.5 *Transit Method*

When a planet crosses (transits) in front of its parent star's disk, then the observed visual brightness of the star drops a small amount. The amount the star dims depends on the relative sizes of the star and the planet. Most transits depths are $<0.1\%$.

The transit method relies upon carefully monitoring the brightness of a star. In order to measure the mass of a planet, and rule out other phenomena that can mimic the presence of a planet transiting a star, candidate transiting planets are followed up with the radial velocity method of detecting extrasolar planets. For example, in the case of planet HD 209458, the star dims 1.7%. Recall that Venus Transit of Sun occurred on June 5, 2012. The next transit of Venus that can be observed from the Earth occurs after a little over 100 years. The first transiting planet was discovered in 2002.

There are two major disadvantages for this method. Planetary transits are only observable for planets whose orbits happen to be perfectly aligned from the Earth's vantage point. About 10% of planets with small orbits have such alignment, and the fraction decreases for planets with larger orbits. For a planet orbiting a Sun-sized star at 1 AU, the chance of a random alignment producing a transit is 0.47%. Additional astronomical observations are necessary to reduce false positive rates due to non-planetary signals.

While the method cannot guarantee that any particular star is not a host to planets, by scanning large areas of the sky containing thousands or even hundreds of thousands of stars at once, transit surveys can find extrasolar planets at a rate that exceeds that of the radial-velocity method. This premise led to the launch of Kepler Space Telescope by NASA on March 7, 2009 to scan a large patch of the sky to discover Earth-like planets orbiting other stars in our galaxy.

3 NASA's Kepler Mission

The Kepler mission is named after the 17th century German mathematician and astronomer Johannes Kepler, best known for his laws of planetary motion and a contemporary of Galileo Galilei. The Kepler satellite was designed to monitor a patch of our galaxy for planetary systems. One of the aims is to detect Earth-size planets in orbits around their host stars that are likely to support life, and estimate the fraction of stars in our galaxy that host such planets. Kepler is designed to stare at hundreds of thousands of stars continuously to monitor for any dips in the light for possible transits by planets. The mission has already established that our solar system is vastly different from the many planetary systems with multiple planets in the Galaxy.

During its four-year prime mission from 2009 to 2013, the Kepler space telescope simultaneously and continuously measured the brightness of more than 150,000 stars, looking for the telltale periodic dimming that would indicate the presence of an orbiting planet. From these dimmings, or transits, and information about the parent star, researchers could determine a planet's size (radius), the time it takes to orbit its star (period), and the amount of energy received from the host star. By folding the light curves so that transits of each planet line up, one can measure the transit depth and duration, providing information about the size and orbits of these planets. All Kepler data is public as of October 2012.

The Kepler team reports ∼3000 highly-probable planets, of which a few (most larger than Earth) are in Habitable Zones. On July 14, 2012, one of the spacecraft's four reaction wheels used for pointing the spacecraft stopped functioning, and then, on May 11, 2013, a second reaction wheel failed, disabling the collection of science data. In May 2014, the Kepler spacecraft began a new mission, *K2*, utilizing the remaining good reaction wheels to observe parts of the sky along the ecliptic plane, the orbital path of the Earth about the Sun where the familiar constellations of the zodiac lie. This new mission provides scientists with an opportunity to search for even more exoplanets. The spacecraft continues to collect data in its new mission. Though the Kepler mission was initially planned for a life of 3.5 years, it lasted more than 7 years and still continuing. Since Kepler launched in 2009, 21 planets less than twice the size of Earth have been discovered in the habitable zones of their stars.

As of March 30, 2017, the number of exoplanets confirmed is 3472, of which 2331 were from Kepler mission and community. Also 581 multi-planet systems were discovered by various methods. As of March 23, 2017, the number of exoplanet candidates (likely discovery, but still needs to be verified) were put at 4,496 by the Kepler project. K2 mission confirmed 147 exoplanets out of 520 K2 candidates.

4 Statistical Methods

Statistical analysis is essential to every type of exoplanet detection (except perhaps microlensing) as the planetary signal is only a tiny fraction (10^{-3}–10^{-9}) of the stellar signal and is often overwhelmed by uninteresting instrumental or stellar effects. Concerted efforts at developing advanced statistical methods for exoplanet discovery goes back to at least 2006, when the first semester long astrostatistics program was organized by the author at Statistical and Applied Mathematical Sciences Institute in Research Triangle Park, North Carolina.

Bayesian methodology was used for RV and transit photometry, strongly interacting RV systems, and microlensing data, since 2006. Bayesian statistics for Kepler data for triple star and multi-planet systems have been used since 2011. Since 2014 Hierarchical Bayesian modeling for exoplanet populations have been in use.

Statistical extrapolation to estimate the planets missing from Kepler's survey (e.g. misaligned orbits, brightness dip too weak to detect) give a spectacular inference: Most stars have multi-planet systems. Very few Earth like planets are directly discovered with Kepler, mainly due to the short duration of the observations (4 years), and extrapolation to exoplanets of small mass and large orbits are quite uncertain. Estimates of the fraction of stars with an Earth-like planet in the Habitable Zone range from ∼22% (Petigura et al., PNAS 2013) to <5% (Forman-Mackey et al., ApJ 2014). The latter estimate is based on hierarchical Bayesian modeling that incorporates observational uncertainties and detection efficiency.

Many stars are magnetically active (starspots, flares, etc.) which produce radial velocity and photometric variations that prevent faint planetary signals from being efficiently detected. Recent collaboration by astronomers and statisticians started investigating whether autoregressive moving average (ARMA) modeling of stellar variations can improve sensitivity to transits in Kepler light curves. ARMA models can also assist in reducing False Positives: sometimes the eye can see repeated patterns in a light curve, or a peak appears in a noisy periodogram, due only to autoregressive processes and noise, but it may turn out that we are sampling many cycles of a periodic variation.

A periodogram calculates the significance of different frequencies in time-series data to identify any intrinsic periodic signals. The Lomb-Scargle periodogram (Scargle, ApJ 1982) is similar to the Fourier Transform, but is optimized for unevenly time-sampled data. Other periodograms, such as the Box Least-Squares algorithm (Kovacs et al., A&A 2002), are adapted for different shapes in periodic signals. Unevenly sampled data is particularly common in astronomy, where the target might rise and set over several nights, or spacecraft observations stop to download the data. Many different frequencies and candidate periodic signals are evaluated. The statistical significance of each frequency is difficult to complete in the presence of autoregressive noise, irregular observations, and non-sinusoidal shapes. Many variations of autoregressive modeling that is common in econometrics, such as ARIMA and ARFIMA, can be useful in analyzing the Kepler light-curves.

5 Conclusions

The discovery of exoplanets has a confused history. Several premature exoplanet claims were made during 1855–1991, often due to inadequate statistical evaluation. Astronomers learned that they should be more careful in their analysis. The first exoplanet discoveries built on simplistic statistics and strong paranoia were made in 1989, 1992, and 1995. Since the mid-1990s, a scientific revolution has occurred. There was rapid progress with RV and frequentist methods with less paranoia in late 1990s to early 2000s. But astronomers started worrying about correlated noise in 2006. Serious population analyses of Kepler data started around 2010. It is still difficult to detect the smallest planets, or to decide whether 3 or 4 planets orbit a star.

An exoplanet, a potentially habitable planet orbiting Proxima Centauri, a star closest to our Sun that sits just 4.24 light-years away was reported on 24 August 2016. This was found by radial velocity methods. Even though the planet's signal was detected earlier, it required a nightly follow-up campaign to produce a convincing dataset. By its nature, the exoplanet data is noisy. The data together with host star's own activity may create a false conclusion of a non-existent planet signature. Without digging deeper, few astronomers monitoring Alpha Centauri B, prematurely announced the discovery of Earth-mass planet in 2012, but subsequent studies repudiated the claim.

Most stars host orbiting planetary systems, and a few percent are calculated to have Earth-like planets in Earth-like orbits. This implies there are hundreds of millions or billions of Earths in the Milky Way Galaxy. However, these are statistical inferences from much sparser datasets where planetary signals are 10^{-4}–10^{-6} times the host star signal and are detected in only a tiny fraction of stars. Massive surveys with extremely accurate measurements, accompanied by careful statistical analysis, are needed to detect exoplanets and make inferences about their cosmic population. Uninteresting variations of the host star mask the tiny effects of the orbiting planet. Autoregressive modeling, Gaussian Processes regression, wavelet analyses, and other techniques are tapped to mitigate these effects so new planets can be discovered.

References

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. (2014). Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs. *Astrophysical Journal*, *795*(1), article id. 64, 12 p.

Kovacs, G., Zucker, S., & Mazeh, T. (2002). A box-fitting algorithm in the search for periodic transits. *Astronomy & Astrophysics*, *391*, 369–377.

Petigura, E. A., Howard, A. W., & Marcy, G. W. (2013). Prevalence of Earth-size planets orbiting Sun-like stars. *Proceedings of the National Academy of Sciences*, *110*(48), 19273–19278

Scargle, J. D. (1982). Studies in astronomical time series analysis. II—Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, Part 1, *263*, 835–853

A Connection Between the Observed Best Prediction and the Fence Model Selection Method

Thuan Nguyen and Jiming Jiang

Abstract Following our presentation at the Platinum Jubilee on “Observed Best Prediction (OBP) for Small Area Counts”, we discuss a connection between OBP and predictive model selection using the fence methods.

Keywords Fence methods · Observed best prediction
Predictive model selection · Small area estimation

1 Introduction

The fence methods (Jiang et al. 2008; Jiang 2014; Jiang and Nguyen 2015) is a recently developed class of strategies for model selection. The methods fit particularly well in non-conventional and complex model selection problems with practical considerations. The idea involves a procedure to isolate a subgroup of what are known as correct models, of which the optimal model is a member. This is accomplished by constructing a statistical fence, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model is selected from amongst those within the fence according to a criterion which can be made flexible. In particular, the criterion of optimality can incorporate consideration of practical interest, thus making model selection a real life practice.

The fence is typically constructed via an inequality such that a candidate model, M, is in the fence if

$$Q(M) - Q(M_*) \leq c, \tag{1}$$

T. Nguyen (✉)
Oregon Health and Science University, Portland, OR, USA
e-mail: nguythua@ohsu.edu

J. Jiang
University of California, Davis, CA, USA

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_9

where $Q(\cdot)$ is a measure of lack-of-fit, M_* is a baseline model that minimizes Q, such as a full model in many cases, and c is a tuning constant that can be chosen adaptively (e.g., Jiang 2014, Sect. 2). A feature of flexibility of the fence is in the choice of Q. In particular, if prediction is of primary interest, such as in small area estimation (SAE; e.g., Jiang and Lahiri 2006), Q may be chosen by taking into account the objective of prediction. Note that the standard measures of lack-of-fit, such as the negative log-likelihood, and the residual sum of squares, are estimation-based rather than prediction-based.

Recently, Jiang et al. (2011) proposed a new method for SAE, known as the observed best prediction (OBP). A main feature of the OBP is that it is more robust, in terms of predictive performance, to model misspecification compared to the traditional empirical best linear unbiased prediction (EBLUP; e.g., Rao and Molina 2015). Chen et al. (2015) extended OBP to Poisson mixed models for count data. In this short note, we describe a general approach to deriving a predictive measure of lack-of-fit that is motivated by OBP.

Let us first consider a general problem of linear mixed model (LMM) prediction (e.g., Robinson 1991; Jiang 2007, Sect. 2.3). The assumed model is

$$y = X\beta + Zv + e, \tag{2}$$

where X, Z are known matrices; β is a vector of fixed effects; v, e are vectors of random effects and errors, respectively, such that $v \sim N(0, G)$, $e \sim N(0, \Sigma)$, and v, e are uncorrelated. An important issue for model-based statistical inference is the possibility of model misspecification. To take the latter into account, suppose that the true underlying model is

$$y = \mu + Zv + e, \tag{3}$$

where $\mu = \mathrm{E}(\mathbf{y})$. Here, E represents expectation with respect to the true distribution of y, which may be unknown but is not model-dependent. So, if $\mu = X\beta$ for some β, the model is correctly specified; otherwise, the model is misspecified. Our interest is prediction of a vector of mixed effects that can be expressed as

$$\theta = F'\mu + R'v, \tag{4}$$

where F, R are known matrices.

A special case of the above general LMM is the Fay-Herriot model (Fay and Herriot 1979). The model was introduced to estimate the per-capita income of small places with population size less than 1,000:

$$y_i = x_i'\beta + v_i + e_i, \quad i = 1, \dots, m,$$

where x_i is a vector of known covariates, β is a vector of unknown regression coefficients, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that v_i's, e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. The variance A is unknown, but the sampling variances D_i's are assumed known. The assumed model can be expressed as (2) with $X = (x_i')_{1 \le i \le m}$, $Z = I_m$, $G = AI_m$ and $\Sigma = \text{diag}(D_1, \dots, D_m)$. The problem of interest is estimation of the small area means. Let $\mu_i = \text{E}(y_i)$. Then, the small area means can be expressed as $\theta_i = \text{E}(y_i|v_i) = \mu_i + v_i$, under the true underlying model (3). Here, again, E represents the true conditional expectation rather than conditional expectation under the assumed model. Thus, the quantity of interest can be expressed as (4) with $\theta = (\theta_i)_{1 \le i \le m}$, and $F = R = I_m$.

For simplicity, assume that both G and Σ are known. Then, under the assumed model, the best predictor (BP) of θ, in the sense of minimum mean squared prediction error (MSPE), is the conditional expectation,

$$\begin{aligned}\text{E}_M(\theta|y) &= F'\mu + R'\text{E}_M(v|y)\\ &= F'X\beta + R'GZ'V^{-1}(y - X\beta),\end{aligned} \tag{5}$$

where $V = \Sigma + ZGZ'$ and β is the true vector of fixed effects (e.g., Jiang 2007, p. 75). The E_M in (5) denotes conditional expectation under the assumed model, (2), rather than the true model (3). Although model (2) may be subject to model misspecification, it is usually (much) simpler and utilizes the available covariates, X. On the other hand, even if model (3) is correct, or close to be correct, it is too broad to be useful; furthermore, it does not make use of any of the available covariates, which is often practically unacceptable. For these reasons, the assumed model, (2), is always the one of main interest. In other words, one cannot abandon the assumed model; all one could do is to try to do the best under the assumed model, that is, to find the best way to estimate the parameters, in this case β, under the assumed model. The question then is: What is the role that the true model, (3), plays in this business? The answer is that the true model can help to determine the best way to estimate the parameters so that it is more robust to model misspecification. More specifically, the true model is used in evaluating the predictive performance of the BP, (5), to make sure that the evaluation is fair and not model-dependent. This is why, intuitively, the resulting estimator of β is more robust to model misspecification (Jiang et al. 2011). The idea introduced here is particularly important when dealing with model selection, because, obviously, the measure of lack-of-fit has to be objective, or "fair", to all of the candidate models.

To derive the best estimator of β, write $B = R'GZ'V^{-1}$ and $\Gamma = F' - B$. Let $\tilde{\theta}$ denote the right side of (5), where β is understood as a parameter vector to be determined. The predictive performance of $\tilde{\theta}$ is typically measured by the MSPE, defined as $\text{MSPE}(\tilde{\theta}) = \text{E}(|\tilde{\theta} - \theta|^2)$. Here, again, E denotes expectation under the true model. It can be shown (Jiang et al. 2011) that the MSPE can be expressed, alternatively, as

$$\text{MSPE}(\tilde{\theta}) = \text{E}\{(y - X\beta)'\Gamma'\Gamma(y - X\beta) + \cdots\}, \tag{6}$$

where $\cdots$ does not depend on β. Here comes another key point: Unlike $\mathrm{E}(|\tilde{\theta}-\theta|^2)$, the expression inside the expectation on the right side of (6) involves a function of the observed data and β (and nothing else), and something unrelated to β. Therefore, it is natural to estimate β by minimizing the expression inside the expectation on the right side of (6), which is equivalent to minimizing the expression without $\cdots$. This leads to what we call the best predictive estimator, or BPE, of β, given by

$$\hat{\beta}_{\mathrm{BPE}} = (X'\Gamma'\Gamma X)^{-1}X'\Gamma'\Gamma y, \tag{7}$$

assuming that $\Gamma'\Gamma$ is nonsingular and X is full rank. As a comparison, the ML estimator (MLE) of β, under the assumed model, is given by

$$\hat{\beta}_{\mathrm{MLE}} = (X'V^{-1}X)^{-1}X'V^{-1}y, \tag{8}$$

assuming nonsingularity of V. See Jiang et al. (2011) for discussion on the difference between the BPE and MLE in the case of Fay-Herriot model. When the β in the BP is substituted by the BPE, the result is called the observed best predictor, or OBP (Jiang et al. 2011). The latter authors showed that the OBP generally outperforms the EBLUP when the underlying model is misspecified.

To develop a fence method that takes into account the particular interest of mixed model prediction, we can define the measure of lack-of-fit, $Q(M)$, as the minimizer, over β, of the expression without $\cdots$ inside the expectation on the right side of (6). Clearly, this measure is designed specifically for the mixed model prediction problem. Also, when it comes to model selection, it is important that the measure of lack-of-fit is "fair" to every candidate model. The above measure $Q(M)$ has this feature, because the expectation in (6) is under an objective true model. Once we have the measure Q, we can use it in (1) for the fence.

The assumption that G and Σ are known can be relaxed, to some extent, especially for G. In fact, it can be shown that essentially the same derivation as the above goes through, and the resulting measure of lack-of-fit, $Q(M)$, is the minimizer of $(y-X\beta)'\Gamma'\Gamma(y-X\beta) - 2\mathrm{tr}(\Gamma'\Sigma)$, over the parameters assuming that Σ is known.

We now consider another situation. In many cases, the data for the response variables are counts (e.g., Münnich et al. 2009). For simplicity, suppose that the responses are counts, denoted by y_i, and that, in addition, a vector of covariates, x_i, is also available. The model of interest assumes that, given the random effects, v_i, y_i has a Poisson distribution with mean μ_i, such that

$$\log(\mu_i) = x_i'\beta + v_i. \tag{9}$$

The BP of μ_i, under the assumed model, can be expressed as

$$\mathrm{E}_{M,\psi}(\mu_i|y) = g_i(\psi, y_i), \tag{10}$$

where $\mathrm{E}_{M,\psi}$ denotes conditional expectation under the assumed model, M, and parameter vector, ψ, under M, and $g_i(\cdot,\cdot)$ is a known function which does not have an analytic expression. Nevertheless, the g function can be evaluated numerically fairly easily. Following a similar idea, we evaluate the performance of the BP under a broader model, which states that, conditioning on μ_i, y_i is Poisson(μ_i), but the expression (9) is not assumed. In other words, under the broader model, the μ_i's are completely unspecified. Write $\mu = (\mu_i)_{1\le i\le m}$, and $\tilde{\mu} = (\tilde{\mu}_i)_{1\le i\le m}$, where $\tilde{\mu}_i$ is the right side of (10) when ψ is considered as an unknown parameter vector. Consider

$$\begin{aligned}\text{MSPE} &= \mathrm{E}(|\tilde{\mu} - \mu|^2)\\ &= \sum_{i=1}^{m} \mathrm{E}\{g_i(\psi, y_i) - \mu_i\}^2\\ &= \mathrm{E}\left\{\sum_{i=1}^{m} g_i^2(\psi, y_i)\right\} - 2\sum_{i=1}^{m} \mathrm{E}\{g_i(\psi, y_i)\mu_i\} + \sum_{i=1}^{m} \mathrm{E}(\mu_i^2)\\ &= I_1 - 2I_2 + I_3, \end{aligned} \tag{11}$$

where E denotes expectation under the broader model. Note that I_3 does not involve ψ, even though it may be completely unknown. It can be shown that

$$\mathrm{E}\{g_i(\psi, y_i)\mu_i\} = \sum_{k=0}^{\infty} g_i(\psi, k)(k+1)\mathrm{E}\{1_{(y_i=k+1)}\}, \tag{12}$$

where 1_A is the indicator of even A (= 1 if A occurs, and 0 otherwise). Thus, if we define $g_i(\psi, -1) = 0$, we have

$$\mathrm{E}\{g_i(\psi, y_i)\mu_i\} = \mathrm{E}\left\{\sum_{k=0}^{\infty} g_i(\psi, k)(k+1)1_{(y_i=k+1)}\right\} = \mathrm{E}\{g_i(\psi, y_i - 1)y_i\},$$

$$\text{MSPE} = \mathrm{E}\left\{\sum_{i=1}^{m} g_i^2(\psi, y_i) - 2\sum_{i=1}^{m} g_i(\psi, y_i - 1)y_i + \cdots\right\}, \tag{13}$$

where $\cdots$ does not depend on ψ. A predictive measure of lack-of-fit, $Q(M)$, is the expression inside the expectation in (13), but without $+\cdots$, minimized with respect to ψ.

References

Chen, S., Jiang, J., & Nguyen, T. (2015). Observed best prediction for small area counts. *Journal Survey Statistical Method*, *3*, 136–161.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistical Association*, *74*, 269–277.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. New York: Springer.

Jiang, J. (2014) The fence methods. *Advances in Statistics*, *2014*, 1–14. Hindawi Publishing Corp.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, *15*, 1–96.

Jiang, J., & Nguyen, T. (2015). *The fence methods*. Singapore: World Scientific.

Jiang, J., Nguyen, T. & Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, *106*, 732–745.

Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, *36*, 1669–1692.

Münnich, R., Burgard, J. P., & Vogt, M. (2009) Small area estimation for population counts in the German Census 2011. In *Section on survey research methods*. Washington, D.C.: JSM.

Rao, J. N. K. & Molina, I. (2015). Small Area Estimation, 2nd ed., Wiley.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, *6*, 15–51.

A New Approximation to the True Randomization-Based Design Effect

Siegfried Gabler, Matthias Ganninger and Partha Lahiri

Abstract It is generally difficult or even impossible to obtain a closed-form randomization-based true design effect formula for a nonlinear estimator under a complex sample design. A model that captures different salient features of the sample design is often used to approximate the randomization-based true design effect. Our simulation results show that the usual model-based design effect for the sample mean could significantly differ from the randomization-based true design effect for different replications of the finite population, even when different replicates of the finite population are generated by the same model used to derive the model-based design effect formula. We propose a new model-assisted design effect formula obtained from an appropriate model-based design effect formula when we replace the model intra-cluster correlation by an ANOVA "estimator" if observations for all units of the finite population were available. For one-stage cluster sampling with equal cluster size, we examine the accuracy of the new model-assisted design effect formula analytically and by a Monte carlo simulation. This new model-assisted design effect tracks the true randomization-based design effect much better than the corresponding model-based design effect formulae and the approximation to the true randomization-based design effect proposed by Kish (1965). The main advantage of the new model-assisted design effect is that it can be readily extended to more complex estimators and/or complex designs where the Kish's approximation is unavailable. Our proposed model-assisted design effect is generally much closer to

S. Gabler
GESIS, B2,1, 68159 Mannheim, Germany
e-mail: siegfried.gabler@gesis.org

M. Ganninger
F. Hoffmann-La Roche Ltd, Grenzacherstr. 124, 4070 Basel, Switzerland
e-mail: matthias.ganninger@roche.com

P. Lahiri (✉)
Joint Program in Survey Methodology and Department of Mathematics,
University of Maryland, 1218 LeFrak Hall, College Park, MD 20742, USA
e-mail: plahiri@umd.edu

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_10

the randomization-based design effect formula than the corresponding model-based design effect, even when model used to obtain the model-based design effect holds for the finite population.

Keywords Design effects · Model-based · Design-based · Intra-class correlation

1 Introduction

In large scale sample surveys, inferences are usually based on the standard randomization principle in survey sampling for estimating characteristics of large populations. Under such an approach, responses are treated as fixed and the randomness is assumed to solely come from the probability mechanism that generates the sample. For example, in simple random sampling (SRS), the sample mean $\bar{y}$ is unbiased under the sample design. The randomization-based variance of $\bar{y}$ is given by

$$\mathrm{Var}_{SRS}(\bar{y}) = (1-f)\frac{S^2}{n},$$

where $n,\ N,\ f = \frac{n}{N}$, and S^2 denote the sample size, population size, sampling fraction, and finite population element variance with divisor $N-1$, respectively. Usually f is negligible and can be dropped from the formula. The sample mean $\bar{y}$ remains an unbiased estimator of the population mean $\bar{Y}$ under the usual randomization approach if the sample design is *epsem*, i.e. each sampling unit of the finite population has the same chance f of being selected. However, $\mathrm{Var}_{SRS}(\bar{y})$ usually underestimates the true randomization variance of $\bar{y}$ under a cluster sample design, denoted by $\mathrm{Var}_C(\bar{y})$. To account for this underestimation, Kish (1965) proposed the following variance inflation factor, commonly known as the *design effect*:

$$\mathrm{Deff}_{\mathrm{R}} = \frac{\mathrm{Var}_C(\bar{y})}{\mathrm{Var}_{SRS}(\bar{y})}. \tag{1}$$

There are several potential uses of design effects. First, design effects are routinely used in determining sample size of a complex survey from the knowledge of sample size requirement for a SRS design. For this use, in the absence of any direct survey data on the response variables, historical and similar survey data are used in conjunction with the available information on different design features such as average cluster size, number of clusters, etc. (see European Social Survey 2011; Häder et al. 2003). Second possible use of design effects is in the variance computation from complex surveys (see Lohr 1999, p. 241) in situations where standard variance estimation techniques cannot be applied due to unavailability of appropriate software, especially in developing countries, or due to the time restriction to compute variance estimates for a large number of statistics by sophisticated variance estimation software or unavailability of actual cluster identifiers to protect the confidentiality of survey respondents. Design effects are also conveniently used in adjusting standard statistical procedures in order to make them appropriate for complex survey data (e.g. Rao and Scott 1981).

For a complex sample design, it is generally difficult or even impossible to obtain a closed-form formula for Deff_R. A model that captures different salient features of the sample design is generally used to approximate Deff_R. Since Deff_R is used in randomization-based inference, it is important to assess the accuracy of the usual model-based design effect approximations. In Sect. 2, we examine the accuracy of the usual model-based approximation analytically and by a Monte Carlo study for one-stage cluster sampling with equal cluster size. We argue that the standard model-based approximation to the true randomization-based design could be subject to substantial error, depending on the realization of the finite population from the assumed model. We then suggest a new model-assisted design effect formula, which works much better than the corresponding model-based approximation or the well-known Kish's approximation. In Sect. 3, we compare the new model-assisted design effect with the corresponding model-based design effect for the ratio mean in the context of one-stage cluster sampling with unequal cluster sizes. We notice that, for a fixed finite population, the accuracy of the model-based approximation depends on the variation of the cluster sizes and the choice of the model. In Sect. 4, we provide a formula for our model-assisted design effect for a general sample design. Finally, in Sect. 5, we make some concluding remarks. Throughout the paper, we use the subscripts R, M and MA to denote randomization-based, model-based and model-assisted design effects, respectively.

2 One-Stage Cluster Sampling: The Case of Equal Cluster Size

Let y_{ij} denote the value of a characteristic of interest for the jth unit in the ith cluster $(i = 1, \ldots, N;\ j = 1, \ldots, M)$. Define $Y_i = \sum_{j=1}^{M} y_{ij}$, the ith cluster total, $\bar{Y}_i = \frac{Y_i}{M}$, the ith cluster mean, $Y = \sum_{i=1}^{N} M\bar{Y}_i$, the finite population total, and $\bar{Y} = \frac{Y}{MN}$, the finite population mean.

Consider estimation of the finite population mean using one-stage cluster sampling, where n clusters, each of equal size M, are drawn by SRS. In this case, the sample mean $\bar{y} = \frac{\sum_{i\in s} Y_i}{nM}$ remains unbiased for the population mean $\bar{Y}$ under the sampling design. The randomization-based true design effect is given by

$$\begin{aligned}\text{Deff}_\text{R} &= \tfrac{NM-1}{M(N-1)}[1 + (M-1)\rho_\text{K}] \\ &= \text{Deff}_\text{K} + O(N^{-1}),\end{aligned} \tag{2}$$

where

$$\text{Deff}_\text{K} = 1 + (M-1)\rho_\text{K}, \tag{3}$$

design effect formula given in Kish (1965), and

$$\rho_{\mathrm{K}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M(M-1)}\sum_{j\neq j'}^{M}(y_{ij}-\bar{Y})(y_{ij'}-\bar{Y})}{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2}, \tag{4}$$

the finite population intra-cluster correlation (see Kish 1965, p. 171). Since $NM - 1 > M(N-1)$, $\mathrm{Deff}_{\mathrm{K}}$ underestimates $\mathrm{Deff}_{\mathrm{R}}$, but the order of underestimation is $O(N^{-1})$.

Define the following sum of squares in the ANOVA table for the finite population:

$$SSB = \sum_{i=1}^{N} M(\bar{Y}_i - \bar{Y})^2, \text{ between cluster sum of squares,} \tag{5}$$

$$SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2, \text{ within cluster sum of squares,} \tag{6}$$

$$SST = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y})^2, \text{ total sum of squares.} \tag{7}$$

It is easy to see (Lohr 1999, p. 139) that

$$\rho_{\mathrm{K}} = 1 - \frac{M}{M-1}\frac{SSW}{SST} = \frac{1}{M-1}\left(M\frac{SSB}{SST} - 1\right). \tag{8}$$

Let $df_b = N - 1$ and $df_w = N(M-1)$ be the degrees of freedom associated with SSB and SSW, respectively. Let $MSB = \frac{SSB}{df_b}$, the between mean square, and $MSW = \frac{SSW}{df_w}$ the within mean square.

It is generally difficult or even impossible to derive exact randomization-based design effect formulae for complex estimators and/or complex sample designs. In this section, we propose a new formula motivated from a model that is used to obtain the model-based design effect formula. Let M_C and M_{SRS} be two different models that capture salient features of the one-stage cluster sampling and SRS, respectively. For example, Gabler et al. (1999) used the following models:

- M_{SRS}: All the sample observations are uncorrelated with equal variance σ^2.
- M_C: All the sample observations have the same variance σ^2. However, only the observations from different clusters are uncorrelated; observations within the same cluster are equally correlated, common correlation being ρ.

The following alternative design effect definition was considered in Gabler et al. (1999):

$$\begin{aligned}\text{Deff}_{\text{M}} &= \frac{\text{Var}_{M_C}(\bar{y})}{\text{Var}_{M_{SRS}}(\bar{y})} \\ &= 1 + (M-1)\rho.\end{aligned} \tag{9}$$

where $\text{Var}_{M_C}(\bar{y})$ and $\text{Var}_{M_{SRS}}(\bar{y})$ are the variances of $\bar{y}$ under models M_C and M_{SRS}, respectively. Skinner (1989) provided an alternative definition of a model-based design effect as

$$\text{Deff}_{\text{Skinner}} = \frac{\text{Var}_{M_C}(\bar{y})}{E_{M_C}\left[\dfrac{SST}{nM(NM-1)}\right]}.$$

and showed that $\text{Deff}_{\text{Skinner}} \approx 1 + (M-1)\rho$.

The formulas of Deff_{K} and Deff_M, respectively, may lead one to erroneously conclude that ρ and ρ_{K} are identical. It is interesting to note that if model M_C of Skinner et al. (1989) holds for the finite population, we have

$$E_{M_C}[\rho_{\text{K}}] \approx \rho,$$

for large N.

Proof From (8) we have

$$\rho_{\text{K}} = \frac{M}{M-1}\frac{SSB}{SST} - \frac{1}{M-1}.$$

Using Taylor series expansion

$$E_{M_C}[\rho_{\text{K}}] \approx \frac{M}{M-1}\frac{E_{M_C}[SSB]}{E_{M_C}[SST]} - \frac{1}{M-1}. \tag{10}$$

Using Lemma of Ghosh and Lahiri (1987), we have

$$\begin{aligned}E_{M_C}[SSB] &= (N-1)[1+(M-1)\rho]\sigma^2, \\ E_{M_C}[SSW] &= N(M-1)(1-\rho)\sigma^2.\end{aligned}$$

Thus, the right hand side of (10) reduces to

$$\rho - \frac{[1+(M-1)\rho](1-\rho)}{M(N-1)+(M-1)(1-\rho)} \approx \rho.$$

This completes the proof. □

Remark: Suppose that y-values are normally distributed, under model M_C. We can find an interval (L, U) such that $P\,[\rho_K \in (L, U)] = 1 - \alpha$ as

$$P\left(\frac{L + \dfrac{1}{M-1}}{\dfrac{N-1}{N(M-1)}(1-L)} \leq F \leq \frac{U + \dfrac{1}{M-1}}{\dfrac{N-1}{N(M-1)}(1-U)}\right) = 1 - \alpha,$$

where $F = \dfrac{MSB}{MSW}$ has a F distribution with degrees of freedom $N - 1$, $N(M - 1)$. The above result can be used to construct a $100 \cdot (1 - \alpha)\%$ confidence interval for $\mathrm{Deff_R}$-$\mathrm{Deff_M}$.

We also note the following relationship:

$$\frac{E_{M_C}[\mathrm{Var}_C(\bar{y})]}{E_{M_{SRS}}[\mathrm{Var}_{SRS}(\bar{y})]} = \mathrm{Deff}_M.$$

Proof Noting that

$$\bar{y} = \frac{1}{nM}\sum_{i \in s} Y_i = \frac{1}{n}\sum_{i \in s} \bar{Y}_i,$$

we have

$$\mathrm{Var}_C(\bar{y}) = \frac{(1-f)}{n}\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2 = \frac{(1-f)}{n}\frac{SSB}{M(N-1)},$$

and

$$\mathrm{Var}_{SRS}(\bar{y}) = \frac{1-f}{nM}\frac{SST}{MN-1}.$$

The result now follows from Lemma of Ghosh and Lahiri (1987) and algebra. □

The standard ANOVA estimator of ρ is given by

$$\rho_{\mathrm{MA}} = \frac{\dfrac{SSB}{N-1} - \dfrac{SSW}{N(M-1)}}{\dfrac{SSB}{N-1} + \dfrac{SSW}{N}} = \frac{MSB - MSW}{MSB + (M-1)MSW} = \frac{F-1}{F+M-1}. \quad (11)$$

It is easy to check that ρ_{MA} is a consistent estimator of ρ, under model M_C, for large N. Moreover,

$$\rho_{\mathrm{K}} \leq \rho_{\mathrm{MA}}, \quad (12)$$

which implies

$$\begin{aligned}\mathrm{Deff}_{\mathrm{K}} = 1 + (M-1)\rho_{\mathrm{K}} &\leq \frac{NM-1}{M(N-1)}\,(1+(M-1)\rho_{\mathrm{K}}) = \mathrm{Deff}_{\mathrm{R}} \\ &\leq \frac{NM-1}{M(N-1)}\,(1+(M-1)\rho_{\mathrm{MA}}) = \mathrm{Deff}_{\mathrm{MA}},\end{aligned} \tag{13}$$

that is, $\mathrm{Deff}_{\mathrm{MA}}$ is a more conservative design effect formula than $\mathrm{Deff}_{\mathrm{K}}$ or $\mathrm{Deff}_{\mathrm{R}}$.

Proof For $SSB = 0$ or $SSW = 0$, but not both, we have $\rho_{\mathrm{K}} = \rho_{\mathrm{MA}}$. If $SSB > 0$ and $SSW > 0$, then

$$f(\alpha) = \frac{SSB - \frac{\alpha}{M-1}SSW}{SSB + \alpha SSW}. \tag{14}$$

Note that $f(\alpha)$ is a monotonically decreasing function in α since

$$f'(\alpha) = -\frac{M}{M-1}\frac{SSB \cdot SSW}{(SSB + \alpha SSW)^2} < 0, \tag{15}$$

for positive SSW and SSB. Since

$$f(1) = \rho_{\mathrm{K}} \tag{16}$$

and

$$f\left(\frac{N-1}{N}\right) = \rho_{\mathrm{MA}},$$

we have $\rho_{\mathrm{K}} < \rho_{\mathrm{MA}}$. □

2.1 Simulation

We compare ρ_{K}, ρ_{MA} and different design effects formulae using a Monte Carlo simulation study. For this, $N \times M$ realizations, y, of a random variable, Y are generated for selected combinations of N (number of clusters), M (cluster size) and intra-class correlation ρ using a common mean model (Valliant et al. 2000, p. 249). The data generating process is replicated 10 000 times to generate different statistics of interest for each combination of N, M and ρ.

The simulation results are displayed in Table 1. In the table, $\bar{\rho}_{\mathrm{K}}$ and $\bar{\rho}_{\mathrm{MA}}$ represent averages of ρ_{K} and ρ_{MA} over the 10 000 replications, respectively; P and P^* denote percentages of cases with $\rho_{\mathrm{K}} < \rho$ and $\rho_{\mathrm{MA}} < \rho$, respectively. It is clear that the mean of ρ_{MA} always is greater than that of ρ_{K}, which is in line with inequality (12), and is closer to the true value of ρ than ρ_{K} is. This is especially true when N is small and M is large.

Table 1 Results from the simulation study–equal cluster size

Case	ρ	N	M	P	P^*	$\bar{\rho}_{\text{K}}$	$\bar{\rho}_{\text{MA}}$	$\overline{\text{Deff}}_{\text{R}}$	$\overline{\text{Deff}}_{\text{K}}$	$\overline{\text{Deff}}_{\text{MA}}$	Deff_{M}
1	0.20	100	10	0.54	0.51	0.20	0.20	2.80	2.77	2.82	2.80
2	0.20	10	100	0.65	0.56	0.18	0.19	20.64	18.59	22.40	20.80
3	0.10	100	10	0.54	0.51	0.10	0.10	1.90	1.88	1.91	1.90
4	0.10	10	100	0.65	0.56	0.09	0.10	10.86	9.78	11.91	10.90
5	0.04	100	10	0.55	0.52	0.04	0.04	1.36	1.35	1.37	1.36
6	0.04	10	100	0.65	0.56	0.03	0.04	4.92	4.43	5.43	4.96

This effect is also seen in the percentage of cases with $\rho_{\text{K}} < \rho$, denoted by P, and $\rho_{\text{MA}} < \rho$, denoted by P^*, respectively. The boxplots (a) to (d) of Fig. 1 illustrate this behavior graphically. In Fig. 1, we see that both, ρ_{K} and ρ_{MA}, vary considerably around the model parameter ρ implying that the model-based design effect could be very different from the randomization-based design effect. Both, ρ_{K} and ρ_{MA}, however, underestimate ρ on the average.

Tables 2, 3 and 4 display the values of Deff_{K}, Deff_{MA} and Deff_{M} for selected quantiles of the 10 000 realizations of Deff_{R}. One can see that Deff_{MA} is very close to Deff_{R}. In contrast, Deff_{M} could substantially deviate from Deff_{R}. The boxplots given in Fig. 2 highlight the above graphically.

It is obvious that the differences Deff_{R}-Deff_{K} and Deff_{R}-Deff_{MA} are much smaller than Deff_{R}-Deff_{M}. Due to scaling, differences between Deff_{R}-Deff_{K} and Deff_{R}-Deff_{MA} cannot be seen in this plot. For this reason, we omit Deff_{R}-Deff_{M} in boxplots given in Fig. 3. Again, we can see that Deff_{MA} is much closer to Deff_{R} than Deff_{K}.

In accordance with Eq. (13), we see from the above plots that the difference Deff_{R}-Deff_{MA} is always negative whereas the difference Deff_{R}-Deff_{K} is always positive.

Let us now consider a contamination case. In this case, the following model holds:

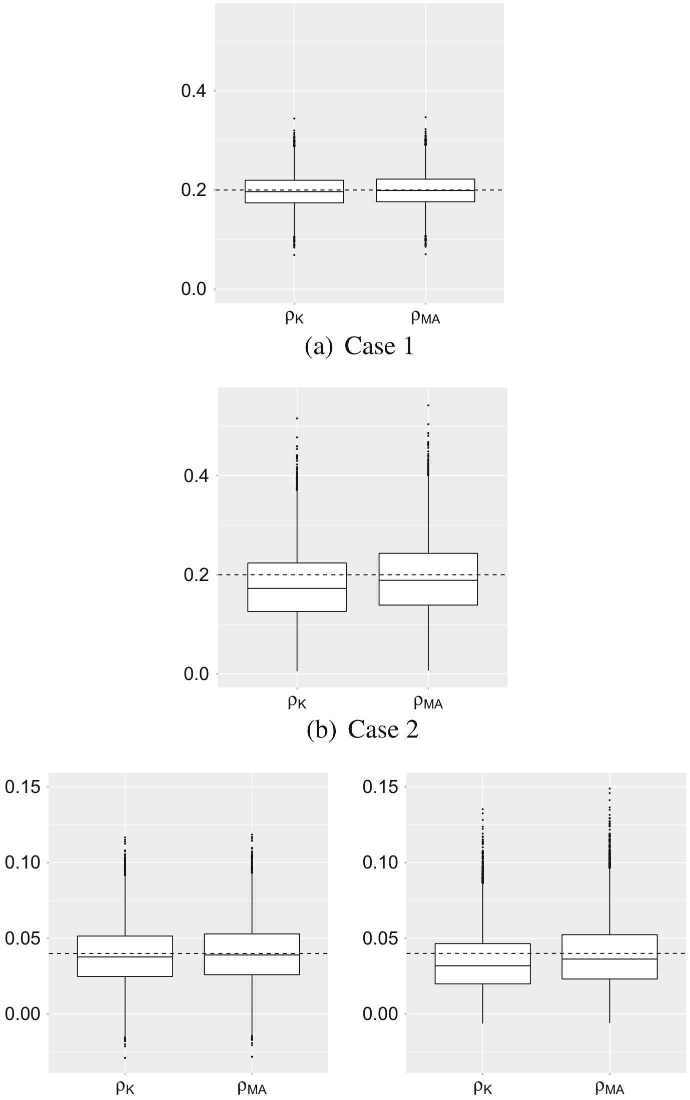
- M_{C^*}: All the observations are normally distributed and have the same variance σ^2. The observations from different clusters are uncorrelated; observations within the same cluster are equally correlated, common correlation being ρ, $\rho \sim \text{U}(\varrho, \tau^2)$.

Note that we contaminate the M_C model by drawing ρ in each cluster from a uniform distribution.

To present a concrete example, a population with $N = 8$; $M = 10$ is given, where the ρ values in the eight clusters are realized as 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8. In the following, we assume, however, the model M_C with parameter $\rho = \bar{\rho} = 0.45$. The sample is of size $nM = 40$ with $n = 4$ clusters being drawn at the first stage. Table 5 provides an overview of the results.

In order to obtain the expected value and variance of Deff_{R}, we use Taylor series expansion and get

$$E[\text{Deff}_{\text{R}}] \approx \frac{NM-1}{N-1}\frac{E_\rho E_{M_{C^*}}[SSB|\rho]}{E_\rho E_{M_{C^*}}[SST|\rho]}. \tag{17}$$



(a) Case 1

(b) Case 2

(c) Case 5

(d) Case 6

Fig. 1 Boxplots of the simulated distribution of ρ_K and ρ_MA for selected cases (the dashed line indicates the respective model parameter ρ)

Table 2 Values of Deff_K, Deff_MA and Deff_M for selected quantiles of 10 000 realizations for Deff_R—Cases 1 and 2: $\rho = 0.20$

	Case 1 ($N = 10$; $M = 100$)				Case 2 ($N = 100$; $M = 10$)			
	Deff_R	Deff_K	Deff_MA	Deff_M	Deff_R	Deff_K	Deff_MA	Deff_M
Min.	1.6341	1.6194	1.6479	2.8000	1.6728	1.5070	1.8556	20.8000
Q25	2.5903	2.5670	2.6097	2.8000	14.9458	13.4647	16.3617	20.8000
Median	2.7949	2.7697	2.8152	2.8000	20.0671	18.0785	21.8578	20.8000
Q75	3.0037	2.9766	3.0249	2.8000	25.7044	23.1571	27.8440	20.8000
Max.	4.1352	4.0980	4.1598	2.8000	57.7633	52.0390	60.6732	20.8000

Table 3 Values of Deff_K, Deff_MA and Deff_M for selected quantiles of 10 000 realizations for Deff_R—Cases 3 and 4: $\rho = 0.10$

	Case 3 ($N = 10$; $M = 100$)				Case 4 ($N = 100$; $M = 10$)			
	Deff_R	Deff_K	Deff_MA	Deff_M	Deff_R	Deff_K	Deff_MA	Deff_M
Min.	1.0845	1.0747	1.0943	1.9000	1.0269	0.9251	1.1398	10.9000
Q25	1.7390	1.7233	1.7535	1.9000	7.4565	6.7176	8.2236	10.9000
Median	1.8927	1.8756	1.9082	1.9000	10.3491	9.3235	11.3811	10.9000
Q75	2.0455	2.0271	2.0620	1.9000	13.6309	12.2801	14.9416	10.9000
Max.	2.8414	2.8158	2.8620	1.9000	32.7768	29.5286	35.2617	10.9000

Table 4 Values of Deff_K, Deff_MA and Deff_M for selected quantiles of 10 000 realizations for Deff_R—Cases 5 and 6: $\rho = 0.04$

	Case 5 ($N = 10$; $M = 100$)				Case 6 ($N = 100$; $M = 10$)			
	Deff_R	Deff_K	Deff_MA	Deff_M	Deff_R	Deff_K	Deff_MA	Deff_M
Min.	0.7454	0.7387	0.7524	1.3600	0.4060	0.3657	0.4509	4.9600
Q25	1.2340	1.2228	1.2449	1.3600	3.2956	2.9690	3.6498	4.9600
Median	1.3518	1.3397	1.3637	1.3600	4.6072	4.1506	5.0956	4.9600
Q75	1.4769	1.4636	1.4896	1.3600	6.2124	5.5967	6.8600	4.9600
Max.	2.0683	2.0497	2.0849	1.3600	15.9701	14.3875	17.4653	4.9600

After some algebraic transformations, we obtain

$$\begin{aligned} E[SSB] &= E_\rho E_{M_{C^*}}[SSB|\rho] = (N-1)\,(1+(M-1)\varrho)\,\sigma^2, \\ E[SSW] &= E_\rho E_{M_{C^*}}[SSW|\rho] = N(M-1)\,(1-\varrho)\,\sigma^2, \\ E[SST] &= E_\rho E_{M_{C^*}}[SST|\rho] = (NM-1-(M-1)\varrho)\,\sigma^2, \end{aligned}$$

and thus

$$E[\text{Deff}_\text{R}] \approx \frac{1+(M-1)\varrho}{1-\dfrac{M-1}{NM-1}\varrho} \approx 1+(M-1)\varrho. \tag{18}$$

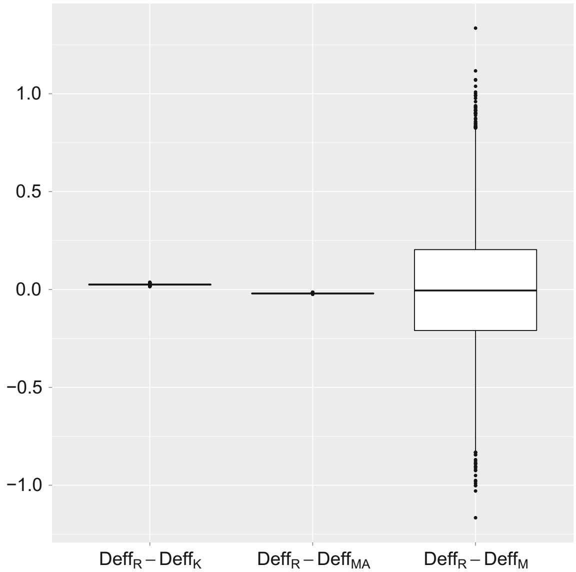


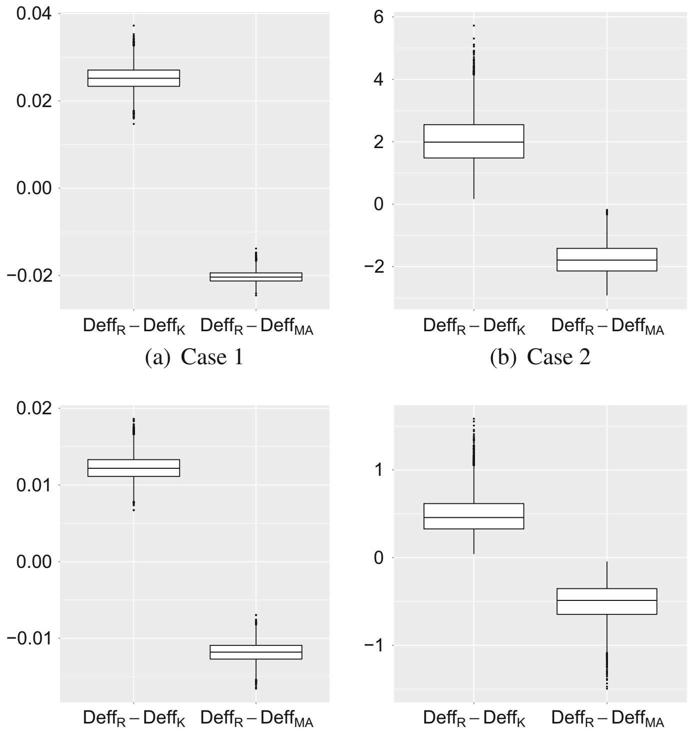
Fig. 2 Boxplots of the differences Deff_R-Deff_K, Deff_R-Deff_MA and Deff_R-Deff_M for Case 1

Again, using Taylor series expansion, the variance of Deff_R can be computed by

$$\text{Var}[\text{Deff}_\text{R}] \approx \left(\frac{NM-1}{N-1}\right)^2 \left[\frac{E_\rho E_{M_{C^*}}[SSB^2|\rho]}{E_\rho E_{M_{C^*}}[SST^2|\rho]} - \left(\frac{E_\rho E_{M_{C^*}}[SSB2|\rho]}{E_\rho E_{M_{C^*}}[SST|\rho]}\right)^2\right] \quad . \tag{19}$$

Define $x = 1 + (M-1)\varrho$. Assuming normality under model M_{C*} and after some algebra, we get the following results

$$\begin{aligned}
E[SSB] &= (N-1)x\sigma^2, \\
E[SST] &= (NM-x)\sigma^2, \\
E[SSB^2] &= \left(\frac{3(M-1)^2(N-1)^2}{N}\tau^2 + (N^2-1)x^2\right)\sigma^4, \\
E[SST^2] &= \left[\frac{N^2M^2(2+(M-1)N)}{(M-1)N} + \right. \\
&\quad \frac{(M-1)^2\left(2N^2(2M-1) - 3(M-1)(2N-1)\right)}{(M-1)N}\tau^2 - \\
&\quad \left. \frac{2MN^2(M+1)}{(M-1)N}x + \frac{N(2MN-M+1)}{(M-1)N}x^2\right]\sigma^4,
\end{aligned} \tag{20}$$



(a) Case 1

(b) Case 2

(c) Case 5

(d) Case 6

Fig. 3 Boxplots of the differences Deff_R-Deff_K and Deff_R-Deff_{MA} for selected cases

Table 5 Values of ρ_{MA} and Deff for the sample mean

Case	ρ	Deff_R	Deff_M	Deff_K	Deff_{MA}
1	0.45	3.4151	5.0500	3.0260	3.7412

which lead to

$$\text{Var}[\text{Deff}_R] \approx \left(\frac{NM-1}{N-1}\right)^2 \left[\frac{E[SSB^2]}{E[SST^2]} - \left(\frac{E[SSB]}{E[SST]}\right)^2\right]. \tag{21}$$

As the number of clusters N increases, we get $\text{Var}[\text{Deff}_R] \to 0$.

3 One-Stage Cluster Sampling: The Case of Unequal Cluster Sizes

Define Y_i, $\bar{Y}_i$, Y, and $\bar{Y}$ as in Sect. 2 with M replaced by M_i, denoting the cluster size in the ith cluster. As before, first consider simple random sampling of n clusters. Consider the sample mean $\bar{y} = \frac{\sum_{i \in s} Y_i}{\sum_{i \in s} M_i}$ as an estimator of $\bar{Y}$. Note that $\bar{y}$ is now a ratio estimator and no longer unbiased under the randomization approach. The derivation of an exact formula of Deff_R is difficult due to the nonlinear nature of the ratio estimator.

Although, in the randomization approach, $\bar{y}$ is a ratio of two random variables, in the model-based framework it is a simple linear estimator. Thus, the calculation of Deff_M is straightforward and is given by

$$\text{Deff}_\text{M} = 1 + (\bar{M}_{ws} - 1)\rho,$$

where

$$\bar{M}_{ws} = \frac{\sum_{i \in s} M_i^2}{\sum_{i \in s} M_i}$$

is a weighted average of the cluster sizes in the sample with weights proportional to the cluster sizes.

We get the following model-assisted design effect formula if we replace ρ in the model-based design effect formula by the following ANOVA "estimator" of ρ if all observations in the finite population were available:

$$\text{Deff}_\text{MA} = \frac{N(M_0 - 1)}{(N-1)M_0}\left(1 + \left[\bar{M}_{ws} - 1\right]\rho_\text{MA}\right),$$

where $M_0 = \sum_{i=1}^N M_i$ and ρ_MA is given by

$$\rho_\text{MA} = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + \left(\frac{M_0 - \bar{M}_w}{N-1} - 1\right)\text{MSW}}, \tag{22}$$

with

$$\text{MSB} = \frac{\sum_i^N M_i \left(\bar{Y}_i - \bar{y}\right)^2}{N-1}, \quad \text{MSW} = \frac{\sum_i^N \sum_{i=1}^{M_i} \left(y_{ij} - \bar{Y}_i\right)^2}{M_0 - N}, \quad \bar{M}_w = \frac{\sum_i^N M_i^2}{\sum_i^N M_i}.$$

Let us now consider a *non-epsem* one-stage cluster sampling. Let $w_i = \pi_i^{-1}$ be the sampling weight for all units in the ith cluster, where π_i denotes the inclusion probability if the ith cluster $(i \in s)$. Thus, we consider the following ratio estimator to estimate the population mean:

$$\hat{\bar{Y}} = \frac{\sum\limits_{i \in s} \sum\limits_{j=1}^{M_i} w_i y_{ij}}{\sum\limits_{i \in s} \sum\limits_{j=1}^{M_i} w_i} = \frac{\sum\limits_{i \in s} w_i \sum\limits_{j=1}^{M_i} y_{ij}}{\sum\limits_{i \in s} w_i M_i}.$$

In this case, an appropriate model-based design effect formula under the models M_C and M_{srs} is given by:

$$\mathrm{Deff_M} = \mathrm{Deff_P} \times \mathrm{Deff_C} \tag{23}$$

where

$$\mathrm{Deff_P} = \frac{\sum\limits_{i \in s} M_i \sum\limits_{i \in s} w_i^2 M_i}{\left(\sum\limits_{i \in s} w_i M_i\right)^2}, \quad \mathrm{Deff_C} = 1 + (\bar{M}_s - 1)\rho, \quad \bar{M}_s = \frac{\sum\limits_{i \in s} w_i^2 M_i^2}{\sum\limits_{i \in s} w_i^2 M_i}.$$

Proof Let $\bar{y}$ be the unweighted sample mean of size $\sum_{i \in s} M_i$. Then we have

$$\begin{aligned}
\mathrm{Deff_M} &= \frac{\mathrm{Var}_{M_C}(\hat{\bar{Y}})}{\mathrm{Var}_{M_{SRS}}(\bar{y})} \\
&= \frac{\sum\limits_{i \in s} M_i \sum\limits_{i \in s} w_i^2 M_i \left(1 + (M_i - 1)\rho\right)}{\left(\sum\limits_{i \in s} w_i M_i\right)^2} \\
&= \frac{\sum\limits_{i \in s} M_i \sum\limits_{i \in s} w_i^2 M_i}{\left(\sum\limits_{i \in s} w_i M_i\right)^2} \cdot \left[1 + \left(\frac{\sum\limits_{i \in s} w_i^2 M_i^2}{\sum\limits_{i \in s} w_i^2 M_i} - 1\right)\rho\right].
\end{aligned}$$

□

This is similar to the formula proposed by Kish (1965). Note that for the *epsem* case $\mathrm{Deff}_P = 1$. Gabler et al. (1999) justified the Kish's formula as a model-based design effect. They interpreted ρ as the model intra-cluster correlation under a model proposed by Skinner et al. (1989) and provided alternative formulas for the average cluster size $\bar{M}$. See Lynn and Gabler (2005) for a comparison of different average cluster size formulas. They concluded that the choice of a particular formula does make a difference.

We obtain a model-assisted design effect formula from (23) when we replace ρ by ρ_{MA} defined in (22).

4 The General Case

Let s_i denote the ith ultimate cluster (Kalton 1983, p. 35) in the sample s. Further, let y_{ik} and w_{ik} denote the value of the characteristic of interest and the associated survey weight for the kth unit in the ith ultimate cluster of size m_i ($i = 1, \ldots, n$; $k \in s_i$). In this case the population mean $\bar{Y}$ is estimated by

$$\hat{\bar{Y}} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{k \in s_i} w_{ik} y_{ik}}{\sum\limits_{i=1}^{n} \sum\limits_{k \in s_i} w_{ik}}.$$

We assume model M_C on the ultimate cluster. Then, using algebra similar to Sect. 3, the model-based design effect is obtained as

$$\mathrm{Deff_M} = \mathrm{Deff_P} \times \mathrm{Deff_C}$$

where

$$\mathrm{Deff_P} = \frac{\left(\sum\limits_{i=1}^{n} m_i\right)\left(\sum\limits_{i=1}^{n} \sum\limits_{k \in s_i} w_{ik}^2\right)}{\left(\sum\limits_{i=1}^{n} \sum\limits_{k \in s_i} w_{ik}\right)^2}, \quad \mathrm{Deff_C} = 1 + \left(\frac{\sum\limits_{i=1}^{n} m_i \sum\limits_{k \in s_i} w_{ik}^2}{\sum\limits_{i=1}^{n} \sum\limits_{k \in s_i} w_{ik}^2} - 1\right)\rho.$$

To obtain a model-assisted design effect, we simply replace ρ by (22), where the cluster needs to be interpreted as ultimate cluster. Furthermore, M_i must be replaced by m_i in the formula for $\bar{M}_w$ and N is the number of ultimate clusters in the population.

5 Discussion

It is difficult or even impossible to derive a simple closed-form formula for randomization-based design effect due to the complexity in the sample design or nonlinearity of the estimator. In this paper, we have demonstrated that standard model-based approximations to the randomization-based design effect may not perform well in certain situations. To circumvent the problem, we have proposed a new model-assisted approach that provides a better approximation to the randomization-based design effect than the corresponding model-based design effect formula. Our approach is quite flexible to handle complex designs and nonlinear estimators. Our focus has been on the evaluation of the new model-assisted design effect formula rather than its estimation from a sample survey data. We will address the estimation of design effect in a future paper.

Acknowledgements The authors thank an anonymous referee for a few constructive suggestions that led to improvement of an earlier version of the article. The research of the third author was supported in part by the National Science Foundation Grant Number SES-1534413.

References

European Social Survey. (2011). *Round 6 specification for participating countries*. London: Centre for Comparative Social Surveys, City University London.

Gabler, S., Häder, S., & Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, *25*(1), 105–106.

Ghosh, M., & Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, *82*(400), 1153–1162.

Häder, S., Gabler, S., Laaksonen, S. & Lynn, P. (2003). *ESS 2002/2003: Technical Report*, ESS, chapter The Sample. http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=199&itemid=80

Kalton, G. (1983). *Introduction to survey sampling*. Thousand Oaks.

Kish, L. (1965). *Survey sampling*. Wiley.

Lohr, S. (1999). *Sampling: Design and analysis*. Duxbury Press.

Lynn, P., & Gabler, S. (2005). Approximations to b* in the prediction of design effects due to clustering. *Survey Methodology*, *31*(2).

Rao, J., & Scott, A. (1981). The analysis of categorical data from complex surveys: chi-square tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, *76*, 221–230.

Skinner, C. (1989). *Analysis of complex surveys* (pp. 23–58). Wiley (chapter 2).

Skinner, C., Holt, D. & Smith, T. (1989). *Analysis of complex surveys*. Wiley.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite population sampling and inference*. Wiley.

Confounded Factorial Design with Partial Balance and Orthogonal Sub-Factorial Structure

Madhura Mandal and Premadhis Das

Abstract In this paper, the contrasts belonging to any effect are divided into a number of subsets and factorial designs are proposed such that these subsets (called sub-effects) are orthogonally estimated with balance. Such designs have been called 'partially balanced design with orthogonal sub-factorial structure'. These designs are important in the sense that these allow more flexibility in the choice of the designs retaining desirable properties such as orthogonality and partial balance and also provide more insight into the nature of the contrasts belonging to any factorial effect.

Keywords Factorial design · Partial balance · Orthogonal sub-factorial structure
Generalised extended group divisible design

1 Introduction

Let $F_1, F_2, \ldots, F_m$ be m factors with $s_1, s_2, \ldots, s_m$ levels respectively. The $v = \prod_{i=1}^{m} s_i$ level combinations $(j_1, j_2, \ldots, j_m)$, $0 \le j_i \le s_i - 1,\ 1 \le i \le m$ are considered as v treatments and the treatment effect at this level combination is given by $t(\mathbf{j}) = t(j_1, j_2, \ldots, j_m), 0 \le j_i \le s_i - 1, 1 \le i \le m$.

The vector $\mathbf{t}^{v\times 1}$ defined as

$$\mathbf{t} = [\ldots t(j_1, j_2, \ldots, j_m) \ldots]' \tag{1}$$

where $(j_1, j_2, \ldots, j_m)$ are arranged lexicographically, is known as the vector of treatment effects.

M. Mandal
Bethune College, Kolkata, India
e-mail: mandal.madhura@gmail.com

P. Das (✉)
University of Kalyani, Kalyani, India
e-mail: dasp50@yahoo.co.in

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_11

A treatment contrast is defined as

$$\pi = \sum_{j_1}\sum_{j_2}\cdots\sum_{j_m} l(j_1 j_2 \ldots j_m)t(j_1, j_2 \ldots, j_m) \tag{2}$$

where $l(j_1, j_2, \ldots, j_m)$'s are real numbers, not all zero, such that

$$\sum_{j_1}\sum_{j_2}\cdots\sum_{j_m} l(j_1 j_2 \ldots j_m) = 0. \tag{3}$$

A treatment contrast is said to belong to the g-factor interaction $F_{i_1}F_{i_2}\ldots F_{i_g}$, $1 \le i_1 < \cdots < i_g \le m$, $1 \le g \le m$ if

(i) $l(j_1, j_2, \ldots, j_m)$ depends only on $j_{i_1}, j_{i_2}, \ldots, j_{i_g}$ and
(ii) the sum of $l(j_1, j_2, \ldots, j_m)$ on any of the arguments $j_{i_1}, j_{i_2}, \ldots, j_{i_g}$ is zero.

If $g = 1$, the 1-factor interactions are called main effects. In all there are $c(m, 1) + c(m, 2) + \cdots + c(m, m) = 2^m - 1$ interactions. The interactions are also called factorial effects. It follows that there are $\prod_{j=1}^{g}(s_{i_j} - 1)$ independent contrasts belonging to the g-factor interaction $F_{i_1}F_{i_2}\ldots F_{i_g}$. The number of independent contrasts belonging to an interaction is called the degrees of freedom (d.f.) carried by the interaction. There is an one-to-one correspondence between the $(2^m - 1)$ factorial effects and the $(2^m - 1)$ non-null binary vectors

$$\Omega = (10\ldots0, 01\ldots0, \ldots, 11\ldots1). \tag{4}$$

Any factorial effect can be denoted by

$$F^{\mathbf{x}} = F_1^{x_1}F_2^{x_2}\ldots F_m^{x_m} \tag{5}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and $\mathbf{x} \in \Omega$.

The d.f. carried by the factorial effect $F^{\mathbf{x}}$ is given by

$$\alpha(\mathbf{x}) = \prod_{i=1}^{m}(s_i - 1)^{x_i}. \tag{6}$$

Let $\pi_1 = \mathbf{l}_1'\mathbf{t}$ and $\pi_2 = \mathbf{l}_2'\mathbf{t}$ be two contrasts belonging to the factorial effect $F^{\mathbf{x}}$, $\mathbf{x} \in \Omega$. Then these contrasts are said to orthonormal contrasts if $\mathbf{l}_1'\mathbf{l}_1 = \mathbf{l}_2'\mathbf{l}_2 = 1, \mathbf{l}_1'\mathbf{l}_2 = 0$.

We shall consider here only orthonormal contrasts belonging to any factorial effect.

Let us define the following matrices

$$\left.\begin{aligned}\mathbf{P}_i^{x_i} &= \mathbf{P}_i \quad \text{if } x_i = 1\\ &= \frac{\mathbf{1}_i'}{\sqrt{s_i}} \quad \text{if } x_i = 0\end{aligned}\right\};\quad 1 \le i \le n \tag{7}$$

such that

$$\overline{\mathbf{O}} = \begin{pmatrix} \frac{\mathbf{1_i}'}{\sqrt{s_i}} \\ \mathbf{P}_i \end{pmatrix} \tag{8}$$

is an $s_i \times s_i$ orthogonal matrix, $\mathbf{1}_i' = s_i \times 1$ vector containing 1's only. Note that the set of $\prod_{i=1}^{m}(s_i - 1)^{x_i} = \alpha(\mathbf{x})$ orthonormal contrasts belonging to the factorial effect $F^{\mathbf{x}}$ is given by $(\mathbf{P}_1^{x_1} \otimes \mathbf{P}_2^{x_2} \otimes \cdots \otimes \mathbf{P}_m^{x_m})\mathbf{t} = \mathbf{P}^{\mathbf{x}}\mathbf{t}$ where $\otimes$ denotes the Kronecker Product of matrices. The set is denoted by the interaction itself i.e.

$$F^{\mathbf{x}} = \mathbf{P}^{\mathbf{x}}\mathbf{t},\ \ \mathbf{x} \in \Omega. \tag{9}$$

Let d be a design with b blocks of sizes $k_1, k_2, \ldots, k_b$. The $v = \prod_{i=1}^{m} s_i$ level combinations are randomly allocated to the plots of the design d such that the ith level combination is replicated r_i times, $1 \le i \le v$. We assume that the design d is connected (we shall work here with connected designs only) so that all the $(v - 1)$ contrasts are estimable. The best linear unbiased estimator (BLUE) of $F^{\mathbf{x}}$ is given by $P^{\mathbf{x}}\hat{\mathbf{t}}$, where $\hat{\mathbf{t}}$ is any solution of the reduced normal equation

$$\mathbf{Ct} = \mathbf{Q} \tag{10}$$

where $\mathbf{C} = \mathbf{R} - \mathbf{NK}^{-\delta}\mathbf{N}'$, $\mathbf{R} = \text{Diag}(r_1, r_2, \ldots, r_v)$, $\mathbf{K} = \text{Diag}(k_1, k_2, \ldots, k_b)$, $\mathbf{N} = (n_{ih})$ = incidence matrix of d and $\mathbf{Q}$ is the vector of adjusted treatment totals.

Definition 1 The design d is said to be balanced with OFS (Orthogonal Factorial Structure) if

$$\text{Cov}(\hat{F}^{\mathbf{x}}, \hat{F}^{\mathbf{y}}) = \mathbf{0}\ \forall\ \mathbf{x} \neq \mathbf{y} \in \Omega \tag{11}$$

$$\text{Disp}(\hat{F}^{\mathbf{x}}) = \sigma^2\rho(\mathbf{x})\mathbf{I}_{\alpha(\mathbf{x})}\ \mathbf{x} \in \Omega \tag{12}$$

where $\hat{F}^{\mathbf{x}} = \mathbf{P}^{\mathbf{x}}\hat{\mathbf{t}}$, $I_{\alpha(\mathbf{x})}$ = identity matrix of order $\alpha(\mathbf{x})$, $\rho(\mathbf{x})$ is a positive constant depending on $\mathbf{x}$ and σ^2 is the error variance.

Characterizations of balanced designs with OFS under different situations are considered in Nair and Rao (1948), Shah (1958, 1960), Kurkjian and Zelen (1963), Kshirsagar (1966) and Gupta (1983). A comprehensive discussion can be found in Gupta and Mukerjee (1989).

It is to be noted that, for balance, the BLUEs of all the $\alpha(\mathbf{x})$ orthonormal treatment contrasts belonging to $F^{\mathbf{x}}$ are required to have the same variance $\sigma^2\rho(\mathbf{x}),\ \mathbf{x} \in \Omega$. The requirement is restrictive and we can bring in some flexibility, if we can divide the $\alpha(\mathbf{x})$ contrasts belonging to $F^{\mathbf{x}},\ \mathbf{x} \in \Omega$, into a number of subsets and can propose designs which allow orthogonal estimation of these sets of contrasts with the same variance maintaining orthogonality of the BLUEs of these subsets.

Definition 2 (*Partially Balanced Design (PBD) with Orthogonal Sub-Factorial Structure(OSFS)*) Let the $\alpha(\mathbf{x})$ orthonormal contrasts belonging to $F^{\mathbf{x}}$ be divided into $p(\mathbf{x})$ sub-sets $F_1^{\mathbf{x}}, F_2^{\mathbf{x}}, \ldots, F_{p(\mathbf{x})}^{\mathbf{x}}, \mathbf{x} \in \Omega$ such that

$$\left.\text{Cov}(\hat{F}_i^{\mathbf{x}}, \hat{F}_j^{\mathbf{y}}) = \mathbf{0},\ \text{if } \mathbf{x} \neq \mathbf{y},\ \text{or } i \neq j\ \text{ if } \mathbf{x} = \mathbf{y};\quad \text{Disp}(\hat{F}_i^{\mathbf{x}}) = \sigma^2\rho_i(\mathbf{x})\mathbf{I}_{\alpha_i(\mathbf{x})}\right\} \tag{13}$$

where $\hat{F}_i^{\mathbf{x}}$ is the BLUE of $F_i^{\mathbf{x}}$, $\rho_i(\mathbf{x})$ is a positive real number and $\alpha_i(\mathbf{x})$ is the number of orthonormal contrasts belonging to $F_i(\mathbf{x})$, $i = 1, 2, \ldots, p(\mathbf{x}), \mathbf{x} \in \Omega$. Then the factorial design is called Partially Balanced Design (PBD) with Orthogonal Sub-Factorial Structure (OSFS).

$[F_i^{\mathbf{x}}]$ may be called sub-effects of $F^{\mathbf{x}}$. Partially Balanced Design (PBD) with Orthogonal Sub-Factorial Structure (OSFS) is considered in Das and Chatterjee (1999) where the orthonormal contrasts belonging to $F^{\mathbf{x}}$ are divided into subsets by grouping the levels of each factor into a number of subsets and considering the between group and within group contrasts. Also in Das (2003) another kind of PBD with OSFS is considered when s_i's are all equal to s, which is a prime or prime power. Here the pencilwise division (cf Bose 1947) of the contrasts belonging to $F^{\mathbf{x}},\ x \in \Omega$ is considered. In this chapter, we shall consider the orthonormal contrast vectors in the rows of P_i in (7) as the orthonormal vectors obtained from orthogonal polynomials (cf. Fisher and Yates Table 1943) and accordingly introduce a division among the contrasts belonging to $F^{\mathbf{x}},\ \mathbf{x} \in \Omega$, when each s_i is an even integer. With such division of the factorial effects into subfactorial effects, we shall characterize partially balanced designs where these sub-factorial effects are orthogonally estimable.

2 Division of the Factorial Effects When Each s_i is an Even Number, $1 \leq i \leq m$

Let s be an even positive integer, such that $s = 2p$, p being any positive integer. Also let, $\mathbf{D}_{s\times s}$ be the orthogonal matrix obtained from the s orthogonal polynomials defined over s arguments $0, 1, 2, \ldots, s-1$ (cf. Fisher and Yates Table 1943). We write D as

$$\mathbf{D} = (\boldsymbol{\xi}(0), \boldsymbol{\xi}(1), \ldots, \boldsymbol{\xi}(s-1)) \tag{14}$$

where

$$\boldsymbol{\xi}'(\alpha) = (d_0(\alpha), d_1(\alpha), \ldots, d_{s-1}(\alpha)), \tag{15}$$

$d_j(\alpha)$ is the normalized value of the αth degree orthogonal polynomial defined over the argument j, $j = 0, 1, 2, \ldots, (s-1)$. It follows from the properties of the elements of **D** that

$$d_j(0) = 1,\ d_j(\alpha) = (-1)^{\alpha} d_{s-1-j}(\alpha),\ \text{and} \sum_{j=0}^{s-1} d_j(\alpha) d_j(\alpha') = 0 \tag{16}$$

for $0 \leq j \leq (s-1), 0 \leq \alpha \neq \alpha' \leq (s-1)$. From (16) it follows that

$$\left.\begin{array}{ll} d_j(\alpha) = -d_{s-1-j}(\alpha) & \text{if } \alpha = \text{odd}, \\ d_j(\alpha) = d_{s-1-j}(\alpha) & \text{if } \alpha = \text{even}. \end{array}\right\} \tag{17}$$

Also from (16) and (17) it follows that

$$\left.\begin{array}{l} \sum_{j=0}^{p-1} d_j(\alpha) d_j(\alpha') = 0 \ \text{ if } \alpha = \text{odd},\ \alpha' = \text{even or } \alpha = \text{even},\ \alpha' = \text{odd} \\ \qquad\qquad\qquad = \frac{1}{2} \ \text{ otherwise.} \end{array}\right\} \tag{18}$$

Define a $p \times 1$ vector $\boldsymbol{\xi}^*(\alpha)$ from $\boldsymbol{\xi}(\alpha)$ as

$$\boldsymbol{\xi}^{*'}(\alpha) = (d_0(\alpha), d_1(\alpha), \ldots, d_{p-1}(\alpha)) \tag{19}$$

and also a $p \times (p-1)$ matrix **A** as

$$\mathbf{A} = (\boldsymbol{\xi}^*(2), \boldsymbol{\xi}^*(4), \ldots, \boldsymbol{\xi}^*(2p-2)) = \begin{pmatrix} d_0(2) & d_0(4) & \ldots & d_0(2p-2) \\ d_1(2) & d_1(4) & \ldots & d_1(2p-2) \\ \vdots & \vdots & \vdots & \vdots \\ d_{p-1}(2) & d_{p-1}(4) & \ldots & d_{p-1}(2p-2) \end{pmatrix}. \tag{20}$$

From (19) and (20) it easily follows that $\mathbf{A}^* = \frac{1}{\sqrt{2}} \left(\frac{1_p}{\sqrt{p}}, \sqrt{2}\mathbf{A} \right)$ is a $p \times p$ orthogonal matrix where $\mathbf{1}_p$ is a $p \times 1$ vector with all elements unity. Similarly, considering the values corresponding to the odd order polynomials over the first $\frac{s}{2} = p$ arguments, we define a $p \times p$ matrix **B** as

$$B = \begin{pmatrix} d_0(1) & d_0(3) & \ldots & d_0(2p-1) \\ d_0(1) & d_1(3) & \ldots & d_1(2p-1) \\ \vdots & \vdots & \vdots & \vdots \\ d_{p-1}(1) & d_{p-1}(3) & \ldots & d_{p-1}(2p-1) \end{pmatrix}. \tag{21}$$

Again from (18) it follows that $\sqrt{2}\mathbf{B}$ is a $p \times p$ orthogonal matrix. So by rearranging the arguments as $(0, 1, \ldots, p-1, 2p-1, 2p-2, \ldots, p)$ and the orders of the polynomials as $(0, 2, \ldots, 2p-2, 1, 3, \ldots, 2p-1)$, the orthogonal matrix $\mathbf{D}$ can be transformed to another orthogonal matrix $\bar{\mathbf{O}}^*$ as

$$\bar{\mathbf{O}}^* = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{\mathbf{1}_p}{\sqrt{p}} & \sqrt{2}\mathbf{A} & \sqrt{2}\mathbf{B} \\ \frac{\mathbf{1}_p}{\sqrt{p}} & \sqrt{2}\mathbf{A} & -\sqrt{2}\mathbf{B} \end{pmatrix} \tag{22}$$

where $\mathbf{1}_{\boldsymbol{p}}$ is the $p \times 1$ vector with all elements unity. From (20) it follows that the orthogonal matrix $\bar{\mathbf{O}}^*$ can be written as the Khatri-Rao product (see Rao (1974, p. 30) where it is termed as new product) of two matrices. For ready reference, the definition is given below.

Definition 3 Suppose there are two partitioned matrices A and B, each with the same number of partitions as

$$\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k) \text{ and } \mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k).$$

Then Khatri-Rao product of $\mathbf{A}$ and $\mathbf{B}$ denoted as $\mathbf{A} * \mathbf{B}$ is defined as

$$\mathbf{A} * \mathbf{B} = (\mathbf{A}_1 \otimes \mathbf{B}_1, \mathbf{A}_2 \otimes \mathbf{B}_2, \ldots, \mathbf{A}_k \otimes \mathbf{B}_k) \tag{23}$$

where $\otimes$ denotes the Kronecker product (see Rao 1974, p. 29).

From (22) and (23) it follows that $\bar{\mathbf{O}}^*$ can be expressed as

$$\bar{\mathbf{O}}^* = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix} * \begin{pmatrix} \frac{\mathbf{1}_p}{\sqrt{2p}} & \mathbf{A} & \mathbf{B} \end{pmatrix} \tag{24}$$

where $*$ denotes the Khatri-Rao product. Again, $\bar{\mathbf{O}}^*$ being an orthogonal matrix, it follows from (24) that

$$\bar{\mathbf{O}}^*\bar{\mathbf{O}}'^* = \begin{pmatrix} \mathbf{A}^* & \mathbf{B} \\ \mathbf{A}^* & -\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{*'} & \mathbf{A}^{*'} \\ \mathbf{B}' & -\mathbf{B}' \end{pmatrix} = \begin{pmatrix} \mathbf{A}^*\mathbf{A}^{*'} + \mathbf{B}\mathbf{B}' & \mathbf{A}^*\mathbf{A}^{*'} - \mathbf{B}\mathbf{B}' \\ \mathbf{A}^*\mathbf{A}^{*'} - \mathbf{B}\mathbf{B}' & \mathbf{A}^*\mathbf{A}^{*'} + \mathbf{B}\mathbf{B}' \end{pmatrix}$$

where $A^* = \begin{pmatrix} \frac{\mathbf{1}_p}{\sqrt{2p}} & \mathbf{A} \end{pmatrix}$. As $\bar{\mathbf{O}}^*$ is an orthogonal matrix, then

$$\mathbf{A}^*\mathbf{A}^{*'} + \mathbf{B}\mathbf{B}' = \mathbf{I}_p, \quad \mathbf{A}^*\mathbf{A}^{*'} - \mathbf{B}\mathbf{B}' = \mathbf{0} \tag{25}$$

$$\Rightarrow \mathbf{A}^*\mathbf{A}^{*'} = \frac{1}{2}\mathbf{I}_p \text{ and } \mathbf{B}\mathbf{B}' = \frac{1}{2}\mathbf{I}_p. \tag{26}$$

Equation (26) also alternatively implies that $\sqrt{2}\mathbf{A}^*$ and $\sqrt{2}\mathbf{B}$ are orthogonal matrices. From (26) it also follows that $\begin{pmatrix} \frac{\mathbf{1}_p}{\sqrt{2p}} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \frac{\mathbf{1}_p}{\sqrt{2p}} & \mathbf{A} \end{pmatrix}' = \frac{1}{2}\mathbf{I}_p$

i.e.

$$\mathbf{A}\mathbf{A}' = (\mathbf{I}_p - \frac{\mathbf{J}_p}{p})\frac{1}{2} \tag{27}$$

where $\mathbf{J}_p = p \times p$ matrix with all elements unity.

Let us consider a $s_1 \times s_2 \times \cdots \times s_m$ factorial set-up where each of the s_i's is even. Let for a positive integer p_i, $s_i = 2p_i \; \forall \; i = 1, 2, \ldots, m$ and we define the following matrices

$$\mathbf{P}_i^{x_i} = \begin{cases} \frac{\mathbf{1}'_{i0}}{\sqrt{s_i}} & \text{if } x_i = 0 \\ \mathbf{P}_{ie} = \mathbf{P}_i^1 & \text{if } x_i = 1 \\ \mathbf{P}_{i0} = \mathbf{P}_i^2 & \text{if } x_i = 2 \end{cases} \quad 1 \le i \le m \tag{28}$$

where $\mathbf{1}'_{i0} = 1 \times s_i$ matrix with all elements unity,

$$\mathbf{P}_{ie} = \begin{pmatrix} \boldsymbol{\xi}'(2) \\ \boldsymbol{\xi}'(4) \\ \vdots \\ \boldsymbol{\xi}'(2p_i - 2) \end{pmatrix} \tag{29}$$

is a $(p_i - 1) \times 2p_i$ matrix with row vectors corresponding to the even order orthogonal polynomials $(\alpha \neq 0)$,

and

$$\mathbf{P}_{i0} = \begin{pmatrix} \boldsymbol{\xi}'(1) \\ \boldsymbol{\xi}'(3) \\ \vdots \\ \boldsymbol{\xi}'(2p_i - 1) \end{pmatrix} \tag{30}$$

is a $p_i \times 2p_i$ matrix with row vectors corresponding to the odd order orthogonal polynomials.

It is to be noted from (19), (20) and (28) that

$$\mathbf{P}_i^1 = \mathbf{P}_{ie} = (\mathbf{A}'_i, \mathbf{A}'_i) \tag{31}$$

and

$$\mathbf{P}_i^2 = \mathbf{P}_{i0} = (\mathbf{B}'_i, -\mathbf{B}'_i). \tag{32}$$

We define the following set of all non-null ternary vectors

$$\Omega^* = (10\ldots0, 20\ldots0, \ldots, 22\ldots2). \tag{33}$$

For each $\mathbf{x} \in \Omega^*$, we define a sub-factorial effect $F^{\mathbf{x}} = F_1^{x_1} F_2^{x_2} \ldots F_m^{x_m}, x_i = 0, 1, 2, \;\; 1 \le i \le m$ in the following manner

$$F^{\mathbf{x}} = (\mathbf{P}_1^{x_1} \otimes \mathbf{P}_2^{x_2} \otimes \cdots \otimes \mathbf{P}_m^{x_m})\mathbf{t} \tag{34}$$

where $\mathbf{P}_i^{x_i}$ is given by (28).

We define

$$\alpha_i(x_i) = \begin{cases} 1 & if\ x_i = 0 \\ (p_i - 1) & if\ x_i = 1 \\ p_i & if\ x_i = 2 \end{cases} \tag{35}$$

It is seen that $F^{\mathbf{x}}$ contains

$$\alpha(\mathbf{x}) = (\alpha_1(x_1)\alpha_2(x_2)\ldots\alpha_m(x_m)) \tag{36}$$

orthonormal contrasts, where $x_i = 0, 1, 2;\ 1 \leq i \leq m$.

Note that $[F^{\mathbf{x}}|\mathbf{x} \in \Omega^*]$ are sub-effects of the factorial effect $[F^{\mathbf{x}}|\mathbf{x} \in \Omega]$ defined in (9) because $\mathbf{P}_i$ in (8) is partitioned as

$$\mathbf{P}_i = \begin{pmatrix} P_{ie} \\ P_{io} \end{pmatrix}, 1 \leq i \leq m. \tag{37}$$

Definition 4 Let a factorial design $d(v = \prod_{i=1}^{m} s_i, b,\ k_1, k_2 \ldots, k_b, r_1, r_2, \ldots, r_v)$ denote a connected block design which allows estimation of the sub-factorial effects $\{F^{\mathbf{x}}|\mathbf{x} \in \Omega^*\}$ defined in (34) such that

$$Cov(\hat{F}^{\mathbf{x}}, \hat{F}^{\mathbf{y}}) = 0\ \ \forall\, \mathbf{x} \neq \mathbf{y} \in \Omega^*, \quad Disp(\hat{F}^{\mathbf{x}}) = \sigma^2\rho^*(\mathbf{x})\mathbf{I}_{\alpha(\mathbf{x})}, \mathbf{x} \in \Omega^* \tag{38}$$

where $s_i = 2p_i, i = 1, 2, \ldots, m$, and $\rho^*(\mathbf{x})$ is a positive constant, then d is called a partially balanced design(PBD) with orthogonal sub-factorial structure (OSFS).

3 Characterization of PBD with OSFS

With respect to the Sub-Effects defined in (34), let $v = \prod_{i=1}^{m} s_i$ where $s_i = 2p_i, 1 \leq i \leq m$ level combinations be experimented in a connected block design with b blocks of sizes $k_1, k_2, \ldots, k_b$ such that the ith level combination occurs r_i, $1 \leq i \leq v$ times. Let $N = (n_{ih})$ be the incidence matrix of the designs, where $n_{ih}(\geq 0)$ indicates the number of times the ith level combination occurs in the hth block, $h = 1, 2, \ldots, b$. The observations are assumed to follow the usual intrablock model with no (treatment × block) interaction and are independent and homoscedastic. The reduced normal equations for the treatment vector $\mathbf{t}$ is given in (10).

For $x_i = 0, 1, 2$, let us define the following matrices

$$\mathbf{M}_i^{x_i} = \mathbf{P}_i^{x_i'}\mathbf{P}_i^{x_i} \tag{39}$$

where $P_i^{x_i}$ are given by (28)–(32).

Therefore from (26), (27), and (39) we get

$$\mathbf{M}_i^0 = \mathbf{P}_i^{0\prime}\mathbf{P}_i^0 = \begin{pmatrix} \frac{\mathbf{1}_i}{\sqrt{2p_i}} \\ \frac{\mathbf{1}_i}{\sqrt{2p_i}} \end{pmatrix} \begin{pmatrix} \frac{\mathbf{1}_i'}{\sqrt{2p_i}} & \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \frac{\mathbf{J}_i}{p_i} & \frac{\mathbf{J}_i}{p_i} \\ \frac{\mathbf{J}_i}{p_i} & \frac{\mathbf{J}_i}{p_i} \end{pmatrix} \tag{40}$$

$$\mathbf{M}_i^1 = \mathbf{P}_i^{1\prime}\mathbf{P}_i^1 = \begin{pmatrix} \mathbf{A} \\ \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{A}' & \mathbf{A}' \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \mathbf{I}_i - \frac{\mathbf{J}_i}{p_i} & \mathbf{I}_i - \frac{\mathbf{J}_i}{p_i} \\ \mathbf{I}_i - \frac{\mathbf{J}_i}{p_i} & \mathbf{I}_i - \frac{\mathbf{J}_i}{p_i} \end{pmatrix} \tag{41}$$

$$\mathbf{M}_i^2 = \mathbf{P}_i^{2\prime}\mathbf{P}_i^2 = \begin{pmatrix} \mathbf{B} \\ -\mathbf{B}' \end{pmatrix} \begin{pmatrix} \mathbf{B}' & -\mathbf{B}' \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \mathbf{I}_i & -\mathbf{I}_i \\ -\mathbf{I}_i & \mathbf{I}_i \end{pmatrix} \tag{42}$$

where $\mathbf{I}_i = p_i \times p_i$ identity matrix and $J_i = p_i \times p_i$ matrix with all elements unity.

Now we prove the following theorem

Theorem 1 *A factorial design $d(v, b, r_1, r_2, \ldots, r_v, k_1, k_2, \ldots, k_b)$ involving m factors $F_1, F_2, \ldots, F_m$ is partially balanced with OSFS with respect to the partition (34) if and only if its $\mathbf{C}$-matrix is given by*

$$\mathbf{C} = \sum_{\substack{x_1,\ldots,x_m \\ \mathbf{x}\in\Omega^*}} \rho(x_1 \ldots x_m)(\mathbf{M}_1^{x_1} \otimes \mathbf{M}_2^{x_2} \otimes \cdots \otimes \mathbf{M}_m^{x_m}) \tag{43}$$

where $\rho(x_1, x_2, \ldots, x_m)$ is a real number and $\mathbf{M}_i^{x_i}$'s are given by (40)–(42).

Proof **'If' part**: Let $\mathbf{C}$ be as in (43). Then for any $\mathbf{y} \in \Omega^*$, defined in (33)

$$\mathbf{P^y C} = \sum_{\mathbf{x}\in\Omega^*} \rho(\mathbf{x})(\mathbf{P}_1^{y_1}\mathbf{M}_1^{x_1} \otimes \mathbf{P}_2^{y_2}\mathbf{M}_2^{x_2} \otimes \cdots \otimes \mathbf{P}_m^{y_m}\mathbf{M}_m^{x_m}).$$

From the definitions of $\mathbf{P}_i^{y_i}$ and $\mathbf{M}_i^{x_i}$ given in (28) and (40)–(42) respectively, we get
for $(x_i, y_i) = (0, 0)$ that

$$\mathbf{P}_i^{y_i}\mathbf{M}_i^{x_i} = \begin{pmatrix} \frac{\mathbf{1}_i'}{\sqrt{2p_i}} & \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \end{pmatrix} \frac{1}{2}\begin{pmatrix} \frac{\mathbf{J}_i}{p_i} & \frac{\mathbf{J}_i}{p_i} \\ \frac{\mathbf{J}_i}{p_i} & \frac{\mathbf{J}_i}{p_i} \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{1}_i'}{\sqrt{2p_i}} & \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \end{pmatrix} = \mathbf{P}_i^0.$$

Again for $(x_i, y_i) = (1, 1)$,

$$\mathbf{P}_i^{y_i}\mathbf{M}_i^{x_i} = \begin{pmatrix} \mathbf{A}' & \mathbf{A}' \end{pmatrix} \frac{1}{2}\begin{pmatrix} \mathbf{I}_p - \frac{\mathbf{J}_p}{p} & \mathbf{I}_p - \frac{\mathbf{J}_p}{p} \\ \mathbf{I}_p - \frac{\mathbf{J}_p}{p} & \mathbf{I}_p - \frac{\mathbf{J}_p}{p} \end{pmatrix} = \begin{pmatrix} \mathbf{A}' & \mathbf{A}' \end{pmatrix} = \mathbf{P}_i^1.$$

Similarly for any $y_i = 0, 1, 2$ and $x_i = 0, 1, 2$ we can prove that

$$\mathbf{P}_i^{y_i}\mathbf{M}_i^{x_i} = \mathbf{P}_i^{y_i}, \text{ if } x_i = y_i; = 0 \text{ otherwise}, 1 \le i \le m. \tag{44}$$

"if $y_i = x_i$; $= \mathbf{0}$ if $y_i \neq x_i$", then continue with "$1 \le i \le m$".

So

$$\mathbf{P}_i^{y_i}\mathbf{M}_i^{y_i}\mathbf{P}_i^{y_i\prime} = \mathbf{I}_{\alpha_i(y_i)}, 1 \le i \le m. \tag{45}$$

where $\alpha_i(y_i)$ is 1 or $(p_i - 1)$ or p_i when $y_i = 0$, 1 or 2 respectively, $1 \le i \le m$.

Again, for $y_i = 0, 1, 2$ and $z_i = 0, 1, 2$, $1 \leq i \leq m$ we have

$$\mathbf{P}_i^{y_i} \mathbf{M}_i^{y_i} \mathbf{P}_i^{z_i\,\prime} = \mathbf{P}_i^{y_i} \mathbf{P}_i^{z_i\,\prime} = \mathbf{I}_{\alpha_i(y_i)} \text{ if } y_i = z_i; \quad = 0 \text{ otherwise}, \ \forall i. \tag{46}$$

Therefore, from (46)

$$\mathbf{P}^{y} \mathbf{M}^{y} \mathbf{P}^{z\prime} = \begin{cases} \mathbf{I}_{\alpha(\mathbf{y})} & \text{if } \boldsymbol{y} = \boldsymbol{z} \in \Omega^* \\ 0 & \text{if } \boldsymbol{y} \neq \boldsymbol{z} \in \Omega^* \end{cases} \tag{47}$$

where $\alpha(\boldsymbol{y}) = \alpha_1(\boldsymbol{y}_1)\alpha_2(\boldsymbol{y}_2)\ldots\alpha_m(\boldsymbol{y}_m)$.

So from (43), (44) and (47) it follows that for all $\mathbf{x}$, $\mathbf{y}$ belonging to Ω^*

$$\mathbf{P}^{\mathbf{y}} \mathbf{C} \mathbf{P}^{\mathbf{z}\prime} = \begin{cases} 0 & \text{if } \boldsymbol{y} \neq \boldsymbol{z} \\ \rho(\boldsymbol{y}) I_\alpha(\boldsymbol{y}) & \text{if } \boldsymbol{y} = \boldsymbol{z}. \end{cases} \tag{48}$$

Following the same lines of proof of Lemma 3.1.3(b) of Gupta and Mukerjee (1989), it can be proved using (48) that $\rho(\mathbf{y}) > 0\ \forall \boldsymbol{y} \in \Omega^*$. Therefore from (43) and (44), it follows that

$$\left.\begin{array}{r} \mathbf{P}^{\mathbf{y}}\mathbf{C} = \rho(\mathbf{y})\mathbf{P}^{\mathbf{y}} \quad (i) \\ \Rightarrow \mathbf{P}^{\mathbf{y}} = \rho(\mathbf{y})^{-1}\mathbf{P}^{\mathbf{y}}\mathbf{C} \quad (ii) \\ \Rightarrow \mathbf{P}^{\mathbf{y}}\hat{\mathbf{t}} = \rho(\mathbf{y})^{-1}\mathbf{P}^{\mathbf{y}}\mathbf{C}\hat{\mathbf{t}} \quad (iii) \\ \Rightarrow \mathbf{P}^{\mathbf{y}}\hat{\mathbf{t}} = \hat{F}^{\mathbf{y}} = \frac{\mathbf{P}^{\mathbf{y}}\mathbf{Q}}{\rho(\mathbf{y})}. \quad (iv) \end{array}\right\} \tag{49}$$

Equation (49) follows from the normal equations (10) and it gives the BLUE for the factorial sub-effect $F^{\mathbf{y}} = \mathbf{P}^{\mathbf{y}}\mathbf{t}$, $\boldsymbol{y} \in \Omega^*$. It is known that

$$\text{Disp}(\mathbf{Q}) = \sigma^2\mathbf{C} \tag{50}$$

where σ^2 is the error variance. Therefore, for any two sub-effects $F^{\mathbf{y}}$ and $F^{\mathbf{z}}$, $\boldsymbol{y}, \ \boldsymbol{z} \in \Omega^*$ it follows from (48) and (49) that

$$\begin{aligned} \text{Cov}(\hat{F}^{\mathbf{y}}, \hat{F}^{\mathbf{z}}) = \text{Cov}(\mathbf{P}^{\mathbf{y}}\hat{\mathbf{t}}, \mathbf{P}^{\mathbf{z}}\hat{\mathbf{t}}) &= [\rho(\mathbf{y})\rho(\mathbf{z})]^{-1}\text{Cov}(\mathbf{P}^{\mathbf{y}}\mathbf{Q}, \mathbf{P}^{\mathbf{z}}\mathbf{Q}) \\ &= [\rho(\mathbf{y})\rho(\mathbf{z})]^{-1}\sigma^2\mathbf{P}^{\mathbf{y}}\mathbf{C}\mathbf{P}^{\mathbf{z}\prime} \\ &= 0 \text{ if } \mathbf{y} \neq \mathbf{z} \in \Omega^*. \end{aligned} \tag{51}$$

Equation (51) implies that d has orthogonal sub-factorial structure (OSFS). Again from (48), (49), (50), (51) it follows that

$$\text{Disp}(\hat{F}^{\mathbf{y}}) = \text{Disp}(\mathbf{P}^{\mathbf{y}}\hat{\mathbf{t}}) = \sigma^2[\rho^2(\mathbf{y})]^{-1}(\mathbf{P}^{\mathbf{y}}\mathbf{C}\mathbf{P}^{\mathbf{y}\prime}) = \frac{\sigma^2}{\rho(\boldsymbol{y})} I_{\alpha(\boldsymbol{y})}\ \forall \boldsymbol{y} \in \Omega^*. \tag{52}$$

Therefore, (52) implies that the BLUE of each of $\alpha(\mathbf{y})$, orthonormal contrasts belonging $F^{\mathbf{y}}$, $\mathbf{y} \in \Omega^*$, has equal variance. So the design d is partially balanced with respect to partition $[F^{\mathbf{y}}]$, $\mathbf{y} \in \Omega^*$.

'Only If' part: Let us define a matrix $\mathbf{P}^{(v-1)\times v}$ as

$$\mathbf{P} = \left(\mathbf{P}^{10\ldots0\prime}\ \mathbf{P}^{20\ldots0\prime}\ \ldots\ \mathbf{P}^{22\ldots2\prime}\right)' = (\mathbf{P}^{\mathbf{y}}|\mathbf{y} \in \Omega^*). \tag{53}$$

See that $\begin{pmatrix} \frac{\mathbf{1}'_v}{\sqrt{v}} \\ \mathbf{P} \end{pmatrix}$ is a $v \times v$ orthogonal matrix, where $\mathbf{1}'_v = (1, 1, \ldots, 1)^{1\times v}$. Therefore it follows that

$$\left(\tfrac{\mathbf{1}_v}{\sqrt{v}}\ \mathbf{P}'\right)\begin{pmatrix} \frac{\mathbf{1}'_v}{\sqrt{v}} \\ \mathbf{P} \end{pmatrix} = \mathbf{I}_v \ \Rightarrow\ \frac{\mathbf{J}_v}{v} + \mathbf{P}'\mathbf{P} = \mathbf{I}_v \ \Rightarrow\ \mathbf{P}'\mathbf{P} = \mathbf{I}_v - \frac{\mathbf{J}_v}{v}$$

where $\mathbf{J}_v = v \times v$ matrix with all element unity and $\mathbf{I}_v = v \times v$ identity matrix.

Again, since C has all row and column sums zero and $P'P = I_v - \frac{J_v}{v}$ it follows that

$$\mathbf{P}'\mathbf{P}\mathbf{C} = \mathbf{C}\mathbf{P}'\mathbf{P} = \mathbf{C}. \tag{54}$$

As $\mathbf{C}$ is positive semi-definite of rank $(v-1)$ it can be easily proved that (following the lines of proof of Lemma 3.1.3(a) of Gupta and Mukerjee 1989) $(\mathbf{PCP}')^{(v-1)\times(v-1)}$ is positive definite. Therefore from (54), we can write that

$$\begin{aligned} (\mathbf{PCP}')\mathbf{P} &= \mathbf{PC} \\ \Rightarrow \mathbf{P} &= (\mathbf{PCP}')^{-1}\mathbf{PC} \\ \Rightarrow \mathbf{P}\hat{\mathbf{t}} &= (\mathbf{PCP}')^{-1}\mathbf{PQ} \\ \Rightarrow \text{Disp}(\mathbf{P}\hat{\mathbf{t}}) &= \sigma^2(\mathbf{PCP}')^{-1}. \end{aligned} \tag{55}$$

Now

$$\mathbf{P}\hat{\mathbf{t}} = (\ldots\mathbf{P}^{\mathbf{x}}\hat{\mathbf{t}}\ldots|\mathbf{x} \in \Omega^*) = (\ldots\hat{F}^{\mathbf{x}}\ldots|\mathbf{x} \in \Omega^*). \tag{56}$$

Therefore

$$\text{Disp}(\mathbf{P}\hat{\mathbf{t}}) = \text{Disp}\left(\ldots\ \hat{F}^{\mathbf{x}}\ \ldots\right)' = \sigma^2(\mathbf{PCP}')^{-1}. \tag{57}$$

So for OSFS, we must have $\text{Cov}(\hat{F}^{\mathbf{x}}, \hat{F}^{\mathrm{z}}) = 0\ \ \forall\mathbf{x} \neq \mathbf{z} \in \Omega^*$ and hence $(\mathbf{PCP}')^{-1}$ must be a block diagonal matrix i.e.

$$\sigma^2(\mathbf{PCP}')^{-1} = \text{Diag}(\ldots\text{Disp}(\hat{F}^{\mathbf{x}})\ldots|\mathbf{x} \in \Omega^*) = \text{Diag}(\ldots\sigma^2\mathbf{D}_{\mathbf{x}}\ldots|\mathbf{x} \in \Omega^*) \tag{58}$$

where $\text{Disp}(\hat{F}^{\mathrm{x}}) = \sigma^2\mathbf{D}_{\mathbf{x}},\ \mathbf{x} \in \Omega^*$.

Again for partial balance, $\mathbf{D}_{\mathbf{x}}$ must be proportional to $\mathbf{I}_{\alpha(\mathbf{x})}$, $\mathbf{x} \in \Omega^*$ i.e. for $d_{\mathbf{x}} > 0 \ \forall \ \mathbf{x} \in \Omega^*$

$$\sigma^2(\mathbf{PCP}')^{-1} = \text{Diag}(\ldots \sigma^2 d_{\mathbf{x}}^{-1} \mathbf{I}_{\alpha(\mathbf{x})} | \mathbf{x} \in \Omega^*) \tag{59}$$

$$\begin{aligned}
&\Rightarrow \mathbf{PCP}' = \text{Diag}(\ldots d_{\mathbf{x}} \mathbf{I}_{\alpha(\mathbf{x})} \ldots | \mathbf{x} \in \Omega^*) \\
&\Rightarrow \mathbf{P}'\mathbf{PCP}'\mathbf{P} = \mathbf{P}'\text{Diag}(\ldots d_{\mathbf{x}} \mathbf{I}_{\alpha(\mathbf{x})} \ldots)\mathbf{P} \\
&\Rightarrow \mathbf{C} = \sum_{\mathbf{x} \in \Omega^*} d(\mathbf{x}) \mathbf{P}^{\mathbf{x}\prime} \mathbf{P}^{\mathbf{x}} \ \ (\text{as } \mathbf{P}'\mathbf{PCP}'\mathbf{P} = \mathbf{C}) \\
&\qquad = \sum_{\mathbf{x} \in \Omega^*} d(\mathbf{x}) \mathbf{M}^{\mathbf{x}}.
\end{aligned}$$

Thus the ‘only if’ part is proved. □

For $i = 1, 2, \ldots, m$, let us define the following matrices

$$\mathbf{Z}_i^{x_i} = \begin{pmatrix} \mathbf{J}_{p_i} & \mathbf{J}_{p_i} \\ \mathbf{J}_{p_i} & \mathbf{J}_{p_i} \end{pmatrix} \text{ if } x_i = 0 \text{ or } \begin{pmatrix} \mathbf{I}_{p_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p_i} \end{pmatrix} \text{ if } x_i = 1, \text{ or } \begin{pmatrix} \mathbf{0} & \mathbf{I}_{p_i} \\ \mathbf{I}_{p_i} & \mathbf{0} \end{pmatrix} \text{ if } x_i = 2. \tag{60}$$

We see from (40)–(42) and (60) that

$$\mathbf{M}_i^0 = \frac{\mathbf{Z}_i^0}{2p_i}, \ \ \mathbf{M}_i^1 = \frac{1}{2}[(\mathbf{Z}_i^1 + \mathbf{Z}_i^2) - \frac{\mathbf{Z}_i^0}{p_i}], \ \ \mathbf{M}_i^2 = \frac{1}{2}[\mathbf{Z}_i^1 - \mathbf{Z}_i^2]. \tag{61}$$

Conversely,

$$\mathbf{Z}_i^0 = 2p_i \mathbf{M}_i^0, \ \ \mathbf{Z}_i^1 = \mathbf{M}_i^0 + \mathbf{M}_i^1 + \mathbf{M}_i^2, \ \ \mathbf{Z}_i^2 = (\mathbf{M}_i^0 + \mathbf{M}_i^1) - \mathbf{M}_i^2. \tag{62}$$

So from Theorem 1, (61) and (62) we get the following theorem.

Theorem 2 *A factorial design d is partially balanced with OSFS if and only if*

$$\mathbf{C} = \sum_{(x_1, x_2, \ldots, x_m) \in \Omega^{**}} h(x_1, x_2, \ldots, x_m)(\mathbf{Z}_1^{x_1} \otimes \mathbf{Z}_2^{x_2} \otimes \cdots \otimes \mathbf{Z}_m^{x_m}) \tag{63}$$

where $h(x_1, x_2, \ldots, x_m)$ *are real numbers and* $\Omega^{**} = \Omega^* \cup (0, 0, \ldots, 0)$.

Proof **‘Only if’ part** From Theorem 1 it is known that d is partially balanced with OSFS only if

$$\mathbf{C} = \sum_{(x_1, x_2, \ldots, x_m) \in \Omega^*} \rho(x_1, x_2, \ldots, x_m)(\mathbf{M}_1^{x_1} \otimes \mathbf{M}_2^{x_2} \otimes \cdots \otimes \mathbf{M}_m^{x_m}).$$

Now by replacing $\mathbf{M}_i^{x_i}$ by $\mathbf{Z}_i^{x_i}$’s from (61) we get the relation (63). So the ‘only if’ part is proved.

‘If’ part We suppose that the relation (63) is given. Then by replacing the $M_i^{x_i}$’s by $Z_i^{x_i}$’s from (62) we can get the following equation

$$\left.\begin{aligned} \mathbf{C} = &\sum_{(x_1,x_2,\ldots,x_m)\in\Omega^{**}} h(x_1, x_2, \ldots, x_m)(\mathbf{M}_1^{x_1} \otimes \mathbf{M}_2^{x_2} \otimes \cdots \otimes \mathbf{M}_m^{x_m}) \\ &= h(0, 0, \ldots 0)(\mathbf{M}_1^0 \otimes \mathbf{M}_2^0 \otimes \cdots \otimes \mathbf{M}_m^0)+ \\ &\sum_{(x_1,x_2,\ldots,x_m)\in\Omega^{*}} h(x_1, x_2, \ldots x_m)(\mathbf{M}_1^{x_1} \otimes \mathbf{M}_2^{x_2} \otimes \cdots \otimes \mathbf{M}_m^{x_m}). \end{aligned}\right\} \tag{64}$$

Now for every $(y_1, y_2, \ldots y_m) \in \Omega^*$
$(\mathbf{M}_1^{y_1} \otimes \cdots \otimes \mathbf{M}_m^{y_m})(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0) = \mathbf{0}$
and also
$(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0)(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0) = (\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0) \neq \mathbf{0}.$

Again $\mathbf{C}(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0)$=$\mathbf{0}$ as the row sums of $\mathbf{C}$ are zeros. So by post-multiplying (64) by $(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0)$ we get

$$\left.\begin{aligned} &\mathbf{C}(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0) = h(0, 0, \ldots, 0)(\mathbf{M}_1^0 \otimes \cdots \otimes \mathbf{M}_m^0) \\ &\Rightarrow h(0, 0, \ldots, 0) = 0. \end{aligned}\right\} \tag{65}$$

So $\mathbf{C}$ takes the form

$$\mathbf{C} = \sum_{(x_1,x_2,\ldots,x_m)\in\Omega^{*}} h(x_1, x_2, \ldots, x_m)(\mathbf{M}_1^{x_1} \otimes \cdots \otimes \mathbf{M}_m^{x_m}). \tag{66}$$

Therefore, from Theorem 1, it follows that d has partial balance with OSFS. □

Let us define, for $x_i = 0, 1,\ 1 \le i \le m$

$$\mathbf{Z}_i^{x_i} = \begin{cases} \mathbf{I}_{i0} & \text{if } x_i = 0 \\ \mathbf{J}_{i0} & \text{if } x_i = 1 \end{cases} \tag{67}$$

where $\mathbf{I}_{i0} = s_i \times s_i$ identity matrix, $\mathbf{J}_{i0} = s_i \times s_i$ matrix with all elements unity.

Then if the C-matrix of a design has the structure

$$\mathbf{C} = \sum_{(x_1,x_2,\ldots,x_m)\in\Omega^{*}} h(x_1, x_2, \ldots, x_m)(\mathbf{Z}_1^{x_1} \otimes \ldots \mathbf{Z}_m^{x_m}) \tag{68}$$

then $\mathbf{C}$ is said to have property A (cf. Gupta and Mukerjee 1989), where $h(x_1, x_2, \ldots x_m)$s are real numbers and $\mathbf{Z}_i^{x_i}$s are as in (67).

$\mathbf{Z}_i^{x_i}$s considered in (60) are of more general nature than those given in Gupta and Mukerjee (1989) and have three alternative forms depending on the values of x_i. If a $\mathbf{C}$-matrix has the structure of (63), we define that the $\mathbf{C}$-matrix has ‘generalized A’ property. So, the Theorem 2 can be alternatively stated as

Theorem 3 *A factorial design d is partially balanced with OSFS with respect to the partition (34) if and only if its* **C***-matrix has ‘generalised A’ property.*

Let us consider a binary, equi-replicate and proper design d. Then its C-matrix is given by

$$\mathbf{C} = r\mathbf{I}_v - \frac{\mathbf{NN}'}{k} \Rightarrow \mathbf{NN}' = k(r\mathbf{I}_v - \mathbf{C}). \tag{69}$$

If d has property 'generalized A' then from (69) it follows that $\mathbf{NN}'$ has property 'generalized A' as

$$\mathbf{I}_v = \mathbf{Z}_1^0 \otimes \mathbf{Z}_2^0 \otimes \cdots \otimes \mathbf{Z}_m^0.$$

So from Theorem 3 and above discussion we get the following result.

Theorem 4 *A binary, proper and equi-replicate design d is partially balanced with orthogonal sub-factorial structure with respect to the partition (34) if and only if the* $\mathbf{NN}'$ *matrix has 'generalized A' property.*

4 Combinatorial Characterization of Partially Balanced Design with Orthogonal Sub-Factorial Structure

We divide the $2p_i$ levels of the factor F_i, into p_i groups each containing 2 levels where

$$G_{i1} = (0, 2p_i - 1),\ G_{i2} = (1, 2p_i - 2),\ \ldots, G_{ip_i} = (p_i - 1, p_i),\ 1 \le i \le m. \tag{70}$$

We define a group-divisible association scheme (cf. Raghavarao 1971) among the $2p_i$ levels of F_i as: if any two treatments belong to the same group, they are first associates; otherwise they are second associates.

We have considered the $2p_i$ levels in the order as $G_i = (0, 1, \ldots, p_i - 1, 2p_i - 1, 2p_i - 2, \ldots, p_i)$, which can alternatively be written as $G_i = (00, 01, \ldots, 0p_i - 1; 10, 11, \ldots, 1p_i - 1) = (G_i^1, G_i^2),$ (say); $1 \le i \le m$.

The groups of levels of (70) can be written as

$$G_{i1} = (00, 10),\ G_{i2} = (01, 11), \ldots, G_{ip_i} = (0p_i - 1, 1p_i - 1). \tag{71}$$

With these new notations of the levels the association scheme(a.s) over the levels of F_i can be reformulated as
(i) level (i, j) and level (i', j') are 0th associate if $(i = i', j = j')$; (ii) level (i, j) and level (i', j') are 1st associate if $(i \neq i', j = j')$; (iii) level (i, j) and level (i', j') are 2nd associate otherwise.

Therefore, the association matrices can be represented as

$$\mathbf{B}_i^0 = \begin{pmatrix} \mathbf{I}_i & \mathbf{0}_i \\ \mathbf{0}_i & \mathbf{I}_i \end{pmatrix}, \tag{72}$$

$$\mathbf{B}_i^1 = \begin{pmatrix} \mathbf{0}_i & \mathbf{I}_i \\ \mathbf{I}_i & \mathbf{0}_i \end{pmatrix}, \tag{73}$$

$$\mathbf{B}_i^2 = \begin{pmatrix} \mathbf{J}_i - \mathbf{I}_i & \mathbf{J}_i - \mathbf{I}_i \\ \mathbf{J}_i - \mathbf{I}_i & \mathbf{J}_i - \mathbf{I}_i \end{pmatrix} \tag{74}$$

where

$\mathbf{0}_i$, $\mathbf{J}_i$ and $\mathbf{I}_i$ are respectively $p_i \times p_i$ null, sum and identity matrices.

Now we prove the following result.

Result 1 *$[\mathbf{Z}_i^{x_i}]$ are linear combinations of $[\mathbf{B}_i^{x_i}]$ and conversely, $x_i = 0, 1, 2;\ 1 \le i \le m$.*

Proof From (60) and (72)–(74) we see that

$$\mathbf{Z}_i^0 = \mathbf{B}_i^0 + \mathbf{B}_i^1 + \mathbf{B}_i^2;\ \mathbf{Z}_i^1 = \mathbf{B}_i^0,\ \mathbf{Z}_i^2 = \mathbf{B}_i^1;\quad 1 \le i \le m. \tag{75}$$

Again,

$$\mathbf{B}_i^0 = \mathbf{Z}_i^1,\ \mathbf{B}_i^1 = \mathbf{Z}_i^2,\ \mathbf{B}_i^2 = \mathbf{Z}_i^0 - (\mathbf{Z}_i^1 + \mathbf{Z}_i^2);\quad 1 \le i \le m. \tag{76}$$

□

We now define a PBIB association scheme called ‘generalized extended group divisible’ (GEGD) association scheme between the $v = \prod_{i=1}^{m} s_i$ level combinations of the m factors $F_1, F_2, \ldots, F_m$.

Definition 5 Consider any two treatment combinations $t = [(i_1, j_1), (i_2, j_2), \ldots, (i_m, j_m)]$ and $t' = [(i'_1, j'_1), (i'_2, j'_2), \ldots, (i'_m, j'_m)]$, $i_u,\ i'_u = 0, 1$ and $j_u,\ j'_u = 0, 1, \ldots, p_u - 1$, $1 \le u \le m$. Then t and t' are defined to be $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^{\text{th}}$ associate where x_u is given by

$$x_u = \begin{cases} 0 \text{ if } i_u = i'_u,\ j_u = j'_u \\ 1 \text{ if } i_u \ne i'_u,\ j_u = j'_u \\ 2 \text{ if } i_u = i'_u,\ j_u \ne j'_u \text{ or } i_u \ne i'_u,\ j_u \ne j'_u \end{cases} \qquad 1 \le u \le m. \tag{77}$$

The above defines a PBIB association scheme with all possible $(3^m - 1)$ associate classes. Nair and Rao (1948) and Shah (1958, 1960) used a $(2^m - 1)$-class association scheme for combinatorial characterization of balanced design with OFS for 2^m experiment. Hinkelman and Kempthorne (1963) called such scheme as ‘extended group divisible’ (EGD) association scheme and Paik and Federer (1973) called such scheme as ‘binary number association scheme’. In this scheme two treatment combinations $(j_1, j_2, \ldots, j_m)$ and $(j'_1, j'_2, \ldots, j'_m)$, $0 \le j_i, j'_i \le s_i - 1, i = 1, 2, \ldots, m$ are said to be $y = (y_1, y_2, \ldots y_m)^{\text{th}}$ associate where

$$y_u = \begin{cases} 0 \text{ if } j_u = j'_u \\ 1 \text{ if } j_u \ne j'_u \end{cases} \quad 1 \le u \le m.$$

It can be proved that the 2^m association matrices (including the $(00\dots0)^{\text{th}}$ association matrices) of the EGD association scheme are given by (Gupta (1988); see also Gupta and Mukerjee (1989))

$$\mathbf{B}^{*\mathbf{y}} = \mathbf{B}_1^{*y_1} \otimes \mathbf{B}_2^{*y_2} \otimes \cdots \otimes \mathbf{B}_m^{*y_m} \tag{78}$$

where

$$\mathbf{B}_i^{*y_i} = \begin{cases} \mathbf{I}_{i0} & \text{if } y_i = 0 \\ (\mathbf{J}_{i0} - \mathbf{I}_{i0}) & \text{if } y_i = 1 \end{cases} \tag{79}$$

and

$\mathbf{I}_{i0} = s_i \times s_i$ identity matrix

$\mathbf{J}_{i0} = s_i \times s_i$ matrix with all elements unity.

Note that $[\mathbf{B}_i^{*y_i}]$ are the association matrices of the BIB association scheme defined over the s_i levels of the ith factor F_i, $1 \le i \le m$.

It can be verified that the $(3^m - 1)$ association matrices of the association scheme defined in (77) can be written as

$$\mathbf{B}^{\mathbf{x}} = \mathbf{B}_1^{x_1} \otimes \mathbf{B}_2^{x_2} \otimes \cdots \otimes \mathbf{B}_m^{x_m} \tag{80}$$

where $[\mathbf{B}_i^{x_i}]$ are given by (72)–(74) and are association matrices of a GD association scheme. Note that this association scheme is a generalization of the EGD association scheme in the same way as GD association scheme is a generalization of BIB association scheme in the one-factor case. This scheme reduces to the EGD association scheme if 1th and 2th associations of the levels of F_i are coaleased to a single association, $1 \le i \le m$. For this, we call such association scheme as 'generalized EGD' (GEGD) association scheme.

Now we consider PBIBDs based on the GEGD association scheme introduced in Definition 5 where any two treatments which are xth associate, occur together in $\lambda(\mathbf{x})$ blocks, $\mathbf{x} \in \Omega^*$. So if $\mathbf{N}^{v \times b}$ is the incidence matrix and r is the common replication number, then obviously, for the GEGDD we have

$$\mathbf{NN}' = r\mathbf{I}_v + \sum_{\mathbf{x} \in \Omega^*} \lambda(\mathbf{x})\mathbf{B}^{\mathbf{x}}. \tag{81}$$

We have seen that [Eqs. (75) and (76)] that $\{\mathbf{B}_i^{x_i}\}$s can be expressed in terms of $[\mathbf{Z}_i^{x_i}]$ and conversely i.e. $\mathbf{NN}'$ in (81) has 'generalized A property'. Conversely if $\mathbf{NN}'$ has generalised A property, then it can be expressed as the $\mathbf{NN}'$ matrix of a GEGD design given in (81). So we get the following theorem.

Theorem 5 *A binary, proper and equi-replicate design d is partially balanced with orthogonal sub-factorial structure with respect to the partition (34) if and only if it is a GEGDD.*

Using (81) the C-matrix of a GEGD design can be written as

$$\mathbf{C} = r\mathbf{I}_v - \frac{\mathbf{NN}'}{k} = r\mathbf{I}_v - \frac{r\mathbf{I}_v}{k} - \sum_{\mathbf{x}\in\Omega^*} \lambda(\mathbf{x})\mathbf{B}^{\mathbf{x}} = \frac{r(k-1)}{k}\mathbf{I}_v - \sum_{\mathbf{x}\in\Omega^*} \lambda(\mathbf{x})\mathbf{B}^{\mathbf{x}}. \quad (82)$$

Let $\mathbf{P}^{\mathbf{y}} = (\mathbf{P}_1^{y_1} \otimes \mathbf{P}_2^{y_2} \otimes \cdots \otimes \mathbf{P}_m^{y_m})$, where $[\mathbf{P}_i^{y_i}],\ 1 \le i \le m$ are given by (28). Then

$$\mathbf{P}^{\mathbf{y}}\mathbf{C}\mathbf{P}^{\mathbf{y}'} = \frac{r(k-1)}{k}\mathbf{I}_{\alpha(y)} - \sum_{\mathbf{x}\in\Omega^*} \lambda(\mathbf{x})\mathbf{P}^{\mathbf{y}}\mathbf{B}^{\mathbf{x}}\mathbf{P}^{\mathbf{y}'}. \quad (83)$$

Now taking $\{\mathbf{B}_i^{x_i}\}$s from (72)–(74) and $\{\mathbf{P}_i^{y_i}\}$s from (28) we compute $\mathbf{P}_i^{y_i}\mathbf{B}_i^{x_i}\mathbf{P}_i^{y_i'}$ for different values of y_is and x_is. For example

(i) if $y_i = 0,\ x_i = 0$, then

$$\mathbf{P}_i^0\mathbf{B}_i^0\mathbf{P}_i^{0'} = \begin{pmatrix} \frac{\mathbf{1}_i'}{\sqrt{2p_i}} & \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \end{pmatrix} \begin{pmatrix} \mathbf{I}_i & \mathbf{0}_i \\ \mathbf{0}_i & \mathbf{I}_i \end{pmatrix} \begin{pmatrix} \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \\ \frac{\mathbf{1}_i'}{\sqrt{2p_i}} \end{pmatrix} = 1 = 1.I_{\alpha_i(0)};$$

(ii) if $y_i = 1, x_i = 2$, then

$$\mathbf{P}_i^1\mathbf{B}_i^2\mathbf{P}_i^{1'} = \begin{pmatrix} \mathbf{A}_i' & \mathbf{A}_i' \end{pmatrix} \begin{pmatrix} \mathbf{J}_i - \mathbf{I}_i & \mathbf{J}_i - \mathbf{I}_i \\ \mathbf{J}_i - \mathbf{I}_i & \mathbf{J}_i - \mathbf{I}_i \end{pmatrix} \begin{pmatrix} \mathbf{A}_i \\ \mathbf{A}_i \end{pmatrix} = (-2)\mathbf{I}_{p_i-1} = (-2)\mathbf{I}_{\alpha_i(1)}.$$

In this way, we construct the following table.
where $\alpha_i(0) = 1, \alpha_i(1) = p_i - 1, \alpha_i(2) = p_i;\ 1 \le i \le m$.

Therefore, from the entries of Table 1 we can write $\mathbf{P}^{\mathbf{y}}\mathbf{B}^{\mathbf{x}}\mathbf{P}^{\mathbf{y}'},\ \mathbf{x},\ \mathbf{y} \in \Omega^*$, as

$$\mathbf{P}^{\mathbf{y}}\mathbf{B}^{\mathbf{x}}\mathbf{P}^{\mathbf{y}'} = \otimes_{i=1}^m \mathbf{P}_i^{y_i}\mathbf{B}_i^{x_i}\mathbf{P}_i^{y_{i'}} = [\prod_{i=1}^m u_i(y_i, x_i)]\mathbf{I}_{\alpha(y)} \quad (84)$$

Table 1 Computation of $\mathbf{P}_i^{y_i}\mathbf{B}_i^{x_i}\mathbf{P}_i^{y_i'};\ y_i,\ x_i = 0, 1, 2;\ 1 \le i \le m$

y_i	x_i	$P_i^{y_i}B_i^{x_i}P_i^{y_i'}$
0	0	$1.I_{\alpha_i(0)} = u_i(0,0)I_{(\alpha_i(0))}$
0	1	$1.I_{\alpha_i(0)} = u_i(0,1)I_{(\alpha_i(0))}$
0	2	$2.(p_i-1)I_{\alpha_i(0)} = u_i(0,2)I_{(\alpha_i(0))}$
1	0	$1.I_{\alpha_i(1)} = u_i(1,0)I_{\alpha_i(1)}$
1	1	$1.I_{\alpha_i(1)} = u_i(1,1)I_{\alpha_i(1)}$
1	2	$(-2)I_{\alpha_i(1)} = u_i(1,2)I_{\alpha_i(1)}$
2	0	$1.I_{\alpha_i(2)} = u_i(2,0)I_{\alpha_i(2)}$
2	1	$1.I_{\alpha_i(2)} = u_i(2,1)I_{\alpha_i(2)}$
2	2	$0.I_{\alpha_i(2)} = u_i(2,2)I_{\alpha_i(2)}$

$$\left.\begin{aligned}\mathbf{P}^{\mathbf{y}}\mathbf{C}\mathbf{P}^{\mathbf{y}'} &= \tfrac{r(k-1)}{k}\mathbf{I}_{\alpha(\mathbf{y})} - \tfrac{1}{k}\sum_{\mathbf{x}\in\Omega^*}\lambda(\mathbf{x})[\prod_{i=1}^{m}u(y_i, x_i)]I_{\alpha(\mathbf{y})}\\ &= [\tfrac{r(k-1)}{k} - \tfrac{1}{k}\sum_{\mathbf{x}\in\Omega^*}\lambda(\mathbf{x})(\prod_{i=1}^{m}u(y_i, x_i))]\mathbf{I}_{\alpha(y)}.\end{aligned}\right\} \tag{85}$$

where $\alpha(\mathbf{y}) = \alpha_1(y_1)\alpha_2(y_2)\ldots\alpha_m(y_m)$ and $u_i(y_i, x_i)$s are given in Table 1

So comparing (48) and (85) we have, for $y \in \Omega^*$

$$\rho(\mathbf{y}) = \frac{r(k-1)}{k} - \sum_{\mathbf{x}\in\Omega^*}\lambda(\mathbf{x})[\prod_{i=1}^{m}u(y_i, x_i)]. \tag{86}$$

Again from (52), we can write for a connected GEGD that

$$\text{Disp}_{\text{GEGD}}(\hat{\mathbf{F}}^{\mathbf{y}}) = \text{Disp}(\mathbf{P}^{\mathbf{y}}\hat{\mathbf{t}}) = \sigma^2\frac{\mathbf{I}_{\alpha(\mathbf{y})}}{\rho(\mathbf{y})},\quad \mathbf{y} \in \Omega^*. \tag{87}$$

Also, on the other hand, if a randomized complete block design with r replications were used, then we would have

$$\text{Disp}_{(\text{RBD})}(\hat{\mathbf{F}}^{\mathbf{y}}) = \sigma^2\frac{\mathbf{I}_{\alpha(\mathbf{y})}}{r}. \tag{88}$$

So from (87), (88) we get the efficiency of GEGD with respect to the sub-factorial effect F^y, $y \in \Omega^*$ as

$$\epsilon(\mathbf{y}) = \frac{\rho(\mathbf{y})}{r} \tag{89}$$

where $\rho(\mathbf{y})$ is given by (86).

Example 1 Let there be two factors F_1 and F_2 with $s_1 = 4,\ s_2 = 6$. The $4 \times 6 = 24$ treatment combinations are written as $[(j_1, j_2),\ j_1 = 0, 1, 2, 3;\ j_2 = 0, 1, 2, 3, 4, 5]$. The 16 blocks are shown below. The level combinations are arranged as

Blocks															
B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}
00	00	01	02	30	30	31	32	00	00	01	02	10	10	11	12
01	04	05	05	31	34	35	35	01	04	05	05	11	14	15	15
02	03	03	04	32	33	33	34	02	03	03	04	12	13	13	14
10	10	11	12	20	20	21	22	20	20	21	22	30	30	31	32
11	14	15	15	21	24	25	25	21	24	25	25	31	34	35	35
12	13	13	14	22	23	23	24	22	23	23	24	32	33	33	34

(0, 1, 3, 2) ⊙ (0, 1, 2, 5, 4, 3) = (00, 01, 02, 05, 04, 03, |10, 11, 12, 15, 14, 13, |30, 31, 32, 35, 34, 33, |20, 21, 22, 25, 24, 23), where ⊙ denotes a direct product.

The incidence matrix of the design is given by

$$\begin{array}{c} \\ 00\\01\\02\\05\\04\\03\\10\\11\\12\\15\\14\\13\\30\\31\\32\\35\\34\\33\\20\\21\\22\\25\\24\\23 \end{array}
\left(\begin{array}{cccc|cccc|cccc|cccc}
1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16\\ \hline
1&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0\\
1&0&1&0&0&0&0&0&1&0&1&0&0&0&0&0\\
1&0&0&1&0&0&0&0&1&0&0&1&0&0&0&0\\
0&0&1&1&0&0&0&0&0&0&1&1&0&0&0&0\\
0&1&0&1&0&0&0&0&0&1&0&1&0&0&0&0\\
0&1&1&0&0&0&0&0&0&1&1&0&0&0&0&0\\
1&1&0&0&0&0&0&0&0&0&0&0&1&1&0&0\\
1&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0\\
1&0&0&1&0&0&0&0&0&0&0&0&1&0&0&1\\
0&0&1&1&0&0&0&0&0&0&0&0&0&0&1&1\\
0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&1\\
0&1&1&0&0&0&0&0&0&0&0&0&0&1&1&0\\
0&0&0&0&1&1&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0\\
0&0&0&0&1&0&0&1&0&0&0&0&1&0&0&1\\
0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&1\\
0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&1\\
0&0&0&0&0&1&1&0&0&0&0&0&0&1&1&0\\
0&0&0&0&1&1&0&0&1&1&0&0&0&0&0&0\\
0&0&0&0&1&0&1&0&1&0&1&0&0&0&0&0\\
0&0&0&0&1&0&0&1&1&0&0&1&0&0&0&0\\
0&0&0&0&0&0&1&1&0&0&1&1&0&0&0&0\\
0&0&0&0&0&1&0&1&0&1&0&1&0&0&0&0\\
0&0&0&0&0&1&1&0&0&1&1&0&0&0&0&0
\end{array}\right)$$

The incidence matrix $\mathbf{N}$ can be expressed as $\mathbf{N} = \mathbf{N}_1 \otimes \mathbf{N}_2$ where $\mathbf{N}_1$ and $\mathbf{N}_2$ are given respectively by

$$\mathbf{N}_1 = \begin{array}{r} \text{Levels}\\ 0\\1\\3\\2 \end{array}
\begin{pmatrix} 1&0&1&0\\ 1&0&0&1\\ 0&1&0&1\\ 0&1&1&0 \end{pmatrix}
\quad
\mathbf{N}_2 = \begin{array}{c} \text{Levels}\\ 0\\1\\2\\5\\4\\3 \end{array}
\begin{pmatrix} 1&1&0&0\\ 1&0&1&0\\ 1&0&0&1\\ 0&0&1&1\\ 0&1&0&1\\ 0&1&1&0 \end{pmatrix}.$$

It follows that $\mathbf{N}\mathbf{N}' = \mathbf{N}_1\mathbf{N}_1' \otimes \mathbf{N}_2\mathbf{N}_2' = \left(\begin{array}{cc|cc} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ \hline 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{array}\right) \otimes \left(\begin{array}{ccc|ccc} 2 & 1 & 1 & 0 & 1 & 1 \\ 1 & 2 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 1 & 1 & 0 \\ \hline 0 & 1 & 1 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 & 2 & 1 \\ 1 & 1 & 0 & 1 & 1 & 2 \end{array}\right)$

$$= (2\mathbf{B}_1^0 + 0\mathbf{B}_1^1 + 1\mathbf{B}_1^2) \otimes (2\mathbf{B}_2^0 + 0\mathbf{B}_2^1 + 1\mathbf{B}_2^2) = 4B^{00} + 2B^{02} + 2B^{20} + 1B^{22} \tag{90}$$

where $B_i^{x_i}$s are given by (72)–(74), $x_i = 0, 1, 2;\ i = 1, 2$.

It follows that $r = 4, \lambda(02) = \lambda(20) = 2, \lambda(22) = 1$ and $\lambda(x_1, x_2) = 0$ for $(x_1, x_2) = (01, 10, 11, 12, 21)$. From (86) the ρ's are given by

$$\begin{aligned}
\rho(10) &= \tfrac{r(k-1)}{k} - \tfrac{1}{k}[\lambda(20)(u_1(12)u_2(00)) + \lambda(02)(u_1(10)u_2(02)) + \lambda(22)(u_1(12)u_2(02))] = 4\\
\rho(20) &= \tfrac{r(k-1)}{k} - \tfrac{1}{k}[\lambda(20)(u_1(22)u_2(00)) + \lambda(02)(u_1(20)u_2(02)) + \lambda(22)(u_1(22)u_2(02))] = 2\\
\rho(11) &= \tfrac{r(k-1)}{k} - \tfrac{1}{k}[\lambda(20)(u_1(12)u_2(10)) + \lambda(02)(u_1(10)u_2(12)) + \lambda(22)(u_1(12)u_2(12))] = 4\\
\rho(12) &= \tfrac{20}{6} - \tfrac{1}{6}[\lambda(20)u_1(12)u_2(20) + \lambda(02)u_1(10)u_2(22) + \lambda(22)u_1(12)u_2(22)] = 4\\
\rho(21) &= \tfrac{20}{6} - \tfrac{1}{6}[\lambda(20)u_1(22)u_2(10) + \lambda(02)u_1(20)u_2(12) + \lambda(22)u_1(22)u_2(12)] = 4\\
\rho(22) &= \tfrac{20}{6} - \tfrac{1}{6}[\lambda(20)u_1(22)u_2(22) + \lambda(02)u_1(20)u_2(22) + \lambda(22)u_1(22)u_2(22)] = \tfrac{10}{3}\\
\rho(02) &= \tfrac{20}{6} - \tfrac{1}{6}[\lambda(20)u_1(02)u_2(20) + \lambda(02)u_1(00)u_2(22) + \lambda(22)u_1(02)u_2(22)] = \tfrac{16}{6}\\
\rho(01) &= \tfrac{20}{6} - \tfrac{1}{6}[\lambda(20)u_1(02)u_2(10) + \lambda(02)u_1(00)u_2(12) + \lambda(22)u_1(02)u_2(12)] = 4
\end{aligned}$$

Again from (86)

$\epsilon(F^{10}) = \epsilon(10) = 1,\ \epsilon(20) = \frac{1}{2},\ \epsilon(11) = 1,\ \epsilon(12) = 1,\ \epsilon(21) = 1,\ \epsilon(22) = \frac{5}{6},$ $\epsilon(02) = \frac{2}{3}, \epsilon(01) = 1$.

Remark 1 The partially balanced design with orthogonal sub-factorial structure derived here have some apparent similarity with the designs proposed in Das and Chatterjee (1999). The coefficient vectors of the contrasts belonging to the sub-factorial effects in Das and Chatterjee (1999) are obtained by introducing groups among the levels of the factors and within-group and between-group contrasts are considered there. The within-group contrasts defined there do not involve the effects of all the levels of the factors; they only involve the effects of the levels contained in the relevant groups. But here the contrasts are chosen differently involving the effects of all the levels. So here the effect contrasts give more insight into the explanation of the factorial effects. Also here, as the coefficient vectors are chosen from the values of orthogonal polynomials, so they are helpful in understanding the linear, quadratic, cubic etc. effects of the factorial effects. Moreover the derivations here are more elegant and the results come in more neat form as the contrasts are divided into different groups using the properties of the orthonormal vectors obtained from orthogonal polynomials.

Remark 2 The advantages of factorial designs with partial balance and orthogonal sub-factorial structure are discussed in Sect. 1 (Introduction). Also the works done in this direction are mentioned therein (viz. Das and Chatterjee 1999 and Das 2003). In Das and Chatterjee (1999) the treatment contrasts representing factorial effects are divided by introducing groups among the levels of the factors while the pencilwise division of the effect contrasts of Bose (1947) is used in Das (2003). There are scopes of other kind of divisions of the effect contrasts and introduction of partially balanced factorial designs accordingly. There is scope for unification also. In particular, it may be interesting to investigate if other PBIB association matrices may be used to construct such designs in the binary, proper and equireplicate case.

References

Bose, R. C. (1947). Mathematical theory of symmetrical factorial design. *Sankhyā*, **8**, 107–166.

Das, P. (2003). On designs with pencilwise orthogonality and balance. *Electronic notes in Discrete Mathematics*, *15*, 72–74.

Das, P., & Chatterjee, B. (1999). On partial balance with orthogonal sub-factorial structure. *Calcutta Statistical Association Bulletin*, *49*, 195–196.

Fisher, R. A., & Yates, F. (1943). *Statistical tables for biological, agricultural and medical research*. Edinburg: Oliver and Boyd.

Gupta, S. (1983). A basic lemma and the analysis of of block and kronecker product designs. *Journal of Statistical Planning and Inference*, *7*, 407–416.

Gupta, S. (1988). The association matrices of extended group divisible scheme. *Journal of Statistical Planning and Inference*, *20*, 115–120.

Gupta, S., & Mukerjee, R. (1989). *A calculus of factorial arrangements* (Vol. 59). Lecture notes in statistics. Springer.

Hinkelman, K., & Kempthorne, O. (1963). Two classes of group divisible partial diallel crosses. *Biometrika*, *50*, 281–291.

Kshirsagar, A. M. (1966). Balanced factorial designs. *Journal of the Royal Statistical Society. Series B*, **28**, 559–569.

Kurkjian, B., & Zelen, M. (1963). Applications of the calculus of factorial arrangements I. Block and direct product designs. *Biometrika*, *50*, 63–73.

Nair, K. R., & Rao, C. R. (1948). Confounding in asymmetrical factorial experiments. *Journal of the Royal Statistical Society. Series B*, *10*, 109–131.

Paik, U. B., & Federer, W. T. (1973). Partially balanced designs and properties A and B. *Communications in Statistics: Theory and Methods*, *1*, 331–350.

Raghavarao, D. (1971). *Construction and combinatorial problems in design of experiments*. Wiley.

Rao, C. R. (1974). *Linear statistical inference and its applications* (2nd ed.). New Delhi: Wiley Eastern private limited.

Shah, B. V. (1958). On balancing in factorial experiments. *Annals of Mathematical Statistics*, *29*, 766–779.

Shah, B. V. (1960). Balanced factorial experiments. *Annals of Mathematical Statistics*, *31*, 502–514.

Beyond the Bayes Factor, A New Bayesian Paradigm for Handling Hypothesis Testing

Christian P. Robert

Abstract This note is a discussion on the perceived shortcomings of the classical Bayesian approach to testing, and an alternative approach advanced by Kamary et al. (Testing hypotheses via a mixture estimation model, 2014) as a solution to this quintessential inference problem.

Keywords Testing of hypotheses · Bayesian inference · Bayes factor Evidence · Decision theory · Consistency · Mixtures of distributions

1 Introduction

Testing hypotheses is undoubtedly a central issue for Statistics and in particular for Bayesian analysis. From the early days of the discipline (Stigler 1986), there has been proposals and divisions as how to conduct the evaluation of hypotheses and the subsequent decisions, including withing the Bayesian framework. One cannot consider that the current state of the field has reached a stationary stage, even though the default Bayesian solution remains the Bayes factor, as exemplified by the applied literature. As discussed below, this solution is however constrained by its adherence to an artificial decision framework, as set by J. Neyman and E. Pearson in the 1930s. We argue in Kamary et al. (2014) that time is ripe for a paradigm shift about testing and that an alternative based on a mixture encompassing model can be defended as a generic solution to this quintessential inference problem.

Christian P. Robert, Université Paris-Dauphine, PSL Research University, UMR CNRS 7534, and Department of Statistics, University of Warwick xian@ceremade.dauphine.fr. Research partly supported by a Institut Universitaire de France senior chair 2016–2021.

C. P. Robert (✉)
Université Paris-Dauphine, PSL, CEREMADE, Paris, France
e-mail: robert@ensae.fr

C. P. Robert
University of Warwick, Coventry, England

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_12

2 Limitations of the Bayes Factor

The Bayes factor is indeed at the core of the standard approach to Bayesian testing and model selection. (Although one may consider both problems as fundamentally different, we will treat them here as one single problem, based on the argument that the selection of one of two possible choices must be followed with further inference on the selected side, which thus operates as a new model.) As proposed and defended by Jeffreys (1939), this procedure is defined as the ratio of integrated likelihoods. Given two models

$$\mathfrak{M}_1 : x \sim f_1(x|\theta_1)\,,\ \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x \sim f_2(x|\theta_2)\,,\ \theta_2 \in \Theta_2\,,$$

to be compared, with respective priors

$$\theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

the respective marginal likelihoods

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\,\pi_1(\theta_1)\,\mathrm{d}\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\,\pi_1(\theta_2)\,\mathrm{d}\theta_2$$

are compared through the Bayes factor

$$\mathfrak{B}_{12} = \frac{m_1(x)}{m_2(x)}\,.$$

Since this construction is equivalent to the derivation of the posterior probability that the model is $\mathfrak{M}_1$ or $\mathfrak{M}_2$, there is nothing to criticise at this stage. However, the use of the Bayes factor as a *decision tool* is more debatable. If one follows the Neyman-Pearson formalism, the Bayes factor is to be compared to a bound that summarises both the prior weights put on both models and the penalties put upon the wrong choice of each model. Depending on those quantities, the bound can be anything between zero and infinity. A major difficulty with this formalism is that the bound is impossible to determine in all practical settings, hereby making the formalism impossible to implement. If instead one follows Jeffreys (1939) and proceeds from his scale on the strength of evidence, taking a value of one as the boundary between both models, this bound amounts to specific choices in the first perspective, while lacking in calibration, the scales proposed by Jeffreys being qualitative and failing to provide the uncertainty behind the decision of chosing one model versus the other.[1]

Furthermore, exploiting the Bayes factor $\mathfrak{B}_{12}$ as a qualitative measure of how strongly the data support one model versus the other often has the undesirable effect

[1] This criticism is actually much more about the approach to aim for a model selection free of a loss function, than about the Bayes factor per se, which, once more, appears as a natural transform of the posterior probabilities of the model.

of turning it into a p-value. Indeed, it is quite simple to turn $\mathfrak{B}_{12}$ into a probability

$$\mathfrak{p}_{12} = \frac{\mathfrak{B}_{12}}{1 + \mathfrak{B}_{12}}$$

by assuming the "natural" division of prior weights, ($^1/_2$, $^1/_2$), and to fail to calibrate this probability $\mathfrak{p}_{12}$ by considering it naturally scales against the (0, 1) interval, just as a classical p-value would do. As discussed by Fraser (2011), there is no (mathematical) reason to treat this probability as producing a frequentist degree of confidence, at least for a finite sample size (Lindley 1961; Welch and Peers 1963), and there is no Bayesian equivalent to the frequentist property that the p-value should be uniformly distributed under the null hypothesis (in the simplest cases).[2] Among other points raised in Robert (2016), let me point out two more issues: one is the long-lasting impact of the prior density on the numerical value of the marginal likelihood, meaning the choice of the prior distribution on the parameter of a given model determines this numerical value by its tail behaviour[3] and another is the lack of mathematical justification in using improper priors (DeGroot 1982) since their lack of normalisation renders them inappropriate in most testing situations, leading to many alternative if ad hoc solutions, where data is either used twice or split in learning-testing partitions that are not altogether consistent.

3 Estimating a Mixture Model

An alternative to testing via Bayes factors has been proposed in Kamary et al. (2014), namely by a reformulation of both the problem and its resolution into a framework that accounts for uncertainty and returns a posterior distribution instead of a single number (like the Bayes factor) or a decision. As shown in Kamary et al. (2014), this new approach offers a convergent and naturally interpretable solution to the testing problem.

The core idea to this alternative approach is to work through a representation of the testing problem as a two-component mixture estimation problem, when the mixture weights are formally equal to 0 or 1. This encompassing model can then be estimated as any other mixture model. The motivation for this approach follows from the consistency results of Rousseau and Mengersen (2011) on overfitted mixtures, that is, mixture models where the data is generated from a mixture distribution with a *lesser* components.

Hence, given two models under comparison,

$$\mathfrak{M}_0 : x \sim f_0(x|\theta_0)\,,\ \theta_0 \in \Theta_0 \quad \text{and} \quad \mathfrak{M}_1 : x \sim f_1(x|\theta_1)\,,\ \theta_1 \in \Theta_0\,,$$

[2] The need for recalibrating Bayesian posterior probabilities towards a relative scale is actually overlooked in the literature.

[3] This difficulty is actually compatible with the consistency property of the Bayes factor.

an encompassing mixture model is

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_0(x|\theta_0) + (1-\alpha) f_1(x|\theta_1)\,, \tag{1}$$

where the mixture weight $0 \leq \alpha \leq 1$ is introduced in addition to the original parameters of the model. This means that inference proceeds as if each term in the original iid sample behind the test or the model comparison is considered as being generated from $\mathfrak{M}_\alpha$. While this artificial model encompasses both $\mathfrak{M}_0$ and $\mathfrak{M}_1$ as two special cases, namely when $\alpha = 0$ and $\alpha = 1$, respectively, a standard Bayesian analysis of this mixture model leads to an estimate of the weight α, relying on an equally artificial prior distribution $\pi(\alpha)$ with support the entire (0, 1) interval. For simplicity reasons, we use a Beta $\mathcal{B}e(a_0, a_0)$ distribution.

As a result of this modelling, a Bayesian processing of the model returns a posterior distribution on the weight α as well as on the other parameters of the mixture (1). The proposal is to use the posterior on α as the basis for deciding (and calibrating the evidence) in favour of one model versus the other. For instance, when the mass of this posterior is primarily concentrated near zero, the data supports more strongly $\mathfrak{M}_1$ than $\mathfrak{M}_0$. Clearly, this alternative paradigm no longer returns a value in the binary set $\{0, 1\}$ as a more traditional testing strategy would do. Instead, the decision or the evidence need be based on the entire distribution. Therefore, our mixture representation moves away from making a hard choice between both models (or hypotheses) or even from computing a posterior probability of $\mathfrak{M}_0$ or $\mathfrak{M}_1$. Inference within this mixture representation thus bypasses the difficulties with the original Neyman-Pearson framework, as it prevents from incorporating the decision within the statistical analysis. It ends up as a more genuine approach to testing, while not expanding on the total number of parameters of the model. Further arguments can be found in Kamary et al. (2014), including consistency.

From a practical perspective–in the sense of the approach being used for solving real life problems—, the implementation of this principle does not induce major computational difficulties since mixtures are rather straightforward to estimate (Lee et al. 2009). The major shift stands with analysing the posterior distribution on the weight α. Shying away from p-values and equivalents, we advocate calibrating the concentration of this distribution near the boundaries, 0 and 1, in absolute terms if this is possible, but also relative to the corresponding concentration of a similar posterior distribution associated with pseudo-data simulated from each model.

4 Conclusion

While the above is but an introduction to this perspective on testing, we are currently working on extensions to non-iid data, multiple testing, and non-parametric alternatives. Early criticisms have foscussed on the slow convergence of the posterior distribution (and of the posterior median) of α to one of the boundaries, as well as the dependence of the outcome on the prior on this weight: Such criticisms only apply to a frequentist interpretation of probabilities, i.e., to a comparison of the posterior distribution with a Uniform distribution and to a Uniform scaling of probabilities on the unit interval. We argue that, instead, the dependence on the prior is a natural feature of Bayesian analysis, which means that the impact of the data on the distribution of the mixture weight need be assessed relative to the prior distribution and calibrated against pre-posteriors, that is, posteriors derived from pseudo-samples generated by the prior predictive for the mixture and for each model under comparison. We genuinely think that moving away from the rudimentary binary decision framework can only contribute to make testing better informed and in fine help to overcome the current testing crisis (Wasserstein and Lazar 2016).

Acknowledgements I am quite grateful to the Department of Statistics of the University of Calcutta both for its kind invitation at the 75th (Platinum) anniversary of its creation and for its warm welcome.

References

DeGroot, M. (1982). Discussion of Shafer's 'Lindley's paradox'. *Journal of the American Statistical Association*, *378*, 337–339.

Fraser, D. (2011). Is Bayes posterior just quick and dirty confidence? *Statistical Science*, *26*(3), 299–316. (With discussion).

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford: The Clarendon Press.

Kamary, K., Mengersen, K., Robert, C., & Rousseau, J. (2014). Testing hypotheses as a mixture estimation model. arXiv:1412.2044.

Lee, K., Marin, J.-M., Mengersen, K., & Robert, C. (2009). Bayesian inference on mixtures of distributions. In N. N. Sastry, M. Delampady, & B. Rajeev (Eds.), *Perspectives in mathematical sciences I: Probability and statistics* (pp. 165–202). Singapore: World Scientific.

Lindley, D. (1961). The use of prior probability distributions in statistical inference and decision. In *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 453–468). University of California Press.

Robert, C. (2016). The demise of the Bayes factor. *Journal of Mathematical Psychology*, *72*, 3–37.

Rousseau, J., & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, *73*(5), 689–710.

Stigler, S. (1986). *The history of statistics*. Cambridge: Belknap.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Welch, B., & Peers, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B*, *25*, 318–329.

No Calculation When Observation Can Be Made

Tommy Wright

Abstract For a long time, when data were needed by a nation, a census was often carried out in an attempt to make measurements on every household or every person in that nation. The move from conducting a census to conducting a sample survey where only a subset of households or persons was measured was met with opposition. Sampling requires that weights be calculated and applied to the observations in the sample, while observations of all units in a census do not require such calculation. Wanting to preserve censuses, the opposition to sampling was best expressed in the translated firm phrase, "No calculation when observation can be made". With advances in probability sampling theory and applications, this resistance movement eventually failed. Societies have become more complex and people more mobile; data collection costs have risen; and the public's response rates to government sample surveys continue to fall, though not as sharply as with non-government sample surveys. Other sources of data (e.g., administrative records, big data, commercial data) to measure human behavior and condition are being investigated for increased use in production of official statistics. Though there are technical concerns (e.g., representativeness, data quality, privacy), other features of big data (e.g., relatively inexpensive; largely digital measurement; minimize respondent burden; lot of data; rich, complex, diverse, flexible) and the growing demand for data gathering and managing tools suggest a very useful source of data. However, with big data and the implicit promise of observation of everything, there is a risk that users may move to question the need for sampling or statistical calculation. Indeed, the new phrase might be "No statistical calculation when big data observation can be made", a phrase lacking scientific merit. In this paper, we take a brief look at the move from censuses to samples. Specifically, we (i) highlight Kiaer (1895, 1897) nonrandom *representative method* which laid seeds for the current use of survey sampling methodology to

Tommy Wright is Chief of the Center for Statistical Research and Methodology, U. S. Census Bureau, Washington, D.C. 20233 (E-mail: tommy.wright@census.gov) and adjunct faculty in statistics at Georgetown University. The views expressed are the author's and not necessarily those of the U. S. Census Bureau.

T. Wright (✉)
U.S. Bureau of the Census, Washington, DC, USA
e-mail: tommy.wright@census.gov

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_13

make measurements for official statistics; (ii) highlight Neyman's 1934 contribution with stratified random sampling and optimal sample allocation to achieve representativeness, (iii) note some recent developments (Wright 2012, 2014, 2016, 2017) to improve Neyman's optimal allocation, and (iv) offer some personal observations looking forward.

Keywords Censuses · Sample surveys · Alternative data sources
Official statistics

1 Introduction

For use in connection with the general and complete information that would be known from a full census, Kiaer (1895, 1897) presents a nonrandom purposive "representative method" for sampling from a finite population to provide "...more penetrating, more detailed, and more specialized surveys..." Kiaer did not use probability in his selection.

Many view Kiaer's method as laying seeds for the use of current sampling methods in producing official social and economic statistics. Kiaer attempted to select a sample that had distributions on some variables that were similar to the distributions observed from a census for the same variables. There seems to be evidence that he calculated weights to help improve the sample's representativeness (Kiaer 1897). Kiaer had few early followers and much opposition, one of his strongest critics, saying (a translation), "...no calculation when observation can be made."

With clarity and elegant style, Neyman (1934) brought probability to this representative method using stratified random sampling. Among several noteworthy statistical contributions in his paper, Neyman presents details for an optimal allocation of the fixed sample size among the various strata to minimize sampling variance. However, there are concerns with Neyman's result: (i) it does not give integer solutions; (ii) rounding does not always give minimum variance, and (iii) it permits a sample size in a stratum that exceeds the stratum size, which is not feasible. The conservative approach of always rounding up can be costly, especially when there are many strata as often occurs with samples of businesses.

Wright (2012) improves on Neyman's result with a simple derivation obtaining exact results that always yield integer sample size allocations while minimizing sampling variance. Wright (2014, 2016, 2017) generalizes and extends his results when there are mixed constraints on sample sizes for each stratum, desired precision constraints, and cost constraints.

In Sect. 2, we discuss and illustrate Kiaer's method. In Sect. 3, we highlight Neyman's result for optimal sample allocation and note some of its defficiencies. In Sect. 4, we note some exact optimal sample allocation results of Wright. In Sects. 5 and 6, we conclude by making personal observations and comments about some current challenges in data collection and by calling on the translated phrase "...no calculation when observation can be made" to muse about current world-wide considerations of using data from alternative non-probability sources (e.g., big data) to produce official statistics.

2 Kiaer's Contribution: Representative Method

Though there are earlier examples of the selection of samples to produce official statistics (e.g., in 1802, Laplace used sampling in combination with administrative records on the number of births to estimate the number of persons in France for Napoleon (Cochran 1997)), many trace the seeds of current survey sampling methodology to Anders Nicolai Kiaer (1838–1919) who served as the first Director of the Norwegian Central Bureau of Statistics (1877–1913). Bellhouse (1988) says, "Kiaer's contribution was to provide a framework under which sampling became a reasonable activity...". Prior to Kiaer, most government statistics were produced by censuses. At the Berne International Statistical Institute Meeting (1895), Kiaer argues that a partial investigation (i.e., a sample) could provide useful information based on what he called the "representative method", which aimed to produce a sample which was a "miniature of the population". Characteristics of Kiaer's representative method include:

(i) Conduct in connection with a census.
(ii) Obtain more penetrating, more detailed and specialized data which would be impractical to collect from all units.
(iii) Spread the sample out or distribute it over the entire population.

To actually implement his representative method for a sample in the context of a census, Kiaer would (Bellhouse 1988), without use of probability or randomization, (1) select the sample at various stages, e.g., districts, then towns/cities, then parts of towns/cities, then streets, then houses, then families, and finally individuals; (2) select large sample sizes at each stage; and (3) spread the sample out so that distributions of variables for the sample would match the observed distributions of the same variables from the census. For example, if he needed more sailors in a certain geographical area for the sample, he would put more sailors in the sample from that area using the observed census results.

Kiaer (1897) discusses some details of selection of the sample for the survey on *Personal Income and Property* in Norway conducted in direct connection with the general census of population in 1891. In 128 selected rural local government districts and 23 selected towns and cities from 1891 Census forms, a sample of the male population was needed to study in more detail using variables not measured in the census. Kiaer believed that "...these districts have a sufficient geographic distribution [spread] over the whole country...[to] at least approximately provide a correct representation of the whole country." Those males to be included in the sample were those (1) who in 1890 reached the ages of 17, 22, 27, 32, 37, etc. at 5 year intervals, and (2) those whose surname started with certain letters (in rural areas and smaller towns, he selected males whose surnames start with A, B, L, M,

Table 1 Compares sample distribution by occupations with 1891 census distribution

	1891 Census	Representative sample
I. Rural districts		
Farmers	21.3	20.7
Sons of farmers employed on family farm	9.4	8.4
Fishermen	8.2	7.4
Servants	4.4	4.9
Other farm workers	5.2	4.4
Factory workers	4.3	5.6
Tenants	5.7	5.5
II. Urban districts		
Craftsmen	7.0	6.8
Workers employed in craft	16.8	18.1
Factory workers	11.5	13.1
Sailors, deck-hands, etc.	8.4	7.6
Ship's officers	4.8	4.6
Businessmen, shipowners, factory owners	6.6	7.0
Employees in trade and commerce	5.9	5.4
Public employees and civil servants	4.6	4.8

or N while in the nine largest towns, he selected males whose surnames start with L, M, or N).

The total number of representative forms for males in the sample was:

From 128 Rural Districts and Small Towns	7, 164
From 23 Towns and Cities	4, 262
TOTAL	11, 426

The resulting sample contained 1.54% and 3.1% of total males population in all rural and urban districts, respectively. So Kiaer prepared tables for the whole country by calculating double weights for rural districts. In Kiaer's translated words, "In preparing the tables for the whole country the figures for the rural districts have been given double weight" (Kiaer 1897). The calculation of these weights for the sample are not needed with a census.

To check the representative nature of his sample, some examples of comparing sample distributions with observed census distributions are given in Tables 1, 2 and 3 (Kiaer 1897).

Kiaer (1897) states, "...the crux of the problem, namely whether we are justified in trusting the accuracy of the results from representative surveys in cases where no

Table 2 Compares distributions by Age (15+ Years) and marital status of males (Towns)

	Males in Towns [137,589 Males in 1891 Census]					Males in Towns [4262 Males in Sample]			
Age	Unmarried	Married	Widowers	Total	Age	Unmarried	Married	Widowers	Total
15–20	15.6	–	–	15.7	17	15.5	0.0	–	15.5
20–25	11.5	1.6	–	13.1	22	11.6	1.6	–	13.2
25–30	6.1	5.6	–	11.7	27	5.9	6.8	–	12.8
30–35	2.9	7.9	0.2	11.0	32	3.0	7.9	0.2	11.1
35–40	1.5	8.1	0.3	9.9	37	1.34	8.2	0.2	9.8
40–50	1.7	13.3	0.8	15.8	42,47	1.74	13.1	0.6	15.5
50–60	1.0	9.0	1.1	11.1	52,57	1.2	8.55	0.8	10.6
60+	0.9	7.8	3.0	11.7	62+	0.9	8.2	2.3	11.5
Total	41.2	53.3	5.5	100.0	Total	41.2	54.35	4.2	100.0

Table 3 Compares distributions by age (15+ Years) and marital status of males (Rural Districts)

	Males in rural districts [459,267 Males in 1891 Census]					Males in rural districts [7164 Males in Sample]			
Age	Un-married	Married	Widowers	Total	Age	Unmarried	Married	Widowers	Total
15–20	15.3	–	–	15.4	17	15.6	0.1	–	15.8
20–25	9.8	1.2	–	11.0	22	10.1	1.4	–	11.5
25–30	5.45	4.05	0.1	9.6	27	6.3	4.16	0.1	10.6
30–35	2.74	6.1	0.15	9.0	32	2.74	6.7	0.3	9.75
35–40	1.6	6.7	0.23	8.6	37	1.3	6.96	0.2	8.5
40–50	1.85	12.3	0.62	14.8	42,47	1.6	12.5	0.6	14.8
50–60	1.3	10.3	1.0	12.6	52,57	1.1	10.5	1.0	12.6
60+	1.4	13.0	4.4	19.0	62+	1.1	11.5	4.1	16.6
Total	39.44	53.65	6.5	100.0	Total	39.84	53.72	6.3	100.15

other information is available, when it is evident that the sample results agree at points where they might be compared with those from complete counts... In my opinion, there is a high degree of likelihood that information derived from representative returns has the same degree of accuracy when it comes to new fields for which the return provide information, as the fields where sample results can be checked in the manner mentioned...”.

Kiaer received *great opposition*, the strongest from von Mayr (1895, Bavarian statistician) saying that partial surveys were of limited value to only the area studied; that partial surveys could never replace *complete* statistical surveys; and “Il faut rester ferme et dire: pas de calcul là où l’observation peut être faite”. [*Translation*: “No calculation when observation can be made!”] The opposition was international (Bethlehem 2009): Bodio (Italian statistician) supported von Mayr’s criticism; Rauchberg (Austrian statistician) said that further discussion of the matter was unnecessary! Milliet (Swiss statistician) demanded that incomplete (sample) surveys should not be granted a status equal to censuses.

Kiaer continued to argue for his “representative method”. Bellhouse (1988) notes some developments: randomization is proposed for use in sample selection (1903);

Bowley presents a central limit theorem for random sampling (1906); Bowley uses systematically chosen sample of houses to study poverty in Reading, England (1912); and Bowley provides a theoretical monograph on random selection and purposive selection (Kiaer's representative method) in 1926. Henceforth, we use "purposive selection" to mean the same as "Kiaer's representative method". In addition to several other ideas, the monograph contains a development of stratified sampling with proportional allocation and a theoretical development of purposive selection through correlation between control variables and the variable of interest.

A noted failure of purposive selection by the Italians (1928–1929) created new reservations about Kiaer's method (Kruskal and Mosteller 1980). Corrado Gini and Luigi Galvani describe the selection of a sample from the 1921 Italian Census where the sample was "balanced" on seven important variables and made a purposive selection of 29 out of 214 administrative units in Italy. The resulting sample showed wide discrepancies with the census counts on other variables (Wright 2001).

Based on the work of a commission to study the application of the representative method, the 1925 International Statistical Institute's meeting in Rome adopts a resolution which gives acceptance to certain methods both by random selection and by purposive selection.

3 Neyman's Contribution: Optimal Allocation

Neyman (1934) discusses the desire to have a "representative method" when sampling from a finite population; specifically, he considers the "...two different aspects of the representative method...(1) the method of random sampling and (2) the method of purposive selection." Neyman argues in favor of stratified random sampling. Under stratified random sampling, a simple random sample is selected independently from each stratum.

In addition to the use of stratification, Neyman addressed the question of how to allocate fixed n optimally among the H strata. Neyman (1934) shows the allocation of fixed n that minimizes $Var(\hat{T}_Y)$ is

$$n_h = \left(\frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \right) n \qquad h = 1, 2, 3, ..., H, \tag{1}$$

subject to the constraint $n = \sum_{h=1}^{H} n_h$, where N_h is the number of units in stratum h, S_h is the standard deviation of the Y values of the N_h units in stratum h, n_h is the number of sample units from stratum h, $\bar{y}_h$ is the sample mean for stratum h, and $Var(\hat{T}_Y)$ is the design-based sampling variance of the estimator $\hat{T}_Y = \sum_{h=1}^{H} N_h \bar{y}_h$ of the population total for the $N = \sum_{h=1}^{H} N_h$ units. (*NOTE*: Tschuprow (1923) had obtained the result in (1) over a decade earlier.) Because the n_h in (1) are almost never positive integers, the n_h need to be rounded in just about every case.

4 Exact Optimal Sample Allocation

Wright (2012) gives an example to illustrate that rounding does not guarantee minimum sampling variance. Supporting what is known, Wright (2016) gives an example illustrating that the optimal n_h can sometimes exceed N_h. Is it possible to get an exact allocation of fixed n that minimizes $Var(\hat{T}_Y)$ where all n_h are positive integers and $n = \sum_{h=1}^{H} n_h$? In Algorithm I, Wright (2012) shows that the answer is yes.

Exact Optimal Allocation Algorithm I [$n_h \geq 1$] (Wright, 2012)

Step 1: First, assign one unit to be selected for the sample from each stratum.

Step 2: Assume $N_1S_1 \geq N_2S_2 \geq \cdots \geq N_HS_H$ and compute the array of *priority values*:

$$\begin{array}{lcccc}
\text{Stratum } 1 & \frac{N_1S_1}{\sqrt{1\cdot 2}} & \frac{N_1S_1}{\sqrt{2\cdot 3}} & \frac{N_1S_1}{\sqrt{3\cdot 4}} & \cdots \\
 & & \vdots & & \\
\text{Stratum } h & \frac{N_hS_h}{\sqrt{1\cdot 2}} & \frac{N_hS_h}{\sqrt{2\cdot 3}} & \frac{N_hS_h}{\sqrt{3\cdot 4}} & \cdots \\
 & & \vdots & & \\
\text{Stratum } H & \frac{N_HS_H}{\sqrt{1\cdot 2}} & \frac{N_HS_H}{\sqrt{2\cdot 3}} & \frac{N_HS_H}{\sqrt{3\cdot 4}} & \cdots
\end{array}$$

Step 3: Pick the $n - H$ largest priority values from the above array in Step 2 along with the associated strata. Each stratum is allocated an additional sample unit each time one of its priority values is among the $n - H$ largest values.

Subsequently, Wright (2016) has observed that the key to Algorithm I and many other algorithms is the observation that the sampling variance has a very simple decomposition which we give below for the case $H = 3$.

$$
\begin{aligned}
Var(\hat{T}_Y) = \sum_{h=1}^{3} N_h(N_h-1)S_h^2 \\
&- \frac{N_1^2S_1^2}{1\cdot 2} - \frac{N_1^2S_1^2}{2\cdot 3} - \frac{N_1^2S_1^2}{3\cdot 4} - \cdots - \frac{N_1^2S_1^2}{(n_1-1)(n_1)} \\
&- \frac{N_2^2S_2^2}{1\cdot 2} - \frac{N_2^2S_2^2}{2\cdot 3} - \frac{N_2^2S_2^2}{3\cdot 4} - \cdots - \frac{N_2^2S_2^2}{(n_2-1)(n_2)} \\
&- \frac{N_3^2S_3^2}{1\cdot 2} - \frac{N_3^2S_3^2}{2\cdot 3} - \frac{N_3^2S_3^2}{3\cdot 4} - \cdots - \frac{N_3^2S_3^2}{(n_3-1)(n_3)}
\end{aligned} \tag{2}
$$

where $\sum_{h=1}^{3} N_h(N_h-1)S_h^2$ is the sampling variance $Var(\hat{T}_Y)$ when $n_h = 1$ for each h. Each subtraction in Eq. (2) shows how much $\sum_{h=1}^{3} N_h(N_h-1)S_h^2$ decreases each time we increase the sample size in each stratum by one additional unit until we have $n = n_1 + n_2 + n_3$. The quantities being subtracted are the square of the priority values that appear in the array of Algorithm 1.

In addition to determining optimal n_h for fixed n, we may add additional constraints on n_h such as

$$a_h \le n_h \le b_h \tag{3}$$

where a_h and b_h are any reals such that $1 \le a_h \le b_h \le N_h$ to obtain an Algorithm 3 (Wright 2014, 2017). He also obtains an Algorithm 4 which gives optimal allocation when n is not given but precision requirements are. It is clear how one can obtain yet other algorithms by combining Algorithms 3 and 4 for optimal results (Wright 2016, 2017). Extensions when there are costs and a fixed budget are also possible.

So, what might the future hold for representativeness and official statistics?

5 Representativeness and Official Statistics

National statistical agencies have a long history of collecting and providing official statistics to inform those who make decisions as they govern. As behavior of units in populations change, there are unending questions about what the measurements should be and how they should be made. We mention six personal thoughts for consideration, in no particular order, which do not necessarily reflect views of the U. S. Bureau of the Census.

5.1 Pressure Mounting for Change in Data Collection Methods

The pressures on statistical agencies for change are real, and key reasons include: (1) increasingly complex society more challenging to measure with current methodology, (2) rising costs in data collection and limited budgets, and (3) growing concerns about privacy and confidentiality.

Relative to societal structures in the first part of the twenty-first century, societal structures in Kiaer's Norway at the last part of the nineteenth century were much less complex. Measurement of the 1890s Norwegian population was labor-intensive and 100% field work. Life in Norway involved: farming, fishing, carpentry, animal power, factories, sailing, handmade tools, etc. Life today is characterized by instant communications (smart phones, email, twitter, air travel), automobiles, driverless cars, instant information (google), "viv", the Internet of Things, increased life expectancy, etc. Measurement methods and what is being measured will change.

5.2 Rising Nonresponse (Unit/Item) to Censuses and Sample Surveys

Current methods of data collection permit increases in nonresponse to government sample surveys, though not nearly as high as in nongovernment efforts. Unfortunately, the data collection effort of government statistical agencies (e.g., in the United States) is an activity in which potential respondents are reluctant to engage. Given this reluctance and the fact that almost all government sample surveys are not mandatory, it is surprising that the response rates are as high as they are. Also, it is worth noting who is paid and who is not paid in the sample survey (also census) model. Table 4 makes it clear.

The most important participant in the data collection enterprise, the respondent, is the one not receiving monetary payment in most sample surveys. To be sure, this is by no means a call to pay respondents, but it is certainly worth noting. There is a need to let a nation's people know the value of such data in determining the outcomes of their daily lives.

5.3 Limited Future for Periodic Censuses Every 5 or 10 Years

The current periodic national censuses around the world will likely be replaced with continuous updating of listings of units in the target population or sampling frame. Availability of data from several sources (government and private) will help make this possible.

Table 4 Who is paid in the data collection enterprise?

Participant	Paid?
Planner	Yes
Developer/Maintainer of sampling frame	Yes
Designer of questionnaire/Data collection instrument	Yes
Designer/Selector of sample	Yes
Data collector	Yes
Data processor (Editing)	Yes
Data analyzer	Yes
Producer of data products	Yes
Reviewer of data products	Yes
Data disseminator	Yes
Respondent	No

5.4 Increasing Use of Administrative Records

Administrative records are records that governments maintain to help administer government programs. These programs can be extensive with varying objectives: improve employment; improve health; provide food against hunger; improve transportation; come from tax collection records; improve job training; improve education; improve living condition; decrease crime; improve communications; provide social security benefits; provide birth/death records; medical records, etc. Advantages of administrative records include that they are available, cheap, and can be used to improve statistical model construction. Well designed sample surveys can help calibrate administrative records to underlying truth. Disadvantages include quality variability, lack of uniformity (time period, definitions,...), lack of coverage of the entire population, representativeness doubt; challenges linking them, and the public's concerns about privacy and confidentiality.

5.5 From Censuses to Sample Surveys to Big Data

A nation measures its people's *behavior* and *condition*: Who they are. How they live. Where they live. What they do. What they produce. Big data can help provide rough indicators of behavior or condition (at lower levels of geography; for smaller subpopulations; more frequent data releases; often captured digitally,...), but there are limitations. Sources of big data include

- Google Searches
- Other Purchases (e.g., Groceries)
- Credit Reports
- Hospital/Medical Records
- Tax Records
- Housing Sales
- Facebook
- Cell Phone Usage
- Tweets
- Utility Records
- Medical Purchases
- Property Taxes
- Drivers' Licenses
- Monitoring by Cameras ...
- Credit Card Purchases
- Bank Accounts
- Travel Tickets
- Insurance Claims
- Migration Records
- Real Estate Listings

These data exist naturally, and in many cases, they are available as by-products of some primary activity other than data gathering.

Assume a well-defined target population about which we want to know (or estimate) some characteristic(s). At a very high level and building from Kish (1979), Table 5 compares data from three different sources [censuses, sample surveys, big data (could include administrative records)] relative to various criteria, where "*" indicates a potential advantage for a source over another. Clearly the contents of Table 5 are debatable, and one can cite examples to contradict every line of Table 5.

To view Table 5, we assume there exists a defined target population about which one wants to know the value of some population characteristic which is unknown. This is the primary objective we have in mind when using the various criteria in comparing results from a census, a sample, or big data.

Hence a carefully executed census with excellent coverage of the target population or a well-designed and controlled probability sample that permits statistical inference about the target population have the advantage in terms of the "representativeness" criterion over using results from big data in many applications where representation of the results is too often in doubt. On the other hand, the "high frequency" criterion gives the advantage to results from big data over results from either a census or a sample. Typically, the data collection and data processing associated with a census or sample permit release of official results that are often monthly, quarterly, or annual; in the case of a census, the release of results is typically every five or ten years. Depending on the specific application, releases of results from big data can be daily, or more frequent.

To illustrate our thoughts on the two criteria "representativeness" and "high frequency", we consider the unemployment rate of persons in the labor force of the United States. We only consider a sample compared to the use of big data. Each month, a national representative probability sample of approximately 70,000 households is contacted to determine the unemployment status of each person in the labor force. The official unemployment rate is released each month with a measure of uncertainty. That is, the sample permits valid statistical inferences. Alternately, investigators have released estimates of an unemployment rate based on counting the number of google searches of certain expressions relative to the total number of google searches. The expressions in the google searches include: unemployment benefits, unemployment office, unemployment claim, unemployment compensation, unemployment insurance, apply for unemployment, applying for unemployment, filing for unemployment, unemployment online, unemployment office location, unemployment eligibility, uninsured benefits, and unemployment benefit. It is difficult

Table 5 Comparison of censuses, sample surveys, and big data

Criterion	Census	Sample	Big data
Representativeness	*	*	–
Measures of uncertainty	*	*	–
Rich, complex, diverse, flexible	–	*	*
High quality	–	*	–
Relatively inexpensive	–	*	*
Timely, seasonal	–	*	*
Inclusive (Large and Complete)	*	–	–
Credible, P.R.	*	–	–
Units asked to report data	*	*	–
Units not asked to report data	–	–	*
Explicit/Implicit consent obtained	*	*	–
Lot of data	–	–	*
High frequency	–	–	*
Supplementary/Complementary	–	*	*
Minimize respondent burden	–	*	*

to determine the quality of the google searches; how many are made by persons not in the labor force; how many persons make more than one search; is each search actually made by someone who is unemployed; etc. So we are unsure about the representativeness of the big data results. However, it is easy to see that the use of big data as described permits monthly, weekly, daily, hourly,... estimates of unemployment rate with relatively little effort. Hence the big data has the potential for high frequency releases because there is the potential for high frequency of data capture.

Rather than taking one source among the three (census, sample, big data), we see that all three can work together (Kish 1979; Capps and Wright 2013).

5.6 With "Data Science" and "Data Analytics", One Can Sense a Move Towards Measuring Everything, Everywhere, All the Time

With big data, one can also sense a growing potential for users to ignore concern about uncertainty, variability, and data properties (representativeness, how obtained, quality, etc.). Indeed, there is a risk for the birth of a movement that one need not worry about statistical structure in gathering data as long as there is a lot of data. It brings to mind the phrase

No Calculation When Observation Can Be Made, but now it might come in the form

No Statistical Calculation When Big Data Observation Can Be Made

which is a near quote once made by a senior science administrator several decades ago when referring to the analysis of massive data sets collected by astronomers.

Technology is bringing strong streams of billions of pieces of data. Three examples help make this point, where the focus is on managing existing data rather than collecting it.

Example 1 The Billion Prices Project (bpp.mit.edu/page/2/) is an academic initiative that uses prices collected (webscraping) from online retailers around the world on a daily basis to conduct economic research. Data sets (bpp.mit.edu/data.sets) provide prices with coverage into many countries around the world! For the United States, The Billion Prices Project frequently produces a visual comparison of its *daily* Online Price Index (based on scraped data from online retailers) with the *monthly* Consumer Price Index (based on a probability sample) produced by the U. S. Bureau of Labor Statistics.

Example 2 JP Morgan Chase and Co launched the JP Morgan Chase and Co Institute (https://www.jpmorganchase.com/corporate/institute/institute.htm) on May 21, 2015 as a global think tank that will deliver data, analyses, and expert insights designed to address global economic challenges. Data in the following report come from the credit and debit card transactions (over 12 billion) of the nearly 50 million JPMorgan Chase customers and provide profiles of local consumer commerce data for 15 cities in the United States over a 34 month period. The front cover of the report, *Profiles of Local Consumer Commerce (December 2015)* highlights that the contents are based on "Insights from 12 Billion Transactions"! Page 9 of the report presents a visual that compares a time series based on its retail sales data (JPMCI-LCC) with a time series based on retail sales (MRTS) produced by the U. S. Bureau of the Census (based on a probability sample). The front cover of another report, *How Falling Gas Prices Fuel the Consumer* highlights that its contents are based on "Evidence from 25 Million People"! Both reports are found on the Institute's website.

Example 3 Uber is launching the Uber Movement Project which will provide data based on its billions of rides, initially for four (e.g., Washington, D.C.) cities around the world. The public website will provide data showing the time it takes to travel between neighborhoods in various cities (https://www.wired.com/2017/01/uber-movement-traffic-data-tool).

In the first two examples, the huge size of the available data seems to minimize any need or desire by some to present statistical uncertainty measures and discussions which often come in technical appendices. Additionally and in both cases, the quality and credibility of the product from big data seems to also come by comparing with similar statistical products that are based on well-established and highly regarded government sample surveys. This practice seems similar to Kiaer's construction of samples that have distributions similar to the distributions from credible censuses.

Kiaer argued that results from his samples were valid because on some variables they matched census results and hence it is reasonable to assume they are valid for other variables of interest that are not available from the census. Thus, one may be led to assume big data results are valid because on some measures, they match results from well-established government sample surveys, not to mention the huge amount of data. How would one measure uncertainty in the big data results in the absence of the results from government sample surveys for comparison?

6 Concluding Remarks

We see evidence for probability sampling and big data being used together to improve representativeness and official statistics. For example, Capps and Wright (2013) call for use of big data to supply variables for modeling in small area estimation. Where one has a disadvantage, the other may offer an advantage. The greatest advantage of probability sampling is the greatest disadvantage of big data; representativeness. The greatest disadvantage of probability sampling is that of the nonresponse that results when questions are asked; there is no asking with big data because the measurements are mostly already available as a secondary consequence from respondents who almost unknowingly provide these measurements while receiving some other primary benefit.

The move from censuses to sample surveys was initially resisted with the phrase "No calculation when observation can be made". Big data offers new opportunities for those who collect and provide data. Technological advances will lead to more big data and some may be tempted by the amount of data to move from sample surveys to big data with the phrase "No statistical calculation when big data observation can be made". Such a move would lack scientific merit and should be resisted because we are limited in making statements about how good the results are and there seems to be no measures of uncertainty for these results. Though not perfect, survey sampling methodology, the current variant of Kiaer's representative method, is grounded in over a century of successful theoretical and practical development and successful application. In addition to having learned much about controlling and decreasing sampling error, much has been gained in understanding and compensating for nonsampling error over this time period, as well.

Probability theory helped bring good measures of uncertainty to Kiaer's representative method leading to probability sampling and current practice. If we measure everything, everywhere, all the time, as seems to be the ultimate goal with big data, will there be need for measures of uncertainty for the data products? The answer is yes.

Even if one could measure everything, everywhere, all the time, it would be impossible to digest it all (in detail) and some sort of data reduction would be required. That is, one would need to take a sample. As Stephan (1948) notes, "All scientific observation whether statistical or not is based on sampling." Thus there are two options:

Option 1: Measure everything, everywhere, all the time, and *then* select a sample from the measurements.

Option 2: Select a sample, and *then* measure everything for the sample units, everywhere for the sample units, and measure the sample units all the time.

Acknowledgements The author is grateful to a referee and his colleagues John Eltinge, Mary Mulry, Lauren Emanuel, and Michael Leibert for providing helpful comments on earlier drafts of this paper.

References

Bellhouse, D. R. (1988). A brief history of random sampling methods. In P. R. Krishnaiah, & C. R. Rao (Eds.) *Handbook of statistics* (Vol. 6, pp. 1–14).

Bethlehem, J. (2009). The rise of survey sampling. In *Discussion paper (09015)*. The Hague/Heerlen: Statistics Netherlands.

Capps, C., & Wright, T. (2013). Toward a vision: Official statistics and big data. In *AMSTAT news*, No. 434. American Statistical Association, Alexandria, Virginia, 9.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Cochran, W. G. (1978). Laplace's ratio estimator. In H. A. David (Ed.) *Contributions to survey sampling and applied statistics* (pp. 3–10). New York: Academic Press.

Kiaer, A. N. (1895). Observations et expériences concernant des dénombrements représentative. *Bulletin of the International Statistical Institute, IX, Book*, *2*, 176–183.

Kiaer, A. N. (1897). The representative method of statistical surveys (1976 English Translation of the Original Norwegian). In Central Bureau of Statistics of Norway, Oslo. [Kiaer, A. N. (1897). Sur les Mèthods Représentatives ou Typologiques Appliquées à la Statistique, *Bulletin of the International Statistical Institute, XI*, 180–189].

Kish, L. (1979). Samples and censuses. *International Statistical Review*, *47*, 99–109.

Kruskal, W., & Mosteller, F. (1980). Representative sampling IV: The history of the concept in statistics. 1895–1939. *International Statistical Review*, *48*(2), 169–195.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*, 558–606.

Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association* , *43*(241), 12–39.

Tschuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, *2*(461–493), 646–683.

Wright, T. (2001). Selected moments in the development of probability sampling: Theory and practice. In *Survey research methods section newsletter: American statistical association* (pp. 1–6).

Wright, T. (2012). The equivalence of Neyman optimum allocation for sampling and equal proportions for apportioning the U. S. House of representatives. *The American Statistician*, *66*(4), 217–224.

Wright, T. (2014). A simple method of exact optimal sample allocation under stratification with any mixed constraint patterns. In *Research report series (Statistics #2014–07)*. Washington, D.C.: Center for Statistical Research & Methodology, U. S. Bureau of the Census.

Wright, T. (2016). Two optimal exact sample allocation algorithms: Sampling variance decomposition is key. In *Research report series (Statistics #2016–03)*. Washington, D.C.: Center for Statistical Research & Methodology, U. S. Bureau of the Census.

Wright, T. (2017). Exact optimal sample allocation: More efficient than Neyman. *Statistics and Probability Letters*, *129*, 50–57.

Design Weighted Quadratic Inference Function Estimators of Superpopulation Parameters

Sumanta Adhya, Debanjan Bhattacharjee and Tathagata Banerjee

Abstract Using information from multiple surveys to produce better pooled estimators is an active research area in recent days. Multiple surveys from same target population is common in many socioeconomic and health surveys. Often all the surveys do not contain same set of variables. Here we consider a standard situation where responses are known for all the samples from multiple surveys but the same set of covariates (or auxiliary variables) is not observed in all the samples. Moreover, in our case we consider a finite population set up where samples are drawn from multiple finite populations using same or different probability sampling designs. Here the problem is to estimate the parameters (or superpopulation parameters) of underlying regression model. We propose quadratic inference function estimator by combining information related to the underlying model from different samples through design weighted estimating functions (or score functions). We did a small simulation study for comprehensive understanding of our approach.

Keywords Model-design based approach · Multiple surveys · Superpopulation
Quadratic inference function

1 Introduction

Drawing inference on super population parameters by combining data from different surveys is of considerable recent interest (Citro 2014; Kim and Rao 2012; Gelman et al. 1998) to the survey practitioners. For an up to date and comprehensive review of the methods, we refer to Lohr and Raghunathan (2016). The central idea behind any

S. Adhya (✉)
Department of Statistics, West Bengal State University, West Bengal, Barasat, India
e-mail: sumanta.adhya@gmail.com

D. Bhattacharjee
Department of Mathematics, Utah Valley University, Orem, UT, USA

T. Banerjee
Indian Institute of Management, Gujarat, Ahmedabad, India

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_14

such method is to use information from different sources effectively for enhancing the efficiency of the estimators. In this paper, we propose a method for combining data based on quadratic inference function (QIF) (Lindsay and Qu 2003) in the context of linear regression analysis. To the best of our knowledge, use of QIF has not been considered before in the survey sampling literature.

For the methodological development in this paper, we consider model-design-based randomization approach to inference discussed in Roberts and Binder (2009), Graubard and Korn (2002), and Godambe and Thompson (1986). Specifically, we consider two finite populations $\mathcal{P}_1 = \{(y_i, x_{1i}, x_{2i}) : i \in U_1\}$ and $\mathcal{P}_2 = \{(y_i, x_{1i}) : i \in U_2\}$ of sizes N_1 and N_2, respectively, where U_1 and U_2 are index sets of the population units in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Notice that $\mathcal{P}_1$ and $\mathcal{P}_2$ can be considered as random samples from a superpopulation. We assume:

(i) The study variables in each finite population are independent realizations of the random variables (y, x_1, x_2), where x_1 and x_2 are exogenous, and y is a continuous endogenous variable. Also, given x_1 and x_2, y is generated by a linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where ϵ is the error term independent of x_1 and x_2, and has mean 0 and variance σ^2. However, in $\mathcal{P}_2$ observations on x_2 are missing.

(ii) A probability sample is selected from each resulting finite population using either the same or different sampling designs.

The above theoretical set-up may represent an important practical situation that often arises in survey sampling. Suppose in a survey with a relatively small sample size, the data are collected on a comprehensive set of exogenous variables; whereas in a different survey from the same population with a considerably larger sample size, the data are collected on a smaller subset of the same set of exogenous variables. The problem is to combine these independent samples effectively to get a better estimator.

Clearly, the problem stated above may be considered as a missing data problem where for some units in the bigger sample the data on one or more exogenous variables are missing. Multiple imputation is an often used method (Rendall et al. 2013; Gelman et al. 1998; Rubin 1986) in such situation, but how does it tide over the omitted variable bias is not quite clear. On the other contrary, the QIF based methodology that we propose here, recognizes and takes into account the omitted variable bias explicitly. Although the proposed methodology is applicable for combining data from any number of surveys in the set-up described above, we restrict our discussion to two surveys simply for ease of exposition.

The paper is organized as follows. In Sect. 2, we briefly discuss the estimation methodology based on QIF in a general setting, keeping in view the context of our application. In Sect. 3, we propose design-weighted QIF estimators of the regression coefficients using data from multiple surveys. Our methodology explicitly takes into account the omitted variable bias. In Sect. 4, we report the results of a limited simulation study. As expected, the simulation results show that the design-weighted QIF estimators based on the combined sample are substantially more efficient than the standard least squares estimators based on the sample with more covariates. Concluding remarks are given in Sect. 5.

2 Quadratic Inference Function

In this section we briefly introduce QIF based estimation methodology in a general setting. Suppose $\mathbf{b}(x, \boldsymbol{\theta}) = (b_1(x, \boldsymbol{\theta}), b_2(x, \boldsymbol{\theta}), ..., b_q(x, \boldsymbol{\theta}))^T$ is a q-dimensional vector of distinct score functions, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ is a p-dimensional vector of parameters. The score functions are also called estimating functions and moment conditions in statistics and economics literature, respectively. Application of QIF based estimation methodology makes sense only if q is greater than p.

Suppose $\mathcal{F}_\theta$ is the semi- parametric model defined by the parameter $\boldsymbol{\theta}$ and the score equations

$$E_F \mathbf{b}(X, \boldsymbol{\theta}) = 0, \tag{1}$$

such that if a distribution $F \in \mathcal{F}_\theta$, then (1) is satisfied and vice versa. On the other hand, if the true $F \notin \mathcal{F}_\theta$, and $E_F \mathbf{b}(X, \boldsymbol{\theta}) = \delta(\boldsymbol{\theta}) \neq 0$, where $\delta(\boldsymbol{\theta})$ is said to represent the vector of discrepancy between the model and the true distribution F.

The quadratic distance function (QDF) between the true distribution F and the semi-parametric model $\mathcal{F}_\theta$ as determined through the basic scores is then defined as

$$d(F, \mathcal{F}_{\boldsymbol{\theta}}) = \delta(\boldsymbol{\theta})^T \Sigma_{\boldsymbol{\theta}}^{-1} \delta(\boldsymbol{\theta}), \tag{2}$$

where $\Sigma_{\boldsymbol{\theta}} = Var(\mathbf{b}(X, \boldsymbol{\theta}))$. For an arbitrary F, the value of $\boldsymbol{\theta}$ for which the basic scores are closest to mean 0 is then given by

$$\boldsymbol{\theta}(F) = \arg min_{\boldsymbol{\theta}}\, d(F, \mathcal{F}_{\boldsymbol{\theta}}). \tag{3}$$

For making data based inference on $\boldsymbol{\theta}$, the QDF in (3) needs to be replaced by its empirical analogue, called quadratic inference function. Suppose $X_1, X_2, ..., X_n$ are independently and identically distributed random variables following the distribution F, then a natural estimator of $E_F \mathbf{b}(X, \boldsymbol{\theta}) = \delta(\boldsymbol{\theta})$ is $\bar{\mathbf{b}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n b(X_i, \boldsymbol{\theta})$. Suppose further, $\hat{\Sigma}$ is a suitably chosen estimator of $Var(\bar{\mathbf{b}}(\boldsymbol{\theta}))$, the QIF is then given by

$$Q(\boldsymbol{\theta}) = \bar{\mathbf{b}}(\boldsymbol{\theta})^T \hat{\Sigma}^{-1} \bar{\mathbf{b}}(\boldsymbol{\theta}). \tag{4}$$

The choice of $\hat{\Sigma}^{-1}$ is an important issue. We refer to Lindsay and Qu (2003) for a detailed discussion on it. The QIF estimator of is given by

$$\hat{\boldsymbol{\theta}} = \arg min_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}). \tag{5}$$

If $F \in \mathcal{F}_\theta$, $\hat{\boldsymbol{\theta}}$ is consistent for the true value of $\boldsymbol{\theta}$, otherwise it is consistent for the nonparametric functional $\boldsymbol{\theta}(F)$ (cf.(3)). For a discussion on the optimum properties of $\hat{\boldsymbol{\theta}}$, we refer to Lindsay and Qu (2003).

3 Design-Weighted QIF Estimator

Let us now consider the estimation of the regression parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ of the superpopulation model introduced in Sect. 1. First, we introduce some important notations. Suppose $\mathbf{S}_1 = \{(y_i, x_{i1}, x_{i2}) : i \in I_1 \subset U_1\}$ and $\mathbf{S}_2 = \{(y_i, x_{i1}, x_{i2}) : i \in I_2 \subset U_2\}$ represent the probability samples of sizes $n_1(< N_1)$ and $n_2(< N_2)$ drawn from the populations $\mathcal{P}_1$ and $\mathcal{P}_2$ using sampling designs $p_1(.)$ and $p_2(.)$, respectively, where I_1 and I_2 are index sets of selected sample units.

As stated at the outset, we adopt the model-design based randomization approach (Roberts and Binder 2009) to the estimation of the superpopulation parameters. Like Chen and Sitter (1999), we propose a two-step design weighted QIF estimator of $\boldsymbol{\beta}$ that could be used for complex surveys. First, we define QIF of $\boldsymbol{\beta}$, say, $Q_U(\boldsymbol{\beta})$, assuming $\mathcal{P}_1$ and $\mathcal{P}_2$ to be known. At the second step, we estimate $Q_U(\boldsymbol{\beta})$ by replacing the population based entities with its design-based estimators based on the samples. We denote it by $\widetilde{Q}_U(\boldsymbol{\beta})$. Finally, the estimator of $\boldsymbol{\beta}$ is obtained by minimizing $\widetilde{Q}_U(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. We now describe the two steps in detail.

Assuming $\mathcal{P}_1$ to be known, and represents a random sample from the superpopulation, the basic score vector for $\boldsymbol{\beta}$ is given by:

$$\mathbf{b}_1(y, \mathbf{x}, \boldsymbol{\beta}) = (Y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)\mathbf{x}, \tag{6}$$

where $\mathbf{x} = (1, x_1, x_2)^T$. Also, the assumed regression model of y given x_1 and x_2 entails $E_{\boldsymbol{\beta}}\mathbf{b}_1(Y, \mathbf{X}, \boldsymbol{\beta}) = 0$. However, for $\mathcal{P}_2$, the basic score function for $\boldsymbol{\beta}^{(1)} = (\beta_0, \beta_1)^T$is given by:

$$\mathbf{b}_2^*(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}^{(1)}) = (Y - \beta_0 - \beta_1 x_1)\mathbf{x}^{(1)}, \tag{7}$$

where $\boldsymbol{\beta}^{(1)} = (\beta_0, \beta_1)^T$ and $\mathbf{x}^{(1)} = (1, x_1)^T$. But omitted variable bias leads to $E_{\boldsymbol{\beta}}\mathbf{b}_2^*(Y, \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)}) = \boldsymbol{\delta}(\boldsymbol{\beta_2})$, where $\boldsymbol{\delta}(\boldsymbol{\beta_2}) = (0, \beta_2\sigma_{12})^T$, and $\sigma_{12} = Cov(x_1, x_2)$. Assuming σ_{12} to be known for the time being, we define a modified score function for $\boldsymbol{\beta}$ that explicitly takes into account the omitted variable bias as follows:

$$\mathbf{b}_2(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}) = (y - \beta_0 - \beta_1 x_1)\mathbf{x}^{(1)} - \boldsymbol{\delta}(\beta_2). \tag{8}$$

Thus, by definition, we have $E_{\boldsymbol{\beta}}\mathbf{b}_2(Y, \mathbf{X}^{(1)}, \boldsymbol{\beta}) = 0$. The population version of QIF are thus based on the basic score functions given by (6) and (8).

Let us define $\bar{\mathbf{b}}_1(\boldsymbol{\beta}) = N_1^{-1}\sum_{i\in U_1}\mathbf{b}_1(y_i, \mathbf{x}_i, \boldsymbol{\beta})$, $\bar{\mathbf{b}}_2(\boldsymbol{\beta}) = N_2^{-1}\sum_{i\in U_2}\mathbf{b}_2(y_i, \mathbf{x}_i^{(1)}, \boldsymbol{\beta})$, and $\bar{\mathbf{b}}(\boldsymbol{\beta}) = (\bar{\mathbf{b}}_1(\boldsymbol{\beta}), \bar{\mathbf{b}}_2(\boldsymbol{\beta}))^T$. Let $\hat{\Sigma}_{1\boldsymbol{\beta}}$, $\hat{\Sigma}_{2\boldsymbol{\beta}}$, and $\hat{\Sigma}_{\boldsymbol{\beta}}$ be suitable finite population based estimators of $Var(\mathbf{b}_1(Y, \mathbf{X}, \boldsymbol{\beta})) = \Sigma_{1\beta}$, $Var(\mathbf{b}_2(Y, \mathbf{X}^{(1)}, \boldsymbol{\beta})) = \Sigma_{2\beta}$ and $Var(\mathbf{b}(Y, \mathbf{X}, \boldsymbol{\beta})) = \Sigma_{\beta}$, respectively, where $\mathbf{b}(y, \mathbf{x}, \boldsymbol{\beta}) = (\mathbf{b}_1(y, \mathbf{x}, \boldsymbol{\beta}), \mathbf{b}_2(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}))^T$.

Then the first-step QIF of $\boldsymbol{\beta}$ is given by

$$Q_U(\boldsymbol{\beta}) = W_1\bar{\mathbf{b}}_1(\boldsymbol{\beta})^T\hat{\Sigma}_{1\boldsymbol{\beta}}^{-1}\bar{\mathbf{b}}_1(\boldsymbol{\beta}) + W_2\bar{\mathbf{b}}_2(\boldsymbol{\beta})^T\hat{\Sigma}_{2\boldsymbol{\beta}}^{-1}\bar{\mathbf{b}}_2(\boldsymbol{\beta}), \tag{9}$$

where, $W_k = N_k N^{-1}, k = 1, 2$, and $N = N_1 + N_2$.

Let us now define the second step QIF, $\widetilde{Q}_U(\boldsymbol{\beta})$, an estimator of $Q_U(\boldsymbol{\beta})$, based on the samples $\mathbf{S}_1$ and $\mathbf{S}_2$. Suppose $\pi_{ik} = P_k(i \in I_k | i \in U_k)(> 0)$ denotes the inclusion probability of the $i-th$ unit of the $k-$th population in the sample $\mathbf{S}_k$, where $P_k(.)$ is the probability measure corresponding to the sampling design $p_k(.)$ for $i = 1, 2, ..., N_k, k = 1, 2$. The design weights are then given by $d_{ik} = \frac{\pi_{ik}^{-1}}{\sum_{i \in S_k} \pi_{ik}^{-1}}$, for $i \in I_k, k = 1, 2$. Defining, $\widetilde{\mathbf{b}}_{i1}(\boldsymbol{\beta}) = \mathbf{b}_1(y_i, \mathbf{x}_i, \boldsymbol{\beta})$ for $i \in I_1$, $\widetilde{\mathbf{b}}_{i2}(\boldsymbol{\beta}) = \mathbf{b}_1(y_i, \mathbf{x}_i^{(1)}, \boldsymbol{\beta})$ for $i \in I_2$, $\widetilde{\mathbf{b}}_1(\boldsymbol{\beta}) = \sum_{i \in I_1} d_{i1} \widetilde{\mathbf{b}}_{i1}(\boldsymbol{\beta})$, $\widetilde{\mathbf{b}}_2(\boldsymbol{\beta}) = \sum_{i \in I_2} d_{i2} \widetilde{\mathbf{b}}_{i2}(\boldsymbol{\beta})$, and $\widetilde{\Sigma}_{k\boldsymbol{\beta}} = \sum_{i \in I_k} d_{ik} (\widetilde{\mathbf{b}}_{ik}(\boldsymbol{\beta}) - \widetilde{\mathbf{b}}_k(\boldsymbol{\beta}))(\widetilde{\mathbf{b}}_{ik}(\boldsymbol{\beta}) - \widetilde{\mathbf{b}}_k(\boldsymbol{\beta}))^T$ for $k = 1, 2$, we obtain

$$\widetilde{Q}_U(\boldsymbol{\beta}) = W_1 \widetilde{\mathbf{b}}_1(\boldsymbol{\beta})^T \widetilde{\Sigma}_{1\boldsymbol{\beta}}^{-1} \widetilde{\mathbf{b}}_1(\boldsymbol{\beta}) + W_2 \widetilde{\mathbf{b}}_2(\boldsymbol{\beta})^T \widetilde{\Sigma}_{2\boldsymbol{\beta}}^{-1} \widetilde{\mathbf{b}}_2(\boldsymbol{\beta}). \tag{10}$$

The design-weighted QIF estimator of $\boldsymbol{\beta}$ is then given by

$$\hat{\boldsymbol{\beta}} = \arg min_{\boldsymbol{\beta}}\, \widetilde{Q}(\boldsymbol{\beta}). \tag{11}$$

Notice that throughout the development we assume σ_{12} to be known. It may be a reasonable assumption if the information on x_1 and x_2 are available at the population level while the values of (y, x_1, x_2) are known for the sample only. In this case, the design-weighted QIF estimators lead to a huge improvement over the standard least squares estimators. In case, it is not known, we plug in its estimate from the sample in $\widetilde{Q}_U(\boldsymbol{\beta})$. The latter also shows some improvement as is evident from the numerical studies reported in the next section.

4 Numerical Studies

We present the results of a limited simulation study comparing the performances of design-weighted quadratic inference function estimator (QIFE) with that of design-weighted least square estimator (LSE).

Suppose the covariate vector $(x_1, x_2)^T$ has a bivariate normal distribution with mean vector $(0, 0)^T$ and covariance matrix $\boldsymbol{\Sigma}(2 \times 2)$. Given (x_1, x_2), y has a normal distribution with mean $1 + 0.5x_1 + 0.25x_2$ and variance 0.25. We consider two superpopulation models $M1$ and $M2$ corresponding to two choices of $\boldsymbol{\Sigma}$, say, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, where

$$\Sigma_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

and

$$\Sigma_1 = \begin{pmatrix} 0.5 & 0.14 \\ 0.14 & 1.0 \end{pmatrix}.$$

Notice that for model $M1$ the correlation coefficient between x_1 and x_2 is 0.7 while for $M2$, it is 0.2.

Following are the steps of the simulation study:

Step 1: We generate finite populations U_1 and U_2 of sizes N_1 and N_2 using the above superpopulation model. First, we randomly generate a value of $\mathbf{x} = (x_1, x_2)^T$, and then generate a value of y given $\mathbf{x}$ using the conditional distribution of y given $\mathbf{x}$. The finite populations U_1 and U_2 then comprise N_1 and N_2 such observations on (y, x_1, x_2) generated independently. Next, by simple random sampling without replacement (SRSWOR), we select L samples of sizes $n_1(= f_1 N_1)$ and $n_2(= f_2 N_2)$ from U_1 and U_2, respectively, where f_1 and f_2 are the sampling fractions. The selected samples from U_1 and U_2 are denoted by $\mathbf{S}_1^{(l)}$ and $\mathbf{S}_2^{(l)}$, $l = 1, 2, ..., L$ respectively.

Step 2: Based on $\mathbf{S}_1$ we compute usual design-weighted LSE of $\boldsymbol{\beta}$. Also based on $\mathbf{S}_1$ and $\mathbf{S}_2$, we compute design-weighted QIFE from (11).

Step 3: We repeat the Step 1 R times. At the r-th ($r = 1, 2, ..., R$) replication, let the populations generated be $U_1^{(r)}$ and $U_2^{(r)}$. For each r, the selected samples from $U_1^{(r)}$ and $U_2^{(r)}$ are denoted by $\mathbf{S}_1^{(rl)}$ and $\mathbf{S}_2^{(rl)}$, $l = 1, 2, ..., L$, respectively. For each r and l, following Step 2, we compute the LSE and QIFE of β_j, $j = 0, 1, 2$, say, $\hat{\beta}_{j(LS)}^{(rl)}$ and $\hat{\beta}_{j(QIF)}^{(rl)}$, respectively.

Step 4: For each estimator of β_j, say, $\hat{\beta}_j^{(rl)}$ (a generic notation) we compute the relative bias (RB) ($[(RL)^{-1} \sum_{r,l} \hat{\beta}_j^{(rl)} - \beta_j]/|\beta_j|$) and relative root mean squared error (RRMSE) ($\sqrt{(RL)^{-1} \sum_{r,l} (\hat{\beta}_j^{(rl)} - \beta_j)^2}/|\beta_j|$).

For our simulation study, we consider (N_1, N_2): (1000, 2000), (1000, 5000), $R = L = 100$ and $f_1 = f_2 = 0.10$. In Table 1, we report the RRMSE values for the LSE's and QIFE's of β_j, $j = 0, 1, 2$. The RB values are not shown. However, it has been observed that for $n_1 = 100, n_2 = 500$, i.e., when the second sample size is relatively large compared to the first, the relative biases of both the estimators are comparable. For $n_1 = 100, n_2 = 200$ the relative bias of QIFE is slightly higher than LSE. This is expected as LSE is unbiased while QIFE is not. What is interesting to observe, that with increase in the relative magnitude of N_2 compared to N_1, the performances of QIFE's of β_j, $j = 0, 1$ improve over the LSE's substantially. Also the improvement is more if the correlation between x_1 and x_2 increases. The performances of QIFE and LSE of β_2 are more or less same.

5 Concluding Remarks

In this article we propose quadratic inference function estimator of the superpopulation parameters using information from multiple samples from the same superpopulation that incorporates the design weights. For illustrative purpose, in this paper,

Table 1 RRMSE of the least squares (LS) and quadratic inference function (QIF) estimators of the superpopulation parameters for models **M1** and **M2**

Regression coefficient	Model M1		Model M2	
	LSE	QIFE	LSE	QIFE
N1 = 1000 N2 = 2000				
β_0	502	316	507	313
β_1	2004	1810	1470	1051
β_2	2845	2902	2063	2092
N1 = 1000 N2 = 5000				
β_0	507	223	511	226
β_1	2046	1712	1485	928
β_2	2827	2898	2113	2147

we have considered linear regression superpopulation model. Our design-adjusted QIF estimator is appealing in the sense that it can be applied for complex survey designs. The simulation study shows encouraging results in situations where size of the sample containing observations on subset of covariates is very high. In future we plan to investigate the asymptotic properties of the proposed QIF estimator under complex survey designs.

References

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, *9*, 385–406.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*, 137–161.

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Mulitple imputation for multiple surveys. *Journal of the American Statistical Association*, *93*, 847–857.

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, *54*, 127138.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*, 73–96.

Kim, J. K., & Rao, J. N. K. (2012). Combining data from two independent surveys: A model assisted approach. *Biometrika*, *99*, 85–100.

Lindsay, B. G., & Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science*, *18*, 394–410.

Lohr, S. L., & Raghunathan, T. E. (2016). Combining survey data with other data sources. *Statistical Science*, *32*, 293–312.

Rendall, M. S., Ghosh-Dastidar, B., Weden, M. M., Baker, E. H., & Nazarov, Z. (2013). Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. *Social Methods and Research*, *42*, 483–530.

Roberts, G., & Binder, D. (2009). Analyses based on combining similar information from multiple surveys. In *JSM: Section on Survey Methods*.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, *21*, 6573.

Detecting a Fake Coin of a Known Type

Jyotirmoy Sarkar and Bikas K. Sinha

Abstract We are given $c \geq 2$ coins which are otherwise identical, except that there may be exactly (or at most) one fake coin among them which is known to be slightly lighter than the other genuine coins. Using only a two-pan weighing balance, we weigh subsets of coins sequentially in order to identify the counterfeit coin (or declare that all coins are genuine) using the fewest weighings on average. We find a formula for the smallest expected number of weighings, and another formula which determines an optimal number of coins to place on each pan during the first (and hence during each successive) weighing.

Keywords Numerical algorithm · Recursive relation
Single counterfeit coin problem · Sequential weighing design

1 The Known Type Fake Coin Problem

The celebrated single counterfeit coin problem (SCCP), with many variations, has appeared in many mathematical magazines and mathematical quiz articles beginning with Dyson (1946). See also Levitin and Levitin (2011), Martelli and Gannon (1997), Ward (1996), Smith (1947). For a history of the problem, we refer the reader to Guy and Nowakowski (1995) and the references therein. An amusing verse, which solves the simultaneous weighing design to detect at most one counterfeit coin of an **unknown type** from among 12 coins and declare its type (whether lighter or heavier than a genuine coin) or declare that all coins are genuine, is found in Descartes (1950), wherein Cedric Smith writes under the pseudonym Blanche Descartes:

J. Sarkar
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

B. K. Sinha (✉)
Retired Faculty, Indian Statistical Institute, Kolkata, India
e-mail: bikassinha1946@gmail.com

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_15

```
F AM NOT LICKED
MA DO LIKE
ME TO FIND
FAKE COIN
```

The first line names the 12 coins with distinct letters, and each of the next three lines gives the first four coins to be placed on the left pan, and the next four on the right pan (and the rest are to be set aside). These weighings are declared simultaneously at the outset. A general solution and a geometric visualization of the simultaneous weighing design that solves the SCCP of an unknown type when there are $c \geq 3$ coins is given in Sarkar and Sinha (2016).

The focus of this paper is to discover sequential weighing designs that minimize the expected number of weighings needed to detect a single fake coin of a **known type** from among $c \geq 2$ coins, or to declare that all coins are genuine. The precise statement of the problem is given below. Here, sequential weighing means that the result(s) of the previous weighing(s) is (are) available to determine which coins to place on each pan during the next weighing.

> **Version 1 (One Lighter Fake Coin)**: There are $c \geq 2$ otherwise identical coins, except that one of these coins is counterfeit, and it is *known* to be slightly lighter than the genuine coins. Which sequential weighing design minimizes the expected number of weighings needed to identify the fake coin with a two-pan balance scale without using any known weight measures?

This problem is symmetric to the problem in which it is known that the counterfeit coin is heavier than a genuine coin (except that the heavier fake coin will be in the lower pan).

1.1 A Motivating Example that Seeks Optimal Sequential Weighing Designs

Let us borrow from Sarkar and Sinha (2016) two unsolved problems. We will describe in Sect. 2.2 the partial solutions given in Sarkar and Sinha (2016), but at present let them serve as motivating examples. Later in this paper we will solve them completely.

A consignment of 10,000 identical pills are to be put in 100 bottles each containing 100 pills. While the 100th bottle is being filled, one extra pill is found. The clerk can take one of two actions: (1) Report that one of the previous 99 bottles is missing a pill; (2) Simply place the extra pill in the last bottle and not report the discovery. In this second case, had the supervisor intentionally placed an extra pill in the consignment to test the clerk, she would know one of the 100 bottles contains 101 pills. In either case, the supervisor can recount and correct the mistake, if any. However, if a large two-pan balance is available, then instead of recounting pills, she can weigh the bottles. How can we minimize the expected number of weighings needed to identify the offending bottle? Note that in case (1), the offending bottle is lighter among 99 bottles; and in case (2), it is heavier among 100 bottles.

2 Identifying the Lighter Fake Coin

If $c = 1$, there is no need to weigh any coin as the given coin must be the fake lighter coin. To document the identification of the fake coin of a known type from among $c \geq 2$ coins we proceed as follows: First, we give explicit weighing designs for small values of c up to 9; and provide some useful commentaries on these designs. For instance, (i) the designs for $c = 3, 9$ easily generalize to the case of $c = 3^w$; (ii) there are multiple designs for $c = 6$; and (iii) the minimum expected number of weighings for $2, 3, \ldots, 9$ exhibit non-monotonicity in c. Next, we develop a recursive relation, which we utilize to construct a numerical algorithm that determines how many coins (which we denote by $n_1(c)$) should be placed on each pan during the first weighing in order to minimize the expected number of weighings needed to identify the lighter fake coin from among c coins. We show that for some values of c a multiplicity of choices for $n_1(c)$ are possible, each attaining the same minimum expected number of weighings (which we denote by $w(c)$). Finally, we present an analytic formula for $w(c)$, and another analytic formula for determining one particular value of $n_1(c)$.

2.1 Finding the Lighter Fake Coin from Among $2 \leq c \leq 9$ *Coins*

As Sarkar and Sinha (2016) explain, the SCCP involving $c = 2, 3$ coins is easily solved in one weighing by placing one coin on each pan (an setting aside the third coin, if $c = 3$). Also, the case of $c = 9$ is solved by splitting the coins into three equal groups of three coins each, identifying which group contains the lighter coin in one weighing, and then identifying which member of the lighter group is the lighter coin during the second weighing. This strategy is called the "method of trisection." We leave it to the reader(s) to verify that three trials suffice to identify the single lighter fake coin from among 27 coins.

In general, the "method of trisection" works perfectly for $c = 3^k$ coins, one of which is lighter, requiring exactly k weighings to detect the lighter fake coin. But when we start with c coins, not a power of 3, then the exact number of weighing required to identify the fake coin depends randomly on the position occupied by the fake coin. If $3^{k-1} < c \leq 3^k$, then **at most** k weighings suffice to identify the fake lighter coin. This is justified by the following argument: Let us extend "trisection" to mean breaking up c coins into three group of sizes as close to one another as possible. Then by mathematical induction it is easily seen that the "method of extended trisection" provides a solution to the problem of detection of the lighter fake coin in at most k weighings. That is, two weighings suffice to detect the single fake lighter coin from among 4–9 coins. Likewise, it takes at most 3 weighings to detect a fake lighter coin from among 10–27 coins; etc.

Using the method of extended trisection, we can see that exactly k weighings are needed to detect the single fake lighter coin from among $c = 3^k - 1$ coins. But when $3^{k-1} < c < 3^k - 1$, the number of weighings needed to identify the lighter fake coin may be random. Therefore, a natural question is which weighing strategy minimizes the expected number of weighings? Here, we assume that the lighter fake coin is equally likely to be in any position to begin with. We call a weighing design **optimal** if it minimizes the expected number of weighings needed to identify the lighter fake coin from among $c \geq 2$ coins. We are interested in finding all optimal weighing designs.

For instance, suppose that we are given 4 coins with one fake lighter coin among them. If we place one coin on each pan and leave two coins aside, then the pans may balance or not. If they balance, then the fake coin is among the two coins set aside; and we need another weighing to detect the fake coin out of these two candidate coins. On the other hand, if the pans do not balance, then we have already detected the fake lighter coin—it is on the higher pan. Assuming that the fake coin is equally likely to be any one of the candidate coins, the chance of the pans balancing or not are respectively 1/2 each. Hence, the expected number of weighings to detect the lighter fake coin from among the initial $c = 4$ coins is $(1/2) * 2 + (1/2) * 1 = 1.5$. Is this the smallest possible expected number of weighings?

Let us see what would happen if we implemented another weighing design in which we place two coins on each pan during the first weighing. Surely for this second weighing design, the fake coin will be among the two coins on the higher pan. Thereafter, a second weighing is necessary to identify the lighter fake coin from among these two candidate coins. Hence, this second weighing design necessarily requires two weighings. Indeed, since these are the only two weighing designs for $c = 4$, it follows that (i) one weighing is not guaranteed to detect the lighter fake coin, and (ii) the first weighing design (putting $n_1(4) = 1$ coin on each pan as suggested by the "method of extended trisection"), minimizes the expected number of weighings to detect the lighter fake coin. The minimum expected number of weighings is $w(4) = 1.5$.

Likewise, if we are given 5 coins with one lighter fake coin among them, we have two possible weighing designs—put one coin or two coins on each pan (leaving aside 3 coins or 1 coin, respectively). For the first weighing design, the pans balance with probability 3/5, and a second weighing identifies the lighter fake coin from among the three coins set aside; and the pans do not balance with probability 2/5, and the higher pan contains the lighter fake coin. Therefore, the expected number of weighing is $(3/5) * 2 + (2/5) * 1 = 1.6$. For the second weighing design, which puts two coins on each pan, the probability that the pans balance is 1/5, in which case the single coin set aside is the lighter fake coin; and the probability that the pans do not balance is 4/5, in which case the higher pan contains the lighter fake coin, which can be detected in a second weighing. Therefore, the expected number of weighing is $(1/5) * 1 + (4/5) * 2 = 1.8$. Hence, the optimal weighing design for $c = 5$ coins puts $n_1(5) = 1$ coin on each pan and sets aside 3 coins, and $w(5) = 1.6$. Note that this optimal weighing design is **not** obtained by the "method of extended trisection." What method is it then? We will answer it in Remark 1 at the end of Sect. 2.

Continuing on to the case of $c = 6$ coins, we notice that there are multiple optimal weighing designs. In fact, each of the three possible number of coins on each pan during the first weighing leads to the same expected number of weighings, namely $w(6) = 2.00$. (i) If we put one coin on each pan, then the pans balance with probability 4/6, the lighter fake coin is among the 4 coins set aside, and can be identified with expected number of weighings 1.5; and the pans do not balance with probability 2/6, and the lighter fake coin is already identified on the higher pan. So the expected number of weighings is $(4/6) * (1 + 1.5) + (2/6) * 1 = 2.0$. Moreover, the number of weighings is random taking values 1, 2 or 3 with probabilities 1/3 each. (ii) If we put two coins on each pan (as the method of trisection suggests), then the lighter fake coin is among the 2 coins set aside if the pans balance; or it is among the two coins on the higher pan if the pans do not balance. In either case, the lighter fake coin can be identified during a second weighing. So for this weighing design exactly 2 weighings are needed. (iii) If we put three coins on each pan, then the lighter fake coin is on the higher pan, and it can be identified during a second weighing. Again, for this third design also exactly 2 weighings are needed. Thus, the case of $c = 6$ coins exhibits a multiplicity of solutions to the optimal design; that is, $n_1(6) \in \{1, 2, 3\}$.

Next, if we are given $c = 7$ coins, the eventual expected number of weighings, if we put one or two or three coins on each pan during the first weighing and follow up with optimal choices during the successive weighings, are seen to be $(2/7) * 1 + (5/7)(1 + 1.6) = 15/7, 2$ and $(6/7) * 2 + (1/7) * 1 = 13/7$, respectively. Therefore, the optimal weighing design places $n_1(7) = 3$ coins on each pan and sets aside only one coin. Again, this optimal design is not what the method of extended trisection suggests. Surprisingly, the minimum expected number of weighings $w(7) = 13/7$ for the case of $c = 7$ coins is smaller than the minimum expected number of weighings for the case of 6 coins! That is, $w(c)$ is **not** monotonic in c.

If $c = 8$ coins are given with one lighter fake coin among them, then the method of extended trisection shows that exactly two weighing are needed to identify the lighter fake. In fact, by mathematical induction, we can see that the single lighter fake coin among $c = 3^k - 1$ coins can be identified using exactly $w(3^k - 1) = k$ weighings.

2.2 *Partial Solution to the Pill Counting Problem*

Let us return to the pill counting problem borrowed from Sarkar and Sinha (2016) and stated in Sect. 1.1.

In Scenario (1) (one lighter bottle among 99 bottles), at most five sequential weighings will suffice to identify the lighter bottle, since $3^4 < 99 < 3^5$. In Sarkar and Sinha (2016), suggested a weighing design; and demonstrated that the actual number of weighings is either 4 or 5 with probabilities 7/11 and 4/11 respectively.

Hence, the expected number of weighings is $4 + 4/11$. Thereafter, they considered an alternative weighing design, which takes on average $4 + 3/11$ weighings, Hence, the alternative design is slightly better than the first weighing design. Figure 1 depicts this second sequential design and shows the computation of the expected number of weighings.

In Sarkar and Sinha (2016), invited the reader to discover, if possible, another weighing design that requires a lower expected number of weighings to detect the lighter bottle from among 99 bottles, or prove that this second weighing design attains the minimum expected number of weighings. In this paper we resolve the issue, proving that the weighing design of Fig. 1 is indeed an optimal weighing design that minimizes the expected number of weighings. In fact, we show in Sect. 3 that there are 148 distinct optimal designs each attaining the same expected number of weighings.

Next, in Scenario (2) (one heavier bottle among 100 bottles), in order to identify the heavier bottle, surely a sequential weighing design involving five weighings suffices. As in the design shown in Fig. 1, we first weigh 27 bottles on each pan, leaving 46 bottles aside. There is a .54 chance the suspected heavier bottle will be in a pool of 27 bottles and a .46 chance it will be in the pool of 46 bottles. In the former case, exactly 3 more weighings are needed to identify the heavier bottle from among the suspected pool of $27 = 3^3$ bottles. In the latter case, during the second weighing, we place 9 bottles in each pan, leaving 28 aside. Again, if the pans balance (that is the heavier bottle is in the pool of 28 bottles left aside, which happens with probability 28/46), we weigh 9, 3 and 1 coin on each pan during the next three weighings, thereby needing on average of $3 + 1/14$ more weighings to detect the heavier bottle out of 28; and if the pans do not balance during the second weighing (which happens with probability 18/46) we need only two more weighings to sort through the $9 = 3^2$ suspected bottles! Hence, the expected number of weighings to identify the heavier bottle from among 46 bottles is $1 + (28/46) * (3 + 1/14) +$

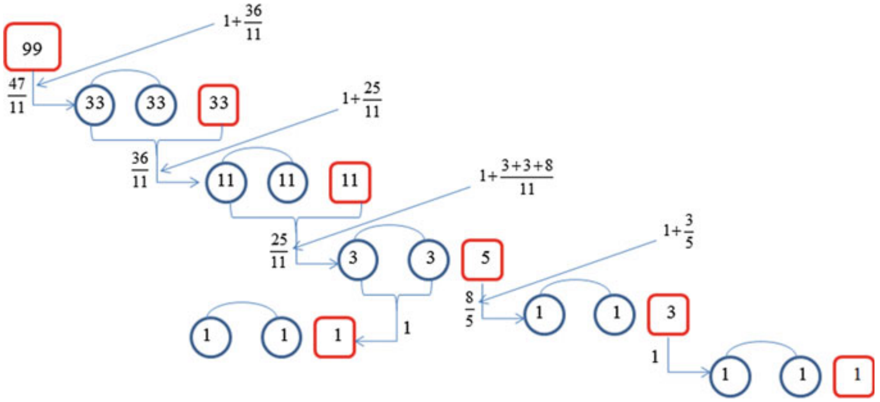


Fig. 1 A proposed weighing design for $c = 99$, and its expected number of weighings

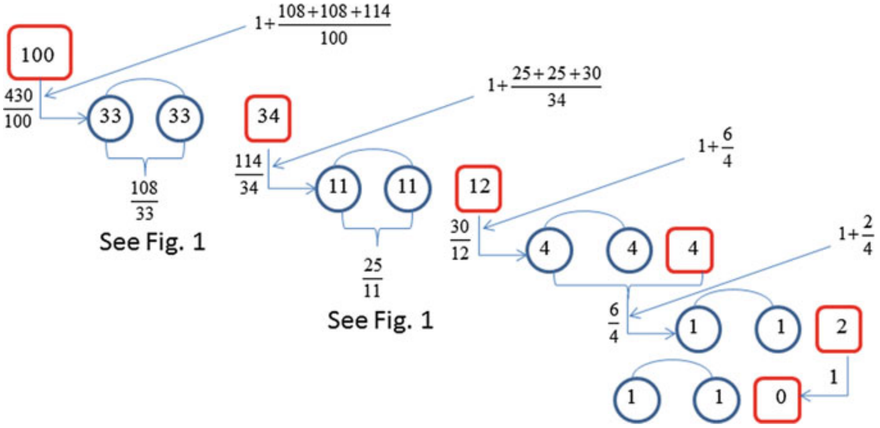


Fig. 2 A proposed weighing design for $c = 100$, and its expected number of weighings

$(18/46) * 2 = 3 + 30/46$. Finally, the expected number of weighings to identify the heavier bottle from among the original 100 bottles is $1 + (.54) * 3 + (.46) * (3 + 30/46) = 4.30$. Figure 2 depicts this sequential design and shows the computation of expected number of weighings.

Again, Sarkar and Sinha (2016), invited the reader to discover, if possible, another weighing design that requires a lower expected number of weighings to detect the heavier bottle from among 100 bottles, or prove the optimality of the above weighing design. In this paper, we resolve the issue by proving the optimality of the weighing design shown in Fig. 2. Furthermore, in Sect. 3 we show that there are 583 optimal designs each of which requires an average of 4.30 weighings to detect the heavier bottle from among 100 bottles.

2.3 A Recursive Relation and a Numerical Algorithm

In the previous Subsection we have directly evaluated

$$\{w(c) : 1 \leq c \leq 9\} = \{0, 1, 1, 1.5, 1.6, 2, 13/7, 2, 2\}$$

and

$$\{n_1(c) : 1 \leq c \leq 9\} = \{0, 1, 1, 1, 1, \{1, 2, 3\}, 3, 3, 3\}$$

Observe the non-monotonicity in $w(c)$ since $w(7) < w(6)$, and multiplicity of solutions (shown within braces) for $c = 6$. Furthermore, in the previous Subsection we proved by mathematical induction that $w(3^k) = k$ using the method of trisection, and $w(3^k - 1) = k$ using the method of extended trisection.

Table 1 The minimum expected number of weighings to detect a single lighter fake coin among c coins

```
> round(ENW, 4)
  [1] 0.0000 1.0000 1.0000 1.5000 1.6000 2.0000 1.8571 2.0000 2.0000 2.2000
 [11] 2.2727 2.5000 2.4615 2.5714 2.6000 2.7500 2.7059 2.7778 2.7895 2.9000
 [21] 2.8571 2.9091 2.9130 3.0000 2.9600 3.0000 3.0000 3.0714 3.1034 3.2000
 [31] 3.1935 3.2500 3.2727 3.3529 3.3429 3.3889 3.4054 3.4737 3.4615 3.5000
 [41] 3.5122 3.5714 3.5581 3.5909 3.6000 3.6522 3.6383 3.6667 3.6735 3.7200
 [51] 3.7059 3.7308 3.7358 3.7778 3.7636 3.7857 3.7895 3.8276 3.8136 3.8333
 [61] 3.8361 3.8710 3.8571 3.8750 3.8769 3.9091 3.8955 3.9118 3.9130 3.9429
 [71] 3.9296 3.9444 3.9452 3.9730 3.9600 3.9737 3.9740 4.0000 3.9873 4.0000
 [81] 4.0000 4.0244 4.0361 4.0714 4.0706 4.0930 4.1034 4.1364 4.1348 4.1556
 [91] 4.1648 4.1957 4.1935 4.2128 4.2211 4.2500 4.2474 4.2653 4.2727 4.3000
[101] 4.2970 4.3137 4.3204 4.3462 4.3429 4.3585 4.3645 4.3889 4.3853 4.4000
```

In general, for any $c \geq 2$, by considering each of the possible number of coins to be placed on each pan during the first weighing, we obtain the following recursive relation:

$$w(c) = \min_{1 \leq j \leq \lfloor c/2 \rfloor} \left\{ 1 + \frac{2j}{c} * w(j) + \left(1 - \frac{2j}{c}\right) * w(c - 2j) \right\} \tag{1}$$

where we have defined $w(0) = 0$. Thereafter, we obtain **an optimal design** by placing on each pan during the first weighing

$$\begin{aligned} n_1(c) &= \text{argmin}_{1 \leq j \leq \lfloor c/2 \rfloor} \left\{ 1 + \frac{2j}{c} * w(j) + \left(1 - \frac{2j}{c}\right) * w(c - 2j) \right\} \\ &= \text{argmin}_{1 \leq j \leq \lfloor c/2 \rfloor} \{2j * w(j) + (c - 2j) * w(c - 2j)\} \end{aligned} \tag{2}$$

coins; and then following up with either (i) an optimal design for the $c - 2j$ coins set aside if the pans balance (which happens with probability $1 - 2j/c$) during the first weighing, or (ii) an optimal design for the j coins on the higher pan if the pans do not balance (with probability $2j/c$) during the first weighing. Since there can be multiple solutions to (2), we use the phrase "an optimal design," rather than "the optimal design."

Defining $w(j; c) = 1 + [2j * w(j) + (c - 2j) * w(c - 2j)]/c$, we see that $w(c) = \min_{1 \leq j \leq \lfloor c/2 \rfloor} w(j; c)$.

The R codes, given in the Appendix, implement the numerical algorithm to evaluate $w(c)$ and to obtain all possible values of $n_1(c)$. We show in Table 1 the values of $w(c)$, and in Table 2 all choices of $n_1(c)$ for $1 \leq c \leq 110$.

Table 2 The number of coins to put on each pan during the first weighing to minimize the expected number of weighings, showing multiple answers within braces and the number of repetitions of the same set of answers within parentheses

```
  [1] 0 1(4) {1-3} 3(5) {3-5} {3,5}(3) {3-5,7} {5,7}(3) {5,7-9}
 [21] {7,9}(3) {7-9} 9(5) {9-11} {9,11}(3) {9-11,13} {9,11,13}(3)
 [38] {9-11,13-15} {9,11,13,15}(3) {9-11,13-15,17} {9,11,13,15,17}(3)
 [46] {9-11,13-15,17-19} {11,13,15,17,19}(3) {11,13-15,17-19,21}
 [51] {13,15,17,19,21}(3) {13-15,17-19,21-23} {15,17,19,21,23}(3)
 [58] {15,17-19,21-23,25} {17,19,21,23,25}(3) {17-19,21-23,25-27}
 [63] {19,21,23,25,27}(3) {19,21-23,25-27} {21,23,25,27}(3)
 [70] {21-23,25-27} {23,25,27}(3) {23,25-27} {25,27}(3) {25-27}
 [79] 27(5) {27-29} {27,29}(3) {27-29,31} {27,29,31}(3)
 [92] {27-29,31-33} {27,29,31,33}(3) {27-29,31-33,35}
 [97] {27,29,31,33,35}(3) {27-29,31-33,35-37} {27,29,31,33,35,37}(3)
[104] {27-29,31-33,35-37,39} {27,29,31,33,35,37,39}(3)
[108] {27-29,31-33,35-37,39-41} {27,29,31,33,35,37,39,41}(3)
```

Let us make a couple of observations on $\{n_1(c)\}$: (1) For $3^k - 2 \le c \le 3^k + 2$ with $k \ge 1$, there is a unique $n_1(c) = k$, (except for $c = 1$, in which case $n_1(1) = 0$). (2) If c is odd, none of the optimal designs ever place an even number of coins on each pan. We leave it to the reader to discover other interesting patterns among the solutions $\{n_1(c)\}$.

Returning to the motivating problem of bottling pills, we note that an optimal weighing design for $c = 99$ allows for multiple choices of $n_1(99)$; namely, 27, 29, 31, 33, 35. Each choice, followed by any optimal design for the suspected pool, leads to the same minimum expected number of weighings $w(99) = 4 + 3/11$. Likewise, $n_1(100)$ is any one element of $\{27, 28, 29, 31, 32, 33, 35, 36, 37\}$. Each of these choices, followed by some optimal design on the suspected pool, will attain the same minimum expected number of weighings $w(100) = 4.30$. Also, we have $w(101) = 4.297030 < 4.30 = w(100)$.

Next, we make the following observations on $\{w(c)\}$ based on the computed values of $w(c)$ for $1 \le c \le 110$. These observations can be proved by mathematical induction on k.

1. For $k \ge 1$, we have $w(3^k) = k$ and $w(3^k - 1) = k$. These are seen by placing 3^{k-1} coins in each pan and setting aside 3^{k-1} or $3^{k-1} - 1$ coins. Thereafter, use $w(2) = 1 = w(3)$, and apply mathematical induction.
2. $w(c)$ is non-monotonic in c. In fact, $w(c) = w(c-1)$ if $c = 3^k$ for some $k \ge 0$; $w(c) < w(c-1)$ if $c = 3^k + 4m$ for some $k \ge 1$, and for some $m \ge 1$; otherwise, $w(c) > w(c-1)$.
3. For $k \ge 1$, we have $w(3^k - 2) = k - 1/(3^k - 2)$. This is seen by placing 3^{k-1} coins in each pan and setting aside $3^{k-1} - 2$ coins. Thereafter, use $w(1) = 1 - 1/1 = 0$, $w(7) = 2 - 1/7 = 13/7$, and apply mathematical induction.

4. For $k \geq 2$, we have $w(3^k - 3) = k$. This is seen by placing 3^{k-1} coins in each pan and setting aside $3^{k-1} - 3$ coins. Thereafter, use $w(6) = 2$, and apply mathematical induction.
5. For $k \geq 0$, we have $w(3^k + 1) = k + 2/(3^k + 1)$. This is seen by placing 3^{k-1} coins in each pan and setting aside $3^{k-1} + 1$ coins. Thereafter, use $w(2) = 1 = 0 + 2/2$, $w(4) = 1.5 = 1 + 2/4$, and apply mathematical induction.

We conclude the paper by proving a general formula for determining an optimal number $n_1(c)$ of coins to be placed on each pan during the first weighing (there are, of course, other optimal choices for $n_1(c)$), and a formula for $w(c)$, the minimum expected number of weighings needed to identify the lighter coin out of c coins. First, we define a special function on integers and note some of its properties:

Definition 1 Define a function on all integers as follows: $g(0) = 0, g(1) = 2$, $g(2) = 3, g(3) = 6$; and extend this definition to all integers using $g(4l + r) = 6l + g(r)$ for all integers l and $r \in \{0, 1, 2, 3\}$. For instance, $g(4) = g(4 + 0) = 6 + g(0) = 6, g(5) = g(4 + 1) = 6 + g(1) = 8, \ldots$; and also $g(-1) = g(-4 + 3) = -6 + g(3) = 0, g(-2) = g(-4 + 2) = -6 + g(2) = -3, \ldots$.

More directly, we can define for all integers x,

$$g(x) = \begin{cases} 3x/2 & \text{if } x = 0\,(\text{mod } 2) \\ (3x + 1)/2 & \text{if } x = 1\,(\text{mod } 4) \\ (3x + 3)/2 & \text{if } x = 3\,(\text{mod } 4) \end{cases}$$

Note the following properties of the $g(\cdot)$ function: (P0) $g(\cdot)$ is non-decreasing; (P1) $2 \leq g(x) - g(x - 2) \leq 4$, (P2) $g(x) - g(x - 4) = 6$, and (P3) $g(2x + y) \leq 2g(x) + g(y)$; with equality if and only if x is even. The verifications of P0, P1 and P2 are straight-forward. To verify P3, we ask the reader to evaluate $g(2x + y)$ and $2g(x) + g(y)$ (using Definition 1) for all combinations of values of $x, y \in \{0, 1, 2, 3\}$.

Theorem 1 *Assume that the lighter fake coin is equally likely to appear in any c positions. Suppose that $3^k \leq c < 3^{k+1}$. Let us write $c = 3^k + 12h + 4i + r$ where $k = \lfloor \log_3 c \rfloor, h = \lfloor (c - 3^k)/12 \rfloor \in \{0, 1, \ldots, \lfloor 3^{k-1}/2 \rfloor\}, i = \lfloor (c - 3^k - 12h)/4 \rfloor \in \{0, 1, 2\}$, and $r = (c - 3^k)$ (mod 4) $\in \{0, 1, 2, 3\}$. Then, $n_1(1) = 0, n_1(2) = 1$ and for $c \geq 3$, one choice of optimal number of coins to be placed on each pan during the first weighing is given by*

$$n_1(c) = 3^{k-1} + 4h + 2i \tag{3}$$

and, using the special function defined in Definition 1, the minimum expected number of weighings needed to identify the lighter coin out of $c \geq 1$ coins is given by

$$w(c) = k + [18h + 6i + g(r)]/c \tag{4}$$

Proof We prove results (3) and (4) using mathematical induction (the strong version) on c. For the base cases, we note that

c	$k\ l\ r$	$n_1(c)$	$w(c)$ using (4)	ENW[c]
$1 = 3^0 + 4 * 0 + 0$	0 0 0	0	$w(1) = 0 + [6 * 0 + g(0)]/1 = 0$	0.0000
$2 = 3^0 + 4 * 0 + 1$	0 0 1	1	$w(2) = 0 + [6 * 0 + g(1)]/2 = 1$	1.0000
$3 = 3^1 + 4 * 0 + 0$	1 0 0	1	$w(3) = 1 + [6 * 0 + g(0)]/3 = 1$	1.0000
$4 = 3^1 + 4 * 0 + 1$	1 0 1	1	$w(4) = 1 + [6 * 0 + g(1)]/4 = 1.5$	1.5000
$5 = 3^1 + 4 * 0 + 2$	1 0 2	1	$w(5) = 1 + [6 * 0 + g(2)]/5 = 1.6$	1.6000
$6 = 3^1 + 4 * 0 + 3$	1 0 3	1	$w(6) = 1 + [6 * 0 + g(3)]/6 = 2$	2.0000
$7 = 3^1 + 4 * 1 + 0$	1 1 0	3	$w(7) = 1 + [6 * 1 + g(0)]/7 = 13/7$	1.8571
$8 = 3^1 + 4 * 1 + 1$	1 1 1	3	$w(8) = 1 + [6 * 1 + g(1)]/8 = 2$	2.0000
$9 = 3^2 + 4 * 0 + 0$	2 0 0	3	$w(9) = 2 + [6 * 0 + g(0)]/9 = 2$	2.0000

Thus, we note that (3) and (4) hold for $1 \leq c \leq 9$. Also note that for $c = 3^k$, we can use either the decomposition $3^k = 3^{k-1} + 4 * \lfloor 3^{k-1}/2 \rfloor + 2$ or $3^k + 4 * 0 + 0$, and get the same value of $w(3^k) = k$. Suppose now that (3) and (4) hold for all values of c in the set $\{1, 2, \ldots, m-1\}$ where $3^k \leq m - 1 < 3^{k+1}$. Next, let $c = m = 3^k + 12h + 4i + r$. We shall show that (3) and (4) hold for $3^k < c = m \leq 3^{k+1}$ by proving the following two lemmas:

Lemma 1 *If we place during the first weighing $j^* = 3^{k-1} + 4h + 2i$ coins on each pan and set aside the remaining $s^* = m - 2j^* = 3^{k-1} + 4h + r$ coins (and thereafter follow-up with an optimal weighing design for $s^* < m$ coins or $j^* < m$ coins according as the pans balance or do not balance), then the expected number of weighings to identify the fake lighter coin is indeed as given on the right hand side of (4).*

Lemma 2 *If we place during the first weighing $1 \leq j \leq \lfloor m/2 \rfloor$ coin(s) on each pan and set aside the remaining $s = m - 2j$ coins (and thereafter follow-up with an optimal weighing design for s coins or j coins according as the pans balance or do not balance), then the expected number of weighings to identify the fake lighter coin is* **at least as big as** *the expression on the right hand side of (4).*

Proof of Lemma 1. Since $j^* = 3^{k-1} + 4h + 2i < m$, by the strong induction hypothesis, if i is even, then writing $2i = 4i/2$, we have

$$w(j^*) = (k-1) + [6(h + i/2)]/j^* = (k-1) + [6h + 3i]/j^*$$

and if i is odd, then writing $2i = 4(i-1)/2 + 2$, we still have

$$w(j^*) = (k-1) + [6h + 6(i-1)/2 + g(2)]/j^* = (k-1) + [6h + 3i]/j^*$$

since $g(2) = 3$. Recall that $m = 3^k + 12h + 4i + r$. Hence, $s^* = m - 2j^* = 3^{k-1} + 4h + r < m$. So, by the strong induction hypothesis, we have

$$w(s^*) = (k-1) + [6h + g(r)]/s^*$$

Therefore, using (1) and the strong induction hypothesis, we have

$$\begin{aligned} w(j^*; m) &= 1 + [2\, j^*\, w(j^*) + s^*\, w(s^*)]/m \\ &= 1 + [2\,\{j^*\,(k-1) + 6h + 3i\} + \{s^*\,(k-1) + 6h + g(r)\}]/m \\ &= 1 + (k-1) + [18h + 6i + g(r)]/m = k + [18h + 6i + g(r)]/m \end{aligned}$$

which agrees with the right hand side of (4). This completes the proof of Lemma 1.

In particular, Lemma 1 implies that

$$w(m) = \min_{1 \le j \le m/2} w(j; m) \le w(j^*; m) = k + [18h + 6i + g(r)]/m$$

To show the opposite inequality, we need Lemma 2.

Proof of Lemma 2. We decompose the proof of Lemma 2 into five cases according as the value (or range of values) of j.

Case 1 ($3^{k-1} \le j \le 3^k$). Then $3^{k-1} \le s \le 3^k$. We write $j = 3^{k-1} + 12\bar{h} + 4\bar{i} + \bar{r}$ and $s = 3^{k-1} + 12\tilde{h} + 4\tilde{i} + \tilde{r}$. Then $m = 2j + s = 3^k + 12(2\bar{h} + \tilde{h}) + 4(2\bar{i} + \tilde{i}) + (2\bar{r} + \tilde{r})$. Using (1) and the strong induction hypothesis, we have

$$\begin{aligned} m\, w(j; m) &= m + 2\, j\, w(j) + s\, w(s) \\ &= m + 2\,\{j\,(k-1) + 18\bar{h} + 6\bar{i} + g(\bar{r})\} + s\,(k-1) + 18\tilde{h} + 6\tilde{i} + g(\tilde{r}) \\ &= m + m(k-1) + 18(2\bar{h} + \tilde{h}) + 6(2\bar{i} + \tilde{i}) + [2\, g(\bar{r}) + g(\tilde{r})] \\ &= mk + 18[2\bar{h} + \tilde{h}] + 6[2\bar{i} + \tilde{i}] + [2\, g(\bar{r}) + g(\tilde{r})] \\ &\ge mk + 18[2\bar{h} + \tilde{h}] + 6[2\bar{i} + \tilde{i}] + g(2\,\bar{r} + \tilde{r}) \\ &= m\, w(j^*; m) \end{aligned}$$

The last inequality above follows from property P3 of the $g(\cdot)$ function.

Thus, no other j (in the range $3^{k-1} \le j \le 3^k$) achieves a lower expected number of weighings than achieved by j^*, though some may achieve the same expectation. The next four cases establish that for values of j below 3^{k-1} and above 3^k the expectation is strictly higher.

Case 2 ($j = 3^{k-1} - 1$). Then $3^{k-1} + 2 \le s$. We write $s = 3^{k-1} + 12\tilde{h} + 4\tilde{i} + \tilde{r}$. Using property P1 of the $g(\cdot)$ function, we have

$$\begin{aligned} m\,w(j;m) &= m + 2\,j\,w(j) + s\,w(s) \\ &= m + 2\,j\,(k-1) + s\,(k-1) + 18\tilde{h} + 6\tilde{i} + g(\tilde{r}) \\ &> m + 2\,(j+1)\,(k-1) + (s-2)\,(k-1) + 18\tilde{h} + 6\tilde{i} + g(\tilde{r}-2) \\ &= m + 2\,(j+1)\,w(j+1) + (s-2)\,w(s) = m\,w(j+1;m) \end{aligned}$$

Hence, instead of placing j coins on each pan and setting aside s coins, it is “better” to place $j+1$ coins on each pan and set aside $s-2$ coins. Here, a weighing design is better if it has a lower expected number of weighings necessary to identify the single fake lighter coin.

Case 3 ($j \le 3^{k-1} - 2$). Then $3^{k-1} + 4 \le s$. Then instead of placing j coins on each pan and setting aside s coins, it is better to place $j+2$ coins on each pan and set aside $s-4$ coins. This is because $w(j;m) > w(j+2;m)$ iff $s\,w(s) - (s-4)\,w(s-4) > 2\{(j+2)\,w(j+2) - j\,w(j)\}$, which holds true since

$$\begin{aligned} &2\,\{(j+2)\,w(j+2) - j\,w(j)\} \\ &= 2\,\{[(j+2)k' + 18\bar{h} + 6\bar{i} + g(\bar{r}+2)] - [jk' + 18\bar{h} + 6\bar{i} + g(\bar{r})]\} \\ &= 2\,\{2k' + g(\bar{r}+2) - g(\bar{r})\} \\ &\le 2\,\{2(k-2) + g(\bar{r}+2) - g(\bar{r})\} \le 4k \end{aligned}$$

The first inequality above uses the fact that $k' \le k-2$, and the second inequality uses property P1 of the $g(\cdot)$ function. Likewise, using property P2, we have

$$\begin{aligned} &s\,w(s) - (s-4)\,w(s-4) \\ &= [s(k-1) + 18\tilde{h} + 6\tilde{i} + g(\tilde{r})] - [(s-4)(k-1) + 18\tilde{h} + 6\tilde{i} + g(\tilde{r}-4)] \\ &= 4(k-1) + g(\tilde{r}) - g(\tilde{r}-4) = 4(k-1) + 6 = 4k + 2 \end{aligned}$$

Case 4 ($j = 3^k + 1$). Then $s \le 3^k - 2$. Then instead of placing $j = 3^k + 1$ coins on each pan and setting aside s coins, it is better to place $j - 1 = 3^k$ coins on each pan and set aside $s+2$ coins. This is because $w(j;m) > w(j-1;m)$ iff $2\{j\,w(j) - (j-1)\,w(j-1)\} > (s+2)\,w(s+2) - s\,w(s)$, which holds true since $2\,\{j\,w(j) - (j-1)\,w(j-1)\} = 2\,\{2k + g(1) - g(0)\} = 4k + 4 > 4\tilde{k} + 4 \ge 4\tilde{k} + g(\tilde{r}+2) - g(\tilde{r}) = (s+2)\,w(s+2) - s\,w(s)$, where $3^{\tilde{k}} \le s < 3^{\tilde{k}+1} - 2$ with $\tilde{k} < k$.

Case 5 ($3^k + 2 \le j \le m/2$). Then instead of placing j coins on each pan and setting aside s coins, it is better to place $j-2$ coins on each pan and set aside $s+4$ coins. This is because $w(j;m) > w(j-2;m)$ iff $(s+4)\,w(s+4) - s\,w(s) < 2\{j\,w(j) - (j-2)\,w(j-2)\}$, which holds true since $2\,\{j\,w(j) - (j-2)\,w(j-2)\} \ge 2\,\{2k + g(\bar{r}) - g(\bar{r}-2)\} \ge 4k + 4 > 4k + 2 = (s+4)\,w(s+4) - s\,w(s)$.

The above five cases together imply that no value of $1 \le j < m/2$ achieves a lower expected number of weighings than that achieved by j^*. That is, $w(m) \ge w(j^*;m)$. This completes the proof of Lemma 2.

In view of Lemmas 1 and 2, $w(m) = w(j^*; m) = k + \{18h + 6i) + g(r)\}]/m$. Indeed, placing $j^* = 3^{k-1} + 4h + 2i$ coins on each pan (and setting aside $s^* = c - 2j^*$ coins) during the first weighing is one of the optimal designs. There may be other optimal designs (putting j coins on each pan for some select values of $3^{k-1} \leq j \leq 3^k$, but never a value of $j < 3^{k-1}$ or $j > 3^k$), each attaining the same minimum expected number of weighings given by (4).

Therefore, (3) and (4) hold for $c = m$; and hence by strong mathematical induction they hold for all $c \geq 1$. This completes the proof of the theorem. Q.E.D.

Remark 1 Expression (3) helps us automate the choice of $n_1(c)$. This choice may be called a "**hybrid trisection-bisection method**," since the first 3^k coins are trisected among the left pan, the right pan and the set aside pool; then from the remaining $(c - 3^k)$ coins, the highest multiple of 12 coins are also trisected; if there are still more coins, the next multiple(s) of 4 coins are bisected between the two pans; and finally any remaining coin(s) (at most 3) is/are set aside. Thus, this hybrid trisection-bisection method always yields a j^* such that $(c - 2j^*) \leq j^* + 3$ and $j^* \leq (c - 2j^*) + 4$. Hence, j^* is an odd number between $\lceil c/3 \rceil - 1$ and $\lfloor (c+1)/3 \rfloor + 1$. Of course, whenever there is a unique value of $n_1(c)$ (for example, when $1 \leq c \leq 5, 7 \leq c \leq 11$, or more generally when $3^k - 2 \leq c \leq 3^k + 2$), we have $n_1(c) = j^* = 3^{k-1}$.

Remark 2 Here is an alternative way to evaluate $w(j)$. The proof of Theorem 1 reveals that $p(j) = j\,w(j)$ can be obtained by taking the cumulative sum of the sequence $\{a_j : j \geq 1\}$ defined as follows: $a_1 = 0, a_2 = 2$, and for $k \geq 1$, if $3^k \leq j = 3^k + 4l + r < 3^{k+1}$, where $r = (j - 3^k) (\text{mod } 4)$; then $a_j = k + f(r)$, where $f(0) = 0$, $f(1) = 2$, $f(2) = 1$, $f(3) = 3$. Thereafter, $w(j)$ is found as $w(j) = p(j)/j$.

3 Discussions

Minimization of the expected number of weighings has been our sole criterion for determining optimal weighing designs. This has led to a multiplicity of optimal weighing designs. Exactly how many optimal weighing designs are there? We say that two weighing designs are distinct if during any one particular weighing either the designs involve different sizes of the suspected pools containing the fake lighter coin, or the number of coins placed on each pan differs across the two designs. Let $D(c)$ denote the number of distinct optimal weighing designs to detect the single fake lighter coin from among c coins. We let the reader work out an exact formula for $D(c)$. Here we only give a recursive relation that we evaluate numerically.

Clearly, for each of $c = 1, 2, 3$, there is exactly one thing to do; namely, for $c = 1$, do nothing; and for $c = 2, 3$, place one coin on each pan. So, $D(1) = 1 = D(2) = D(3)$. Next, for $c = 4$, the two distinct designs are $1 + 1 + 2$(U) and $1 + 1 + 2$(B) $\to 1 + 1 + 0$(U). Here, we are writing the number of coins on the left pan, the right pan and the set aside pool. Also, U denotes the pans are unbalanced

Table 3 Number of distinct optimal weighing designs to identify a single fake lighter coin from among $1 \leq c \leq 110$ coins

```
  [1]   1   1   1   2   2   5   2   2   1   3   3  11   6   7   4
 [16]  18   7   8   7  18   4   7   6  13   3   3   1   4   4  20
 [31]  11  14   8  51  24  28  24  84  26  39  34 100  42  49  38
 [46] 126  48  55  48 126  42  56  49 128  49  56  42 126  48  55
 [61]  48 126  38  49  42 102  34  39  26  88  24  28  24  55   8
 [76]  14  11  24   4   4   1   5   5  32  17  23  13 108  52  62
 [91]  52 232  75 109  94 363 152 182 148 583 224 261 228 762 268
[106] 343 303 964 375 433
```

and B denotes they are balanced during the particular weighing. For $c = 5$, the two distinct designs are $1 + 1 + 3(\text{U})$ and $1 + 1 + 3(\text{B}) \rightarrow 1 + 1 + 1$. For $c = 6$, there are five distinct designs: $1 + 1 + 4(\text{U})$, $1 + 1 + 4(\text{B}) \rightarrow 1 + 1 + 2(\text{U})$, $1 + 1 + 4(\text{B}) \rightarrow 1 + 1 + 2(\text{B}) \rightarrow 1 + 1 + 0(\text{U})$, $2 + 2 + 2 \rightarrow 1 + 1 + 0(\text{U})$, and $3 + 3 + 0(\text{U}) \rightarrow 1 + 1 + 1$.

In general, for $c \geq 4$, we have the following recursive relation:

$$D(c) = \sum_{j \in n_1(c)} \{D(j) + D(c - 2j) * I(c \neq 3j) * I(c \neq 2j)\} \tag{5}$$

To justify (5), we reason as follows: During the first weighing we place $j \in n_1(c)$ coins on each pan and set aside $c - 2j$ coins. The first weighing results in either (i) unbalanced pans, causing us to select an optimal design for j coins on the higher pan, or (ii) balanced pans, causing us to select an optimal design for $c - 2j$ coins from the set aside pool. However, if $j = s = c/3$ we only select an optimal design from the appropriate set of j coins (either on the higher pan or from the set aside pool). Also, we rule out the case when $(j = c/2 \in n_1(c), s = 0)$ (which is applicable only for $c = 6$) since in that case we must impose the convention $D(c - 2j) = D(0) = 0$.

We report the values of $D(c)$ in Table 3, obtained by using the R codes given in the Appendix.

Returning to the pill counting Scenarios (1) and (2), note that $D(99) = 148$ and $D(100) = 583$. While we leave the reader to find a general formula for $D(c)$, we make just one observation: For $k \geq 1$, by mathematical induction on k, we see that $D(3^k) = 1$, $D(3^k - 1) = k = D(3^k - 2)$ and $D(3^k + 1) = k + 1 = D(3^k + 2)$.

To reduce the number of optimal weighing designs, one may wish to impose an additional desirable feature to select a subset of these optimal designs. Such an additional feature may be: (1) minimize the variance of the number of weighings, or (2) minimize a fixed percentile (say, the 80th percentile) of the number of weighings needed to identify the fake lighter coin. We leave these topics to the reader to investigate.

What if there is **at most one** fake coin? Let us allow the possibility that either there is one fake lighter coin or all coins are genuine, but we do not know which is the case. Clearly, if $c = 1$ it is not possible to determine whether this coin is genuine or fake using only a two pan balance. For $c \geq 2$, we consider two approaches: (1) Assume that there is exactly one fake lighter coin, and proceed to identify it using recursive relation (1), but with a possible exception at the very end as described below. (2) As quickly as possible, ascertain whether all coins are genuine or there is a fake coin, by using the method of extended bisection during the first weighing; that is, by splitting the coins equally (if c is even), or nearly equally (if c is odd) on the two pans. Let $\bar{w}_1(c)$ and $\bar{w}_2(c)$ denote the minimum expected number of weighings needed to identify the fake lighter coin or to declare that all coins are genuine in the two approaches.

In Approach (1), three cases are possible, of which only the third case results in a weighing design different from that under the knowledge that there is exactly one fake lighter coin: (i) Should the pans ever be unbalanced, we will know for sure that there is a fake lighter coin on the higher pan. (ii) If the pans always balance and eventually the suspected pool reduces to 2 coins (which happens only if c is even), we will place one coin on each pan, just as we would do so under the knowledge of exactly one fake lighter coin. (iii) If the pans always balance until the suspected pool reduces to 1 coin (which happens with probability $1/c$ only if c is odd), we will have to weigh it against any one of the $c - 1$ genuine coins to determine whether it is lighter or genuine, requiring one more weighing than under the knowledge of exactly one fake lighter coin. Hence, if unbeknownst to us there is exactly one fake coin then $\bar{w}_1(c) = w(c) + I\{c \text{ odd}\}/c$. On the other hand, if unbeknownst to us there is no fake coin then $\bar{w}_1(c) = \lceil \log_3(c + 1) \rceil$.

In Approach (2), when $c \geq 2$ is even, the first weighing determines that all coins are genuine (if the pans balance), or that there is a fake lighter coin on the higher pan containing $c/2$ coins, and it can be detected using $w(c/2)$ more weighings on average. Hence, if unbeknownst to us there is exactly one fake coin then $\bar{w}_2(c) = 1 + w(c/2)$. On the other hand, if unbeknownst to us there is no fake coin then $\bar{w}_2(c) = 1$. Likewise, when $c \geq 3$ is odd, the first weighing reduces the suspected pool to the single coin set aside (if the pans balance), or that there is a fake lighter coin on the higher pan containing $(c - 1)/2$ coins. In the former case, one more weighing determines whether the single set aside coin is genuine or lighter. In the later case, we need $w((c - 1)/2)$ more weighings to identify the fake lighter coin. Hence, if unbeknownst to us there is exactly one fake coin then $\bar{w}_2(c) = 1 + (1 - 1/c)w((c - 1)/2) + 1/c$. On the other hand, if unbeknownst to us there is no fake coin then $\bar{w}_2(c) = 2$.

Since we are dealing with identification of a fake coin, some people would tend to believe that there is a fake coin. However, since we are cautioned that there might be no fake coin at all, some other people would believe in that proposition. Most people would have different degrees of belief on the two propositions. If one believes that with a high probability there is a fake coin, then Approach (1) is preferable. But if one believes that with a high probability there is no fake coin, then Approach (2) is preferable. Thus, a subjective belief influences the choice of optimal design when there is at most one fake lighter coin among $c \geq 2$ coins.

Acknowledgements The first author thanks the Calcutta University Statistics Department for hosting his visit during which time the research was concluded. The second author thanks the audience who participated in a discussion during his talk on this topic at the Calcutta University Statistics Department's Platinum Jubilee International Conference on Applications of Statistics.

Appendix

R codes to compute the minimum expected number of weighings, the number of coins to place on each pan during the first weighing, and the number of distinct optimal designs.

```
N=110 # set the number of coins
NEPL=rep(1,N) # ENW min with this lowest #coins on each pan at 1st weighing
NEPH=rep(1,N) # ENW min with this highest #coins on each pan at 1st weighing
ENW=c(0,rep(1,N-1)) # set ENW[1]=0, ENW[2]=1, if exactly one coin is lighter
# Now recursively compute ENW[c] for c>=3
 for (c in 3:N){
    IPS=(c-2)*ENW[c-2] # min inner product if put 1 coin on each pan
    c2=floor(c/2)      # put i=2..c2 coins on each pan at 1st weighing
    for (i in 2:c2){
      if (c >2*i){    # there are left-over coins during first weighing
        IP=2*i*ENW[i]+(c-2*i)*ENW[c-2*i]  # inner product
        if (IP <IPS){ IPS=IP; NEPL[c]=i } # a lower inner product is found
        if (IP==IPS){ IPS=IP; NEPH[c]=i } } # the same lowest inner product
      if (c==2*i){    # no coin is set aside during first weighing
        IP=2*i*ENW[i]
        if (IP <IPS){ IPS=IP; NEPL[c]=i }
        if (IP==IPS){ IPS=IP; NEPH[c]=i } }
   }
ENW[c]=1+IPS/c  # expected number of weighings to detect the lighter
}
cbind(NEPL, NEPH, ENW) # print w(c)=ENW, solution unique if NEPL=NEPH
#
# How many coins to put on each pan during the first weighing? n_1(c)
# Also calculate the number of distinct optimal designs
D=rep(1,N) # initialize the number of distinct optimal designs
for (n in 4:110){
  m=floor(n/2)
  W=rep(0,m)
  for (j in 1:m){
    if (2*j!=n){W[j]=2*j*ENW[j]+(n-2*j)*ENW[n-2*j]}
    if (2*j==n){W[j]=2*j*ENW[j]}
 }
 smallest=min(W)
    sol=which( W < (smallest+1)*rep(1,length(W)) )
# W is a vector of integers. So, +1 allows truncation error in ENW.
     print(c(n, sol))
D[n]=sum(D[sol]+D[n-2*sol]*(n*rep(1,length(sol))!=3*sol)*
             (n*rep(1,length(sol))!=2*sol))   # explained in Section 3
}
```

References

Descartes, B. (1950). The twelve coin problem. *Eureka*, *13*, 7, 20.

Dyson, F. J. (1946). The problem of the pennies. *The Mathematical Gazette*, *30*, 231–233.

Guy, R. K., & Nowakowski, R. J. (1995). Coin-weighing problems. *The American Mathematical Monthly*, *102*, 164–167.

Halbeisen, L., & Hungerbuhler, N. (1995). The general counterfeit coin problem. *Discrete Mathematics*, *147*, 139–150.

Levitin, A., & Levitin, M. (2011). *Algorithmic puzzles*. Oxford University Press. Puzzle #10.

Martelli, M., & Gannon, G. (1997). Weighing coins: Divide and conquer to detect a counterfeit. *The College Mathematics Journal*, *28*(5), 365–367.

Rao, S. B., Rao, P., & Sinha, B. K. (2005). Some combinatorial aspects of a counterfeit coin problem. *Linear Algebra and its Applications*. Special Issue.

Sarkar, J., & Sinha, B. K. (2016). Weighing designs to detect a single counterfeit coin. *Resonance: The Journal of Science Education*, *21*(2), 125–150. https://doi.org/10.1007/s12045-016-0306-8.

Smith, C. A. B. (1947). The counterfeit coin problem. *The Mathematical Gazette*, *31*(293), 31–39.

Ward, R. L. (1996). Finding one coin of 12 in 3 Steps. *The Math Forum@Drexel: Ask Dr. Math*, http://mathforum.org/library/drmath/view/55618.html.