Statistics and Operations Research Transactions, vol. 37, n. 2 (2013)

SORT 37 (2) July-December 2013, 111-130

Locally adaptive density estimation on Riemannian manifolds

Guillermo Henry[1,2], Andrés Muñoz[1] and Daniela Rodriguez[1,2]

Abstract

In this paper, we consider kernel type estimator with variable bandwidth when the random variables belong in a Riemannian manifolds. We study asymptotic properties such as the consistency and the asymptotic distribution. A simulation study is also considered to evaluate the performance of the proposal. Finally, to illustrate the potential applications of the proposed estimator, we analyse two real examples where two different manifolds are considered.

MSC: 62G07, 62G20

Keywords: Asymptotic results, density estimation, nonparametric, Riemannian manifolds.

1. Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in $\mathbb{R}^d$ and having density function f. A class of estimators of f which has been widely studied since the work of Rosenblatt (1956) and Parzen (1962) has the form

$$f_n(x) = \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right),$$

where $K(u)$ is a bounded density on $\mathbb{R}^d$ and h is a sequence of positive number such that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

If we apply this estimator to data coming from long tailed distributions, with a small enough h to be appropriate for the central part of the distribution, a spurious noise ap-

[1] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

[2] CONICET, Argentina.

ghenry@dm.uba.ar, andreslemm@gmail.com and drodrig@dm.uba.ar

Received: April 2012

Accepted: January 2013

pears in the tails. With a large value of h for correctly handling the tails, we can not see the details occurring in the main part of the distribution. To overcome these defects, adaptive kernel estimators were introduced. For instance, a conceptually similar estimator of $f(x)$ was studied by Wagner (1975) who defined a general neighbour density estimator by

$$\widehat{f_n}(x) = \frac{1}{nH_n^d(x)} \sum_{j=1}^{n} K\left(\frac{x - X_j}{H_n(x)}\right),$$

where $H_n(x)$ is the distance between x and the k-nearest neighbour of x among $X_1, \ldots, X_n$, and $k = k_n$ is a sequence of non–random integers such that $\lim_{n\to\infty} k_n = \infty$. Through this adaptive bandwidth , the estimation in the point x has the guarantee that to be calculated using at least k points of the sample.

However, in many applications, the variables X take values on different spaces than $\mathbb{R}^d$. Usually these spaces have a more complicated geometry than the Euclidean space and this has to be taken into account in the analysis of the data. For example, if we study the distribution of the stars with luminosity in a given range it is natural to think that the variables belong to a spherical cylinder ($S^2 \times \mathbb{R}$) instead of $\mathbb{R}^4$. If we consider a region of the planet M, then the direction and the velocity of the wind in this region are points in the tangent bundle of M, that is a manifold of dimension 4. Other examples could be found in image analysis, mechanics, geology and other fields. They include distributions on spheres, Lie groups, among others, see for example Joshi et al. (2007), Goh and Vidal (2008). For this reason, it is interesting to study an estimation procedure of the density function that takes into account a more complex structure of the variables.

Nonparametric kernel methods for estimating densities of spherical data have been studied by Hall et al. (1987) and Bai et al. (1988). Pelletier (2005) proposed a family of nonparametric estimators for the density function based on kernel weighting when the variables are random objects valued in a closed Riemannian manifold. Pelletier's estimators are consistent with the kernel density estimators in the Euclidean case considered by Rosenblatt (1956) and Parzen (1962).

As we comment above, the importance of local adaptive bandwidth is well known in nonparametric statistics and this is even more true with data taking values in complex spaces. In this paper, we propose a kernel density estimator on a Riemannian manifold with a variable bandwidth defined by k-nearest neighbours.

This paper is organized as follows. Section 2 contains a brief summary of the necessary concepts of Riemannian geometry. In Section 2.1, we introduce the estimator. Uniform consistency of the estimator is derived in Section 3.1, while in Section 3.2 the asymptotic distribution is obtained under regular assumptions. Section 4 contains a Monte Carlo study designed to evaluate the proposed estimator. Finally, Section 5 presents two example using real data. Proofs are given in the Appendix.

2. Preliminaries and the estimator

Let (M,g) be a d-dimensional Riemannian manifold without boundary. We denote by d_g the distance induced by the metric g. With $B_s(p)$ we denote a normal ball with radius s centred at p. The injectivity radius of (M,g) is given by $inj_g M = \inf_{p\in M} \sup\{s \in \mathbb{R} > 0 : B_s(p) \text{ is a normal ball}\}$. It is easy to see that a compact Riemannian manifold has strictly positive injectivity radius. For example, it is not difficult to see that the d-dimensional sphere S^d endowed with the metric induced by the canonical metric g_0 of R^{d+1} has injectivity radius equal to π. If N is a proper submanifold of the same dimension than (M,g), then $inj_{g|_N} N = 0$. The Euclidean space or the hyperbolic space have infinite injectivity radius. Moreover, a complete and simply connected Riemannian manifold with non-positive sectional curvature has also this property.

Throughout this paper, we will assume that (M,g) is a complete Riemannian manifold, i.e. (M,d_g) is a complete metric space. Also we will consider that $inj_g M$ is strictly positive. This restriction will be clear in the Section 2.1 when we define the estimator. For standard result on differential and Riemannian geometry we refer to the reader to Boothby (1975), Besse (1978), Do Carmo (1988) and Gallot, Hulin and Lafontaine (2004).

Let $p \in M$, we denote with 0_p and T_pM the null tangent vector and the tangent space of M at p. Let $B_s(p)$ be a normal ball centred at p. Then $B_s(0_p) = exp_p^{-1}(B_s(p))$ is an open neighbourhood of 0_p in T_pM and so it has a natural structure of differential manifold. We are going to consider the Riemannian metrics g' and g'' in $B_s(0_p)$, where $g' = exp_p^*(g)$ is the pullback of g by the exponential map and g'' is the canonical metric induced by g_p in $B_s(0_p)$. Let $w \in B_s(0_p)$, and $(\bar{U},\bar{\psi})$ be a chart of $B_s(0_p)$ such that $w \in \bar{U}$. We note by $\{\partial/\partial\bar{\psi}_1|_w, \ldots, \partial/\partial\bar{\psi}_d|_w\}$ the tangent vectors induced by $(\bar{U},\psi)$. Consider the matricial function with entries (i,j) are given by $g'\left(\left(\partial/\partial\bar{\psi}_i|_w\right), \left(\partial/\partial\bar{\psi}_j|_w\right)\right)$. The volumes of the parallelepiped spanned by $\{\left(\partial/\partial\bar{\psi}_1|_w\right), \ldots, \left(\partial/\partial\bar{\psi}_d|_w\right)\}$ with respect to the metrics g' and g'' are given by $|\det g'\left(\left(\partial/\partial\bar{\psi}_i|_w\right), \left(\partial/\partial\bar{\psi}_j|_w\right)\right)|^{1/2}$ and $|\det g''\left(\left(\partial/\partial\bar{\psi}_i|_w\right), \left(\partial/\partial\bar{\psi}_j|_w\right)\right)|^{1/2}$ respectively. The quotient between these two volumes is independent of the selected chart. So, given $q \in B_s(p)$, if $w = exp_p^{-1}(q) \in B_s(0_p)$ we can define the volume density function, $\theta_p(q)$, on (M,g) as

$$\theta_p(q) = \frac{|\det g'\left(\left(\partial/\partial\bar{\psi}_i|_w\right), \left(\partial/\partial\bar{\psi}_j|_w\right)\right)|^{1/2}}{|\det g''\left(\left(\partial/\partial\bar{\psi}_i|_w\right), \left(\partial/\partial\bar{\psi}_j|_w\right)\right)|^{1/2}}$$

for any chart $(\bar{U},\bar{\psi})$ of $B_s(0_p)$ that contains $w = exp_p^{-1}(q)$. For instance, if we consider a normal coordinate system (U,ψ) induced by an orthonormal basis $\{v_1,\ldots,v_d\}$ of T_pM then $\theta_p(q)$ is the function of the volume element $d\nu_g$ in the local expression with respect to chart (U,ψ) evaluated at q, i.e.

$$\theta_p(q) = \left|\det g_q\left(\frac{\partial}{\partial\psi_i}\Big|_q, \frac{\partial}{\partial\psi_j}\Big|_q\right)\right|^{\frac{1}{2}},$$

where $\frac{\partial}{\partial \psi_i}|_q = D_{\alpha_i(0)} exp_p(\dot{\alpha}_i(0))$ with $\alpha_i(t) = exp_p^{-1}(q) + tv_i$ for $q \in U$. Note that the volume density function $\theta_p(q)$ is not defined for all the pairs p and q in M, but it is if $d_g(p,q) < inj_g M$.

We finish the section showing some examples of the density function:

i) In the case of $(\mathbb{R}^d, g_0)$ the density function is $\theta_p(q) = 1$ for all $(p,q) \in \mathbb{R}^d \times \mathbb{R}^d$.

ii) In the 2-dimensional sphere of radius R, the volume density is

$$\theta_{p_1}(p_2) = R\,\frac{|\sin(d_g(p_1,p_2)/R)|}{d_g(p_1,p_2)} \quad \text{if } p_2 \neq p_1, -p_1 \text{ and } \theta_{p_1}(p_1) = 1.$$

where d_g induced is given by

$$d_g(p_1,p_2) = R\arccos\left(\frac{\langle p_1,\ p_2\rangle}{R^2}\right).$$

iii) In the case of the cylinder of radius 1 $\mathscr{C}_1$ endowed with the metric induced by the canonical metric of $\mathbb{R}^3$, $\theta_{p_1}(p_2) = 1$ for all $(p_1,p_2) \in \mathscr{C}_1 \times \mathscr{C}_1$, and the distance induced is given by $d_g(p_1,p_2) = d_2((r_1,s_1),(r_2,s_2))$ if $d_2((r_1,s_1),(r_2,s_2)) < \pi$, where d_2 is the Euclidean distance of $\mathbb{R}^2$ and $p_i = (\cos(r_i), \sin(r_i), s_i)$ for $i = 1,2$.

See also Besse (1978) and Pennec (2006) for a discussion on the volume density function.

2.1. The estimator

Consider a probability distribution with a density f on a d-dimensional Riemannian manifold (M,g). Let $X_1, \cdots, X_n$ be i.i.d random object taking values on M with density f. A natural extension of the estimator proposed by Wagner (1975) in the context of a Riemannian manifold is to consider the following estimator

$$\widehat{f}_n(p) = \frac{1}{nH_n^d(p)} \sum_{j=1}^{n} \frac{1}{\theta_{X_j}(p)} K\left(\frac{d_g(p,X_j)}{H_n(p)}\right),$$

where $K : \mathbb{R} \to \mathbb{R}$ is a non-negative function with compact support, $\theta_p(q)$ denotes the volume density function on (M,g) and $H_n(p)$ is the distance d_g between p and the k-nearest neighbour of p among $X_1, \ldots, X_n$, and $k = k_n$ is a sequence of non-random integers such that $\lim_{n\to\infty} k_n = \infty$.

As we mention above, the volume density function is not defined for all p and q. Therefore, in order to guarantee the well definition of the estimator we consider a modification of the proposed estimator. Using the fact that the kernel K has compact

support, we consider as bandwidth $\zeta_n(p) = \min\{H_n(p), inj_g M\}$ instead of $H_n(p)$. Thus, the kernel only considers the points X_i such that $d_g(X_i, p) \leq \zeta_n(p)$ that are smaller than $inj_g M$ and for these points, the volume density function is well defined. Hence, the k-nearest neighbour kernel type estimator is defined as follows,

$$\widehat{f}_n(p) = \frac{1}{n\zeta_n^d(p)} \sum_{j=1}^{n} \frac{1}{\theta_{X_j}(p)} K\left(\frac{d_g(p, X_j)}{\zeta_n(p)}\right), \tag{1}$$

where $\zeta_n(p) = \min\{H_n(p), inj_g M\}$.

Remark 2.1.1. If (M, g) is a compact Riemannian manifold and its sectional curvature is not bigger than $a > 0$, then we know by the Lemma of Klingerberg (see Gallot, Hulin, Lafontaine (2004)) that $inj_g M \geq \min\{\pi/\sqrt{a}, l/2\}$ where l is the length of the shortest closed geodesic in (M, g).

3. Asymptotic results

Denote by $C^\ell(U)$ the set of ℓ times continuously differentiable functions from U to $\mathbb{R}$ where U is an open set of M. We assume that the measure induced by the probability P and by X is absolutely continuous with respect to the Riemannian volume measure dv_g, and we denote by f its density on M with respect to dv_g. More precisely, let $\mathscr{B}(M)$ be the Borel $\sigma-$field of M (the $\sigma-$field generated by the class of open sets of M). The random variable X has a probability density function f, i.e. if $\chi \in \mathscr{B}(M)$, $P(X^{-1}(\chi)) = \int_\chi f dv_g$.

3.1. Uniform consistency

We will consider the following set of assumptions in order to derive the strong consistency results of the estimate $\widehat{f}_n(p)$ defined in (1).

H1. Let M_0 be a compact set on M such that:

  i) f is a bounded function such that $\inf_{p \in M_0} f(p) = A > 0$.

  ii) $\inf_{p,q \in M_0} \theta_p(q) = B > 0$.

H2. For any open set U_0 of M_0 such that $M_0 \subset U_0$, f is of class C^2 on U_0.

H3. The sequence k_n is such that $k_n \to \infty$, $\frac{k_n}{n} \to 0$ and $\frac{k_n}{\log n} \to \infty$ as $n \to \infty$.

H4. $K : \mathbb{R} \to \mathbb{R}$ is a bounded nonnegative Lipschitz function of order one, with compact support $[0, 1]$ satisfying: $\int_{\mathbb{R}^d} K(\|\mathbf{u}\|) d\mathbf{u} = 1$, $\int_{\mathbb{R}^d} \mathbf{u} K(\|\mathbf{u}\|) d\mathbf{u} = 0$ and $0 < \int_{\mathbb{R}^d} \|\mathbf{u}\|^2 K(\|\mathbf{u}\|) d\mathbf{u} < \infty$.

H5. The kernel $K(u)$ verifies $K(uz) \geq K(z)$ for all $u \in (0, 1)$.

Remark 3.1.1. The fact that $\theta_p(p) = 1$ for all $p \in M$ guarantees that H1 ii) holds. The assumption H3 is usual when dealing with nearest neighbor and the assumption H4 is standard when dealing with kernel estimators.

Theorem 3.1.2. Assume that H1 to H5 holds, then we have that

$$\sup_{p \in M_0} |\widehat{f}_n(p) - f(p)| \xrightarrow{a.s.} 0.$$

3.2. Asymptotic normality

To derive the asymptotic distribution of the regression parameter estimates we will need two additional assumptions. We will denote with $\mathscr{V}_r$ the Euclidean ball of radius r centered at the origin and with $\lambda(\mathscr{V}_r)$ its Lebesgue measure.

H5. $f(p) > 0$, $f \in C^2(V)$ with $V \subset M$ an open neighborhood of M and the second derivative of f is bounded.

H6. The sequence k_n is such that $k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$ and there exists $0 \leq \beta < \infty$ such that $\sqrt{k_n n^{-4/(d+4)}} \to \beta$ as $n \to \infty$.

H7. The kernel verifies

i) $\int K_1(\|\mathbf{u}\|)\|\mathbf{u}\|^2 d\mathbf{u} < \infty$ as $s \to \infty$ where $K_1(\mathbf{u}) = K'(\|\mathbf{u}\|)\|\mathbf{u}\|$.

ii) $\|\mathbf{u}\|^{d+1} K_2(\mathbf{u}) \to 0$ as $\|\mathbf{u}\| \to \infty$ where $K_2(\mathbf{u}) = K''(\|\mathbf{u}\|)\|\mathbf{u}\|^2 - K_1(\mathbf{u})$

Remark 3.2.1. Note that $div(K(\|\mathbf{u}\|)\mathbf{u}) = K'(\|\mathbf{u}\|)\|\mathbf{u}\| + d\, K(\|\mathbf{u}\|)$, then using the divergence Theorem, we get that $\int K'(\|\mathbf{u}\|)\|\mathbf{u}\| d\mathbf{u} = \int_{\|\mathbf{u}\|=1} K(\|\mathbf{u}\|)\mathbf{u}\frac{\mathbf{u}}{\|\mathbf{u}\|} d\mathbf{u} - d \int K(\|\mathbf{u}\|) d\mathbf{u}$. Thus, the fact that K has compact support in $[0,1]$ implies that $\int K_1(\mathbf{u}) d\mathbf{u} = -d$.

On the other hand, note that $\nabla(K(\|\mathbf{u}\|)\|\mathbf{u}\|^2) = K_1(\|\mathbf{u}\|)\mathbf{u} + 2K(\|\mathbf{u}\|)\mathbf{u}$ and by $H4$ we get that $\int K_1(\|\mathbf{u}\|)\mathbf{u}\, d\mathbf{u} = 0$.

Theorem 3.2.2. Assume H4 to H7. Then we have that

$$\sqrt{k_n}(\widehat{f}_n(p) - f(p)) \xrightarrow{\mathscr{D}} \mathscr{N}(b(p), \sigma^2(p))$$

with

$$b(p) = \frac{1}{2} \frac{\beta^{\frac{d+4}{d}}}{(f(p)\lambda(\mathscr{V}_1))^{\frac{2}{d}}} \int_{\mathscr{V}_1} K(\|\mathbf{u}\|) u_1^2 \, d\mathbf{u} \sum_{i=1}^{d} \frac{\partial f \circ \psi^{-1}}{\partial u_i \partial u_i}|_{u=0}$$

and

$$\sigma^2(p) = \lambda(\mathscr{V}_1) f^2(p) \int_{\mathscr{V}_1} K^2(\|\mathbf{u}\|) d\mathbf{u}$$

where $\mathbf{u} = (u_1, \ldots, u_d)$ and $(B_h(p), \psi)$ is any normal coordinate system.

In order to derive the asymptotic distribution of $\widehat{f}_n(p)$, we will study the asymptotic behaviour of $h_n^d / \zeta_n^d(p)$ where $h_n^d = k_n/(nf(p)\lambda(\mathscr{V}_1))$. Note that if we consider $\widetilde{f}_n(p) = k_n/(n\zeta_n^d(p)\lambda(\mathscr{V}_1))$, $\widehat{f}_n(p)$ is a consistent estimator of $f(p)$ (see the proof of Theorem 3.1.2). The next Theorem states that this estimator is also asymptotically normally distributed as in the Euclidean case.

Theorem 3.2.3. Assume H4 to H6, and let $h_n^d = k_n/(nf(p)\lambda(\mathscr{V}_1))$. Then we have that

$$\sqrt{k_n}\left(\frac{h_n^d}{\zeta_n^d(p)} - 1\right) \xrightarrow{\mathscr{D}} N(b_1(p), 1)$$

with

$$b_1(p) = \left(\frac{\beta^{\frac{d+4}{2}}}{f(p)\mu(\mathscr{V}_1)}\right)^{\frac{2}{d}} \left\{\frac{\tau}{6d+12} + \frac{\int_{\mathscr{V}_1} u_1^2 d\mathbf{u}\, L_1(p)}{f(p)\mu(\mathscr{V}_1)}\right\}$$

where $\mathbf{u} = (u_1, \ldots, u_d)$, τ is the scalar curvature of (M, g), i.e. the trace of the Ricci tensor,

$$L_1(p) = \sum_{i=1}^{d} \left(\frac{\partial^2 f \circ \psi^{-1}}{\partial u_i u_i}\Big|_{u=0} + \frac{\partial f \circ \psi^{-1}}{\partial u_i}\Big|_{u=0} \frac{\partial \theta_p \circ \psi^{-1}}{\partial u_i}\Big|_{u=0} \right)$$

and $(B_h(p), \psi)$ is any normal coordinate system.

4. Simulations

This section contains the results of a simulation study designed to evaluate the performance of the estimator defined in the Section 2.1. We consider three models in two different Riemannian manifolds, the sphere and the cylinder endowed with the metric induced by the canonical metric of $\mathbb{R}^3$. We performed 1000 replications of independent samples of size $n = 200$ according to the following models:

> **Model 1 (in the sphere)**: The random variables X_i for $1 \leq i \leq n$ are i.i.d. Von Mises distribution $VM(\mu, \kappa)$ i.e.

$$f_{\mu,\kappa}(X) = \left(\frac{\kappa}{2}\right)^{1/2} I_{1/2}(\kappa) \exp\{\kappa X^{\mathsf{T}} \boldsymbol{\mu}\},$$

with $\boldsymbol{\mu}$ is the mean parameter, $\kappa > 0$ is the concentration parameter and $I_{1/2}(\kappa) = \left(\frac{\kappa\pi}{2}\right)\sinh(\kappa)$ stands for the modified Bessel function. This model has many important applications, as described in Jammalamadaka and Sengupta (2001) and Mardia and Jupp (2000). We generate a random sample following a Von Mises distribution with mean $(0,0,1)$ and concentration parameter 3.

Model 2 (in the sphere): We simulate i.i.d. random variables Z_i for $1 \leq i \leq n$ following a multivariate normal distribution of dimension 3, with mean $(0,0,0)$ and covariance matrix equal to the identity. We define $X_i = \frac{Z_i}{\|Z_i\|}$ for $1 \leq i \leq n$, therefore the variables X_i follow a uniform distribution in the two-dimensional sphere.

Model 3 (in the cylinder): We consider random variables $X_i = (\mathbf{y}_i, t_i)$ taking values in the cylinder $S^1 \times \mathbb{R}$. We generated the model proposed by Mardia and Sutton (1978) where,

$$\mathbf{y}_i = (\cos(\theta_i), \sin(\theta_i)) \sim VM((-1,0),5)$$

$$t_i|\mathbf{y}_i \sim N(1+2\sqrt{5}\cos(\theta_i), 1).$$

Some examples of variables with this distribution can be found in Mardia and Sutton (1978).

In all cases, for smoothing procedure, the kernel was taken as the quadratic kernel $K(t) = (15/16)(1-t^2)^2 I(|x| < 1)$. We have considered a grid of equidistant values of k between 5 and 150 of length 20.

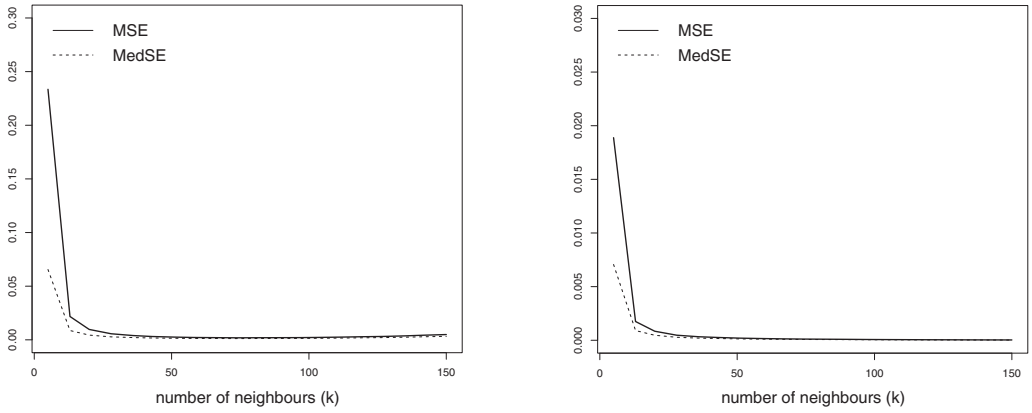
To study the performance of the estimators of the density function f, denoted by $\widehat{f}_n$, we have considered the mean square error (MSE) and the median square error (MedSE), i.e,

$$\mathrm{MSE}(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} [\widehat{f}_n(X_i) - f(X_i)]^2 \,.$$

$$\mathrm{MedSE}(\widehat{f}_n) = \mathrm{median}\, |\widehat{f}_n(X_i) - f(X_i)|^2 \,.$$

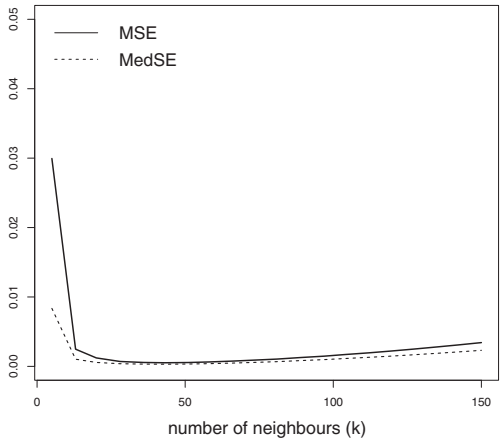
Figure 1 gives the values of the MSE and MedSE of $\widehat{f}_n$ in the sphere model considering different numbers of neighbours, while Figure 2 shows the cylinder model. The simulation study confirms the good behaviour of k-nearest neighbour estimators, under the

different models considered. In all cases, the estimators are stable under large numbers of neighbours. However, as expected, the estimators using a small number of neighbours have a poor behaviour, because in the neighborhood of each point there is a small number of samples.



a) *b)*

Figure 1: *The nonparametric density estimator using different numbers of neighbours, a) the Von Mises model and b) the uniform model.*



c)

Figure 2: *The nonparametric density estimator using different numbers of neighbours in the cylinder.*

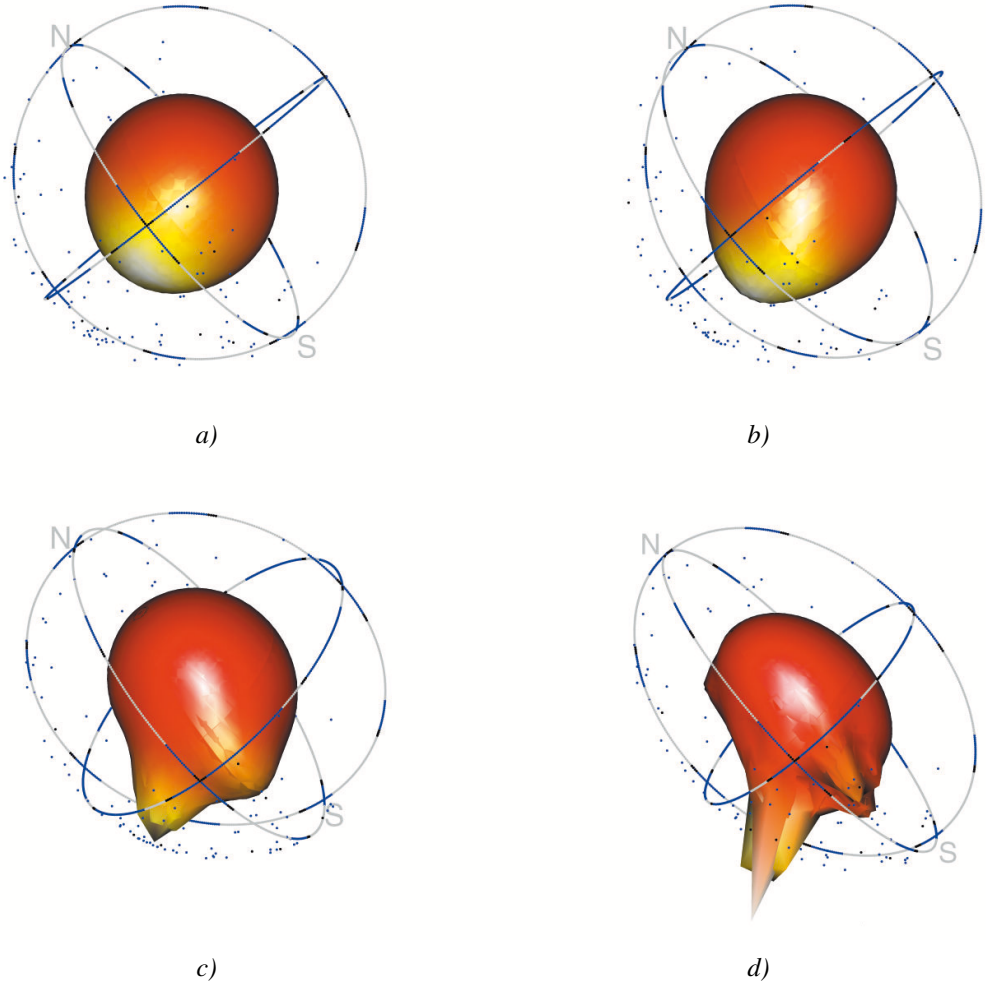


Figure 3: *The nonparametric density estimator using different numbers of neighbours, a) $k = 75$, b) $k = 50$, c) $k = 25$ and d) $k = 10$.*

5. Real Example

5.1. Paleomagnetic data

The need for statistical analysis of paleomagnetic data is well known. Since the work developed by Fisher (1953), the study of parametric families was considered a principal tools to analyse and quantify this type of data (see Cox and Doell (1960), Butler (1992) and Love and Constable (2003)). In particular, our proposal allows to explore the nature of directional dataset that include paleomagnetic data without making any parametric assumptions.

In order to illustrate the k-nearest neighbor kernel type estimator on the two-dimensional sphere, we illustrate the estimator using a paleomagnetic data set studied by Fisher, Lewis, and Embleton (1987). The data set consists of $n = 107$ sites from specimens of Precambrian volcanos with measurements of magnetic remanence. The data set contains two variables corresponding to the directional component on a longitude scale, and the directional component on a latitude scale. The original data set is available in the package sm of R.

To calculate the estimators the volume density function and the geodesic distance were taken as in Section 2 and we considered the quadratic kernel $K(t) = (15/16)$ $(1-t^2)^2 I(|x| < 1)$. In order to analyse the sensitivity of the results with respect to the number of neighbours, we plot the estimator using different bandwidths. The results are shown in Figure 3.

The real data were plotted in blue and with a large radius in order to obtain a better visualization. The Equator line, the Greenwich meridian and a second meridian are in gray while the north and south poles are denoted with the capital letters N and S respectively. The levels of concentration of measurements of magnetic remanence are shown in yellow for high levels and in red for lowest density levels. Also, the levels of concentration of measurements of magnetic remanence were illustrated with relief on the sphere, which emphasizes high density levels and the form of the density function.

As in the Euclidean case a large number of neighbours produces estimators with small variance but high bias, while small values produce more wiggly estimators. This fact shows the need of the implementation of a method to select the adequate bandwidth for this estimator. However, this requires further careful investigation and is beyond the scope of this paper.

5.2. Meteorological data

In this section we consider a real data set collected in the meteorological station "Agüita de Perdiz", located in Viedma, province of Río Negro, Argentine. The data set consists of wind directions and temperatures during January 2011 and contains 1326 observations that were registered with a frequency of thirty minutes. We note that the considered variables belong to a cylinder with radius 1.

As in the previous section, we consider the quadratic kernel and we took the density function and the geodesic distance as in Section 2. Figure 4 shows the result of the estimation, the colour and form of the graphic was constructed as in the previous example.

It is important to remark that the measurement devices of wind direction do not present a sufficient precision to avoid repeated data. Therefore, we consider the proposal given in García-Portugués et al. (2011) to solve this problem. The proposal consists in perturbing the repeated data as follows $\widetilde{r}_i = r_i + \xi\varepsilon_i$, where r_i denotes the wind direction measurements and ε_i, for $i = 1,\ldots,n$ were independently generated from a von Mises distribution with $\mu = (1,0)$ and $\kappa = 1$. The selection of the perturbation scale ξ was taken as $\xi = n^{-1/5}$, as in García-Portugués et al. (2011) where in this case $n = 1326$.

The work of García-Portugués et al. (2011) contains other nice real example where the proposed estimator can be applied. They considered a naive density estimator applied to wind directions and SO_2 concentrations, which allows one to explore high levels of contamination.

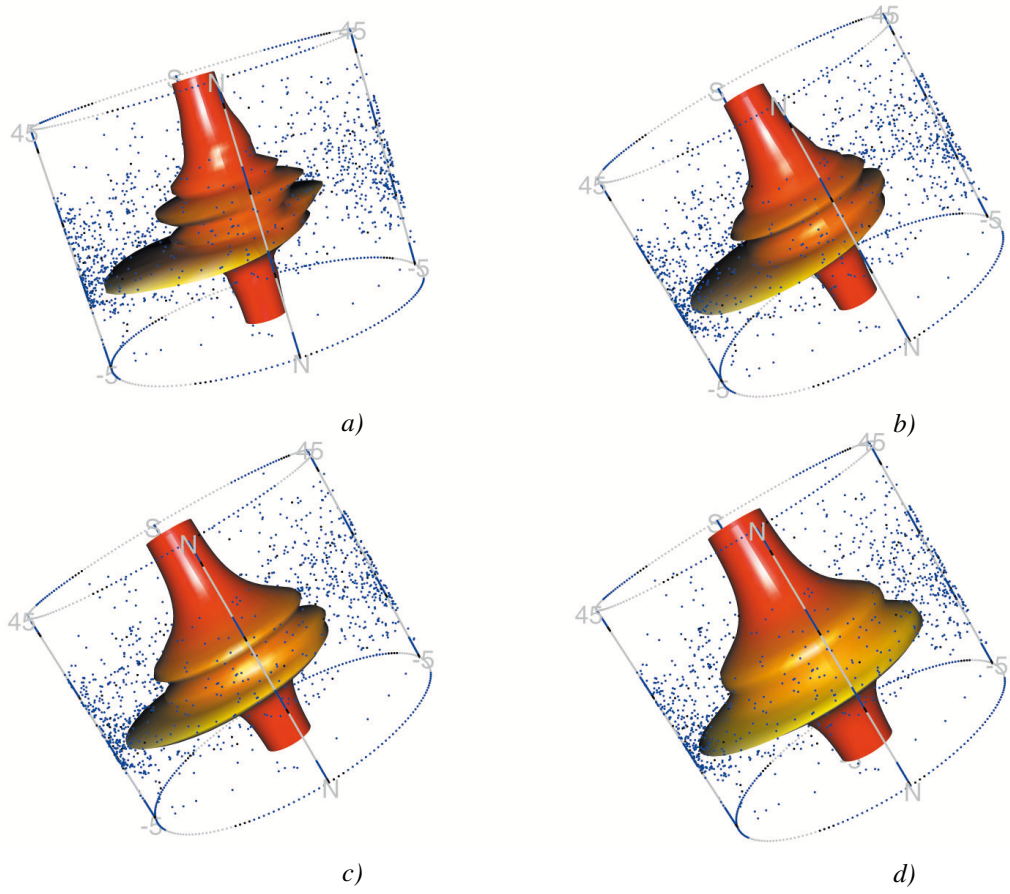


Figure 4: *The nonparametric density estimator using different numbers of neighbours, a)* $k = 75$*, b)* $k = 150$*, c)* $k = 300$ *and d)* $k = 400$*.*

In Figure 4 we can see that the lowest temperatures are more probable when the wind comes from an easterly direction. However, the highest temperature does not seem to have correlation with the wind direction. Also, note that in Figure 4 we can see two modes corresponding to the minimum and maximum daily temperatures.

These examples show the usefulness of the proposed estimator for the analysis and exploration of these type of data set.

Appendix

Proof of Theorem 3.1.2.

Let

$$f_n(p,\delta_n) = \frac{1}{n\delta_n^d}\sum_{i=1}^{n}\frac{1}{\theta_{X_i}(p)}K\left(\frac{d_g(p,X_i)}{\delta_n}\right).$$

Note that if $\delta_n = \delta_n(p)$ verifies $\delta_{1n} \leq \delta_n(p) \leq \delta_{2n}$ for all $p \in M_0$ where δ_{1n} and δ_{2n} satisfy $\delta_{in} \to 0$ and $\dfrac{n\delta_{in}^d}{\log n} \to \infty$ as $n \to \infty$ for $i = 1,2$ then by Theorem 3.2 in Henry and Rodriguez (2009) we have that

$$\sup_{p\in M_0}|f_n(p,\delta_n) - f(p)| \xrightarrow{a.s.} 0 \tag{2}$$

For each $0 < \beta < 1$ we define,

$$f_n^-(p,\beta) = \frac{1}{nD_n^+(\beta)^d}\sum_{i=1}^{n}\frac{1}{\theta_{X_i}(p)}K\left(\frac{d_g(p,X_i)}{D_n^-(\beta)}\right) = f_n^-(p,D_n^-(\beta)^d)\frac{D_n^-(\beta)^d}{D_n^+(\beta)^d}.$$

$$f_n^+(p,\beta) = \frac{1}{nD_n^-(\beta)^d}\sum_{i=1}^{n}\frac{1}{\theta_{X_i}(p)}K\left(\frac{d_g(p,X_i)}{D_n^+(\beta)}\right) = f_n^+(p,D_n^+(\beta)^d)\frac{D_n^+(\beta)^d}{D_n^-(\beta)^d}.$$

where $D_n^-(\beta) = \beta^{1/2d}h_n$, $D_n^+(\beta) = \beta^{-1/2d}h_n$ and $h_n^d = k_n/(n\lambda(\mathscr{V}_1)f(p))$ with $\lambda(\mathscr{V}_1)$ denote the Lebesgue measure of the ball in $\mathbb{R}^d$ with radius r centred at the origin. Note that

$$\sup_{p\in M_0}|f_n^-(p,\beta) - \beta f(p)| \xrightarrow{a.s.} 0 \text{ and } \sup_{p\in M_0}|f_n^+(p,\beta) - \beta^{-1}f(p)| \xrightarrow{a.s.} 0. \tag{3}$$

For all $0 < \beta < 1$ and $\varepsilon > 0$ we define

$$S_n^-(\beta,\varepsilon) = \{w : \sup_{p\in M_0}|f_n^-(p,\beta) - f(p)| < \varepsilon\},$$

$$S_n^+(\beta,\varepsilon) = \{w : \sup_{p\in M_0}|f_n^+(p,\beta) - f(p)| < \varepsilon\},$$

$$S_n(\varepsilon) = \{w : \sup_{p\in M_0}|\widehat{f}_n(p) - f(p)| < \varepsilon\},$$

$$A_n(\beta) = \{f_n^-(p,\beta) \leq \widehat{f}_n(p) \leq f_n^+(p,\beta)\}$$

Then, $A_n(\beta) \cap S_n^-(\beta,\varepsilon) \cap S_n^+(\beta,\varepsilon) \subset S_n(\varepsilon)$. Let $A = \sup_{p\in M_0} f(p)$. For $0 < \varepsilon < 3A/2$ and $\beta_\varepsilon = 1 - \frac{\varepsilon}{3A}$ consider the following sets

$$G_n(\varepsilon) = \left\{ w:\ D_n^-(\beta_\varepsilon) \le \zeta_n(p) \le D_n^+(\beta_\varepsilon) \ \text{ for all } p \in M_0 \right\}$$
$$G_n^-(\varepsilon) = \{ \sup_{p\in M_0} |f_n^-(p,\beta_\varepsilon) - \beta_\varepsilon f(p)| < \frac{\varepsilon}{3} \}$$
$$G_n^+(\varepsilon) = \left\{ \sup_{p\in M_0} |f_n^+(p,\beta_\varepsilon) - \beta_\varepsilon^{-1} f(p)| < \frac{\varepsilon}{3} \right\}.$$

Then we have that $G_n(\varepsilon) \subset A_n(\beta_\varepsilon)$, $G_n^-(\varepsilon) \subset S_n^-(\beta_\varepsilon,\varepsilon)$ and $G_n^+(\varepsilon) \subset S_n^+(\beta_\varepsilon,\varepsilon)$. Therefore, $G_n(\varepsilon) \cap G_n^-(\varepsilon) \cap G_n^+(\varepsilon) \subset S_n(\varepsilon)$.

On the other hand, using that $\lim_{r\to 0} V(B_r(p))/r^d \mu(\mathscr{V}_1) = 1$, where $V(B_r(p))$ denotes the volume of the geodesic ball centered at p with radius r (see Gray and Vanhecke (1979)) and similar arguments those considered in Devroye and Wagner (1977), we get that

$$\sup_{p\in M_0} \left| \frac{k_n}{n\lambda(\mathscr{V}_1) f(p) H_n^d(p)} - 1 \right| \xrightarrow{a.s.} 0.$$

Recall that $inj_g M > 0$ and $H_n^d(p) \xrightarrow{a.s.} 0$. Then for straightforward calculations we obtained that $\sup_{p\in M_0} \left| \frac{k_n}{n\lambda(\mathscr{V}_1) f(p) \zeta_n^d(p)} - 1 \right| \xrightarrow{a.s.} 0$. Thus, $I_{G_n^c(\varepsilon)} \xrightarrow{a.s.} 0$ and (3) imply that $I_{S_n^c(\varepsilon)} \xrightarrow{a.s.} 0$. □

Proof of Theorem 3.2.2.

A Taylor expansion of second order gives

$$\sqrt{k_n} \left\{ \frac{1}{n\zeta_n^d(p)} \sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} K\left(\frac{d_g(p,X_j)}{\zeta_n(p)} \right) - f(p) \right\} = A_n + B_n + C_n$$

where

$$A_n = (h_n^d/\zeta_n^d(p)) \sqrt{k_n} \left\{ \frac{1}{nh_n^d} \sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} K\left(\frac{d_g(p,X_j)}{h_n} \right) - f(p) \right\},$$
$$B_n = \sqrt{k_n}((h_n^d/\zeta_n^d(p)) - 1) \left\{ f(p) + \frac{[(h_n/\zeta_n(p)) - 1] h_n^d}{[(h_n^d/\zeta_n^d(p)) - 1]\zeta_n^d(p)} \ \frac{1}{nh_n^d} \sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} K_1\left(\frac{d_g(p,X_j)}{\zeta_n(p)} \right) \right\}$$

and

$$C_n = \sqrt{k_n}((h_n^d/\zeta_n^d(p)) - 1) \frac{[(h_n/\zeta_n(p)) - 1]^2}{2[(h_n^d/\zeta_n^d(p)) - 1]} \ \frac{1}{n\zeta_n^d(p)} \sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} K_2\left(\frac{d_g(p,X_j)}{\xi_n} \right) [\xi_n/h_n]^2$$

with $h_n^d = k_n/nf(p)\lambda(\mathscr{V}_1)$ and $\min(h_n,\zeta_n) \leq \xi_n \leq \max(h_n,\zeta_n)$. Note that H6 implies that h_n satisfies the necessary hypothesis given in Theorem 4.1 in Rodriguez and Henry (2009), in particular

$$\sqrt{nh_n^{d+4}} \to \beta^{\frac{d+4}{d}}(f(p)\lambda(\mathscr{V}_1))^{-\frac{d+4}{2d}}.$$

By the Theorem and the fact that $h_n/\zeta_n(p) \xrightarrow{p} 1$, we obtain that A_n converges to a normal distribution with mean $b(p)$ and variance $\sigma^2(p)$. Therefore it is enough to show that B_n and C_n converges to zero in probability.

Note that $\frac{(h_n/H_n(p))-1}{(h_n^d/\zeta_n^d(p))-1} \xrightarrow{p} d^{-1}$ and by similar arguments those considered in Theorem 3.1 in Pelletier (2005) and Remark 3.2.1 we get that

$$\frac{1}{nh_n^d}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}K_1\left(\frac{d_g(p,X_j)}{\zeta_n(p)}\right) \xrightarrow{p} \int K_1(\mathbf{u})d\mathbf{u}f(p) = -d\, f(p).$$

Therefore, by Theorem 3.2.3, we obtain that $B_n \xrightarrow{p} 0$. As ξ_n/h_n converges to one in probability, in order to concluded the proof, it remains to prove that

$$\frac{1}{n\zeta_n^d(p)}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}|K_2\left(d_g(p,X_j)/\xi_n\right)|$$

is bounded in probability.

By H7, there exits $r>0$ such that $|t|^{d+1}|K_2(t)| \leq 1$ if $|t| \geq r$. Let $C_r = (-r,r)$, then we have that

$$\begin{aligned}\frac{1}{n\zeta_n^d(p)}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}\left|K_2\left(\frac{d_g(p,X_j)}{\xi_n}\right)\right| \;&\leq\; \frac{\sup_{|t|\leq r}|K_2(t)|}{n\zeta_n^d(p)}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}I_{C_r}\left(\frac{d_g(p,X_j)}{\xi_n}\right)\\ &+\; \frac{1}{n\zeta_n^d(p)}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}I_{C_r^c}\left(\frac{d_g(p,X_j)}{\xi_n}\right)\left|\frac{d_g(p,X_j)}{\xi_n}\right|^{-(d+1)}\end{aligned}$$

As $\min(h_n,\zeta_n(p)) \leq \xi_n \leq \max(h_n,\zeta_n(p)) = \widetilde{\xi}_n$ it follows that

$$\begin{aligned}\frac{1}{n\zeta_n^d(p)} &\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}\left|K_2\left(\frac{d_g(p,X_j)}{\xi_n}\right)\right| \leq\\ &\leq\; \left(\frac{h_n}{\zeta_n(p)}\right)^d \sup_{|t|\leq r}|K_2(t)|\,\frac{1}{nh_n^d}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}I_{C_r}\left(\frac{d_g(p,X_j)}{h_n}\right)\\ &+\; \sup_{|t|\leq r}|K_2(t)|\,\frac{1}{n\zeta_n^d(p)}\sum_{j=1}^{n}\frac{1}{\theta_{X_j}(p)}I_{C_r}\left(\frac{d_g(p,X_j)}{\zeta_n(p)}\right)\end{aligned}$$

$$+ \left(\frac{h_n}{\zeta_n(p)}\right)^d \frac{1}{nh_n^d}\sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} I_{C_r^c}\left(\frac{d_g(p,X_j)}{h_n}\right)\left|\frac{d_g(p,X_j)}{h_n}\right|^{-(d+1)}\left|\frac{\widetilde{\xi}_n}{h_n}\right|^{(d+1)}$$

$$+ \frac{1}{n\zeta_n^d(p)}\sum_{j=1}^n \frac{1}{\theta_{X_j}(p)} I_{C_r^c}\left(\frac{d_g(p,X_j)}{\zeta_n(p)}\right)\left|\frac{d_g(p,X_j)}{\zeta_n(p)}\right|^{-(d+1)}\left|\frac{\widetilde{\xi}_n}{\zeta_n(p)}\right|^{(d+1)}$$

$$= C_{n1}+C_{n2}+C_{n3}+C_{n4}.$$

By similar arguments those considered in Theorem 3.1 in Pelletier (2005), we have that $C_{n1} \xrightarrow{p} f(p)\int I_{C_r}(s)ds$ and $C_{n3} \xrightarrow{p} f(p)\int I_{C_r^c}(s)|s|^{-(d+1)}ds$.

Finally, let $A_n^\varepsilon = \{(1-\varepsilon)h_n \le \zeta_n \le (1+\varepsilon)h_n\}$ for $0 \le \varepsilon \le 1$. Then for n large enough $P(A_n^\varepsilon) > 1-\varepsilon$ and in A_n^ε we have that

$$I_{C_r}\left(\frac{d_g(X_j,p)}{\zeta_n(p)}\right) \le I_{C_r}\left(\frac{d_g(X_j,p)}{(1+\varepsilon)h_n}\right),$$

$$I_{C_r^c}\left(\frac{d_g(X_j,p)}{\zeta_n(p)}\right)\left|\frac{d_g(X_j,p)}{\zeta_n(p)}\right|^{-(d+1)} \le I_{C_r^c}\left(\frac{d_g(X_j,p)}{(1-\varepsilon)h_n}\right)\left|\frac{d_g(X_j,p)}{(1-\varepsilon)h_n}\right|^{-(d+1)}\left|\frac{\zeta_n(p)}{(1-\varepsilon)h_n}\right|^{(d+1)}.$$

This fact and analogous arguments those considered in Theorem 3.1 in Pelletier (2005), allow to conclude the proof. □

Proof of Theorem 3.2.3.

Denote $b_n = h_n^d/(1+zk_n^{-1/2})$, then

$$P(\sqrt{k_n}(h_n^d/\zeta_n^d - 1) \le z) = P(\zeta_n^d \ge b_n) = P(H_n^d \ge b_n,\ inj_g M^d \ge b_n).$$

As $b_n \to 0$ and $inj_g M > 0$, there exists n_0 such that for all $n \ge n_0$ we have that

$$P(H_n^d \ge b_n,\ inj_g M^d \ge b_n) = P(H_n^d \ge b_n).$$

Let Z_i such that $Z_i = 1$ when $d_g(p,X_i) \le b_n^{1/d}$ and $Z_i = 0$ elsewhere. Thus, we have that $P(H_n^d \ge b_n) = P(\sum_{i=1}^n Z_i \le k_n)$. Let $q_n = P(d_g(p,X_i) \le b_n^{1/d})$. Note that $q_n \to 0$ and $nq_n \to \infty$ as $n \to \infty$, therefore

$$P\left(\sum_{i=1}^n Z_i \le k_n\right) = P\left(\frac{1}{\sqrt{nq_n}}\sum_{i=1}^n (Z_i - E(Z_i)) \le \frac{1}{\sqrt{nq_n}}(k_n - nq_n)\right).$$

Using the Lindeberg Central Limit Theorem we easily obtain that $(nq_n)^{-1/2}$ $\sum_{i=1}^{n}(Z_i - E(Z_i))$ is asymptotically normal with mean zero and variance one. Hence, it is enough to show that $(nq_n)^{-1/2}(k_n - nq_n) \xrightarrow{p} z + b_1(p)$.

Denote by $\mu_n = n\int_{B_{b_n^{1/d}}(p)} (f(q) - f(p))d\nu_g(q)$. Note that $\mu_n = n\ q_n - w_n$ with $w_n = n\ f(p)V(B_{b_n^{1/d}}(p))$. Thus,

$$\frac{1}{\sqrt{nq_n}}(k_n - nq_n) = w_n^{-1/2}(k_n - w_n)\left(\frac{w_n}{w_n+\mu_n}\right)^{1/2} + \frac{\mu_n}{w_n^{1/2}}\left(\frac{w_n}{w_n+\mu_n}\right)^{1/2}.$$

Let $(B_{b_n^{1/d}}(p), \psi)$ be a coordinate normal system. Then, we note that

$$\frac{1}{\lambda(\mathscr{V}_{b_n^{1/d}})}\int_{B_{b_n^{1/d}}(p)} f(q)d\nu_g(q) = \frac{1}{\lambda(\mathscr{V}_{b_n^{1/d}})}\int_{\mathscr{V}_{b_n^{1/d}}} f\circ\psi^{-1}(\mathbf{u})\theta_p\circ\psi^{-1}(\mathbf{u})d\mathbf{u}.$$

The Lebesgue's Differentiation Theorem and the fact that $\dfrac{V(B_{b_n^{1/d}}(p))}{\lambda(\mathscr{V}_{b_n^{1/d}})} \to 1$ imply that $\dfrac{\lambda_n}{w_n} \to 0$. On the other hand, from Gray and Vanhecke (1979), we have that

$$V(B_r(p)) = r^d\lambda(\mathscr{V}_1)(1 - \frac{\tau}{6d+12}r^2 + O(r^4)).$$

Hence, we obtain that

$$\begin{aligned} w_n^{-1/2}(k_n - w_n) &= \frac{w_n^{-1/2}\ k_n\ z\ k_n^{-1/2}}{1 + zk_n^{-1/2}} + \frac{w_n^{-1/2}\ \tau b_n^{2/d}\ k_n}{(6d+12)(1+zk_n^{-1/2})} + w_n^{-1/2}\ k_n\ O(b_n^{4/d}) \\ &= A_n + B_n + C_n. \end{aligned}$$

It's easy to see that $A_n \to z$ and $w_n^{-1/2}\ b_n^{2/d}\ k_n = \frac{k_n n^{-1/2} b_n^{2/d-1/2}}{(f(p)\lambda(\mathscr{V}_1))^{-2/d}}\left(\frac{b_n\lambda(\mathscr{V}_1)}{V(B_{b_n^{1/d}}(p))}\right)^{1/2}$, since $H6$ we obtain that $B_n \to \tau\ \beta^{(d+4)/d}/(6d+12)\ (f(p)\mu(\mathscr{V}_1))^{-2/d}$. A similar argument shows that $C_n \to 0$ and therefore we get that $w_n^{-1/2}(k_n - w_n) \to z + \beta^{\frac{d+4}{d}}\frac{\tau}{6d+12}(f(p)\lambda(\mathscr{V}_1))^{-d/2}$.

In order to concluded the proof we will show that

$$\frac{\mu_n}{w_n^{1/2}} \to \frac{\beta^{\frac{d+4}{d}}}{(f(p)\lambda(\mathscr{V}_1))^{(d+2)/d}}\int_{\mathscr{V}_1} u_1^2\ d\mathbf{u}\ L_1(p).$$

We use a second-order Taylor expansion that leads to,

$$\int_{B_{b_n^{1/d}}(p)} (f(q)-f(p))d\nu_g(q) = \sum_{i=1}^{d} \frac{\partial f\circ\psi^{-1}}{\partial u_i}|_{u=0} b_n^{1+1/d} \int_{\mathscr{V}_1} u_i\, \theta_p\circ\psi^{-1}(b_n^{1/d}\mathbf{u})\, d\mathbf{u}$$

$$+ \sum_{i,j=1}^{d} \frac{\partial^2 f\circ\psi^{-1}}{\partial u_i \partial u_j}|_{u=0} b_n^{1+2/d} \int_{\mathscr{V}_1} u_i u_j\, \theta_p\circ\psi^{-1}(b_n^{1/d}\mathbf{u})\, d\mathbf{u}$$

$$+ O(b_n^{1+3/d}).$$

Using again a Taylor expansion on $\theta_p\circ\psi^{-1}(\cdot)$ at 0 we have that

$$\int_{B_{b_n^{1/d}}(p)} (f(q)-f(p))d\nu_g(q) = b_n^{1+2/d} \int_{\mathscr{V}_1} u_1^2\, d\mathbf{u}\, L_1(p) + O(b_n^{1+3/d})$$

and by H6 the theorem follows. □

References

Bai, Z. D., Rao, C. and Zhao, L. (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27, 24–39.

Berger, M., Gauduchon, P. and Mazet, E. (1971). *Le Spectre d'une Variété Riemannienne*. Springer-Verlag.

Boothby, W. M. (1975). *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, New York.

Butler, R. (1992). *Paleomagnetism:Magnetic Domains to Geologic Terranes*. Blackwell Scientific Publications.

Cox, A. and Doell, R. (1960). Review of paleomagnetism, *Geological Society of America Bulletin*, 71, 645–768.

Devroye, L. and Wagner, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Annals of Statistics*, 3, 536–540.

Do Carmo, M. (1988). Geometria Riemaniana. *Proyecto Euclides, IMPA*. 2ª edición.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London, Ser. A*, 217, 295–305.

Fisher, N. I., T. Lewis and Embleton, B. J. J. (1987). *Statistical Analysis of Spherical Data*. New York: Cambridge University Press.

Gallot, S., Hulin, D. and Lafontaine, J. (2004). *Riemannian Geometry*. Springer. Third Edition.

García-Portugués, E., Crujeiras, R. and Gonzalez-Manteiga, W. (2011). Exploring wind direction and SO2 concentration by circular–linear density estimation. Prepint.

Goh, A. and Vidal, R. (2008). Unsupervised Riemannian clustering of probability density functions. *Lecture Notes In Artificial Intelligence*, 5211.

Gray, A. and Vanhecke, L. (1979). Riemannian geometry as determined by the volumes of small geodesic balls. *Acta Mathematica*, 142, 157–198.

Hall, P., Watson, G. S. and Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74, 751–762.

Henry, G. and Rodriguez, D. (2009). Kernel density estimation on Riemannian manifolds: asymptotic results. *Journal of Mathematical Imaging and Vision*, 43, 235–639.

Jammalamadaka, S. and SenGupta, A. (2001). Topics in circular statistics. *Multivariate Analysis*, 5. World Scientific, Singapore.

Joshi, J., Srivastava, A. and Jermyn, I. H. (2007). Riemannian analysis of probability density functions with applications in vision. *Proceedings of the IEEE Computer Vision and Pattern Recognition*.

Love, J. and Constable, C. (2003). Gaussian statistics for palaeomagnetic vectors. *Geophysical Journal International*, 152, 515–565.

Mardia, K. and Jupp, P. (2000). *Directional Data*. New York: Wiley.

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

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pelletier, B. (2005). Kernel density estimation on Riemannian manifolds. *Statistics and Probability Letters*, 73, 3, 297–304.

Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25, 127–154.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27, 832–837.

Wagner, T. (1975). Nonparametric estimates of probability densities. *IEEE Transactions on Information Theory IT*, 21, 438–440.

SORT 37 (2) July-December 2013, 131-152

Selection and pattern mixture models for modelling longitudinal data with dropout: An application study

A. Satty and H. Mwambi*

Abstract

Incomplete data are unavoidable in studies that involve data measured or observed longitudinally on individuals, regardless of how well they are designed. Dropout can potentially cause serious bias problems in the analysis of longitudinal data. In the presence of dropout, an appropriate strategy for analyzing such data would require the definition of a joint model for dropout and measurement processes. This paper is primarily concerned with selection and pattern mixture models as modelling frameworks that could be used for sensitivity analysis to jointly model the distribution for the dropout process and the longitudinal measurement process. We demonstrate the application of these models for handling dropout in longitudinal data where the dependent variable is missing across time. We restrict attention to the situation in which outcomes are continuous. The primary objectives are to investigate the potential influence that dropout might have or exert on the dependent measurement process based on the considered data as well as to deal with incomplete sequences. We apply the methods to a data set arising from a serum cholesterol study. The results obtained from these methods are then compared to help gain additional insight into the serum cholesterol data and assess sensitivity of the assumptions made. Results showed that additional confidence in the findings was gained as both models led to similar results when assessing significant effects, such as marginal treatment effects.

MSC: (MSC 2000) 97K80 or 46N30

Keywords: Identifying restrictions, under-identification, selection models, pattern mixture models, sensitivity analysis.

1. Introduction

In most longitudinal studies where data are collected over a sequence of time points, missing data are caused by individuals dropping out of the study prior to the time

* School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg, South Africa. E-mail: alisatty1981@gmail.com

Received: May 2012

Accepted: January 2013

at which the primary endpoint data would be collected. Missingness for longitudinal data often occurs as dropout that is a particular case of missing data. Furthermore, the resulting data obtained from such studies would have a particular type of missing data pattern; that is, a monotone missingness pattern, in which if an individual has missing values for a given time, no data can be obtained for all subsequent times for that individual. In this paper, our focus will be on this type of missing data pattern. Other types of missingness patterns are possible, such as intermittent missingness, but we focus on dropout which occurs most often in longitudinal studies. The mechanisms that lead to missing data are varied. Rubin (1976) and Little and Rubin (1987) classified these mechanisms into three possible categories, namely data missing completely at random, at random, or not at random. For longitudinal data, when data are missed at random or completely at random, available cases analysis, such as mixed models can be used. In contrast, when data are missed not at random, then using a standard mixed model without accounting for the missingness may lead to biased and inconsistent assessment of study results. Standard strategies of analysis currently assess non-random dropout by performing sensitivity analysis using analytical methods that incorporate non-random dropout in longitudinal data with and without a non-random component. Common families of models for data that are subject to dropout are selection and pattern mixture models.

Selection and pattern mixture models are two alternative and important approaches for dealing with longitudinal data when there are dropouts. They make empirically unverifiable assumptions and require extra constraints to identify the parameter estimates. Both models differ in the way the joint distribution of the measurement and dropout processes are factorized. However, other models that drive both the measurement process and dropout process, such shared-parameter models by Wu and Carrol (1988) and Wu and Bailey (1988, 1989) are also available. We restrict ourselves to the selection and pattern mixture models with dropout that falls under the monotone missing data pattern. A selection model factors the joint distribution into the marginal measurement model that describes the distribution of the complete measurements, and the dropout model that describes the conditional distribution of the dropout indicators, given the observed and unobserved measurements (Diggle and Kenward, 1994). However, in many discussions, for example, Diggle and Kenward (1994), Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005), the conclusions obtained from selection models depend on the assumptions made some of which cannot be investigated from the data under analysis. Early reference to such models is found in Heckman (1976) in the econometrics area. The use of pattern mixture models, on the other hand, was originally proposed by Little (1993, 1994) as a viable alternative to selection models. In this approach, models are under-identified; that is, for each dropout pattern the observed data does not provide direct information to identify the distributions for the incomplete patterns. Therefore, to overcome this problem, Little (1993, 1994) solves the under-identification problem through the use of identifying restrictions. Early applications concerning selection and pattern mixture models can be found in Marini et al. (1980) and Glynn et al. (1986).

Selection and pattern mixture models are somewhat opposite to each other. That is, these models exploit the conditional probability rule, but they do so in opposite ways. The marginal estimates in selection models can be derived directly, while pattern mixture models estimate the marginal parameters as a weighted average through pattern specific estimates (Little, 1995).

There are several studies in the literature which provide a comprehensive review of these models. The differences between selection models and pattern mixture models have been discussed in many works, for example, Glynn et al. (1986) and Little (1993, 1994). Little (1995) also made an important distinction between selection and pattern mixture models. A comparison of the conclusions based on the selection model with those based on the pattern mixture models have been discussed in Verbeke et al. (2001) and Michiels et al. (2002). Molenberghs et al. (1998a) contrast selection and pattern mixture models. Further discussion of these models can be found in McArdle and Hamagami (1992), Little and Wang (1996), Hedeker and Gibbons (1997), Hogan and Laird (1997), Kenward and Molenberghs (1999), Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005), Molenberghs and Kenward (2007) and Daniels and Hogan (2008). However, the approach by Daniels and Hogan (2008) is Bayesian based, which is not the focus of the current study.

This paper is primarily concerned with two attractive modelling frameworks to account for non-random dropout, namely selection and pattern mixture models. We demonstrate the application of selection and pattern mixture models for handling dropout in longitudinal data where the dependent variable is missing across time. In particular, we illustrate the application and results of analysis with these models. The underidentification in pattern mixture models is addressed through identifying restrictions, while the use of the selection model is based on Diggle and Kenward's (1994) model. We restrict our attention to the situation in which linear models are used and the outcomes are continuous. The primary objectives are to investigate the potential influence that dropout might exert on the dependent measurement on the considered data as well as how to deal with incomplete sequences. We relate the identified restrictions estimates using a pattern mixture model framework to their corresponding estimates using a selection model framework. We apply the methods to a data set arising from a serum cholesterol study. Section 2 describes the notation and general concepts based on the selection and pattern mixture models. In Section 3, we give a discussion of the two families of models that are used in the analysis, namely selection and pattern mixture models. An application study is provided in Section 4 including the description of the serum cholesterol data to which our methods will be applied. In addition, full analysis and results of the application is also given. Section 5 presents concluding remarks and discussion.

2. Notation and concepts

We introduce modelling incompleteness notation which is largely due to Rubin (1976) and Little and Rubin (1987). Let y_{ij} be the response of interest, for the ith study subject, where $i = 1,\ldots,N$, designed to be measured at occasion t_j, where $j = 1,\ldots,n$. In other words, the original intention was to have n observations per individual. However, due to dropout some individuals end up contributing less than n intended observations. Therefore, generally, we can assume that the ith individual is actually observed n_i times. For subject i and occasion j, define R_{ij}=1, if y_{ij} is observed, and 0, if not. We split y_{ij} into two sub-vectors, y_i^o and y_i^m, representing those y_{ij} for which R_{ij}=1, and R_{ij}=0, respectively. In addition, suppose the missing data occur due to dropout, then the measurements for each subject can be recorded up to a certain time point, after which all data are unobserved. In this case, a dropout indicator can then be defined as D_i, given by $D_i = 1 + \sum_{j=1}^{n} R_{ij}$, denoting the occasion at which dropout first occurs. In modelling a missing data process, it is often necessary to consider a joint model for the measurement process together with the dropout process. Therefore, we assume the full data density is given by

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}), \tag{1}$$

where $\boldsymbol{X}_i$ denotes the design matrix for fixed effects, $\boldsymbol{Z}_i$ denotes the design matrix for random effects, while $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ represent the vectors of parametrization for the joint distribution. In considering the above model in expression (1), we can factorize this joint density function in two possible ways that can facilitate modelling. Specifically, the selection and pattern mixture models mentioned earlier are defined by the conditional factorizations of the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$, and both are discussed in more detail in Little (1995) and stated briefly below. A selection model is based on the following factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}), \tag{2}$$

where the first factor in the above factorization represents the marginal density of the measurement process, while the second factor represents the density of the dropout process, conditional on the measurements. An alternative factorization based on the pattern mixture models (Little 1993, 1994) is of the form

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i \mid \boldsymbol{r}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\psi}). \tag{3}$$

This factorized density (3) can be seen as a mixture of the conditional distributions, and the model for the measurements depends on the particular missing data pattern. An excellent review of these models is given in Glynn et al. (1986), Little and Rubin (1987), Little (1993, 1994), Hogan and Laird (1997) and Ekholm and Skinner (1998).

The missing data processes have been developed by Rubin (1976) and Little and Rubin (1987) through the selection model framework. They make distinctions among different missing data processes. These processes can be formulated based on the second factor of model (2), i.e,

$$f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi}). \tag{4}$$

Thus, if the distribution of missingness process is reduced to $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i, \boldsymbol{X}_i, \boldsymbol{\psi})$, i.e., the process is independent of the measurements, then the process is defined as missing completely at random (MCAR). If the missingness probability depends on the observed measurement $\boldsymbol{y}_i^o$, but not on the unobserved measurements $\boldsymbol{y}_i^m$, i.e, $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o, \boldsymbol{X}_i, \boldsymbol{\psi})$, then the process is termed missing at random (MAR). Finally, data are missing not at random (MNAR) or exhibiting an informative process, when the missingness probability depends on the unobserved measurement, $\boldsymbol{y}_i^m$, and possibly on the observed measurement, $\boldsymbol{y}_i^o$, i.e., $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$. In other words, an informative process in expression (4) cannot be reduced.

3. Selection and pattern mixture models for modelling dropout

We consider the comparison between the selection and pattern mixture models concerning the significant characteristics, such as marginal treatment effects since such a comparison is a useful form of a sensitivity analysis. Specifically, we are interested in parametric selection and pattern-mixture models for modelling dropout. In the following, we briefly review these models.

3.1. Selection model

As mentioned above, a selection model factors the joint distribution into two parts: the marginal measurement model that describes the distribution of the complete measurements and the missingness model that describes the conditional distribution of the response indicators given the observed and unobserved measurements. In other words, in a selection model, we first specify a distribution for the measurement, then propose a manner in which the probability of being observed depends on the data. For continuous outcomes, using a selection model formulation as in equation (2), Diggle and Kenward (1994) combine the multivariate Gaussian linear model together with the dropout model. Similarly, we consider the measurement model to be of the linear mixed effects model (Laird and Ware, 1982). Recall that y_{ij} is the response of interest for the ith study subject, where $i = 1, \ldots, N$, at time point j, where $j = 1, \ldots, n_i$. More generally, the model for $\boldsymbol{y}_i$ the $(n_i \times 1)$ vector of responses for the ith subject can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{5}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known $(n_i \times p)$ and $(n_i \times q)$ design matrices for fixed and random effects, respectively, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of fixed effects, $\boldsymbol{b}_i$ is the $(q \times 1)$ vector of the random effects distributed as $N(\boldsymbol{0}, \boldsymbol{G})$, $\boldsymbol{\varepsilon}_i$ is the $(n_i \times 1)$ vector of the residual components distributed independently as $N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, $\boldsymbol{G}$ is the general $(q \times q)$ covariance matrix with (i, j)th element $d_{ij} = d_{ji}$ and $\boldsymbol{\Sigma}_i$ is the $(n_i \times n_i)$ error covariance matrix. Then, marginally, the responses $\boldsymbol{y}_i$ are distributed as independent normal $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{\mathsf{T}} + \boldsymbol{\Sigma}_i)$. Here, $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{H}_i + \tau^2\boldsymbol{I}$, where σ^2 denotes the variance of the serially correlated process, $\boldsymbol{H}_i = (h_{jk}) = (\rho(t_j, t_k))$ denotes the associated correlation matrix, τ^2 pertains to the measurement error variability and $\boldsymbol{I}$ is a $(n_i \times n_i)$ identity matrix.

We assume the missingness is due to dropout only, and that the first measurement $\boldsymbol{y}_{i1}$ is observed for each individual. Again, recall that D_i was defined as the dropout indicator which denote the occasion at which dropout first occurs. Now, let $D_i = d_i$ identify the dropout time for subject i, where $D_i = n+1$, if the sequence of measurement is complete. Therefore, the selection models introduced in equation (2) arise when the joint likelihood of the measurement and dropout processes is factorized as following

$$f(\boldsymbol{y}_i, D_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i, D_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) f(d_i \mid \boldsymbol{y}_i, \boldsymbol{\psi}).$$

The model for dropout process is based on a logistic regression for the probability of dropout at occasion j, given the subject was still in the study at the previous occasion. Let $g_i(y_{ij}, h_{ij})$ denote this probability, where h_{ij} represent the history of the measurement process. Thus, one can assume that $g_i(y_{ij}, h_{ij})$ satisfies the model

$$logit[g(h_{ij}, y_{ij})] = logit[p(D_i = j \mid D_i \geq j, h_{ij}, y_{ij})] = \eta(h_{ij}, y_{ij}), \tag{6}$$

where $\eta(h_{ij}, y_{ij})$ is the linear predictor depending on h_{ij} and y_{ij}. Modelling the dropout mechanism may be simplified in the expression in equation (6) by assuming $\eta(h_{ij}, y_{ij})$ depends only on the current measurement and the previous measurement y_{ij-1}, but not on future measurements or higher order history, with corresponding regression coefficients, ψ_1 and ψ_2. Dependence on future unobserved measurements is not easy to justify therefore it is not modelled here. Higher order history can be included, but we assume first order history for simplicity. This leads to the following logistic expression

$$logit[g(y_{i,j-1}, y_{ij})] = logit[p(D_i = j \mid y_{i,j-1}, y_{ij})] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{7}$$

Note that the linear predictor in equation (7) may depend on other covariates but in the current model we only include the constant ψ_0. According to Little and Rubin's (1987) terminology introduced in the previous section, and based on the expression in equation (7), it is clear that when both parameters ψ_1 and ψ_2 are equal to 0, the dropout

mechanism should be MCAR. However, when ψ_1 is not equal to 0, but ψ_2 equal to 0, the dropout mechanism is referred to as MAR, and finally, when ψ_2 is not equal to 0, dropout mechanism is referred to as MNAR. Here, we note that a likelihood ratio test (LRT) can be used to compare model fit under a model that assumes the missing data due to dropout are MCAR versus MAR (Diggle and Kenward, 1994). The LRT test statistic follows a null asymptotic χ^2_1 distribution. See, Diggle and Kenward (1994) and Molenberghs et al. (1997) for details on the derivation of this statistic. When the LRT test statistic is significant, then it suggests that the least restrictive of the two models is preferred; that is, the model that assumes the dropout is MAR. However, based on the argument of Jansen et al. (2006), we restate that the test for MAR against MNAR is not recommended using the LRT statistic via a model based on the Diggle and Kenward's (1994) type. This is because the behaviour of the LRT statistic for the MNAR parameter ψ_2 is non-standard since the availability of the information on ψ_2 is very rare and interwoven with other features of both measurement and dropout models (Jansen et al., 2006). This is specially true when one considers the model based on Diggle and Kenward type, but it is important to realize that their tests are conditional on the alternative model holding. According to Kenward (1998), such a distinction, between a MAR mechanism or a MNAR mechanism, can only be made using untestable modeling assumptions, such as the distributional form. Molenberghs and Kenward (2007) stated that the assumption giving arise to the dropout in a sample cannot be verified by the observed measurements and any test regarding the dropout process can be invalidated. This can be justified by the fact that parameters of the dropout model are dependent in part on dropout. Furthermore, unless one puts a priori belief in the posited MNAR model, the distinction (MAR/MNAR) is not possible, due to the fact that for any dropout model that assumes dropout are MNAR, there is a MAR model that provides exactly the same fit to the data, but the two models differ in the prediction of what is unobserved (Molenberghs et al., 2008). This problem of model identifiability poses a major complication when considering models for the dropout mechanism. Thus, one recommendation is to conduct a sensitivity analysis of the parameters of the measurement model across models that make different assumptions about the dropout process (see, Molenberghs and Kenward, 2007). Therefore, although the dropout process cannot be known via empirical examination, the analysis can be carried out to study differences in parameters estimates of the measurement process across varying assumptions about the dropout.

3.2. Pattern mixture model

Now, we shift our attention to the pattern mixture models that stratify subjects according to their missingness pattern. Under these models, the thinking is that, a separate model is fit for each pattern and then the results can be combined across the different patterns in order to derive an average estimate of the model parameters. Thus, in these mod-

els the joint distribution of the longitudinal measurements as well as the missing data indicators is divided into response pattern so that the distribution of the longitudinal measurements depends on the pattern of responses. As mentioned earlier, pattern mixture models are under-identified, or possess non-estimable parameters. Therefore, some identifying constraints are required. Little (1993, 1994) proposed the use of the identifying restrictions in which inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers to deal with under-identifiability of these models. In fact, there is an alternative major strategy simplified to deal with the under-identifiability of pattern mixture models, called model specification in which the different pattern allows for sharing of certain parameters so that the missing pattern can borrow information from patterns with more data points (Verbeke and Molenberghs, 2000). The advantage of this strategy is that the number of parameters decreases which is in general an issue with pattern mixture models. Detailed strategies of pattern mixture modelling are given in Verbeke and Molebergh (2000), Molenberghs et al. (2003) and Molenberghs and Kenward (2007).

Our primary concern in this study is to apply a pattern mixture model including the identifying restriction strategy. In doing so, we follow Verbeke and Molenberghs (2000) in illustrating the use of this strategy based on the results obtained by Molenberghs et al. (1998b). We are restricting attention to dropout which is a special case of monotone missingness. Let us assume that there are $t = 1,\ldots,T$ dropout patterns, where the dropout indicator, introduced in section 2, is $d = t + 1$. The complete data density, for pattern t, can be expressed as

$$f_t(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T) = f_t(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t) f_t(\boldsymbol{y}_{t+1},\ldots,\boldsymbol{y}_T \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_t). \tag{8}$$

It is clear from equation (8) that the first factor $f_t(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t)$ is identified from the observed data assuming the first factor is known, and modeled using the observed data. Whereas the second factor is not identifiable from the observed data. In order to identify the second component, the identifying restriction can be applied (Verbeke and Molenberghs, 2000). It is often necessary to base identification on all patterns for which a given component is identified. We denote this component by $\boldsymbol{y}_s$. Thus, this can be described as

$$f_t(\boldsymbol{y}_s \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}) = \sum_{j=s}^{T} \omega_{sj} f_j(\boldsymbol{y}_s \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}), \qquad s = t+1,\ldots,T. \tag{9}$$

We denote the set of ω_{sj} used by the vector $\boldsymbol{\omega}_s$, components of which are typically non-negative. Every $\boldsymbol{\omega}_s$ that sums to 1 provides a valid identification scheme. Hence, by incorporating equation (9) into (8), we have

$$f_t(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T) = f_t(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t)\prod_{s=0}^{T-t-1}\left[\sum_{j=T-s}^{T}\omega_{T-s,j}f_j(\boldsymbol{y}_{T-s}\,|\,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{T-s-1})\right] \tag{10}$$

To establish the complete data density, it is clear in equation (10) whose information can be used to complement the observed data density in pattern t. There are three sets of identifying restrictions associated with such choices of $\boldsymbol{\omega}_s$. Complete case missing values (CCMV) that were proposed by Little (1993) use the following identification

$$f_t(\boldsymbol{y}_s\,|\,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}) = f_T(\boldsymbol{y}_s\,|\,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}), \qquad s = t+1,\ldots,T,$$

corresponding to $\omega_{sT} = 1$ and all others equal 0, which is to say that identification is always done from the completers's pattern. Alternative restrictions are based on so called neighboring case missing values (NCMV). In these restrictions, the nearest identified pattern can be used as follows

$$f_t(\boldsymbol{y}_s\,|\,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}) = f_s(\boldsymbol{y}_s\,|\,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1}), \qquad s = t+1,\ldots,T.$$

The NCMV restriction follows from setting $\boldsymbol{\omega}_s = \mathbf{1}$ and all others equal $\mathbf{0}$. Finally, the third case for equation (10) is the available case missing values (ACMV). With regard to the corresponding $\boldsymbol{\omega}_s$ for ACMV, there always is a unique choice. Molenberghs et al. (1998b) show that the corresponding $\boldsymbol{\omega}_s$ can have the following components

$$\omega_{sj} = \frac{\alpha_j f_j(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1})}{\sum_{\ell=s}^{T}\alpha_\ell f_\ell(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{s-1})}, \qquad j = s,\ldots,T, \tag{11}$$

where α_j is the fraction of observations in pattern j. Clearly, ω_{sj} defined by (11) contains positive components and sum to 1. That is, a valid density function is defined. The selection and pattern mixture families can be connected using this MAR-ACMV link. The ACMV is reserved for a counterpart of MAR in the pattern mixture setting. A specific counterpart to MNAR selection models has been studied by Kenward et al. (2003).

4. Application to the NCGS data

4.1. The data

In this section, we describe the application of the selection and pattern mixtures models to data from the National Cooperative Gallstone Study (NCGS). Further background details of this experiment are given in Schoenfield and Lachin (1981) and in its accompanying discussion. In this study, 103 patients were randomly assigned to three

treatment groups corresponding to two doses; that is, high-dose (750 mg per day), low-dose (375 mg per day) and placebo, and were to be treated for four weeks. The current analysis is based on a subset of the data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups. In the NCGS it was suggested that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result, serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20 and 24 weeks of follow-up. In this experiment, many cholesterol measurements contain missing values because of missed visits, laboratory specimens were lost or inadequate, or patient follow-up was terminated. In addition, all subjects have observed values at time 6. One group of individuals received study treatment (drug and placebo), but dropped out of the study before the scheduled post-baseline time. These individuals dropped out of the study at time point 12. However, other individuals dropped out of the study either at time point 20 or 24. Therefore, the data presents three possible dropout patterns (dropout at time points 12, 20, or 24). All 103 patients are observed at the first occasion, whereas there are 93, 78 and 67 patients seen at the second, third and fourth weeks, respectively. The percentage of patients that are still in the study after each week is tabulated in Table 1 by treatment arm. Figure 1 represents the means across weeks by treatment group. A primary objective of this trial was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones. In what follows, we restrict our attention to examination of more than just this association between treatment and cholesterol. That is, we investigate the potential influence of dropout on the outcome of interest, the serum cholesterol, as well as the interactive effect of dropout with week and treatment-related influences on the serum cholesterol. The focus here will be on the parameter estimates, standard errors and p-values.

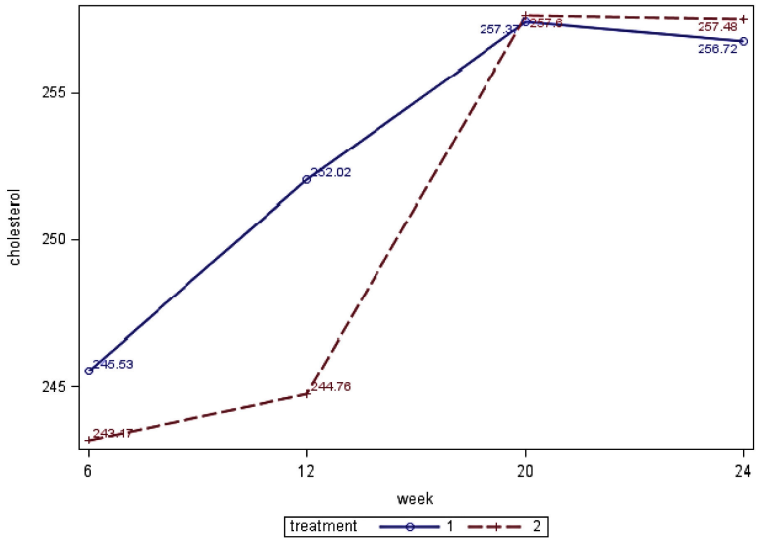


Figure 1: *Serum cholesterol data. Means across weeks by treatment (High dose "1" and Placebo "2").*

Table 1: *NCGS data: Percentage of patients still in study, by treatment arm (Drug = high-dose (750 mg per day)).*

week	drug	placebo
6	100	100
12	45	62
20	57	63
24	46	69

4.2. Fitting selection model

First, we consider fitting the selection model. In line with Diggle and Kenward (1994), we fit the selection models to the serum cholesterol data by combining the measurement model with the logistic regression for dropout model. The combined model for measurement/dropout will be fitted to the serum cholesterol by maximum likelihood using a generic function maximization routine. We use the linear mixed effects model of the form in equation (5) in order to obtain initial values for the parameters of the measurement model. In the fitted model, we assume different intercepts and treatment effects for each of the four time points, with a (4×4) unstructured variance-covariance matrix. Specifically, we consider a multivariate normal model, with unconstrained time trend under placebo and an occasion-specific treatment effect. Since serum cholesterol data consist of 103 subjects ($i = 1, \ldots, 103$) on four time points ($j = 6, 12, 20$ and 24), the model can be written as

$$Y_{ij} = \beta_{j1} + \beta_{j2} G_i + \varepsilon_{ij}, \tag{12}$$

where $G_i = 0$ for placebo and $G_1 = 1$ for active drug. In this way, the parameter estimates and standard errors as well as p-values for the eight mean model parameters can be obtained. To fit this model, we use SAS procedure MIXED with REPEATED statement. Next, we consider the dropout model. The dropout will be allowed to be independent of covariates. We fit the model with an intercept, an effect for previous outcome and an effect for the current unobserved measurement, corresponding to MCAR, MAR and MNAR, respectively. Dependence on future unobserved measurements is theoretically possible, but for simplicity, we model dependence on the current unobserved measurements. The probability of serum cholesterol is assumed to follow the logistic regression model (a commonly used model for dropout process, see, Molenberghs and Kenward, 2007) in equation (7). Therefore, the logistic regression model consists of three parameters; that is, an intercept (ψ_0), the effect of the measurement prior to dropout (ψ_1) and the effect of the measurement at the time of dropout (ψ_2). Consequently, for the four time points, the model can be expressed as follows

$$logit[g(y_{ij-1}, y_{ij})] = logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = \psi_0 + \psi_1 y_{ij-1} + \psi_2 y_{ij}, \quad j = 2,3,4,5. \tag{13}$$

Estimation of a selection model for MNAR can be seen as major complication as the dropout indicators depend on the unobserved measurement. For example, in the selection model mentioned above, the dropout indicators depend in part on the unobserved longitudinal measurements at the time of dropout. This leads to complexity in assessing the likelihood function, however, one that can be handled (Diggle and Kenward, 1994). Virtually, the parameters were estimated using a code written in SAS provided by Dmitrienko et al. (2005) that maximizes the log-likelihood for the model using PROC IML.

Table 2 shows the parameter estimates, standard errors and p-values of the fixed effects for the selection model, including the eight mean model parameters, all into the marginal measurement model as well as in the logistic dropout model. Interestingly, the comparison of the MCAR and MAR produces the same results when compared to those of the complete case analysis, except for negligible differences, as seen in the standard errors. These results are in line with theoretical findings, see, for example, Molenberghs and Kenward (2007). In the context of the assumed model, when examining the statistical significance of the results in the dropout model, the LRT test statistic for comparing the MAR and MCAR models is 17.1. The corresponding tail probability from χ^2 with 1 degree of freedom is $p < 0.001$ which is significant. This indicates that there is a significant evidence for MAR. In other words, dropout completely at random can be ruled out in the context of the assumed model. However, great care has to be taken with such a conclusion (Molenberghs et al., 1997; Molenberghs and Verbeke, 2005). To assess the mechanism that the dropout are MNAR, a problem occurs in that neither an LRT statistic between the models that assume the dropout is MAR against MNAR nor an assessment of ψ_2 relative to its standard error is reliable (Jansen et al., 2006). Consequently, it is not possible to verify the mechanism that the dropout is MNAR (see, Molenberghs, et al., 2008). One of our interests lies in the marginal treatment effect. There is no overall treatment effect and p-values between the three models do not vary too much.

However, the situation is different for the occasion-specific treatment effects considered here. For all weeks, all four p-values for the treatment effects indicate non-significance, whereas for all cases the p-values are certainly highly significance ($p < 0.0001$) for all intercepts. Now, we discuss factors which influence dropout. In doing so, in the full selection models, the logistic regression for dropout is modeled based on (13). As can be seen in Table 2, the maximum likelihood estimates for ψ_1 (0.04) and ψ_2 (-0.16) are not necessarily equal, however, their signs are different. This finding is not surprising. It confirms the argument put forward by Molenberghs et al. (2001a). They pointed out that since two subsequent measurements are usually positively correlated, the dropout model can depend on the increment, i.e., $y_{ij} - y_{i,j-1}$. The dropout estimated from the MNAR model is as follows:

Table 2: *Maximum likelihood for the parameter estimates-Est. (standard errors-s.e.) and p-values, resulting from the selection model under complete cases analysis, MCAR, MAR and MNAR.*

		Complete cases		MCAR		MAR		MNAR	
Effect	Parameter	Est.(s.e.)	p-value	Est.(s.e.)	p-value	Est.(s.e.)	p-value	Est.(s.e.)	p-value
Measurement model									
intercept$_6$	β_{11}	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001	243.17 (6.74)	< 0.0001
intercept$_{12}$	β_{21}	244.93 (6.46)	< 0.0001	244.93 (6.46)	< 0.0001	244.93 (6.45)	< 0.0001	243.98 (6.46)	< 0.0001
intercept$_{20}$	β_{31}	258.92 (6.70)	< 0.0001	258.92 (6.69)	< 0.0001	258.91 (6.70)	< 0.0001	258.13 (6.70)	< 0.0001
intercept$_{24}$	β_{41}	257.08 (8.03)	< 0.0001	257.08 (8.02)	< 0.0001	257.08 (8.03)	< 0.0001	256.28 (8.99)	< 0.0001
treatment effect$_6$	β_{12}	2.36 (8.69)	0.786	2.36 (8.69)	0.784	2.36 (8.68)	0.788	2.36 (8.69)	0.732
treatment effect$_{12}$	β_{22}	6.41 (8.36)	0.445	6.41 (8.32)	0.478	6.41 (8.36)	0.461	6.54 (8.36)	0.441
treatment effect$_{20}$	β_{32}	–5.77 (8.78)	0.512	-5.77 (8.78)	0.535	–5.77 (8.79)	0.427	–5.83 (8.79)	0.456
treatment effect$_{24}$	β_{42}	–2.06 (10.70)	0.847	–2.06 (10.70)	0.883	–2.06 (10.74)	0.856	–2.59 (10.79)	0.912
dropout model process									
intercept	ψ_0			–1.88(0.11)		–1.73(0.14)		–1.64(0.27)	
previous measurement	ψ_1					–0.20(0.05)		0.04(0.02)	
current measurement	ψ_2							–0.16(0.08)	
–2 log-likelihood		3313.3		3346.4		3329.3		3327.7	

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = -1.64 - 0.12y_{i,j-1} - 0.16(y_{ij} - y_{i,j-1}). \quad (14)$$

However, some insight into this fitted model can be obtained by the re-parameterizing the dropout parameters with respect to increment and the sum of the successive measurements. Therefore, we re-parameterize the dropout probabilities from the dropout model as in equation (13) to obtain

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = \vartheta_0 + \vartheta_1(y_{i,j} + y_{i,j-1}) + \vartheta_2(y_{ij} - y_{i,j-1}), \qquad j = 2,3,4,5. \quad (15)$$

Here, $\vartheta_1 = (\psi_1 + \psi_2)/2$ and $\vartheta_2 = (\psi_1 - \psi_2)/2$. These parameters represent dependence on level and increment in the serum cholesterol, and these quantities are likely to be much less strongly correlated than are y_{ij} and $y_{i,j-1}$. Rewriting the fitted MNAR model as in equation (15),

$$logit[p(D_i = j \mid y_{ij-1}, y_{ij})] = -1.64 - 0.06(y_{i,j} + y_{i,j-1}) + 0.10(y_{ij} - y_{i,j-1}), \quad (16)$$

suggests that the probability of dropout increases with larger negative increments. In other words, those patients with a greater increase in the overall level of the serum cholesterol from the previous week have a higher probability of dropping out of the experiment.

4.3. Fitting pattern mixture models

Now, we turn our attention to fitting the pattern mixture models using the strategy outlined in section 3, making CCMV, NCMV and ACMV identifying restrictions. To fit pattern mixture models through identifying restrictions, three steps in the analysis procedure are needed (For details of implementation, see Molenberghs and Kenward, 2007). First, fit the initial model to the observed data within each of the patterns

$$f_t(y_1, \ldots, y_t), \quad (17)$$

where $t = 1, \ldots, T$ indicate the observed dropout times in the data set. In this step, we fit a separate model within each pattern, then the resulting parameter estimates and their estimated variance-covariance matrices were used to extrapolate the patterns. Second, select an identification scheme to determine the conditional distributions of the unobserved measurements, given the observed ones

$$f_t(y_{t+1}, \ldots, y_T \mid y_1, \ldots, y_t). \quad (18)$$

Table 3: *Multiple imputation parameter estimates-Est. (standard errors-s.e.) and p-values resulting from the pattern mixture model using identifying restrictions ACMV, CCMV and NCMV.*

		ACMV		CCMV		NCMV	
Effect	Parameter	Est.(s.e.)	p-value	Est.(s.e.)	p-value	Est.(s.e.)	p-value
intercept$_6$	β_{11}	243.17 (6.74)	—	243.17 (6.74)	—	243.17 (6.74)	—
intercept$_{12}$	β_{21}	245.44 (7.06)	< 0.0001	245.36 (6.51)	< 0.0001	245.86 (6.55)	< 0.0001
intercept$_{20}$	β_{31}	255.78 (6.71)	< 0.0001	255.88 (6.83)	< 0.0001	257.99 (6.78)	< 0.0001
intercept$_{24}$	β_{41}	256.59 (8.10)	< 0.0001	256.99 (8.21)	< 0.0001	256.99 (8.08)	< 0.0001
treatment effect$_6$	β_{12}	2.36 (8.69)	—	2.36 (8.69)	—	2.36 (8.69)	—
treatment effect$_{12}$	β_{22}	6.23 (8.39)	0.540	6.16 (8.37)	0.539	5.41 (6.45)	0.716
treatment effect$_{20}$	β_{32}	–5.98 (8.85)	0.484	–5.13 (8.81)	0.475	–6.73 (8.82)	0.290
treatment effect$_{24}$	β_{42}	–2.18 (11.03)	0.627	–2.12 (11.64)	0.565	–1.87 (10.13)	0.629

As stated earlier, each of such conditional distributions is a mixture of known normal densities for continuous repeated measures. According to the weights w_s introduced in equation (9), an easy way to simulate values from the mixture distribution is to randomly select a component of the mixture and then draw from it. In this regard, we choose an identifying restriction, mentioned earlier, to define the conditional distributions of the unobserved measurements, conditional upon the observed ones. Third, fit a model to the so-augmented data. Multiple imputation (MI) can be used to fit such models by aiding to draw values for the unobserved components, conditional upon the observed outcomes and correct pattern-specific density in model (18). Here, we notice that MI is a simulation-based technique that imputes the missing values multiple times in order to construct multiple complete data sets. For more detail of this technique, we recommend Rubin's (1987) book. Analytically, MI involves three steps, imputation, analysis and combination. Thus, the identifying step corresponds to the so-called imputation step, and the final model corresponds to the analysis step. Finally, the combination step, is where the inferences from a number of imputations are drawn together and combined into a single one. The goal being to pool the simplicity of imputation strategies, without bias in both point estimates and measures of precision. After applying each of the three identifying restrictions, as above introduced, the same model as before being fitted (12) is analyzed. The model is parameterized as follows: different intercepts and treatment effects for each of the four time points, with a 4×4 unstructured covariance matrix for each pattern. We draw multiple imputations five times. The choice of five times imputations is considered adequate as the efficiency of a parameter estimate based on the number of imputations is $(1+\varsigma/M)^{-1}$, where ς is the rate of missing data and M is the number of imputations (Rubin, 1987). Rubin's (1987) simulation studies indicate that the number of imputations can generally be constrained to fewer than 10. Also, many statistical practices tend to support Rubin's heuristics of 3 to 10 imputations. In general, Schafer and Olsen (1998) recommended the use of M=5 before the results are combined. By this rationale, we achieve at least 97% efficiency as in our case the missing data rate is almost 17%. In this way, we ended up with totally five multiply-imputed data sets for each choice of identifying restriction strategy which can be analysed, using several possible models. Once the imputations have been generated, the final analysis model from each completed data sets is fitted and MI inference conducted. The parameter and precision estimates can be obtained using classical MI machinery. In particular, the asymptotic covariance matrix of the form

$$V = W + \left(\frac{M+1}{M}\right) B,$$

where W denotes the average within-imputation variance and B the between-imputation variance (Rubin, 1987). The analysis of identifying restrictions, fitting of imputed data, and a combination of the results into a single inference was implemented using the SAS macro. This SAS macro dealt with the analysis of the three types of identifying

restrictions as follows. First, fit the linear mixed model per pattern using PROCs SORT and MIXED. Second, complete the data using ACMV, CCMV and NCMV restrictions using PROCs IML and MI. Third, analyze the 5 complete data sets using a linear-mixed model using PROC MIXED. Fourth, combine the results from the 5 analyzed models using PROC MIANALYZE.

The results of the three types of identifying restrictions are listed in Table 3. Examining these results we see that the estimates for the corresponding parameters are comparable and their numerical values are indeed very close to each other under the three possible restrictions, namely ACMV, CCMV and NCMV. It can be seen from the analysis that the association p-values for the marginal effect assessments are all nonsignificant, their p-values being all greater than 0.05. However, the association p-values for the intercepts are highly significant ($p < 0.0001$), in line with the p-values obtained from the selection model analysis. In summary, no significant treatment effect is obtained. These findings confirm those obtained from the selection model formulation which gives more weight to this conclusion. These results can be justified by the fact that pattern mixture models using identifying restrictions play a very similar role to the modelling assumptions in the selection model case (Michiels et al., 1999). Furthermore, the parameter estimates and standard errors for the first marginal effect are equal for all the three restrictions CCMV, NCMV and ACMV, see the effects for intercept_6 and treatment_6. Such results should be expected considering the fact that the first outcome contained observed data for all subjects that were considered in the analysis.

As shown in the results in Table 3, the model building using CCMV, NCMV and ACMV restrictions in contrast to selection model did not allow an estimation of whether the dropout process is MNAR or not, because of differences in the modelling assumptions. This agrees with previous studies, see, for example, Molenberghs et al. (1998b), in that the identifying restrictions in a pattern mixture models context can be used only to relate the model to a MAR mechanism. Consequently, an important issue is to equate results for both the ACMV and MAR to make a clear and useful connection between the selection model and the pattern mixture model framework (Verbeke and Molenberghs, 2000; Kenward et al., 2003). With this in mind, the same is true for the selection model, MAR-based ACMV restrictions indicating non-significant treatment effects at all weeks. This can be explained to mean that the treatment effect appear to be independent of the ACMV (MAR) assumption. Although corresponding models include the same effects, the estimates for ACMV are slightly different to those for MAR. These slight differences are to be expected as argued in Kenward et al. (2003) that both models are similar in spirit but not necessarily identical. On the other hand, the parameter estimates and standard errors for the treatment effects obtained by applying NCMV are smaller than those of CCMV and ACMV as seen in some cases. This is to be expected as somewhat CCMV and ACMV pattern mixture models use data from different patterns to multiply impute new values, whereas in NCMV, pattern mixture models take information from the neighboring case patterns only. Further, ACMV and CCMV estimates are closer to each other since many more completers are available than

does NCMV. Therefore, additional variability may be introduced because, depending on the nature of the conditional distributions sampled from, data have been borrowed from more distant patterns.

5. Discussion

In this study, we demonstrated the application of two families of models for analysing incomplete longitudinal data, where the dependent variable is missing across time. In particular, we illustrated the application and compared results of analysis using these models. We focused on the situation in which outcomes are continuous. The models that were considered were the selection model and the pattern mixture model. Many authors have recommended fitting both families of models to be able to gain extra insight into the data to assess sensitivity to the modelling assumptions and to assess the extent of agreement in results as well (see, Molenberghs and Verbeke, 2005). The study focused on the specific cases of selection model and pattern mixture models; that is, a Diggle and Kenward's (1994) model and an identifying restrictions strategy (Little, 1993, 1994), respectively. In applying the selection model, we used logistic regression for modelling dropout, however, a number of other probabilities can be used, for example, using survival analysis techniques, the length of duration on treatment or placebo before dropout can also be modelled. However, in this study, the survival model for dropout cannot be used because the time to event (dropout) is not exactly determined by design. For example, if someone is not seen at week 12, the exact time to dropout could theoretically be any time between week 6 and 12. The objective was to investigate the potential influence that dropout might have or exert on the dependent measurement on the considered data and to deal with incomplete sequences. The results from the pattern mixture models were analogous to those from the selection model to obtain additional insights into the serum cholesterol data. The application was based on an example from a longitudinal clinical trial data.

Findings in general suggested that the conclusion obtained under both modelling frameworks practically coincide. Thus, one can put more confidence in these results as argued by many authors. For example, Michiels et al. (1999, 2002) have argued that greater confidence in a conclusion can be reached when the analysis of joint applications of these models leads to essentially similar inference. Both families of models were compared and noticeable similarities in results were found. Hence, this begs the question as how, depending on the scientific question of interest such as conditional measurement probabilities, to choose between them. Michiels et al. (1999) argued that the selection model can be recommended as a natural choice when the interest is in the population as a whole, i.e., marginal effects. Whereas, pattern mixture models can be considered, when investigating the differences between subgroups that are identified by their measurement patterns, i.e., pattern-specific.

The selection models suggested that the dropout mechanisms were not completely at random. In other words, in the context of the assumed model, there was a lot of evidence in favour of the prevalence of an MAR rather than an MCAR dropout process. However, many authors, Diggle and Kenward (1994) and Molenberghs and Verbeke (2005) for example, stated that careful consideration is necessary with such a conclusion when using only the data under analysis. A problem arises for dealing with dropout that are MNAR. Given this problem in a longitudinal study, it is important to realize that this assumption gives rise to the dropout that is not likely to be known in the application setting. Therefore, any of the different proposed application methods to address dropout that are MNAR cannot easily be verified. For example, one often does not know if the dropout process is accurately captured by a particular method used. Molenberghs and Kenward (2007) suggested that one should apply several approaches to the same data problem. This is the case when the sensitivity of parameters estimates to the different mechanisms about the dropout process may be investigated, for example, by applying models that allow for the dropout to be MNAR. According to Xu and Blozis (2011), if parameter estimates are comparable under different methods, this can indicate that the dropout process may be ignored. However, if different methods give different parameters estimates of the longitudinal model, this can indicate that the dropout process is a vital element for the description of the data in the analysis.

The structure of the selection dropout model adopted that dropout increases with a unit change in the serum cholesterol; that is, the dropout is related to the larger negative increments $(y_{ij} - y_{i,j-1})$ rather than to any actual observation $(y_{ij} + y_{i,j-1})$, which implies that patients with a greater decrease in the overall level of the serum cholesterol from the previous week have a higher probability of dropping out of the experiment. This situation is very common in practice within a model of the Diggle and Kenward type, and we refer to Molenberghs and Kenward (2007), Diggle and Kenward (1994) and Molenberghs et al. (1997) as examples. Under the modeling scheme applied in this study, it can be seen from the analysis that the treatment effects over all weeks under all ACMV, CCMV and NCMV restrictions were non-significant, and the same is true for the selection model analysis. Therefore, it is clear that there is a strong evidence for no significant treatment in the context of serum cholesterol data. It appeared that the non-significant treatment effects were not conditional upon any dropout mechanism holding. As a results, the conclusions obtained from CCMV, NCMV and NCMV restrictions did not differ considerably. As argued in Molenberghs et al. (2008), the choice between them is not always clear. Although they fit the observed data equally well, the difference between them only becomes clear with respect to estimation of the missing data, conditional upon the observed data.

On the other hand, the use of different models in which the data were analysed, can be considered as a sensitivity analysis. In particular, the use of pattern mixture models including identifying restrictions can be seen as a first tool for assessing the sensitivity of the assumptions made. Further, other more complex or flexible sensitivity analysis are also possible, under new models for the probability of dropout. The

analysis conducted here is a typical sensitivity analysis as the serum cholesterol data were analyzed using different assumptions concerning the longitudinal measurements and dropout mechanisms. In particular, both models compared well concerning some aspects, for example, marginal treatment effects. Such comparisons as these can play a vital role in sensitivity analysis by providing additional motivation, for example, when considering the choice between selection and pattern mixture models. In conclusion, because the true model and measurement process as well as dropout process are often unverifiable, the recommendation that in many settings, multiple strategies or models such as selection and pattern mixture models be applied to the same data set in order to investigate the impact of assumption on dropout or missingness is supported.

References

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. CRC: Chapman and Hall.

Diggle, P. J. and Kenward, M. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49–94.

Dmitrienko, A., Offen, W. W., Faries, D., Christy Chuang-Stein, J. L. and Molenberghs G. (2005). *Analysis of Clinical Trial Data Using the SAS System*, Cary, NC: SAS Publishing.

Ekholm, A. and Skinner, C. (1998). The muscatine children's obesity data reanalysed using pattern mixture models. *Applied Statistics*, 47, 251–263.

Glynn, R. J., Laird, N. M. and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In: *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Heckman, J. J. (1976). The common structure of statistical models of trucation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.

Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64–78.

Hogan, J. and Laird, N. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16, 259–272.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G. and Kenward, M. G. (2006). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis*, 50, 830–858.

Kenward, M. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8, 51–83.

Kenward, M., Molenberghs, G. and Thijs, H. (2003). Pattern mixture models with proper time dependence. *Biometrika*, 90, 53–71.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81, 471–483.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. and Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 52, 98–111.

Marini, M. M., Olsen, A. R. and Rubin, D. B. (1980). Maximum likelihood estimation in panel studies with attrition. *Sociology Methodology*, 11, 314–357.

McArdle, J. J. and Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using laten growth structural models. *Experimental Aging Research*, 18, 145–166.

Michiels, B., Molenberghs, G. and Lipsitz, S. R. (1999). Selection models and pattern mixture models for incomplete data with covariates. *Biometrics*, 55, 978–983.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T. and Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, 21, 1023–1042.

Molenberghs, G., Michiels, B. and Kenward, M. (1998a). Pseudo-likelihood for combined selection and pattern-mixture models for missing data problems. *Biometrical Journal*, 40, 557–572.

Molenberghs, G., Michiels, B., Kenward, M. and Diggle, P. J. (1998b). Missing data mechanism and pattern mixture models. *Statistica Neerlandica*, 52, 135–161.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*. New York: Wiley.

Molenberghs, G., Thijs, H., Kenward, M. G. and Verbeke, G. (2003). Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistics Neerlandica*, 57, 112–135.

Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, 84, 33–44.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E. and Kenward, M. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, 37, 93–113.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of Royal Statistical Soceity*, 70, 371–388.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, 33, 545–571.

Schoenfield, L. J. and Lachin, J. M. (1981). The steering committee, and the NCGS group, "Chenodiol (Chenadeoxycholic Acid) for Dissolution of Gallstones: The National Cooperative Gall-stone Study". *Annals of Internal Medicine*, 95, 257–282.

Verbeke, G., Lesaffre, E. and Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, 35, 419–439.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175–188.

Wu, M. C. and Bailey, K. R. (1988). Analysis changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7, 337–346.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45, 939–955.

Xu, S. and Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, 36, 237–256.

SORT 37 (2) July-December 2013, 153-174

A nonparametric visual test of mixed hazard models

Jaap Spreeuw[1], Jens Perch Nielsen[2] and Søren Fiig Jarner[3]

Abstract

We consider mixed hazard models and introduce a new visual inspection technique capable of detecting the credibility of our model assumptions. Our technique is based on a transformed data approach, where the density of the transformed data should be close to the uniform distribution when our model assumptions are correct. To estimate the density on the transformed axis we take advantage of a recently defined local linear density estimator based on filtered data. We apply the method to national mortality data and show that it is capable of detecting signs of heterogeneity even in small data sets with substantial variability in observed death rates.

MSC: 62F10, 62N01, 62N02, 62P05.

Keywords: Mortality data, frailty models, visual inspection.

1. Introduction

There is an increasing use of mortality models to answer a number of pension related questions. Mortality tables and their estimation have always been of importance while calculating appropriate prices of risk products depending on individuals' survival. More recently, mortality models are being used in more complex models assessing the value of financial products incorporating survival in a variety of ways. Financial users of mortality models are therefore not only actuaries nowadays, but also investors looking for opportunities in survival bonds and other packages of survival risks. Different purposes of mortality models lead to different measures of quality.

[1] Faculty of Actuarial Science and Insurance, Cass Business School, City University London, 106 Bunhill Row, London, EC1Y 8TZ, UK. E-mail: j.spreeuw@city.ac.uk

[2] Faculty of Actuarial Science and Insurance, Cass Business School, City University London, 106 Bunhill Row, London, EC1Y 8TZ, UK.

[3] Danish Labour Market Supplementary Pension Fund, Kongens Vænge 8, 3400 Hillerød, Denmark.

Received: October 2012
Accepted: July 2013

In this paper we develop a visualization technique that seems useful for the individual assessment of the quality of a mortality model. One application we are thinking of is forecasting of mortalities that is a basic building block for the financial pricing of survival, but also a useful tool in asset liability management of pension portfolios. Typically, relatively simple parametric mortality models including calendar effects are used as starting point for mortality forecasts. The calendar effect is the explicit tool for the forecast and is often isolated and estimated through standard time series methodology. A perfect historical fit of the past is therefore not always what the mortality modeller is looking for. Often it is more important to have an overall good fit, without too systematic deviations giving reliable and meaningful forecasts. These latter objectives are not easy to generalize to some quantitative model that can be tested. Often simple mortality models are rejected, simply because mortality data often is nationwide and sufficiently abundant to inform relatively complex underlying parametric structures. Therefore, a test rejecting our simple model is often not what we want. We do know that our simple model is not accurate, we do not want an excessive fit, what we want is a good, intuitive and reliable forecast.

When modelling mortality of a population, there is a variety of potentially suitable lifetime data models available. Potential models differ in levels of complexity and they try to capture different features of data. Specific parametric life tables combined with time series forecasts are omnipresent in the actuarial and demographic literature.

The literature about parametric mortality projection has been developing rapidly in the last few years. Recent reviews of mainstream mortality forecasting models can be found in Cairns et al. (2009), Cairns et al. (2011), Dowd et al. (2010a,b) and Haberman and Renshaw (2011). Cairns et al. (2009) compare eight models on the basis of several desirable ex post qualitative properties (like model parsimony, transparency, possibility to generate sample paths, presence (or absence) of cohort effects and ability to achieve a nontrivial correlation structure) and quantitative criteria (consistency with historical data and robustness of parameter estimates). Six of these models are subject of subsequent investigation by Dowd et al. (2010a,b) and Cairns et al. (2011). These include the original Lee-Carter model (Lee and Carter, 1992), the basic age-period-cohort model by Renshaw and Haberman (2006), an alternative age-period-cohort model by Currie (2006), the original Cairns-Blake-Dowd model as launched in Cairns et al. (2006), and two extensions thereof. The six models are the subject of formal goodness-of-fit tests in Dowd et al. (2010a) and backtesting in Dowd et al. (2010b). Cairns et al. (2011) judges these models on the basis of ex ante qualitative aspects like biological reasonableness, plausibility of forecast levels of uncertainty in projections at several ages, and robustness of forecasts. In all these papers, the mortality data applied was confined to those of individuals aged 60 or above. Haberman and Renshaw (2011), concentrating on the key factors of life expectancy and annuity values, first conduct a detailed comparison of the several models at pensioner ages. Apart from the models in the above papers, they also consider special cases of the Renshaw and Haberman (2006) model in their study. Later on, they extend the age range and involve the model by Plat (2009) and several variants thereof.

The stability of the forecast depends crucially on the choice of the parametric form. Generally, a complex model with many parameters is not a good choice even though such models might be selected from classical mathematical statistical model selection designed for in-sample prediction. Models with many parameters generally fit data better than models with fewer parameters. On the other hand, a large number of parameters are harder to forecast than fewer parameters. Forecasting uncertainty increases dramatically with the number of parameters. Thus, to obtain reliable forecasts we want models which describe the key features of data with as few parameters as possible.

The purpose of this paper is to introduce a visual diagnostic tool which can be used to guide us when choosing a parametric model. A good parametric model is a simple model without obvious systematic errors. That model could be chosen by the well informed statistician working with the particular mortality forecast application in mind. Our visual diagnostic tool will be just one helpful tool in the overall mathematical statistical toolbox. Our method is inspired from recent developments in extreme value estimation, where transformations of data give visual information on the quality of the distributional fit in the tail. This recent methodology has found its way into insurance pricing and also the related field of operational risk. For a comprehensive overview of this new transformation methodology in the latter context, see Bolancé et al. (2012a).

The transformation based method can for example compare the performance of several candidate models for a data set at hand. Assume we were told by an oracle what the exact true distribution is, then we would transform our data using this oracle information such that our transformed data would exactly originate from a uniform distribution. Now we do not have access to any oracles. However, if we take some estimated parametrically fitted survival distribution as defining our transformation, then any detectable deviance on the transformed scale from the uniform distribution implies deviances of the parametric distribution used in the transformation step from the underlying true distribution. Our methodology uses a nonparametric smooth kernel estimator on the transformed scale. One difficulty we meet here is that our data is classical survival data that is not independent identically distributed. We therefore use a recent local linear kernel density estimator – specifically the one of Nielsen et al. (2009) – that is adjusted for the truncation and censoring pattern we meet in our data. Comparison between different underlying suggested parametric models are carried out by first estimating these parametric models and then to investigate through visual inspection, whether the density of the transformed data indeed looks uniform.

If the underlying parametric model under investigation would be true, the estimated density should be close to one over the unit interval. Therefore different underlying parametric models can be visualized and compared on the transformed scale. In principle, the densities could also be estimated and compared on the original scale. However, there are several visual and estimational advantages to working on the transformed scale. One of these is that our method makes maximal use of sparse and volatile data and is thus particularly well suited to explore how potential models describe the mortality at ad-

vanced ages where exposure is invariably limited. We test our method using data from nations of different size: USA, United Kingdom, Denmark and Iceland.

Although the main focus of our paper is to model human mortality, it is worthwhile mentioning that our methodology is applicable to any probability density model, whether it concerns human survival or not.

1.1. Mixed hazard models

Frailty theory offers a possible explanation to the presence of an old-age mortality plateau. According to this theory populations are heterogeneous with some people being more frail, i.e. having a higher hazard rate, than other people. Since persons with high hazard rates tend to die sooner than persons with low hazard rates old age groups will be dominated by low frailty persons and this effect reduces the rate of increase at the population level.

Frailty models were introduced in the demographic literature by Vaupel et al. (1979). In a multiplicative frailty model, an individual's hazard rate consists of two parts, namely a certain standard intensity and a certain nonnegative random variable, the frailty, acting multiplicatively on the standard intensity. A Gompertz or Makeham specification is usually taken for the standard intensity, although sometimes a Weibull model can be seen. Frailty is usually assumed to follow a Gamma distribution, which is known to be mathematically very tractable.

A few publications about frailty modelling appeared in the actuarial literature. Wang and Brown (1998) use the Gompertz-Gamma or Perks model to graduate mortality improvement factors in a Society of Actuaries' Life Table. Butt and Haberman (2004) employ Generalized Linear Models to graduate mortality of insured lives. They consider three mixture models, namely i) Perks; ii) modified Perks, and iii) Gompertz-Inverse Gaussian. The authors conclude that the Perks model fits the data best. An overview of heterogeneity models in life insurance is given in Olivieri (2006), while Jones (1998) develops a multiple state model to measure the impact of frailty on the propensity to lapse a policy. Finally, Li et al. (2009) extend the Lee-Carter model by allowing for unobserved heterogeneity within a cell, determined by age and time.

In this paper we illustrate our methodology in the one dimensional case. Most forecasting models operate with a multiplicative relationship between age effect and time effect. To visualize the fit of the age effect, one would then have to divide out the estimated time effect and vice versa to visualize the time effect only.

We are happy to say that our paper – diffused in preliminary versions – already has inspired a number of other works in mathematical and computational statistics. It has for example been cited in the three recent papers Gámiz-Pérez et al. (2013a,b,c).

1.2. Outline

The set-up of this paper is as follows. In Section 2 we present the visual inspection technique in detail. Both the continuous-time framework with transformed counting processes and the implementation with discrete data is discussed. Section 3 discusses frailty models in general and introduces the class of models we will be using. Section 4 presents the numerical application. For four countries varying significantly in size (United States, United Kingdom, Denmark and Iceland), one data set per country (female period 2006 from the Human Mortality Database) and three different frailty specifications, namely Gamma, Inverse Gaussian, and degenerate (no frailty), we show the estimates as well as the visual inspection technique. In particular, we give a thorough analysis of the mortality at advanced ages that can be extracted from the continuous graphs. Section 5 sets out a conclusion.

2. Visual inspection technique

2.1. Sampling scheme of the survival data

Consider a data set with mortality statistics of n lives. Let for each of these n individuals Y_i be an exposure process with value one when the i'th individual is alive and under observation and let N_i be a counting process taking the value one if the i'th individual has died while under observation. Both Y_i and N_i are functions of the age x. Formally, we assume that N_i is a one-dimensional counting process with respect to an increasing right continuous complete filtration $\mathscr{F}_x,\ x \in \mathscr{R}_+$, i.e. one that obeys *les conditions habituelles*, see Andersen et al. (1993, p. 60). We model the intensity as

$$\lambda_i^c(x) = \mu_\theta(x)\, Y_i(x),$$

where θ belongs to the parameter space Θ of the parameters determining the exact mortality and frailty. The estimator $\widehat{\theta}$ of θ is derived from minimizing the log likelihood of Borgan (1984):

$$l(\theta) = \sum_{i=1}^{n} \int \log\{\mu_\theta(x)\}\, dN_i(x) - \sum_{i=1}^{n} \int \mu_\theta(x)\, Y_i(x) dx,$$

that is maximized over the parameter space Θ.

2.2. Visual inspection by transformations

Assume that some oracle has given us the true underlying c.d.f. F_θ. Then consider the transformed counting processes $\overline{N}_i = N_i \circ F_\theta^{-1}$ defined on $[0,1]$. If our oracle really had

told us the truth, then $\overline{N}_i$ would have stochastic intensity

$$\lambda_i(y) = \alpha(y)\overline{Y}_i(y),$$

where $\overline{Y}_i(y) = \left\{Y_i\left(F_\theta^{-1}(y)\right)\right\}$ with $\alpha(y) = 1/(1-y)$ corresponding to the hazard of the uniform distribution with density

$$f(y) = \alpha(y)\exp\left(\int_0^y -\alpha(s)ds\right) = 1,$$

for $y \in [0,1]$.

Another more statistical term for oracle information is prior information. It is that type of information that is external to the data set at hand. In our application below our prior information will always be some parametric specification of the model and our oracle candidate for the true c.d.f will be $F_{\widehat{\theta}}$, where $\widehat{\theta}$ is the estimated parameter in the specified parametric model. If $F_{\widehat{\theta}}$ really is a good description of the true c.d.f. F, then our data should be uniformly distributed after a transformation by $F_{\widehat{\theta}}$.

To be able to inspect the credibility of our oracle information or prior information or parametric assumptions, we estimate the density f based on the filtered survival data $\left(\overline{N}_1,\overline{Y}_1\right),\ldots,\left(\overline{N}_n,\overline{Y}_n\right)$ on $[0,1]$ and see whether it looks flat. This density estimator should have good boundary correction because it is defined on the transformed axis $[0,1]$. We suggest to use the natural weighted local linear density estimator of Nielsen et al. (2009):

$$\widehat{f}(y) = \sum_{i=1}^{n}\int \overline{K}_{y,b}(y-s)\overline{Y}_i(s)\widehat{S}(s)d\overline{N}_i(s),$$

where

$$\overline{K}_{y,b}(y-s) = \frac{a_2(y)-a_1(y)(y-s)}{a_0(y)a_2(y)-\{a_1(y)\}^2}K_b(y-s),$$

with

$$K_b(y-s) = \frac{1}{b}K(\frac{y-s}{b}), \tag{1}$$

and

$$a_j(y) = \sum_{i=1}^{n}\int K_b(y-s)(y-s)^j\overline{Y}_i(s)ds,$$

and

$$\widehat{S}(s) = \prod_{t \leq s} \left\{ 1 - d\widehat{\Lambda}(t) \right\},$$

being the Kaplan-Meier estimate of the survival function, with

$$\widehat{\Lambda}(s) = \sum_{i=1}^{n} \int_0^s \left\{ Y^{(n)}(t) \right\}^{-1} d\overline{N}_i(t),$$

where $Y^{(n)}(t) = \sum_{i=1}^{n} \overline{Y}_i(s)$.

2.3. Implementing with discrete data

In most real life applications we only have discretized versions of the stochastic processes Y_i and N_i available. First we need to define the relevant discretized time points $H_1, \ldots, H_K$ and the corresponding differences $h_k = H_k - H_{k-1}$ for $k \in \{1, \ldots, K\}$, with $H_0 = 0$. We define $H_K = \inf\left(t; F_{\widehat{\theta}}(t) = 1\right)$ for any plausible survival function F_θ.

Discretized data are often defined as occurrences and exposures. Let respectively

$$O_k = \sum_{i=1}^{n} \int_{H_{k-1}}^{H_k} dN_i(x)$$

and

$$E_k = \sum_{i=1}^{n} \int_{H_{k-1}}^{H_k} Y_i(x) dx.$$

Now assume that we only observe these discrete occurrences – the O_k's – and exposures – the E_k's. Then a natural approximation of the log likelihood function $l(\theta)$ above to our discrete observations would be

$$l_d(\theta) = \sum_k \left\{ \log \mu_\theta \left(H_k^* \right) \right\} O_k - \sum_k \mu_\theta \left(H_k^* \right) E_k,$$

where $H_k^* = \left(H_{k-1} + H_k \right) / 2$.

Now consider discretized time points on the axis transformed by $F_{\widehat{\theta}}$. Let $\overline{H}_k = F_{\widehat{\theta}}^*(H_k)$. $\overline{h}_k = \overline{H}_k - \overline{H}_{k-1}$ and $\overline{H}_k^* = \left(\overline{H}_{k-1} + \overline{H}_k \right) / 2$ for $k \in \{1, \ldots, K\}$. Note that $\overline{H}_K = 1$. Also note that often the discrete time points are equidistant before the time transformation but not thereafter.

On the transformed axis with time, the series $H_1^*, \ldots, H_K^*$ is transformed into $\overline{H}_1^*, \ldots, \overline{H}_K^*$. We will have occurrences

$$\overline{O}_k = O_k$$

and exposures

$$\overline{E}_k = E_k * \overline{h}_k / h_k.$$

Assume that we were given the true c.d.f. with very large risk exposures E_k. Then on the original axis $O_k \sim \mu_{\widehat{\theta}}(H_k^*) E_k h_k$ while on the transformed axis $\overline{O}_k = O_k \sim \alpha_{\widehat{\theta}}(\overline{H}_k^*) E_k \overline{h}_k$. If the model were the true one, the hazard rates $\overline{O}_k / \overline{E}_k$ on the transformed axis would be equal to $1 / (1 - \overline{H}_k^*)$, and hence the density functions would be constant at 1.

The local linear density estimator on the transformed axis will in the discrete case be defined as

$$\widehat{f}_d(y) = \sum_k \overline{K}_{d,y,b}(y - \overline{H}_k^*) \widehat{S}_d^t(\overline{H}_k^*) \overline{O}_k, \tag{2}$$

where

$$\overline{K}_{d,y,b}(y-s) = \frac{a_{2,d}(y) - a_{1,d}(y)(y-s)}{a_{0,d}(y) a_{2,d}(y) - \{a_{1,d}(y)\}^2} K_b(y-s),$$

$$a_{j,d}(y) = \sum_{k=1}^{K} K_b(y - \overline{H}_k^*)(y - \overline{H}_k^*)^j \overline{E}_k$$

and

$$\widehat{S}_d^t(\overline{H}_k^*) = 0.5 \left\{ \widehat{S}_d^t(\overline{H}_{k-1}) + \widehat{S}_d^t(\overline{H}_k) \right\} = 0.5 \left[\exp\left\{ -\sum_{i=1}^{k-1} \overline{h}_i \frac{\overline{O}_i}{\overline{E}_i} \right\} + \exp\left\{ -\sum_{i=1}^{k} \overline{h}_i \frac{\overline{O}_i}{\overline{E}_i} \right\} \right].$$

The choice of the bandwidth b depends on the availability of data. Large countries have a large risk exposure; then most of the deviation between the density estimate and 1 can be attributed to model uncertainty. In such cases, no or hardly any smoothing is required and b can be small. For not so densely populated countries with small risk exposure, on the other hand, proper smoothing – with a larger bandwidth – is needed to compensate for parameter uncertainty.

3. Mixed hazard models

In an individual frailty model the individual effect for a life's mortality acts multiplicatively. Assume that a cohort consists of n individuals. Then for the ith person of the cohort, the individual effect is represented by the random variable Z_i and the conditional force of mortality at age x, given $Z_i = z_i$, is given by

$$\mu(x, z_i) = z_i \mu(x), \quad i \in \{1, \ldots, n\},$$

with $\mu(x)$ denoting the standard force of mortality at age x – which is the force of mortality of a life with frailty level 1 – and all Z_i independent and identically distributed, with a mean equal to 1.

In this paper we will assume that the individual hazard is of the form

$$\mu(x) = \exp(a_0 + a_1 x + a_2 x^2). \tag{3}$$

In the notation of Forfar et al. (1988) this model is labelled GM(0,3). Note that the special case $a_2 = 0$ leads to the Gompertz model (GM(0,2)). The structure in (3) forms the basis for national and international mortality modelling in Jarner and Kryger (2011).

We have $d\mu(x)/dx = \mu(x)(a_1 + 2a_2 x)$. It is reasonable to assume that mortality is increasing as a function of age. This would imply $a_1 \geq 0$ and $a_2 \geq 0$. Nonnegative estimates of a_1 and a_2 are also obtained in Jarner and Kryger (2011). The relative change of mortality as a function of age x – defined in Horiuchi and Coale (1990) as $k(x) = d \ln \mu(x) / dx$ – is a linear function of age: $k(x) = a_1 + 2a_2 x$.

The cohort mortality at age x is given as $\mu_\theta(x) = E[Z|x] \cdot \mu(x)$, where $E[Z|x]$ denotes the mean frailty of lives surviving to age x. Let L_Z denote the Laplace transform of frailty at birth, i.e. $L_Z(s) = E[\exp(-sZ)]$. It can then be shown, see e.g. Hougaard (1984), that

$$E[Z|x] = -\frac{L_Z'[M(x)]}{L_Z[M(x)]}$$

with

$$M(x) = \int_0^x \mu(s)\, ds.$$

Hence the cohort mortality can be easily calculated for all frailty specifications with known Laplace transform. In the literature, the Gamma distribution has been by far the most popular specification in the frailty model. This is partly due to its mathematical tractability. Abbring and Van den Berg (2007) show that, under mild conditions

regarding regular variation, for several frailty distributions, Gamma is the limiting frailty distribution of survivors. Therefore, Gamma frailty is in some sense a natural choice. We will also be using Gamma frailty as one of the bases of our numerical illustrations in the next section.

Since the model (3) contains a scaling parameter, a_0, we will assume, without loss of generality, that the distribution of frailty at birth has mean 1 and variance σ^2. If frailty is Gamma, then $L_Z(s) = \left(1+\sigma^2 s\right)^{-1/\sigma^2}$, while $E[Z|x] = \left(1+\sigma^2 M(x)\right)^{-1}$.

Another common choice for frailty in the literature on survival models is the Inverse Gaussian distribution. This specification has also been discussed by Butt and Haberman (2004). Under the above assumption, $L_Z(s) = \exp\left[\sigma^{-2}\left(1-\sqrt{1+2\sigma^2 s}\right)\right]$ and $E[Z|x] = \left\{1+2\sigma^2 M(x)\right\}^{-1/2}$.

Disregarding heterogeneity implies $L_Z(s) = e^{-s}$ and $E[Z|x] \equiv 1$. This is a special case of the Gamma distribution, obtained by taking the limit $\sigma^2 \downarrow 0$. Obviously, this case of no frailty should give a worse fit than Gamma. In the numerical application we will show how this transpires by comparing the resulting local linear density estimators on the transformed scale.

4. Application to mortality data

We analyze the mortality data of four countries differing significantly in terms of population size, namely United States (US), United Kingdom (UK), Denmark and Iceland.

For each country, the data set consists of female period mortality data, obtained from the Human Mortality Database, and concerning the calendar year 2006. Since we are primarily interested in adult and old age mortality, only the ages from 40 to 110 are included. The exposed to risk at age 40 (defined before as E_1) are equal to $2,116,995.31$ (US), $478,424.36$ (UK), $43,463.83$ (Denmark) and $2,193.00$ (Iceland). So, roughly, the largest country is about 1,000 times as large as the smallest country.

For Gamma frailty, Inverse Gaussian frailty and no frailty, we estimate the parameters from maximum likelihood following Borgan (1984) and perform a visual inspection of its fit based on the transformation approach. The parameter estimates are shown in Table 1. As basic kernel function, we choose

$$K(x) = \frac{3003}{2048}\left(1-x^2\right)^6 I_{[-1,1]}(x). \tag{4}$$

For further details about this specific kernel, consult Nielsen et al. (2009). We have selected (4) for our example, but the actual choice of basic kernel function is not so important.

Table 1: *Estimates of parameters.*

Country/frailty model	a_0	a_1	a_2	σ^2
United States				
No frailty	−9.0389	0.04975	0.0003014	–
Gamma	−8.1373	0.02014	0.0005402	0.1193
Inverse Gaussian	−9.0388	0.04975	0.0003014	2.3966e-11
United Kingdom				
No frailty	−10.1631	0.06766	0.0002594	–
Gamma	−8.5070	0.01479	0.0006729	0.1632
Inverse Gaussian	−8.5047	0.01374	0.0006914	0.2488
Denmark				
No frailty	−10.1153	0.07338	0.0001991	–
Gamma	−9.7088	0.06034	0.0003021	0.04007
Inverse Gaussian	−9.7390	0.06125	0.0002954	0.04062
Iceland				
No frailty	−10.1777	0.05911	0.0003532	–
Gamma	−10.1136	0.05708	0.0003689	0.005556
Inverse Gaussian	−10.1777	0.05911	0.0003532	4.2043e-08

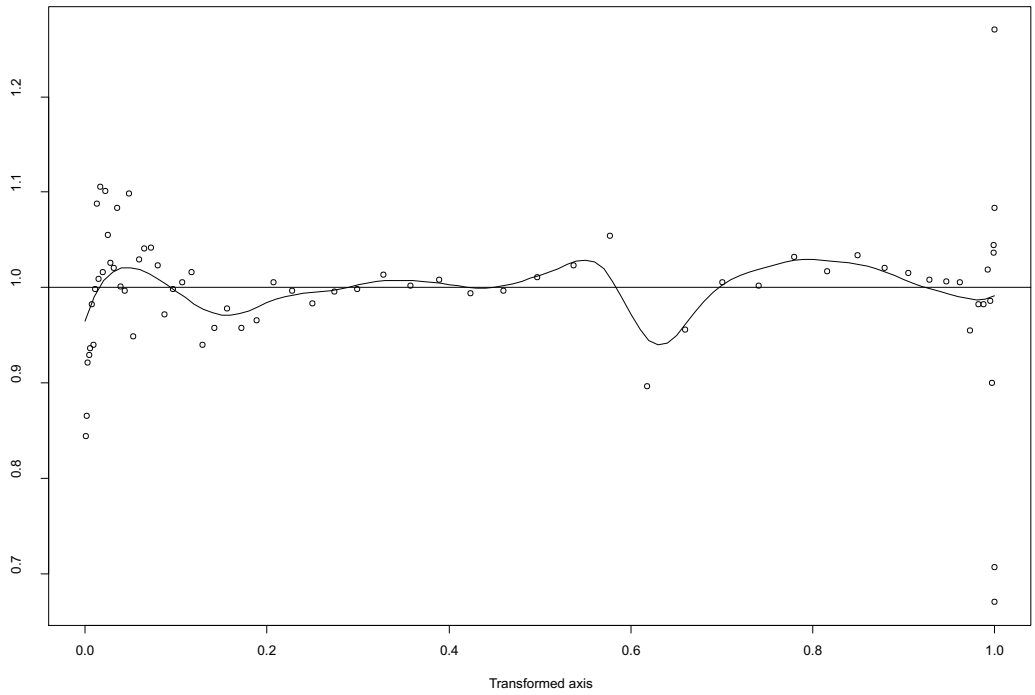


Figure 1: *United Kingdom: Values of density function obtained from the transformed observations (points) and local linear density estimator (curve) as in (2), with $b = 1/9$, all on the transformed scale, using Gamma frailty.*

Now we assume that Gamma frailty is suitable for the data at hand, and we want to test that assumption through our visual inspection technique. We use the UK as an illustrative example. Figure 1 displays the values of the density functions of the 71 transformed observations on the transformed scale, using Gamma frailty. For observation k, with $k \in \{1,\ldots,K\}$, the x-coordinate is equal to $\overline{H}_k^*$, while the corresponding y-coordinate is equal to $\left(1-\overline{H}_k^*\right) * \overline{O}_k / \overline{E}_k$.

As explained before, if the model assumed were the true one and risk exposure were infinite, these y-coordinates would all be equal to unity. Deviations from unity essentially arise from two sources: the extent to which the assumed/estimated model deviates from the true underlying model, and the noise in data caused by the stochastic nature of death. We are interested in assessing the first kind of (systematic) deviations, and for this purpose we want to reduce the second kind of (unsystematic) deviations.

Assume for a moment that all deviations could be taken at face value, i.e. that there were no unsystematic deviations. How would model deviations then manifest themselves on the transformed scale? Assume the model has density f with cdf F, while the true density is g. Then the density of transformed data is $\left(g/f\right)\left(F^{-1}(u)\right)$ for $0 < u < 1$. We are particularly interested in the behaviour in the right tail of the distribution. There are three possibilities:

- If the model overestimates the density of dying old then the density of transformed data will be below 1 in the right tail.
- If the model estimates the density of dying old correctly then the density of transformed data will be close to 1 in the right tail.
- If the model underestimates the density of dying old then the density of transformed data will be above 1 in the right tail.

Small risk exposure and consequently noisy data is a well known problem for very high ages, even for large countries like the US. This is why the observed density points start to deviate significantly from 1 when the x-coordinate approaches 1, bearing in mind that the probability of death before attaining a very advanced age is close to 1. The purpose of the density estimator is to reduce the noise in data to get a clearer picture of the performance of the model. The reason for working on the transformed scale is that the density can be more effectively estimated and with smaller bias on the unit interval than on the original scale.

4.1. Effect of bandwidth

A non-smoothed density estimate is obtained by simply connecting these points, but this will lead to an irregular pattern in all cases. The points are displayed together with the local linear density estimator based on kernel smoothing as in equation (2). In Figure 1, the value for the bandwidth b of $1/9$ has been selected by eye-ball, according to whatever looks best for this particular data set.

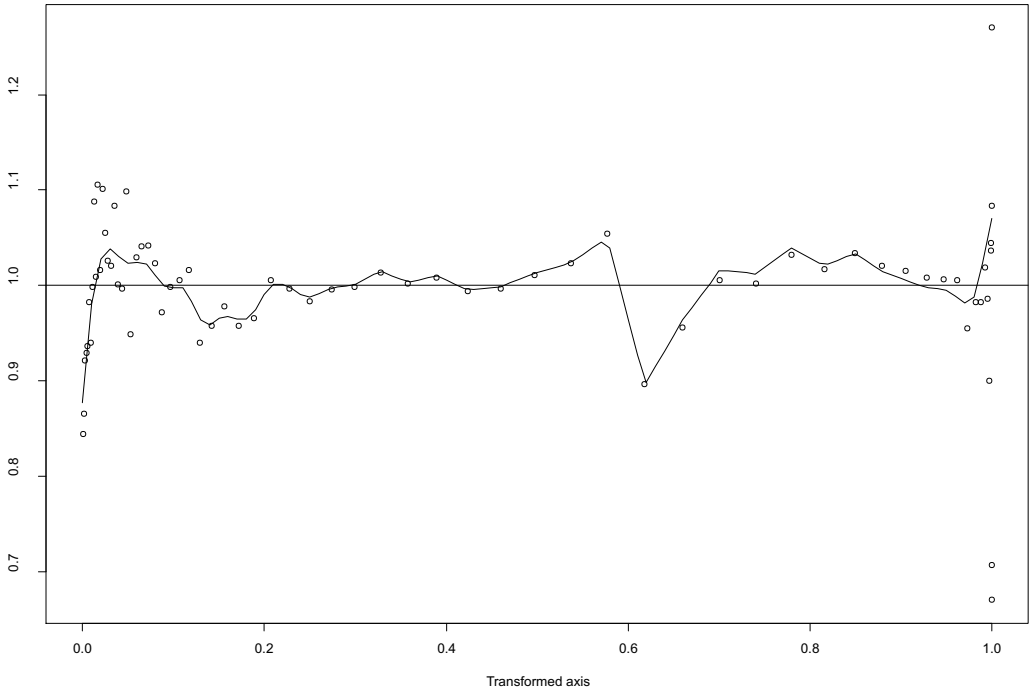


Figure 2: *United Kingdom: Values of density function obtained from the transformed observations (points) and local linear density estimator (curve) as in (2), with* $b = 1/24$*, all on the transformed scale, using Gamma frailty.*

The importance of choosing an appropriate value for the bandwidth is illustrated in the next two diagrams. Figure 2 shows the same points alongside the density estimator with smaller bandwidth $b = 1/24$. The lack of smoothness is evident: the estimator seems to be a set of line segments connecting several points.

Figure 3, on the other hand, gives the density estimator with larger bandwidth $b = 1/3$. This is an example of over-smoothing: features are displayed that do not reflect the characteristics of the data.

However, regardless of the chosen bandwidth the estimated density is close to unity in all three cases, and hence we would judge the model to provide a good description of old age mortality for the UK data.

The bandwidth choices for the other countries are $1/12$ (US), $1/6$ (Denmark) and $1/2$ (Iceland). Note that the bandwidths go up with decreasing population size for reasons as stated above. Observe that the set of observed density points for Iceland contains a lot of outliers, whence the choice of a large bandwidth.

While these bandwidths have been selected by eye-ball, we also reproduced automatic bandwidth selectors. The cross-validation procedure of Nielsen et al. (2009) breaks down and undersmooths way too much in our case. However, we adapted the Do-validation bandwidth selection procedure of Mammen et al. (2011) to our case. Except

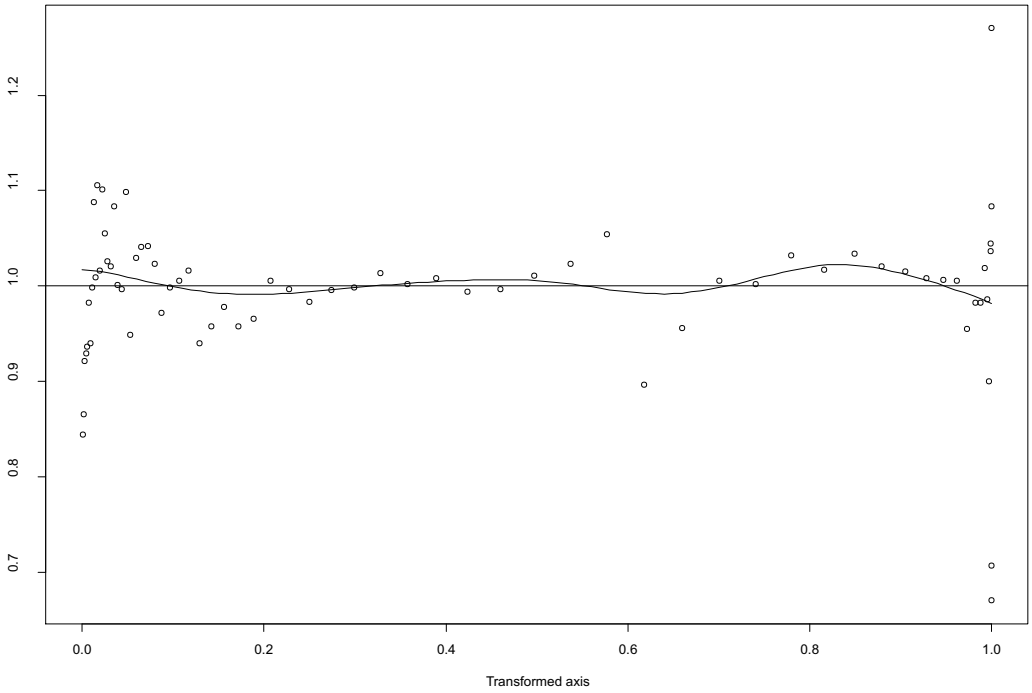


Figure 3: *United Kingdom: Values of density function obtained from the transformed observations (points) and local linear density estimator (curve) as in (2), with $b = 1/3$, all on the transformed scale, using Gamma frailty.*

for Iceland, we got almost exactly the same bandwidths as our eye-ball bandwidths. The automatic bandwidth selector for Iceland was somewhat higher than our eye-ball bandwidth selector. We did, however, in the end like our eye-ball selector for Iceland more and we therefore present that one here along with our other eye-ball selected bandwidths.

4.2. Old age mortality

In Figures 4 to 7, the local linear density estimate for Gamma frailty is compared with the one obtained in case of Inverse Gaussian and no frailty (with the same bandwidth, of course). Note that for the US the curves of Inverse Gaussian and no frailty practically coincide. Obviously this is due to the parameter estimates of a_0, a_1 and a_2 which are virtually the same for both models, while the estimate of the additional parameter σ^2 in Inverse Gaussian is very close to zero.

The overall impression regarding the four countries is best when it comes to the Gamma frailty survival model. In the USA case, the Gamma frailty parametric density is never more than two percent away from one on the transformed scale. This is not bad at all and most forecasters could accept deviances at this scale. Also, the USA data is quite abundant and we can allow ourself to work with a relatively small bandwidth.

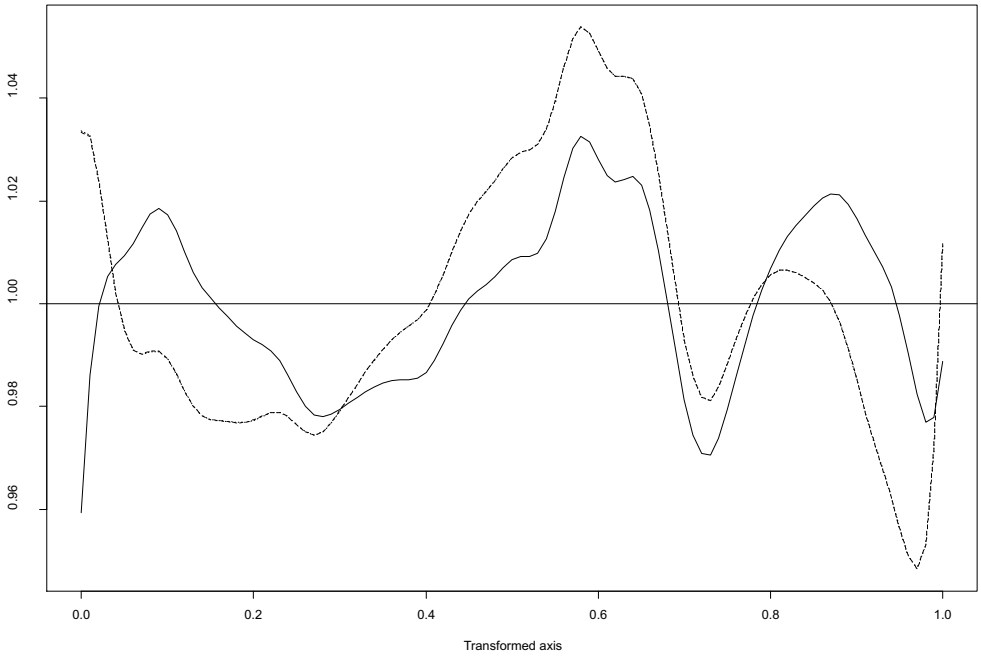


Figure 4: *United States: Local linear density estimator as in (2), with $b = 1/12$, on the transformed scale: Gamma frailty (solid) compared with Inverse Gaussian (dotted) and no frailty (dashed).*

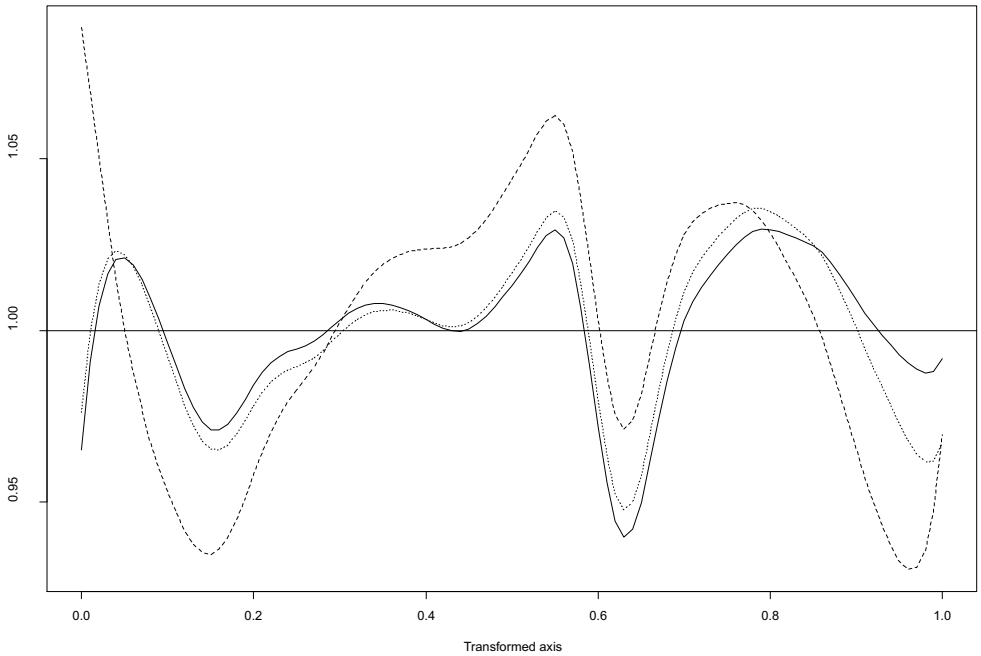


Figure 5: *United Kingdom: Local linear density estimator as in (2), with $b = 1/9$, on the transformed scale: Gamma frailty (solid) compared with Inverse Gaussian (dotted) and no frailty (dashed).*

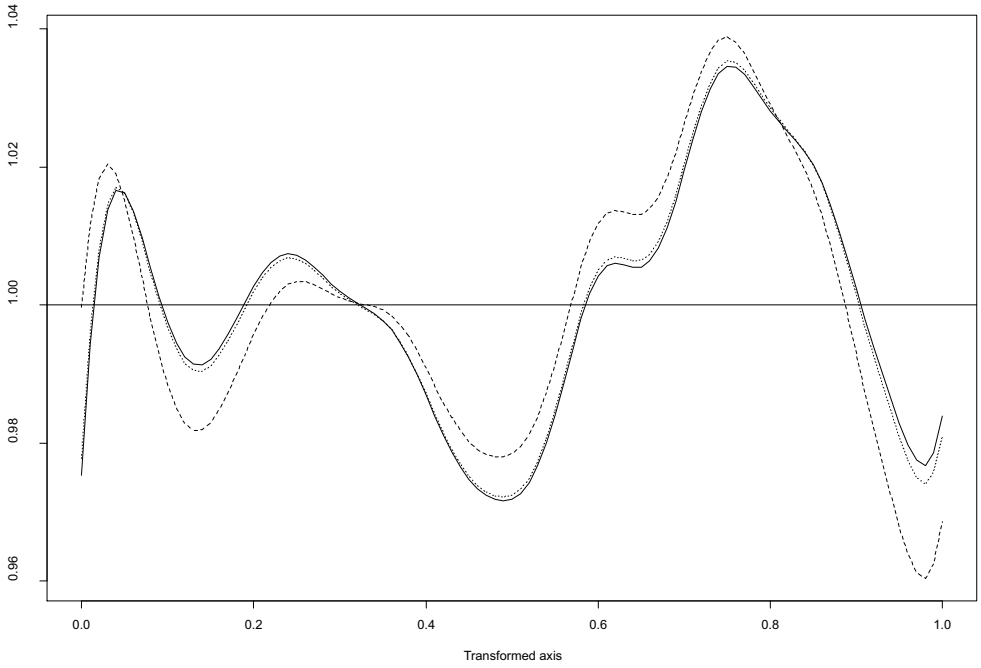


Figure 6: *Denmark: Local linear density estimator as in (2), with $b = 1/6$, on the transformed scale: Gamma frailty (solid) compared with Inverse Gaussian (dotted) and no frailty (dashed).*

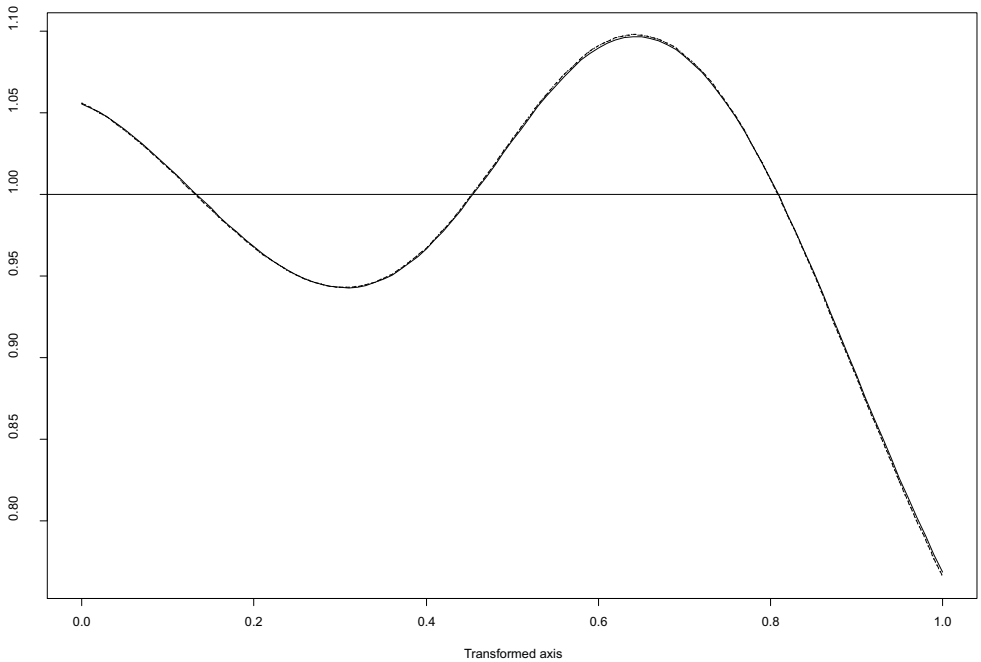


Figure 7: *Iceland: Local linear density estimator as in (2), with $b = 1/2$, on the transformed scale: Gamma frailty (solid) compared with Inverse Gaussian (dotted) and no frailty (dashed).*

Therefore, there is no reason to fear that we have smoothed too much and that should be the reason for the small deviance. This USA study gives us some confidence that the Gamma frailty survival model is working well also for smaller data sets, where bigger fluctuations are to be expected. For the United Kingdom the Gamma frailty survival model also fits relatively well, but now with deviances up to five percent. Surprisingly the Danish Gamma frailty survival density has very small deviances with the biggest being less than three percent. Iceland is another case, deviances up to 20% are found and the two frailty models do not seem to improve the fit compared to having no frailty at all. Overall the conclusion from the graphs is that the Gamma frailty makes the best fit, the Inverse Gaussian less so, but with both frailty models being superior to having no frailty at all. If we take a closer look at the tail of the three fitted Danish survival models at the transformed scale, we can get some further insight into the question posed in the introduction. It is indeed very clear that the flattening out of the Gamma frailty density in the tail helps the fit. The Gamma frailty version is much closer to one around the tail with about half the deviance from one compared to the no-frailty density version. In general, the performance of Inverse Gaussian is somewhat between that of Gamma and no frailty. For Iceland, the curves are almost identical, due to the small estimate of σ^2 and very similar estimates of the other parameters. In other words, for Iceland, the cases of Gamma frailty, Inverse Gaussian frailty and no frailty are nearly the same.

For US and UK the Gamma specification clearly provides the best description of data of the candidates considered. Both the Inverse Gaussian and no frailty alternative deviate substantially more in the right tail than Gamma frailty. These two specifications both overestimate old age mortality substantially, while the Gamma frailty seems to capture the old age mortality plateau evident in data. Moreover, the right tail deviations of Gamma frailty is of the same magnitude as deviations for younger age segments, while the right tail deviations for Inverse Gaussian and no frailty seems to diverge. While the picture is less clear for Denmark, the frailty densities also here improve the description of old age mortality. It also seems that without frailty the deviation diverges in the right tail, but the magnitude of deviation is much smaller than for US and UK. In contrast to US and UK, the Gamma and Inverse Gaussian essentially perform equally well. Thus we conclude that there is enough information in data to indicate the presence of heterogeneity, but not enough information to distinguish between the different kinds of heterogeneity.

Lastly, Iceland has so little exposure and so much uncertainty in data that even with the method derived in this paper we cannot distinguish between the models.

A critical part of the study concerns the performance of the estimator for advanced ages. To this end, for each country we calculate the second largest and largest points of intersection of the estimator with the horizontal line (i.e. the two largest roots of the equation $\widehat{f}_d(y) = 1$) and translate this back into the corresponding ages. For comparability, we have left out in this investigation the late spike of no frailty and the Inverse Gaussian in the USA case. The results are given in Table 2 below. It is quite clear that Gamma frailty densities in all case are having the last crossing point. This indicates

that the Gamma frailty density provides the best description of old age mortality among the considered models. In the USA case, we get almost to the age of 100 before our transformed density drops below one. The lowest last crossing for the gamma frailty is still quite high, namely 93 years. Above this last crossing point on the transformed scale, all the fitted parametric models seem to have too low densities. Thus, above the last crossing point our parametric models are overstating the possibility of dying. In other words, above the last crossing point all models seem to be on the safe side. In particular the Gamma frailty seems well behaved for annuity purposes. The density of very old are a bit too high, but rarely more than two percent, and these two percent are on the safe side when calculating for example annuities. Most of the extra old age mass is taken from the interval between the next last crossing point and the last crossing point, where the underlying parametric densities are overestimated in all cases. Therefore, while none of the densities are making a perfect fit, the Gamma frailty density is very close and with good properties for the annuity forecaster. It is on the safe side for the very old ages, with an overall annuity that seems to be close to the truth, overestimating the density in the very old ages, but compensating for that overestimation in an interval leading up to those old ages. Without frailty the deviations for old ages are substantially larger than with (gamma) frailty. The transformed density is below 1 which indicates that the probability of dying old is overestimated. At first glance this appears to be at odds with the fact that without frailty the old-age hazard is overestimated, cf. Figure 1. The explanation is that while the old-age hazard is overestimated the hazard is underestimated in the age groups below and therefore too many attain the (high) age of 90, say, after which they die too quickly. The model without frailty is on the safe when setting aside reserves for annuities for 40 year-olds, but if we were to use the same model for older age groups it would only be conservative up to a certain point. This clearly is not a desirable feature, and it illustrates the point that overstating the probability of dying old for one cohort is not necessarily a conservative assumption for other cohorts.

Notice that it would be hard to get this kind of detailed information from testing the underlying densities or even from graphical visualization techniques on the original scale. Therefore, our simple transformation technique has enabled us to comfort the statistician forecasting mortality models based on simple underlying parametric survival distributions.

Table 2: *Second largest and largest crossing point of density estimator with horizontal line at 1.*

Country	**Gamma**		**Inverse Gaussian**		**No frailty**	
	Second largest crossing point	Largest crossing point	Second largest crossing point	Largest crossing point	Second largest crossing point	Largest crossing point
US	92.37	98.68	92.14	95.21	92.14	95.21
UK	89.51	96.85	89.15	95.34	87.61	92.10
Denmark	85.53	94.77	85.47	94.67	84.91	93.63
Iceland	84.38	93.02	84.34	93.01	84.34	93.01

5. Conclusions

We have developed a new visual inspection technique of survival models. It generalizes developments of transformation techniques of i.i.d. data, see for example Bolancé et al. (2008, 2012a, 2012b, 2013). The method seems useful in many versions of follow-up studies, see for example Guillén et al. (2012) and Pinquet et al. (2011). We imagine it to be useful when the applied statistician wants the data to guide his intuition. The working methodology could be through running the knowledge loop cycle: Data→Visualization→New Assumption a number of times until the final assumptions seem intuitively reasonable and well behaved also according to more standard statistical techniques.

All the mortality projection models discussed in the Introduction involve both an age and a time dimension. As mentioned in the Introduction one can use our one-dimensional visualization technique for the age effect after having adjusted for the time effect and vice versa when visualizing the time effect. A full multidimensional version of our methodology is also possible. One could use multidimensional density estimation of filtered data to introduce a similar visual inspection technique to assessing the quality of mortality depending on both age and time. See for example Buch-Kromann and Nielsen (2012) for a recent multivariate density estimator that could be used in our visual diagnostic step after having transformed our data with our favourite forecasting mortality model.

Transformations and visual fitting as developed in this paper would also seem relevant in other areas of actuarial science as, for example, reserving, see the recent papers Martínez-Miranda et al. (2012) and Kuang et al. (2011).

Acknowledgement

This project was funded by a research grant from The Actuarial Foundation and the Society of Actuaries.

References

Abbring, J. H. and Van den Berg, G. J. (2007). The unobserved heterogeneity distribution in duration analysis. *Biometrika*, 94 (1), 87–99.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Bolancé, C., Guillen, M. and Nielsen, J. P. (2008). Inverse beta transformation in kernel density estimation. *Statistics and Probability Letters*, 78, 1757–1764.

Bolancé, C., Guillén, M., Nielsen, J. P. and Gustafsson, J. (2012a). *Quantitative Operational Risk Models*. Chapman and Hall/CRC Finance Series, New York.

Bolancé, C., Ayuso, M. and Guillén, M. (2012b). A nonparametric approach to analysing operational risk with an application to insurance fraud. *The Journal of Operational Risk*, 7 (1), 1–16.

Bolancé, C., Guillén, M., Gustafsson, J. and Nielsen, J. P. (2013). Adding prior knowledge to quantitative operational risk models. *The Journal of Operational Risk*, 8 (1), 17–32.

Borgan, O. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, 11, 1–16.

Buch-Kromann T. and Nielsen, J. P. (2012). Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated and censored data. *Annals of the Institute of Mathematical Statistics*, 48 (1), 167–192.

Butt, Z. and Haberman, S. (2004). Application of frailty-based mortality models using Generalized Linear Models. *ASTIN Bulletin*, 34 (1), 175–197.

Cairns, A. J. G., Blake, D. and Dowd, K. (2006). A two factor model for stochastic mortality and parameter uncertainty: theory and calibration. *The Journal of Risk and Insurance*, 73 (4), 687–718.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13 (1), 1–35.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D. and Khalaf-Allah, M. (2011). Mortality density forecasts: an analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48, 355–367.

Currie, I. D. (2006). Smoothing and forecasting mortality rates with P-splines. *Talk given at the Institute of Actuaries*, June 2006. http://www.actuaries.org.uk

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D. and Khalaf-Allah, M. (2010a). Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics*, 47, 255–265.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D. and Khalaf-Allah, M. (2010b). Backtesting stochastic mortality models: an ex post evaluation of multi-period ahead density forecasts. *North American Actuarial Journal*, 14 (3), 281–298.

Forfar, D. O., McCutcheon, J. J. and Wilkie, A. D. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, 115, 1–149.

Gámiz-Pérez, M. L., Martínez-Miranda, M. D. and Nielsen, J. P. (2013a). Smoothing survival densities in practice. *Computational Statistics and Data Analysis*, 58 (1), 368–382.

Gámiz Pérez, M. L, Mammen, E., Martínez Miranda, M. D. and Nielsen, J. P. (2013b). Do-validating local linear hazards. *Submitted preprint.*

Gámiz-Pérez, M. L., Janys, L., Martínez-Miranda, M. D. and Nielsen, J. P. (2013c). Smooth marker dependent hazard estimation in praxis. *Computational Statistics and Data Analysis*, *Forthcoming.*

Guillén, M., Nielsen, J. P., Scheike, T. H. and Pérez-Marín, A. M. (2012). Time-varying effects in the analysis of customer loyalty: A case study in insurance. *Expert Systems with Applications*, 39 (3), 3551–3558.

Haberman, S. and Renshaw, A. E. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48, 35–55.

Horiuchi, S. and Coale, A. J. (1990). Age patterns of mortality for older women: an analysis using the age-specific rate of mortality change with age. *Mathematical Population Studies*, 2 (4), 245–267.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 71 (1), 75–83.

Jarner, S. F. and Kryger, E. M. (2011). Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin*, 41 (2), 377–418.

Jones, B. L. (1998). A model for analyzing the impact of selective lapsation on mortality. *North American Actuarial Journal*, 2 (1), 79–86.

Kuang D., Nielsen, B. and Nielsen, J. P. (2011). Forecasting in an extended chain-ladder-type model. *Journal of Risk and Insurance*, 78 (2), 345–359.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 87 (419), 659–671.

Li, J. S.-H., Hardy, M. R. and Tan, K. S. (2009). Uncertainty in mortality forecasting: an extension of the Lee-Carter approach. *ASTIN Bulletin*, 39 (1), 137–164.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2011). Do-validation for kernel density estimation. *Journal of the American Statistical Association*, 106 (494), 651–660.

Martínez-Miranda, M. D., Nielsen, J. P. and Wüthrich, M. V. (2012). Statistical modelling and forecasting of outstanding liabilities in non-life insurance. *SORT-Statistics and Operations Research Transactions*, 36 (2), 195–218.

Nielsen, J. P., Tanggaard, C. and Jones, M. C. (2009). Local linear density estimation for filtered survival data. *Statistics*, 43 (2), 167–186.

Olivieri, A. (2006). Heterogeneity in survival models-applications to pensions and life annuities. *Belgian Actuarial Bulletin*, 6 (1), 23–39.

Pinquet, J., Guillén, M. and Ayuso, M. (2011). Commitment and lapse behavior in long-term insurance: a case study. *Journal of Risk and Insurance*, 78 (4), 983–1002.

Plat, R. (2009). On stochastic mortality modelling. *Insurance: Mathematics and Economics*, 45, 393–404.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38, 556–570.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.

Wang, S. and Brown, R. L. (1998). A frailty model for projection of human mortality improvements. *Journal of Actuarial Practice*, 6, 221–241.

SORT 37 (2) July-December 2013, 175-188

Quantile estimation of the rejection distribution of food products integrating assessor values and interval-censored consumer data

Klaus Langohr[1], Guadalupe Gómez[1] and Guillermo Hough[2]

Abstract

Fitting parametric survival models with interval-censored data is a common task in survival analysis and implemented in many statistical software packages. Here, we present a novel approach to fit such models if the values on the scale of interest are measured with error. Random effects ANOVA models are used to account for the measurement errors and the likelihood function of the parametric survival model is maximized with numerical methods. An illustration is provided with a real data set on the rejection of yogurt as a function of its acid taste.

MSC: 62N99, 62F99.

Keywords: Interval-censored data; maximization of the likelihood function; parametric survival model; sensory shelf-life data.

1. Introduction

Since the publication of the work of Hough et al. (2003), survival data methods have become a common tool for the analysis of sensory shelf-life data of foods; see applications, among others, in Curia et al. (2005), Araneda et al. (2008) and Østli et al. (2013). The methodology has also been applied to determine consumer acceptance limits of sensory defects (Hough et al., 2004), and to optimize the concentration of food ingredients (Garitta et al., 2006).

[1] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya/BARCELONATECH, Spain. E-mail: klaus.langohr@upc.edu

[2] Departamento de Evaluación Sensorial de Alimentos, Instituto Superior Experimental de Tecnología Alimentaria, Nueve de Julio, Buenos Aires, Argentina.

Received: December 2012
Accepted: July 2013

A typical shelf-life study consists of storing food samples for different lengths of time. For each time, consumers evaluate the product and report their acceptance or rejection. For example, for a yogurt study (Curia et al., 2005), samples were stored for 0, 14, 28, 42, 56, 70, and 84 days. A typical consumer's response would be: accept, accept, accept, reject, reject, reject, and reject, for each one of the respective times. This consumer's data is interval-censored between 28 and 42 days. Another consumer might accept all samples, and in this case the data would be right-censored at 84 days. A left-censored consumer would be one who rejects the sample which has only been stored for 14 days. Thus, data on the acceptance or rejection of a food product are generally interval-censored – including both left and right censoring as particular cases – where the intervals contain the real unknown values of rejection on the scale of interest; for the yogurt example the scale of interest was storage time.

The methodology proposed by Hough et al. (2003) furnishes the estimation of the rejection quantiles of interest for a given parametric model such as the Weibull, loglogistic or lognormal distribution. It assumes that the endpoints of the observed censoring intervals are all measured exactly without any error. Another instance of the application of this methodology is found in Sosa et al. (2008) who estimated the optimum concentration of salt in French-type bread from a consumer's perspective. They prepared samples of bread with 0.6, 1.2, 1.8, 2.4, 3.0, 3.6, and 4.2 g sodium chloride per 100 g of flour. Since the weighing error of these salt quantities could be considered negligible, the values could be taken as exact.

However, the values of the independent variable may not always be free of error. Consider the case of a yogurt manufacturer who has applied survival analysis methodology to establish sensory shelf life of his product as described by Hough et al. (2003). If this manufacturer, in the future, wants to test a formulation change and make sure the sensory shelf life is still valid, he/she would have to assemble approximately 100 consumers (Hough et al., 2007). This is a costly and time-consuming experiment. If the critical descriptor (Hough, 2010) of yogurt from a consumer's perspective is acid taste, it would be of interest to the manufacturer to know how much the acid taste can increase before reaching 50% consumer rejection. If this acid taste cut-off value is known, then for future shelf-life determinations of the yogurt, the manufacturer can assemble a trained panel to measure acid taste instead of assembling the costly consumer panel. In this case the independent variable of the survival analysis experiment would become acid taste. These values are measured on a sensory scale by a sensory panel consisting of trained assessors. Presented with the same stimulus (a sample of yogurt) different assessors can produce different responses on the sensory scale; and the same assessor can produce different responses to sample replicates; thus the measurements are with error.

The objective of this work is to estimate the quantiles of the rejection distribution of a given food product integrating data from trained assessors and from consumers. Trained assessors provide the value of a certain characteristic of the product, such as the acid taste of yogurt. These values are random and subject to two sources of variability, one inherent to the assessor and the other corresponding to the specific acid taste of

the product. Consumers will evaluate the same products as the trained assessors, stating their acceptance or rejection. Consumer data is interval-censored, where the endpoints of the intervals are random variables corresponding to the trained panel's measurements. Unlike other works, the law of the censoring endpoints is taken into account.

The content of the remainder of this work is the following. After describing the data of both trained assessors in Section 2 and consumers in Subsection 3.1, the likelihood function of the model and data under study is derived in Subsection 3.2. In Section 4, we give details on how to maximize this function in the framework of a parametric model and how to estimate the parameters and quantiles of interest. Section 5 presents the application of the estimation proposal to the motivating data set under study and in Section 6 the main results of this work are discussed.

2. Trained assessors: data, model, and analysis

For the sake of a better understanding, throughout the following sections, we use the data on the rejection of yogurt as a function of its acid taste. Yogurt samples were stored different times so that they would develop different levels of acid taste. These samples were given both to a panel of trained assessors and to consumers. Assessors received three replicates of each sample and measured their level of acid taste on a common scale from 0 to 100. Consumers received a single replicate of each sample and judged whether or not they would accept it.

2.1. Data and Model

A panel of J trained assessors are given K replicates of I different samples of yogurt which correspond to I different degrees of acid taste. Acid taste, denoted by X_{ijk}, was measured on a sensory scale from 0 (minimum acid taste) to 100 (maximum value), where k stands for replication ($k = 1,\ldots,K$), j for assessor ($j = 1,\ldots,J$), and i for sample ($i = 1,\ldots,I$). In our motivating example, we have $K = 3$, $J = 13$, and $I = 6$. A graphical representation of all trained assessors' data is shown in Figure 1.

It is assumed that the data of a given sample i, $i = 1,\ldots,I$, come from a one-way random effects ANOVA model:

$$X_{ijk} = \mu_i + \alpha_{ij} + \epsilon_{ijk}, \tag{1}$$

where $\alpha_{ij} \sim \mathcal{N}(0, \sigma^2_{b;i})$ and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_{w;i})$. For sample i, the grand mean μ_i, representing the unknown acid taste of sample i, is the parameter of interest, α_{ij} is the random effect corresponding to assessor j, $j = 1,\ldots,J$, and $\sigma^2_{b;i}$ and $\sigma^2_{w;i}$ denote, respectively, the between and within-assessors variances. Note that $\sigma^2_{b;i}$ is equivalent to the covariance between two observations on the same assessor (Vittinghoff et al., 2005, Chap. 8).

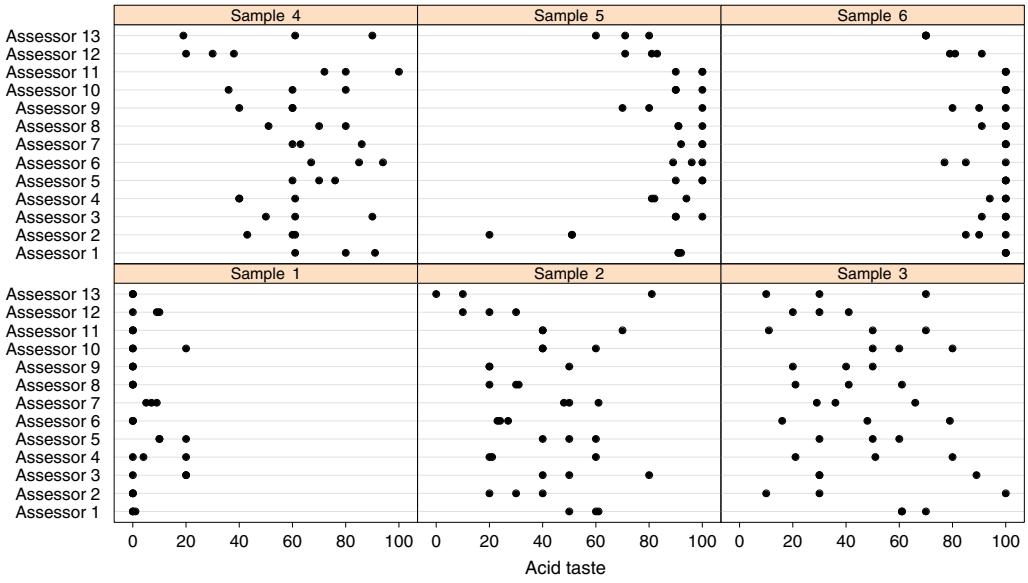


Figure 1: *Estimates of acid taste of yogurt given by 13 trained assessors on three replications of six different samples. Acid taste was measured on a 0 (minimum) to 100 (maximum) sensory scale.*

The model assumes independence among the assessors and between α_{ij} and ϵ_{ijk}. Hence, the overall variance of X_{ijk} is the sum of both variance components, that is, $\text{Var}(X_{ijk}) = \sigma^2_{b;i} + \sigma^2_{w;i}$. In addition, and without loss of generality,

$$0 < \mu_1 < \cdots < \mu_I < 100, \tag{2}$$

where 100 may be substituted by any other value determined to be the maximum of the scale of interest.

2.2. Estimation

The estimator of the grand mean μ_i, $i = 1, \ldots, I$, is given by the overall mean of all $J \cdot K$ measurements given for each sample:

$$\hat{\mu}_i = \bar{X}_i = \frac{1}{J}\frac{1}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk},$$

and its variance is equal to

$$\text{Var}(\bar{X}_i) = \frac{1}{J \cdot K}(\sigma^2_{w;i} + K \cdot \sigma^2_{b;i}).$$

See Appendix A for further details.

Given the normal distribution assumption in model (1), the distribution of $\bar{X}_i$ is

$$\bar{X}_i \sim \mathcal{N}\left(\mu_i, \frac{1}{J \cdot K}(\sigma^2_{w;i} + K \cdot \sigma^2_{b;i})\right), \tag{3}$$

and for $J = 13$ and $K = 3$, the overall mean $\bar{X}_i$ follows a normal distribution with mean μ_i and variance $\frac{1}{39}(\sigma^2_{w;i} + 3 \cdot \sigma^2_{b;i})$.

Several estimators exist for both variance components including the restricted maximum likelihood estimators shown in (4). They are based on the between and within-assessors sum of squares ($\mathrm{SS}_{b;i}$ and $\mathrm{SS}_{w;i}$):

$$\begin{aligned} \hat{\sigma}^2_{w;i} &= \min\left(\frac{\mathrm{SS}_{w;i}}{J(K-1)}, \frac{\mathrm{SS}_{w;i} + \mathrm{SS}_{b;i}}{J \cdot K - 1}\right), \\ \hat{\sigma}^2_{b;i} &= \max\left(0, \frac{1}{K}\left(\frac{\mathrm{SS}_{b;i}}{J-1} - \frac{\mathrm{SS}_{w;i}}{J(K-1)}\right)\right), \end{aligned} \tag{4}$$

where $\mathrm{SS}_{w;i} = \sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X}_{ij})^2$ and $\mathrm{SS}_{b;i} = \sum_{j=1}^{J} K \cdot (\bar{X}_{ij} - \bar{X}_i)^2$. Herein, $\bar{X}_{ij}$ is the mean of assessor j's values for the ith sample. For a detailed discussion on these and other possible estimators, see Chapter 2 in Sahai and Ojeda (2004). In Appendix B, we give some details on computational aspects with R (The R Foundation for Statistical Computing).

Applying the previous formulas to our data set, we obtain sample mean estimates, the between and within-assessors standard deviations as well as the standard errors of $\bar{X}_i$ for all six samples which are shown in Table 1. We observe, for example, that the within-assessors standard deviations for samples 1 and 6 are much smaller then the rest; this is also reflected in Figure 1. When assessors measure samples with very low (sample 1) or very high (sample 6) acidities, they are all in agreement as to how to score these extreme samples. However, when intermediate acidities (samples 2 to 5) are presented, assessors can differ in their scores due to different perceptions and responses. This can be observed in the case of sample 3, where the estimated between-assessors variance is virtually 0, indicating that the variability observed in the estimation of the acid taste of this sample can be attributed entirely to the within-assessors variance.

Table 1: *Estimation results for model* (1) *for all six samples.*

	Sample					
	1	**2**	**3**	**4**	**5**	**6**
$\hat{\boldsymbol{\mu}}_i$	4.2	39.2	46.2	62.7	85.8	93.4
$\hat{\boldsymbol{\sigma}}_{w;i}$	5.9	17.7	23.4	17	8.6	5.6
$\hat{\boldsymbol{\sigma}}_{b;i}$	3.9	8.5	0.0	11.4	15	8.3
$\hat{\boldsymbol{\sigma}}_{\bar{X}_i}$	1.4	3.7	3.7	4.2	4.4	2.5

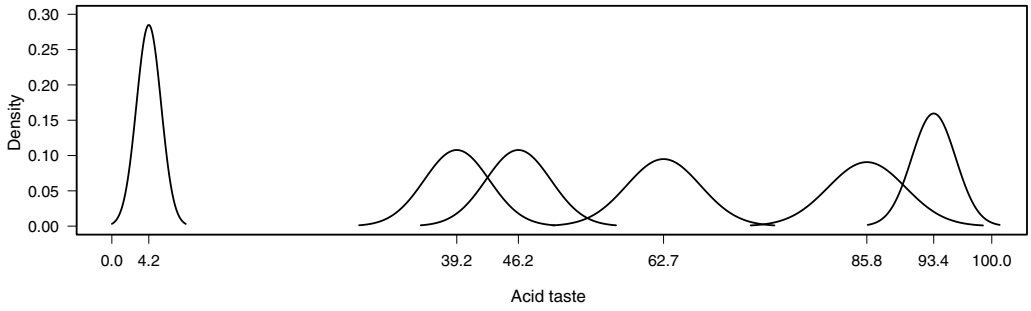


Figure 2: *Density functions of sample mean estimators* $\bar{X}_i$, $i = 1, \ldots, 6$.

In Figure 2, the density functions of all six mean estimators are represented assuming $\sigma^2_{b;i}$ and $\sigma^2_{w;i}$ are equal to the estimates shown in Table 1. In the following section, we will show how the uncertainty in the estimation of μ_i is taken into account in the analysis of the consumers' data.

3. Consumers: Data, rationale, and likelihood function

3.1. Data, rationale, and notation

In Section 1, the typical characteristics of a shelf-life study were presented. It was mentioned that the resulting data from the consumers, who are given the food product under study, are generally interval-censored containing the unknown value of rejection. Note that survival analysis methods can be applied to any positive random variable, for instance, yogurt's acid taste, as it is applied in the study that motivated the present work.

A total of $n = 74$ subjects are presented with $I = 6$ yogurt samples of different acid taste in a random order and have to answer the question whether they would normally consume such a yogurt or not. Based on their answers (acceptance/rejection), intervals of degrees of acid taste are determined that contain the acid taste from which a yogurt would be rejected. The interval for subject m, $m = 1, \ldots, n$, is of either of the two following types, where l_m and r_m indicate the sample number: $(l_m, r_m]$ or, in case of a right-censored observation, (l_m, ∞). In case of a left-censored observation, we define $l_m = 0$. Hence, $l_m \in \{0, \ldots, I\}, \forall m$, and $r_m \in \{1, \ldots, I, \infty\}, \forall m$. We denote the corresponding (unknown) acid tastes on the sensory scale from 0 to 100 by $(X_{l_m}, X_{r_m}]$, $m = 1, \ldots, n$.

In Table 2, the frequency distribution of the intervals obtained is shown. It can be seen that there are no left-censored and two right-censored data. That is, all subjects accepted sample 1, and two subjects did not reject any of the six samples. The fact that apart from the two right-censored observations not all of the remaining intervals are of type $(l_m, l_m + 1]$ is due to certain inconsistencies of the consumers' answers such as a sequence of "accept, reject, accept, reject". In that particular case, the interval obtained

Table 2: *Frequency distribution of intervals that contain rejection value.*

Interval	***n***	***%***
$(1,2]$	5	6.8
$(1,4]$	5	6.8
$(1,5]$	7	9.4
$(2,3]$	10	13.5
$(2,5]$	7	9.4
$(3,4]$	12	16.2
$(3,6]$	1	1.4
$(4,5]$	25	33.8
$(6,\infty)$	2	2.7
Total	74	100

is of type $(l_m, l_m+3]$, for example $(1,4]$; see Hough et al. (2003) for a more detailed discussion.

If the sample numbers were substituted by the corresponding estimated acid tastes shown in Table 1 without taking into account the uncertainty of the estimation, one could apply standard nonparametric methodology such as the Turnbull estimator (Turnbull, 1976) to estimate the quantiles of interest. The resulting graphical representation is shown in Figure 3 indicating, for example, that, according to this estimation, the median lies between 46.2 and 62.7.

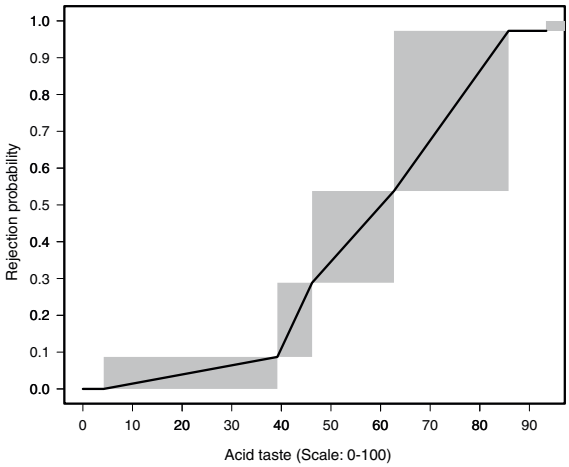


Figure 3: *Turnbull estimator of F if acid tastes were estimated without error.*

3.2. The likelihood function

In the following, we denote the distribution function of the random variable T, the acid taste from which yogurts are rejected, by F_T.

Assuming non-informative censoring (Oller et al., 2004) and if the acid tastes were measured without error, the contribution to the likelihood function of subject m, whose rejection value lies in interval $(x_{l_m}, x_{r_m}]$, would be (Gómez et al., 2009)

$$L_m = F_T(x_{r_m}) - F_T(x_{l_m}). \tag{5}$$

However, the exact acid tastes are unknown and estimates obtained from the panel of the 13 trained assessors are given instead. For this reason, we substitute the unknown acid tastes by these estimates and account for the corresponding uncertainty by integrating over the whole range of $\hat{\mu}_i$, $i = 1,\ldots,I$, which are all real-valued numbers in $[0, 100]$ restricted to $x_{l_m} < x_{r_m}$. Hence, the likelihood contribution in (5) converts into

$$L_m = \int_0^{100} \int_0^{r} \big(F_T(r) - F_T(l)\big)\, dF_{\bar{X}_{l_m}}(l)\, dF_{\bar{X}_{r_m}}(r). \tag{6}$$

Given a sample of size n, $(l_m, r_m]$, $m = 1,\ldots,n$, and assuming independence among the observations, the likelihood function is

$$L = \prod_{m=1}^{n} \int_0^{100} \int_0^{r} \big(F_T(r) - F_T(l)\big)\, dF_{\bar{X}_{l_m}}(l)\, dF_{\bar{X}_{r_m}}(r). \tag{7}$$

In case of left and right-censored observations, that is $l_m = 0$ and $r_m = \infty$, respectively, the likelihood contribution in (6) reduces to the following respective single integrals: $L_m = \int_0^{100} F_T(r)\, dF_{\bar{X}_{r_m}}(r)$ (left censoring) and $L_m = \int_0^{100} \big(1 - F_T(l)\big)\, dF_{\bar{X}_{l_m}}(l)$ (right censoring).

4. Maximization of the log-likelihood function

To maximize the logarithm of the likelihood function (7), following Wang (2010), discrete supports for $\bar{X}_i$, $i = 1,\ldots,I$, with corresponding probability masses have to be chosen. We denote these by

$$S_i = \{s_{i_1},\ldots,s_{i_{p_i}}\} \quad \text{and} \quad \mathbf{\Pi}_i = \{\pi_{i_1},\ldots,\pi_{i_{p_i}}\},\ i = 1,\ldots,I, \tag{8}$$

respectively. Different discrete supports of $\bar{X}_i$ can be thought of. For example, using the notation in (8), the first and last element of each support could be:

- $s_{i_1} = 0$ and $s_{i_{p_i}} = 100$,
- $s_{i_1} = \bar{x}_{i-1}$ and $s_{i_{p_i}} = \bar{x}_{i+1}$ with $\bar{x}_0 = 0$ and $\bar{x}_{I+1} = 100$,
- $s_{i_1} = \max(0, \bar{x}_i - p \cdot \hat{\sigma}_{\bar{x}_i})$ and $s_{i_{p_i}} = \min(100, \bar{x}_i + p \cdot \hat{\sigma}_{\bar{x}_i})$ for some $p \in \mathbb{N}$.

In either case, the mesh size h should be kept constant over the whole support, choosing, for example, $h = 0.1$ or $h = 0.5$.

The resulting expression of the log-likelihood function for the likelihood function given in (7) is as follows:

$$l = \sum_{m=1}^{n} \ln\Big(\sum_{v=1}^{p_{r_m}} \sum_{w=1}^{p_{l_m}} \big(F_T(s_{r_{mv}}) - F_T(s_{l_{mw}})\big)\pi_{l_{mw}}\pi_{r_{mv}} \mathbb{1}\{s_{l_{mw}} < s_{r_{mv}}\}\Big), \tag{9}$$

where both indices, v and w, cover the ranges of the corresponding supports but are restricted to $s_{l_{mw}} < s_{r_{mv}}, \forall v, w$, because of (2).

Given that $\bar{X}_i$ follows a normal distribution according to (3) and defining $\sum_{\Pi_i} = \sum_{l=1}^{p_i} f_{\bar{X}_i}(s_{i_l})$, we propose the following probability masses $\mathbf{\Pi}_i$, which are proportional to the density function of $\bar{X}_i$ evaluated in each point of the support S_i:

$$\pi_{i_v} = f_{\bar{X}_i}(s_{i_v}) \Big/ \sum\nolimits_{\Pi_i}, \; v = 1, \ldots, p_i,$$

where

$$f_{\bar{X}_i}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{\bar{X}_i}} \exp\left(-\frac{1}{2}\Big(\frac{x - \bar{x}_i}{\hat{\sigma}_{\bar{X}_i}}\Big)^2 \right).$$

Hence, the expression of the log-likelihood function (9) becomes:

$$\begin{aligned} l = \sum_{m=1}^{n} \ln\Bigg(\Big(\sum_{v=1}^{p_{r_m}} \sum_{w=1}^{p_{l_m}} \big(F_T(s_{r_{mv}}) - F_T(s_{l_{mw}})\big) \\ \cdot \frac{1}{2\pi\hat{\sigma}_{\bar{X}_{l_m}}\hat{\sigma}_{\bar{X}_{r_m}}} \exp\Big(-\frac{1}{2}\Big(\Big(\frac{s_{l_{mw}} - \bar{x}_{l_m}}{\hat{\sigma}_{\bar{X}_{l_m}}}\Big)^2 + \Big(\frac{s_{r_{mv}} - \bar{x}_{r_m}}{\hat{\sigma}_{\bar{X}_{r_m}}}\Big)^2 \Big)\Big) \mathbb{1}\{s_{l_{mw}} < s_{r_{mv}}\} \Big) \Big/ \sum\nolimits_{\Pi_{l_m}} \cdot \sum\nolimits_{\Pi_{r_m}} \Bigg) \\ = \sum_{m=1}^{n} \Bigg(\ln\Big(\sum_{v=1}^{p_{r_m}} \sum_{w=1}^{p_{l_m}} \big(F_T(s_{r_{mv}}) - F_T(s_{l_{mw}})\big) \\ \cdot \frac{1}{2\pi\hat{\sigma}_{\bar{X}_{l_m}}\hat{\sigma}_{\bar{X}_{r_m}}} \exp\Big(-\frac{1}{2}\Big(\Big(\frac{s_{l_{mw}} - \bar{x}_{l_m}}{\hat{\sigma}_{\bar{X}_{l_m}}}\Big)^2 + \Big(\frac{s_{r_{mv}} - \bar{x}_{r_m}}{\hat{\sigma}_{\bar{X}_{r_m}}}\Big)^2 \Big)\Big) \mathbb{1}\{s_{l_{mw}} < s_{r_{mv}}\} \Big) - \ln\Big(\underbrace{\sum\nolimits_{\Pi_{l_m}} \cdot \sum\nolimits_{\Pi_{r_m}}}_{\blacksquare} \Big) \Bigg). \end{aligned}$$

and since $\blacksquare$ does not depend on F, the log-likelihood function to be maximized is

$$\begin{aligned} l = \sum_{m=1}^{n} \ln\Big(\sum_{v=1}^{p_{r_m}} \sum_{w=1}^{p_{l_m}} \big(F_T(s_{r_{mv}}) - F_T(s_{l_{mw}})\big) \\ \cdot \frac{1}{2\pi\hat{\sigma}_{\bar{X}_{l_m}}\hat{\sigma}_{\bar{X}_{r_m}}} \exp\Big(-\frac{1}{2}\Big(\frac{(s_{l_{mw}} - \bar{x}_{l_m})^2}{\hat{\sigma}^2_{\bar{X}_{l_m}}} + \frac{(s_{r_{mv}} - \bar{x}_{r_m})^2}{\hat{\sigma}^2_{\bar{X}_{r_m}}} \Big)\Big) \mathbb{1}\{s_{l_{mw}} < s_{r_{mv}}\} \Big). \end{aligned} \tag{10}$$

In case of left and right-censored data, the contributions to the log-likelihood function are, respectively:

$$l_m = \ln\Big(\sum_{v=1}^{p_1} F_T(s_{1v})\frac{1}{\sqrt{2\pi}\hat{\sigma}_{\bar{X}_1}}\exp\big(-\frac{1}{2}(\frac{s_{1v}-\bar{x}_1}{\hat{\sigma}_{\bar{X}_1}})^2\big)\Big),$$

$$l_m = \ln\Big(\sum_{w=1}^{p_I} \big(1-F_T(s_{Iw})\big)\frac{1}{\sqrt{2\pi}\hat{\sigma}_{\bar{X}_I}}\exp\big(-\frac{1}{2}(\frac{s_{Iw}-\bar{x}_I}{\hat{\sigma}_{\bar{X}_I}})^2\big)\Big).$$

As pointed out in the introduction, our objective consists of estimating the quantiles of the rejection distribution under different parametric models. That is, we will substitute F by different expressions according to the parametric choices for T as shown in the following section.

5. Quantile estimation for parametric models

Three parametric laws, which are commonly used for shelf-life studies of foods (Hough et al., 2003), are considered for the random variable of interest T:

- Weibull with shape parameter k, scale parameter λ, distribution function given by $F_T(t) = 1-\exp(-(t/\lambda)^k)$, and $t_\alpha = \lambda\cdot\ln(\frac{1}{1-\alpha})^{1/k}$ as the quantile α,
- loglogistic with shape parameter k, scale parameter λ, $F_T(t) = 1-\frac{1}{1+(t/\lambda)^k}$, and $t_\alpha = \lambda(\frac{\alpha}{1-\alpha})^{1/k}$,
- lognormal with parameters μ and σ, $F_T(t) = \Phi\big(\frac{\ln(t)-\mu}{\sigma}\big)$, and $t_\alpha = \exp\big(\mu+\sigma\cdot\Phi^{-1}(\alpha)\big)$.

For sample i, $i = 1,\ldots,I$, we have chosen a discrete support with first element given by $s_{i_1} = \max(0,\bar{x}_i - 3\cdot\hat{\sigma}_{\bar{x}_i})$, last element given by $s_{i_{p_i}} = \min(100,\bar{x}_i + 3\cdot\hat{\sigma}_{\bar{x}_i})$, and with mesh size equal to 0.1. These supports cover intervals on the domain of $\bar{X}_i$ of probability masses larger than 0.99 for each sample. With these choices, the computation time for the maximization of the log-likelihood function takes about 25 seconds with the Intel i7 processor (1.73 GHz) under Windows 7. Technical details on the implementation in R are given in Appendix B.

The maximization of function (10) yields the parameter estimates and five quantiles as shown in Table 3. Whereas the standard errors are returned together with the parameters' estimates, the delta method is used in order to compute the standard errors of the log-transformed quantiles. 95% confidence intervals are computed for $\ln(t_\alpha)$ and the exponential transformation is applied to obtain the confidence intervals for t_α. They are, hence, not symmetric with respect to $\hat{t}_\alpha$.

Table 3: *Estimates obtained under different parametric models: parameter estimates are shown together with standard errors, quantile estimates together with 95% confidence intervals.*

	Weibull	**Loglogistic**	**Lognormal**
$\hat{\boldsymbol{\theta}}$ $(s.e.(\hat{\boldsymbol{\theta}}))$	$\hat{k} = 4.113$ (0.467)	$\hat{k} = 6.510$ (0.805)	$\hat{\mu} = 4.044$ (0.036)
	$\hat{\lambda} = 65.138$ (2.292)	$\hat{\lambda} = 57.426$ (2.153)	$\hat{\sigma} = 0.263$ (0.029)
Quantiles (95%-CI)			
0.1	37.7 ([32.2, 44.2])	41.0 ([36.5, 46.1])	40.7 ([36.5, 45.4])
0.25 (Q1)	48.1 ([43.0, 53.8])	48.5 ([44.4, 53.0])	47.8 ([43.9, 52.1])
0.5 (Median)	59.6 ([55.1, 64.5])	57.4 ([53.3, 61.9])	57.0 ([53.0, 61.3])
0.75 (Q3)	70.5 ([65.9, 75.4])	68.0 ([62.6, 73.8])	68.1 ([63.0, 73.6])
0.9	79.8 ([74.3, 85.7])	80.5 ([72.4, 89.5])	79.9 ([72.5, 88.1])

We can see, for example, that the estimated median under the Weibull model is 59.6 and that the corresponding 95% confidence interval ranges from 55.1 to 64.5. That is, under the Weibull model, 50% of all consumers are expected to reject yogurt with an acid taste above 59.6 and this value would serve as the cut-off value for the yogurt manufacturer if the objective is to produce yogurt whose acid taste is rejected by at most 50% of all consumers. Note that the median estimates are somewhat lower in case of the two other parametric choices (57.4 and 57, respectively) and that all three estimates lie in the interval obtained by the nonparametric estimation shown in Figure 3.

6. Conclusions and discussion

In this work, we have presented an approach to fit parametric models to interval-censored data when the interval limits are not fixed values, but are rather measured with certain error. As stated in the introduction, survival analysis methodology has so far been used to estimate rejection probabilities in food products as a function of variables of interest such as storage time which were measured exactly. However, there are other situations in which the variable of interest is not error-free, such was the case of acid taste in yogurt presented as an example in this work. We have developed a model to take account of the variability in the measurement of the independent variable.

Since the maximization of the likelihood function with such data is not implemented in statistical software, we have accomplished the parameter estimation in R with different functions of contributed packages; see Appendix B. The R code used can be provided on request from the authors.

The results obtained permit us to draw conclusions about the rejection distribution of a given food product based on a scale whose values are estimated by a trained panel. However, from a statistical point of view, our primary interest is the comparison of the obtained results with the ones of the method that ignores the uncertainty of the sample mean estimation. It could be expected that our approach would yield larger standard

errors and confidence intervals, nonetheless, the results (not shown here) are fairly similar. For example, the standard errors of both parameters of the Weibull distribution do only differ in the second decimal place among both methods, whereas they are even virtually the same considering the lognormal distribution. Therefore, the differences of the quantiles obtained with both methods as well as the widths of the corresponding confidence intervals are notably small. The same findings held when we used broader discrete supports for $\bar{X}_i$, $i = 1, \ldots, 6$.

Another approach to estimate the parameters is to use multiple imputations as described in Rubin (1987). For each of B runs, sample mean values would be generated from the normal distributions (3) and the parametric models would be fitted assuming these values were measured error-free. The parameters estimates are then obtained as means over the B estimates obtained. We did this for $B = 1000$ obtaining similar parameter estimates (results not shown) but with larger standard errors (between 18% and 44% larger) reflecting both sources of variances: between and within-imputation variances. We, therefore, do not recommend this approach.

Two aspects of interest, which were not addressed in this work, are the nonparametric estimation of F and methods to judge the goodness-of-fit of a given parametric choice. These are relevant topics for further research.

In summary, final results showed small differences in quantile estimations between our model and the ad hoc calculations that did not consider variability. Whether these small differences will hold for most practical applications is difficult to predict. Our recommendation is for researchers to apply the complete model presented in this work in order to be sure that their quantile estimations are correct.

Acknowledgements

The authors are grateful to the research group *Grup de Recerca en Anàlisi Estadística de la Supervivència* (GRASS) for the fruitful discussions. This research was partially supported by Grant MTM2012-38067-C02-01 of the Ministerio de Economía y Competitividad (Spain). We also thank the reviewers of this paper for their useful comments and suggestions resulting in an improvement of the manuscript.

A. Variance of the sample mean estimator $\bar{X}_i$

In the following, the variance of the sample mean estimators $\bar{X}_i$, $i = 1, \ldots, I$, is derived. Remember that independence is assumed among assessors and that the covariance between two observations on the same assessor is equal to the between-assessors variance:

$$\begin{aligned}
\mathrm{Var}(\bar{X}_i) &= (\frac{1}{J}\frac{1}{K})^2 \mathrm{Var}\Big(\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk}\Big) = \frac{1}{(J\cdot K)^2}\sum_{j=1}^{J}\mathrm{Var}\Big(\sum_{k=1}^{K} X_{ijk}\Big) \\
&= \frac{1}{(J\cdot K)^2}\cdot J\cdot \mathrm{Var}\Big(\sum_{k=1}^{K} X_{ijk}\Big) \\
&= \frac{J}{(J\cdot K)^2}\Big(\sum_{k=1}^{K}\mathrm{Var}(X_{ijk}) + 2\cdot\sum_{k=2}^{K}\sum_{k^*=1}^{k-1}\mathrm{Cov}(X_{ijk}, X_{ijk^*})\Big) \\
&= \frac{J}{(J\cdot K)^2}\big(K\cdot(\sigma^2_{b;i}+\sigma^2_{w;i}) + 2\cdot\frac{1}{2}\cdot(K-1)\cdot K\cdot\sigma^2_{b;i}\big) \\
&= \frac{1}{J\cdot K}(\sigma^2_{w;i} + K\cdot\sigma^2_{b;i}).
\end{aligned}$$

B. Computational Issues

All computations of this work were carried out with R (The R Foundation for Statistical Computing), version 3.0.1. Following, we give some details on the functions used.

The estimates of the one-way random effects ANOVA model shown in Table 1 are obtained by fitting model (1) with function `lme` of package `nlme` (Pinheiro et al., 2013). This function, which uses the restricted maximum likelihood estimators in (4) for the variance components by default, could also handle unbalanced designs with different numbers of replicates among assessors.

The maximization of the log-likelihood function (10) under different parametric models was accomplished with function `mle2` of the contributed package `bbmle` (Bolker and R Development Core Team, 2012). This function returns both the maximum likelihood estimates and their standard errors. As initial values for the parameters to be estimated, which are required by the maximization algorithm, one can choose the parameter estimates that are obtained by fitting the corresponding parametric model under the assumption that sample means were measured error-free.

References

Araneda, M., Hough, G. and de Penna, E. (2008). Current-status survival analysis methodology applied to estimating sensory shelf life of ready-to-eat lettuce (*lactuta sativa*). *Journal of Sensory Studies*, 23(2), 162–170.

Bolker, B. and R Development Core Team. (2012). *bbmle: Tools for general maximum likelihood estimation*. R package version 1.0.5.2.

Curia, A., Aguerrido, M., Langohr, K. and Hough, G. (2005). Survival analysis applied to sensory shelf life of yogurts-I: Argentine formulations. *Journal of Food Science*, 70(7), S442–S445.

Garitta, L., Serrat, C., Hough, G. and Curia, A. (2006). Determination of optimum concentrations of a food ingredient using survival analysis statistics. *Journal of Food Science*, 71, S527–S531.

Gómez, G., Calle, M. L., Oller, R. and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9(4), 259–297.

Hough, G. (2010). *Sensory Shelf Life Estimation of Food Products*. Boca Raton, Florida: CRC Press.

Hough, G., Calle, M. L., Serrat, C. and Curia, A. (2007). Number of consumers necessary for shelf-life estimations based on survival analysis statistics. *Food Quality and Preference*, 18, 771–775.

Hough, G., Garitta, L. and Sánchez, R. (2004). Determination of consumer acceptance limits to sensory defects using survival analysis. *Food Quality and Preference*, 15, 729–734.

Hough, G., Langohr, K., Gómez, G. and Curia, A. (2003). Survival analysis applied to sensory shelf life of foods. *Journal of Food Science*, 68(1), 359–363.

Oller, R., Gómez, G. and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, 32, 315–325.

Østli, J., Esaiassen, M., Garitta, L., Nøstvold, B. and Hough, G. (2013). How fresh is fresh? perceptions and experience when buying and consuming fresh cod fillets. *Food Quality and Preference*, 27(1), 26–34.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team. (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1–109.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley and Sons.

Sahai, H. and Ojeda, M. (2004). *Analysis of Variance for Random Models 1: Balanced Data. Theory, Methods, Applications and Data Analysis*. Boston: Birkhäuser.

Sosa, M., Flores, A., Hough, G., Apro, N., Ferreyra, V. and Orbea, M. (2008). Optimum level of salt in french-type bread. Influence of income status, salt level in daily bread consumption and test location. *Journal of Food Science*, 73, S392–S397.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38, 290–95.

Vittinghoff, E., Glidden, D., Shiboski, S. and McCulloch, C. (2005). *Regression Methods in Biostatistics*. Springer.

Wang, Y. (2010). Maximum likelihood computation for fitting semiparametric mixture models. *Statistics and Computing*, 20(1), 75–86.

SORT 37 (2) July-December 2013, 189-210

An alternative to Kim and Warde's mixed randomized response model

Housila P. Singh and Tanveer A. Tarray*

Abstract

When open or direct surveys are about sensitive matters (e.g. gambling habits, addiction to drug and others intoxicants, alcoholism, proneness to tax invasion, induced abortions, drunken driving, history of past involvement in crimes, and homosexuality), non-response bias and response bias become serious problems because people oftentimes do not wish to give correct information. To reduce non-response and response bias, various alternative approaches have been proposed, for example a randomized response survey technique, or a mixed randomized response model using simple random sampling with a replacement sampling scheme that improves the privacy of respondents, proposed by authors Kim and Warde. In this paper we have suggested an alternative to Kim and Warde's mixed randomized response model to estimate the proportion of qualitative sensitive variable under the conditions presented in both the cases of completely truthful reporting and less than completely truthful reporting by the respondents. Properties of the proposed randomized response model have been studied along with recommendations. We have also extended the proposed model to stratified random sampling. Numerical illustrations and graphs are also given in support of the present study.

MSC: 62D05.

Keywords: Randomized response technique, Dichotomous population, Estimation of proportion, Privacy of respondents, Sensitive characteristics.

1. Introduction

Warner (1965) was first to introduce a randomized response (RR) model to estimate the proportion for sensitive attributes including homosexuality, drug addiction or abortion. Greenberg et al. (1969) proposed the unrelated question RR model that is a variation of Warner's (1965) RR model. Since the work by Warner (1965), a huge literature

* School of Studies in Statistics, Vikram University, Ujjain – 456010 – India. E-mail: tanveerstat@gmail.com
Received: February 2013
Accepted: September 2013

has emerged on the use and formulation of different randomization device to estimate the population proportion of a sensitive attribute in survey sampling. Mention may be made of the work of Tracy and Mangat (1996), Chudhuari and Mukherjee (1988), Ryu et al. (1993), Fox and Tracy (1986), Singh (2003), Singh and Tarray (2012, 2013a, b, c) and the references cited there in.

Mangat et al. (1997) and Singh et al. (2000) pointed out the privacy problem with the Moors (1971) model. Mangat et al. (1997) and Singh et al. (2000) have presented several strategies as an alternative to Moors model, but their models may lose a large portion of data information and require a high cost to obtain confidentiality of the respondents. Kim and Warde (2005) have suggested a mixed randomized response model using simple random sampling which rectifies the privacy problem.

In this paper we have suggested an alternative to Kim and Warde's (2005) mixed randomized response model and its properties are studied in simple random sampling with replacement (SRSWR) and Stratified random sampling in both the cases of completely truthful reporting and less than completely truthful reporting. Numerically we show that the proposed mixed randomized response model is better than Kim and Warde's (2005) estimator.

2. The suggested model

Let a random sample of size n be selected using simple random sampling with replacement (SRSWR). Each respondent from the sample is instructed to answer the direct question "I am a member of the innocuous group". If a respondent answers "Yes" to the direct question, then she or he is instructed to go to randomization device R_1 consisting of the statements (i) "I am the member of the sensitive trait group" and (ii) "I am a member of the innocuous trait group" with respective probabilities P_1 and $(1-P_1)$. If a respondent answers "No" to the direct question, then the respondent is instructed to use the randomization device R_2 consisting of the statements: (i) I belong to the sensitive group, (ii) "Yes" and (iii) "No" with known probabilities P, $(1-P)/2$ and $(1-P)/2$ respectively. For the second and third statements, the respondent is simply to report "Yes" or "No" as observed on the random device R_2 and it has no relevance to his actual status. It is to be mentioned that the randomization device R_2 is due to Tracy and Osahan (1999). The survey procedures are performed under the assumption that both the sensitive and innocuous questions are unrelated and independent in a randomization device R_1. To protect the respondent's privacy, the respondents should not disclose to the interviewer the question they answered from either R_1 or R_2.

Let n be the sample size confronted with a direct question and n_1 and n_2 $(=n-n_1)$ denote the number of "Yes" and "No" answers from the sample. Note that the respondents coming to R_1 have reported a "Yes" to the initial direct question, therefore $\pi_1 = 1$ in R_1, where π_1 is the proportion of "Yes" answers from the innocuous question.

Denote by 'Y' the probability of "Yes" from the respondents using R_1. Then

$$Y = P_1\pi_S + (1-P_1)\pi_1 = P_1\pi_S + (1-P_1), \tag{2.1}$$

where π_S is the proportion of "Yes" answers from the sensitive trait.

An unbiased estimator of π_S, in terms of the sample proportion of "Yes" responses $\hat{Y}$, becomes

$$\hat{\pi}_{a1} = \frac{\hat{Y} - (1-P_1)}{P_1}. \tag{2.2}$$

The variance of $\hat{\pi}_{a1}$ is

$$V(\hat{\pi}_{a1}) = \frac{Y(1-Y)}{n_1 P_1^2} = \frac{(1-\pi_S)\,[\,P_1\pi_S + (1-P_1)]}{n_1 P_1}$$

$$= \frac{1}{n_1}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right]. \tag{2.3}$$

The proportion of "Yes" answers from the respondents using randomization device R_2 follows:

$$X = P\pi_S + \frac{(1-P)}{2} \tag{2.4}$$

An unbiased estimator of π_S, in terms of the sample proportion of "Yes" responses $\hat{X}$, becomes

$$\hat{\pi}_{b1} = \frac{\hat{X} - (1-P)/2}{P}. \tag{2.5}$$

The variance of $\hat{\pi}_{b1}$ is given by

$$V(\hat{\pi}_{b1}) = \frac{X(1-X)}{n_2 P^2} = \left[\frac{\pi_S(1-\pi_S)}{n_2} + \frac{(1-P^2)}{4n_2P^2}\right]. \tag{2.6}$$

The estimator of π_S, in terms of the sample proportions of "Yes" responses $\hat{Y}$ and $\hat{X}$, is

$$\hat{\pi}_t = \frac{n_1}{n}\hat{\pi}_{a1} + \frac{n_2}{n}\hat{\pi}_{b1}$$

$$= \frac{n_1}{n}\hat{\pi}_{a1} + \frac{(n-n_1)}{n}\hat{\pi}_{b1}, \quad \text{for } 0 \;<\; \frac{n_1}{n} \;<\; 1. \tag{2.7}$$

As both $\hat{\pi}_{a1}$ and $\hat{\pi}_{b1}$ are unbiased estimators, the expected value of $\hat{\pi}_t$ is

$$E(\hat{\pi}_t) = E\left[\frac{n_1}{n}\hat{\pi}_{a1} + \frac{n_2}{n}\hat{\pi}_{b1}\right] = \frac{n_1}{n}\pi_S + \frac{(n-n_1)}{n}\pi_S = \pi_S.$$

Thus the proposed estimator $\hat{\pi}_t$ is an unbiased estimator π_S.

Now the variance of $\hat{\pi}_t$ is given by

$$\begin{aligned} V(\hat{\pi}_t) &= \left(\frac{n_1}{n}\right)^2 V(\hat{\pi}_{a1}) + \left(\frac{n_2}{n}\right)^2 V(\hat{\pi}_{b1}) \\ &= \left(\frac{n_1}{n}\right)^2 \frac{1}{n_1}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right] \\ &+ \left(\frac{n_2}{n}\right)^2 \frac{1}{n_2}\left[\pi_S(1-\pi_S) + \frac{(1-P^2)}{4P^2}\right] \\ &= \frac{n_1}{n^2}\left[\pi_S(1-\pi_S) + \frac{(1-\pi_S)(1-P_1)}{P_1}\right] + \frac{n_2}{n^2}\left[\pi_S(1-\pi_S) + \frac{(1-P^2)}{4P^2}\right]. \quad (2.8) \end{aligned}$$

Since our mixed RR model also uses Simmon's (1967) method when $\pi_1 = 1$, we can apply Lanke's (1976) idea to our suggested model. Thus using Lanke's (1976) result for P with $\pi_1 = 1$, we get

$$P = \frac{1}{2-P_1}. \quad (2.9)$$

Putting $P = (2-P_1)^{-1}$ in (2.6), we get

$$\begin{aligned} V(\hat{\pi}_{b1}) &= \frac{\pi_S(1-\pi_S)}{(n-n_1)} + \frac{(1-P_1)\,(3-P_1)}{4(n-n_1)} \\ &= \frac{1}{(n-n_1)}\left[\pi_S(1-\pi_S) + \frac{(1-P_1)(3-P_1)}{4}\right]. \quad (2.10) \end{aligned}$$

Thus we established the following theorem.

Theorem 2.1 *The variance of $\hat{\pi}_t$ is given by*

$$V(\hat{\pi}_t) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)\,[\,4\,\lambda(1-\pi_S) + (1-\lambda)P_1(3-P_1)]}{4nP_1} \quad (2.11)$$

for $n = n_1 + n_2$ and $\lambda = \dfrac{n_1}{n}$.

3. Efficiency comparisons

An efficiency comparison of the suggested model, under completely truthful reporting case, has been done with Kim and Warde's (2005) model.

From Kim and Warde's (2005) model, we have

$$V(\hat{\pi}_{kw}) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)\,[\ \lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2}. \tag{2.12}$$

From (2.11) and (2.12) we have $V(\hat{\pi}_t) < V(\hat{\pi}_{kw})$ if

$$\frac{[4\lambda(1-\pi_S)+(1-\lambda)P_1(3-P_1)]}{4} < \frac{[\lambda P_1(1-\pi_S)+(1-\lambda)]}{P_1}$$

i.e. if $4-3P_1^2+P_1^3 > 0$ which is always true.

Thus the proposed model is always better than Kim and Warde's (2005) model.

An efficiency comparison of the proposed mixed randomized response technique to that of Kim and Warde's, we have computed the percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's estimator $\hat{\pi}_{kw}$ by using the formula:

$$\mathrm{PRE}(\hat{\pi}_t,\ \hat{\pi}_{kw}) = \frac{V(\hat{\pi}_{kw})}{V(\hat{\pi}_t)} \times 100$$

$$= \frac{4\left[\pi_S(1-\pi_S)+\{(1-P_1)/P_1^2\}\{\lambda P_1(1-\pi_S)+(1-\lambda)\}\right]\pi_S(1-\pi_S)}{[4\pi_S(1-\pi_S)+\{(1-P_1)/P_1\}\{4\lambda(1-\pi_S)+(1-\lambda)P_1(3-P_1)\}]} \times 100$$

for different values of P_1, n and n_1.

We have obtained the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t,\ \hat{\pi}_{kw})$ for $\lambda = 0.3, 0.5, 0.7$ and for different cases of π_S, n, n_1 and P_1. Findings are shown in Table 1 and its diagrammatic representation is given in Figure 1.

It is observed from Table 1 and Figure 1 that: The values of percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t,\ \hat{\pi}_{kw})$ are more than 100. We can say that the envisaged estimator $\hat{\pi}_t$ is always efficient than Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$. Figure 1 shows results for $\pi_S = 0.1$ and 0.6, $\lambda = 0.3, 0.5, 0.7$ and different values of P_1, n, n_1.

We note from Table 1 that the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t,\ \hat{\pi}_{kw})$ decrease as the value of P_1 increases. Also the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t,\ \hat{\pi}_{kw})$ increase as the value of λ decrease for fixed values of π_S and P_1.

We further note from the results of Figure 1 that there is large gain in efficiency by using the suggested estimator $\hat{\pi}_t$ over the estimator $\hat{\pi}_{kw}$ when the proportion of stigmatizing attribute is moderately large.

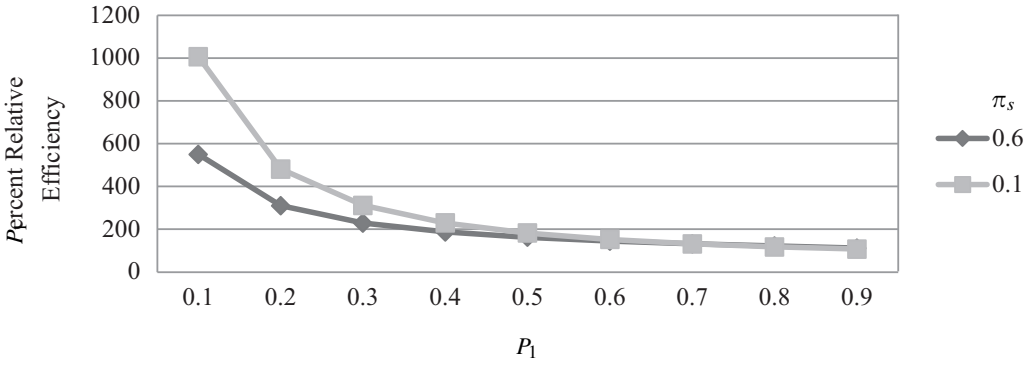


Figure 1: *Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$.*

4. Less than completely truthful reporting

The problem of "Less than completely truthful reporting" in randomized response model has been tackled by several authors including Singh (1993), Mangat (1994 a,b), Tracy and Osahan (1999), Chang and Huang (2001), Kim and Warde (2004), Kim and Elam (2005), Nazuk and Shabbir (2010) and others. We write the proportion of "Yes" answers from the two randomization devices R_1 and R_2, incorporating the probability of truthful reporting. Let T_1 and T_2 be the probabilities of telling the truth regarding the stigmatizing question in the randomization device R_1 and R_2 respectively. The respondents in the innocuous trait have no reason to tell a lie, they may lie for the sensitive trait.

Note that the respondents coming to R_1 have reported a "Yes" to the initial direct question therefore $\pi_1 = 1$ in R_1. The probability of "Yes" answers from the respondents using R_1 is given by

$$Y^* = P_1\pi_S T_1 + (1 - P_1). \tag{4.1}$$

An estimator for the true population proportion π_S of the sensitive trait is given by

$$\hat{\pi}_{a(1)} = \frac{\hat{Y}^* - (1 - P_1)}{P_1}, \tag{4.2}$$

where $\hat{Y}^*$ is the sample proportion of "Yes" response from the randomization device R_1.

Since $\hat{Y}^*$ follows Binomial distribution $\mathrm{B}(n_1, Y^*)$, therefore the bias and variance of the estimator $\hat{\pi}_{a(1)}$ are respectively given by

$$\mathrm{B}(\hat{\pi}_{a(1)}) = \pi_S(T_1 - 1) \tag{4.3}$$

Table 1: *Percent relative efficiency of the proposed estimator $\hat{\pi}_t$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$.*

π_S	$n=1000$		λ	P_1								
	n_1	n_2		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	700	300	0.7	550.05	309.93	228.78	187.50	162.21	144.90	132.03	121.65	112.11
0.1	500	500	0.5	1100.25	547.92	365.40	275.00	221.18	185.40	159.56	139.32	121.34
0.1	300	700	0.3	2201.05	971.19	586.67	406.25	304.10	239.21	194.28	160.44	131.68
0.2	700	300	0.7	596.76	327.10	236.25	190.24	162.21	143.17	129.20	118.24	108.94
0.2	500	500	0.5	1193.27	576.47	374.72	275.91	217.80	179.67	152.64	132.14	115.37
0.2	300	700	0.3	2352.39	1000.00	586.67	396.59	290.90	224.95	180.24	147.72	122.23
0.3	700	300	0.7	656.59	349.44	246.58	194.91	163.77	142.93	127.96	116.62	107.59
0.3	500	500	0.5	1311.90	614.28	389.25	280.64	217.80	177.31	149.27	128.72	112.89
0.3	300	700	0.3	2546.20	1043.68	596.11	394.73	284.93	217.71	173.16	141.75	118.39
0.4	700	300	0.7	735.81	379.31	260.93	202.12	167.16	144.14	127.96	116.07	107.03
0.4	500	500	0.5	1467.73	665.11	410.53	289.83	221.18	177.77	148.34	127.36	111.84
0.4	300	700	0.3	2799.61	1106.50	616.14	400.38	284.93	215.48	170.29	139.15	116.73
0.5	700	300	0.7	845.44	420.79	281.35	213.00	172.97	147.05	129.20	116.39	106.98
0.5	500	500	0.5	1680.67	735.29	441.50	304.87	228.57	181.15	149.58	127.55	111.66
0.5	300	700	0.3	3140.87	1195.65	649.41	414.43	290.90	217.71	170.75	138.88	116.37
0.6	700	300	0.7	1006.85	481.67	311.73	229.72	182.48	152.40	132.03	117.71	107.42
0.6	500	500	0.5	1987.95	836.36	487.37	328.57	241.50	188.28	153.37	129.34	112.29
0.6	300	700	0.3	3620.09	1323.74	701.08	439.39	304.10	224.95	174.68	140.85	117.12

and

$$V(\hat{\pi}_{a(1)}) = \frac{Y^*(1-Y^*)}{n_1 P_1^2} = \frac{(1-\pi_S T_1)[1-P_1(1-\pi_S T_1)]}{n_1 P_1}. \tag{4.4}$$

So the mean square error (MSE) of the estimator $\hat{\pi}_{a(1)}$ is given by

$$\mathrm{MSE}(\hat{\pi}_{a(1)}) = \left\{ \frac{(1-\pi_S T_1)[1-P_1(1-\pi_S T_1)]}{n_1 P_1} + \pi_S^2 (T_1 - 1)^2 \right\}. \tag{4.5}$$

The proportion of "Yes" answers from the respondents using randomization device R_2 is

$$X^* = P\pi_S T_2 + \frac{(1-P)}{2}. \tag{4.6}$$

Thus an estimator of π_S is given by

$$\hat{\pi}_{b(1)} = \frac{\hat{X}^* - (1-P)/2}{P}, \tag{4.7}$$

where $\hat{X}^*$ is the sample proportion of "Yes" responses from the randomization device R_2.

Since $\hat{X}^*$ follows Binomial distribution $\mathrm{B}(n_1, X^*)$, therefore the bias and variance of the estimator $\hat{\pi}_{b(1)}$ are respectively given by

$$\mathrm{B}(\hat{\pi}_{b(1)}) = \pi_S (T_2 - 1) \tag{4.8}$$

and

$$V(\hat{\pi}_{b(1)}) = \frac{X^*(1-X^*)}{n_2 P^2} = \frac{[1-P^2(1-2\pi_S T_2)^2]}{4 n_2 P^2}, \tag{4.9}$$

where $n_1 + n_2 = n$.

Thus the mean square error (MSE) of the estimator $\hat{\pi}_{b(1)}$ is given by

$$\mathrm{MSE}(\hat{\pi}_{b(1)}) = \left\{ \frac{[1-P^2(1-2\pi_S T_2)^2]}{4 n_2 P^2} + \pi_S^2 (T_2 - 1)^2 \right\}. \tag{4.10}$$

Now we propose the weighted estimator of π_S as

$$\hat{\pi}_t^* = \left[\left(\frac{n_1}{n} \right) \hat{\pi}_{a(1)} + \left(\frac{n_2}{n} \right) \hat{\pi}_{b(1)} \right]. \tag{4.11}$$

Since the two randomization devices are independent, we can derive the bias and MSE of $\hat{\pi}_t^*$ respectively as

$$\mathrm{B}\left(\hat{\pi}_t^*\right) = \pi_S\left[\left(\frac{n_1}{n}\right)(T_1-1)+\left(\frac{n-n_1}{n}\right)(T_2-1)\right] \tag{4.12}$$

and

$$\mathrm{MSE}\left(\hat{\pi}_t^*\right) = \left\{\frac{\lambda(1-\pi_S T_1)[1-P_1(1-\pi_S T_1)]}{nP_1}+\frac{(1-\lambda)[1-P^2(1-2\pi_S T_2)^2]}{4nP^2}\right.$$
$$\left.+\pi_S^2[\lambda(T_1-1)+(1-\lambda)(T_2-1)]^2\right\}. \tag{4.13}$$

Putting $P=(2-P_1)^{-1}$ [see Lanke (1976)] in (4.13), we get the MSE of $\hat{\pi}_t^*$ as

$$\mathrm{MSE}\left(\hat{\pi}_t^*\right) = \left\{\frac{\pi_S\{\lambda T_1(1-\pi_S T_1)+(1-\lambda)T_2(1-\pi_S T_2)\}}{n}\right.$$
$$+\frac{(1-P_1)\,[\,4\,\lambda(1-\pi_S T_1)+(1-\lambda)P_1(3-P_1)]}{4nP_1}$$
$$\left.+\pi_S^2[\lambda(T_1-1)+(1-\lambda)(T_2-1)]^2\right\}. \tag{4.14}$$

Proceeding as above in a situation of "Less than completely truthful reporting" one can easily derive the following bias and MSE of Kim and Warde's estimator $\hat{\pi}_{kw}^*$ (say):

$$B(\hat{\pi}_{kw}^*) = \pi_S\left[\left(\frac{n_1}{n}\right)(T_1-1)+\left(\frac{n-n_1}{n}\right)(T_2-1)\right] \tag{4.15}$$

$$\mathrm{MSE}\left(\hat{\pi}_{kw}^*\right) = \left\{\frac{\pi_S\{\lambda T_1(1-\pi_S T_1)+(1-\lambda)T_2(1-\pi_S T_2)\}}{n}\right.$$
$$+\frac{(1-P_1)\,[\,\lambda P_1(1-\pi_S T_1)+(1-\lambda)]}{nP_1^2}$$
$$\left.+\pi_S^2[\lambda(T_1-1)+(1-\lambda)(T_2-1)]^2\right\}. \tag{4.16}$$

From (4.14) and (4.16) we have

$$MSE(\hat{\pi}_{kw}^*)-MSE(\hat{\pi}_t^*) = \frac{(1-P_1)\,(1-\lambda)(4-3P_1^2+P_1^3)}{4nP_1^2} \tag{4.17}$$

which is always positive.

Table 2: *Percent relative efficiency of the proposed estimator $\hat{\pi}_t^*$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}^*$.*

π_S	$n = 1000$					P_1								
	n_1	n_1	T_1	T_2	λ	0.35	0.4	0.45	0.50	0.55	0.60	0.65	0.7	0.75
0.1	700	300	0.6	0.5	0.2	228.01	191.20	166.08	148.36	135.54	126.09	119.02	113.67	108.05
0.1	500	500	0.7	0.6	0.3	251.84	210.96	182.19	161.33	145.87	134.21	125.31	118.45	103.20
0.1	300	700	0.8	0.7	0.4	273.47	231.72	201.04	177.88	160.05	146.09	135.05	126.23	101.48
0.1	700	300	0.9	0.8	0.5	278.84	242.34	214.36	192.31	174.55	159.98	147.86	137.64	102.29
0.2	500	500	0.6	0.5	0.2	139.08	127.19	119.30	113.86	110.02	107.24	105.20	103.69	108.05
0.2	300	700	0.7	0.6	0.3	153.60	137.65	126.93	119.48	114.16	110.29	107.42	105.29	103.20
0.2	700	300	0.8	0.7	0.4	178.79	156.50	141.12	130.19	122.23	116.34	111.92	108.57	101.48
0.2	500	500	0.9	0.8	0.5	220.81	190.76	168.86	152.49	140.01	130.36	122.81	116.87	102.29
0.3	300	700	0.6	0.5	0.2	118.12	112.54	108.86	106.34	104.56	103.29	102.35	101.66	108.05
0.3	700	300	0.7	0.6	0.3	125.82	117.95	112.72	109.13	106.59	104.76	103.42	102.42	103.20
0.3	500	500	0.8	0.7	0.4	141.32	129.02	120.75	115.00	110.90	107.91	105.71	104.06	101.48
0.3	300	700	0.9	0.8	0.5	178.43	156.81	141.70	130.83	122.85	116.89	112.38	108.94	102.29
0.4	700	300	0.6	0.5	0.2	110.35	107.15	105.04	103.60	102.59	101.86	101.33	100.94	108.05
0.4	500	500	0.7	0.6	0.3	114.97	110.36	107.32	105.24	103.77	102.72	101.95	101.38	103.20
0.4	300	700	0.8	0.7	0.4	124.81	117.28	112.27	108.81	106.37	104.60	103.30	102.34	101.48
0.4	700	300	0.9	0.8	0.5	152.61	137.34	126.93	119.62	114.34	110.47	107.59	105.43	102.29
0.5	500	500	0.6	0.5	0.2	106.67	104.61	103.24	102.32	101.66	101.20	100.85	100.60	108.05
0.5	300	700	0.7	0.6	0.3	109.72	106.71	104.74	103.38	102.43	101.75	101.25	100.88	103.20
0.5	700	300	0.8	0.7	0.4	116.40	111.37	108.04	105.76	104.15	102.99	102.14	101.52	101.48
0.5	500	500	0.9	0.8	0.5	136.96	125.94	118.53	113.39	109.72	107.05	105.08	103.62	102.29

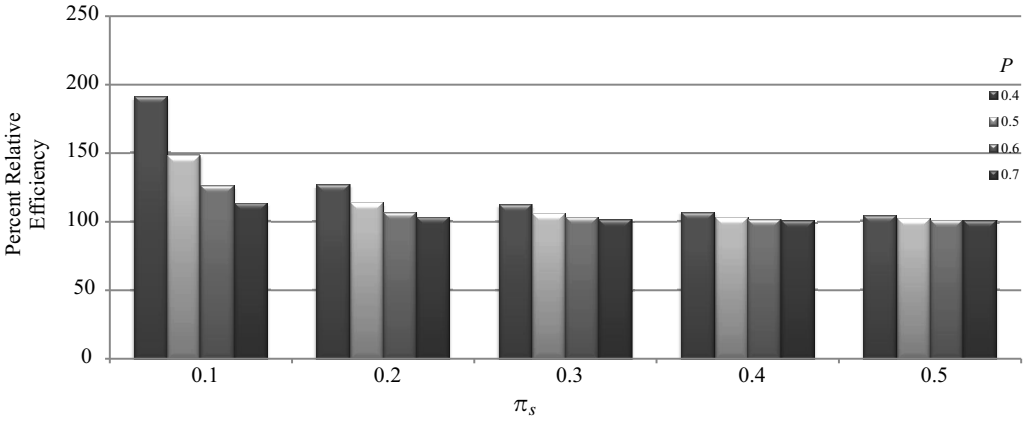


Figure 2: *Percent relative efficiency of the proposed estimator $\hat{\pi}_t^*$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}^*$.*

Thus in the situation of "Less than completely truthful reporting" the proposed estimator $\hat{\pi}_t^*$ is more efficient than Kim and Warde's estimator $\hat{\pi}_{kw}^*$.

To have tangible idea about the performance of the proposed estimator $\hat{\pi}_t^*$ compared to estimator $\hat{\pi}_{kw}^*$, we have computed the percent relative efficiency of the proposed estimator $\hat{\pi}_t^*$ with respect to $\hat{\pi}_{kw}^*$ by using the formula:

$$\mathrm{PRE}(\hat{\pi}_t^*,\ \hat{\pi}_{kw}^*) = \frac{MSE(\hat{\pi}_{kw}^*)}{MSE(\hat{\pi}_t^*)} \times 100$$

We have obtained the values of the percent relative efficiencies $\mathrm{PRE}(\hat{\pi}_t^*,\ \hat{\pi}_{kw}^*)$ for $\lambda = 0.2, 0.3, 0.4, 0.5$, $n = 1000$ and for different cases of π_S, T_1, T_2 and P_1. Findings are shown in Table 2 and its diagrammatic representation is also demonstrated in Figure 2.

It is observed from Table 2 that the values of percent relative efficiencies PRE $(\hat{\pi}_t^*,\ \hat{\pi}_{kw}^*)$ are more than 100. We can say that the proposed estimator $\hat{\pi}_t^*$ is more efficient than Kim and Warde's estimator $\hat{\pi}_{kw}^*$. Figure 2 shows results for $\pi_S = 0.1, 0.2, 0.3, 0.4, 0.5$ and $P = 0.4, 0.5, 0.6, 0.7$, for $T_1 = 0.6$, $T_2 = 0.5$, and $n = 1000$.

Table 2 conceals that the values of the percent relative efficiency of the proposed estimator $\hat{\pi}_t^*$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{kw}^*$ decrease as the value of P_1 increases. Higher gain in efficiency is seen when the sample size n and π_S are small. However, the percent relative efficiency is more than 100 for all parametric values considered here; therefore the proposed estimator $\hat{\pi}_t^*$ is better than Kim and Warde's estimator $\hat{\pi}_{kw}^*$.

5. An alternative mixed randomized response model using stratification

5.1. An alternative to Kim and Warde's (2005) mixed stratified randomized response model

Stratified random sampling is usually obtained by partitioning the population into non-overlapping groups called strata and selecting a simple random sample from each stratum. A randomized response (RR) technique using a stratified random sampling yields the group characteristics associated to each stratum estimator. We also note that stratified sampling protects a researcher from the possibility of obtaining a poor sample. Hong et al. (1994) suggested a stratified RR technique using a proportional allocation. Kim and Warde (2004) suggested a stratified Warner's RR model using an optimal allocation which is more efficient than that using a proportional allocation. Kim and Elam (2005) have applied Kim and Warde's (2004) stratified Warner's RR model to Mangat and Singh's (1990) two-stage RR model. Further Kim and Elam (2007) have given a RR model that combines Kim and Warde's (2004) stratified Warner's RR technique using optimal allocation with the unrelated question randomized response model. Kim and Warde (2005) have suggested a mixed stratified RR model.

In the proposed model, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement in each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. An individual respondent in a sample from each stratum is instructed to answer a direct question "I am a member of the innocuous trait group". Respondents reply the direct question by "Yes" or "No". If a respondent answers "Yes", then she or he is instructed to go to the randomization device R_{j1} consisting of the statements: (i) "I belong to the sensitive trait group" and (ii) "I belong to the innocuous trait group" with pre-assigned probabilities Q_j and $(1-Q_j)$, respectively. If a respondent answers "No", then the respondent is instructed to use the randomization device R_{j2} uses three statements: (i) " I belong to the stigmatizing group", (ii) " Yes" and (iii) "No" with known probabilities P_j, $(1-P_j)/2$ and $(1-P_j)/2$, respectively. For the second and third statements, the respondent is simply to report "Yes" or "No" as observed on the randomization device R_{j2}, and it has no relevance to his actual status. Let m_j denote the number of units in the sample from stratum j and n as the total number of units in samples from all strata. Let m_{j1} be the number of people answering "Yes" when respondents in a sample m_j were asked the direct question and m_{j2} be the number of people answering "No' when respondents in a sample m_j were asked the direct question so that $n = \sum_{j=1}^{L} m_j = \sum_{j=1}^{L} (m_{j1} + m_{j2})$. Under the supposition that these "Yes" or "No" reports are made truthfully, and Q_j and P_j are set by the researcher, then the proportion of "Yes" answers from the respondents using the randomization device R_{j1} will be

$$Y_j = Q_j \pi_{Sj} + (1 - Q_j)\pi_{1_j} \quad \text{for } j = 1,2,\ldots,L, \tag{5.1}$$

where Y_j the probability of "Yes" response in stratum j, π_{Sj} is the proportion of respondents with the sensitive traits in stratum j, π_{1j} is the proportion of respondents with the innocuous trait in stratum j, and Q_j is the probability that a respondent in the sample stratum j is asked a sensitive question.

Since the respondent performing a randomization device R_{j1} answered "Yes" to the direct question of the innocuous trait, if he or she selects the same innocuous question from R_{j1}, then $\pi_{1j} = 1$, see Kim and Warde (2005, p. 217). Thus (5.1) reduces to

$$Y_j = Q_j \pi_{S_j} + (1 - Q_j) \quad \text{for } j = 1,2,\ldots,L. \tag{5.2}$$

An unbiased estimator of π_{Sj} is given by

$$\hat{\pi}_{a_j} = \frac{\hat{Y}_j - (1 - Q_j)}{Q_j} \quad \text{for } j = 1,2,\ldots,L, \tag{5.3}$$

where $\hat{Y}_j$ is the proportion of "Yes" answers in a sample in stratum j and $\hat{\pi}_{a_j}$ is the proportion of respondents with the sensitive trait in a sample from stratum j. The variance of $\hat{\pi}_{aj}$ is given by

$$V(\hat{\pi}_{a_j}) = \frac{(1 - \pi_{Sj})\,[1 - Q_j(1 - \pi_{S_j})]}{m_{j1} Q_j} \quad \text{for } j = 1,2,\ldots,L. \tag{5.4}$$

The proportion of "Yes" responses from the respondents using randomization device R_{j2} will be

$$X_j = P_j \pi_{S_j} + (1 - P_j)/2 \quad \text{for } j = 1,2,\ldots,L, \tag{5.5}$$

where X_j is the probability of "Yes" responses in stratum j. Thus an unbiased estimator of π_{Sj} is given by

$$\hat{\pi}_{b_j} = \frac{\hat{X}_j - (1 - P_j)/2}{P_j} \quad \text{for } j = 1,2,\ldots,L, \tag{5.6}$$

where $\hat{X}_j$ is the proportion of "Yes" responses in a sample from a stratum j and $\hat{\pi}_{bj}$ is the proportion of respondents with the sensitive trait in a sample from stratum j.

The variance of $\hat{\pi}_{b_j}$ is given by

$$V(\hat{\pi}_{b_j}) = \frac{\pi_{S_j}(1-\pi_{S_j})}{(m_j - m_{j1})} + \frac{(1-P_j^2)}{4(m_j - m_{j1})P_j^2} \quad \text{for } j = 1,2,\ldots,L. \tag{5.7}$$

Putting $P_j = (2-Q_j)^{-1}$ [see Lanke (1976)] for $j = 1,2,\ldots,L$ in (5.7) we get

$$V(\hat{\pi}_{b_j}) = \frac{\pi_{S_j}(1-\pi_{S_j})}{(m_j - m_{j1})} + \frac{(1-Q_j)(3-Q_j)}{4(m_j - m_{j1})} \quad \text{for } j = 1,2,\ldots,L. \tag{5.8}$$

Now we develop the unbiased estimator of π_{Sj}, in terms of sample proportion of "Yes" responses $\hat{Y}_j$ and $\hat{X}_j$,

$$\hat{\pi}_{mS_j} = \frac{m_{j_1}}{m_j}\hat{\pi}_{a_j} + \frac{m_j - m_{j1}}{m_j}\hat{\pi}_{b_j} \quad \text{for } 0 < \frac{m_{j1}}{m_j} < 1. \tag{5.9}$$

The variance of $\hat{\pi}_{mSj}$ is given by

$$V(\hat{\pi}_{mS_j}) = \frac{\pi_{S_j}(1-\pi_{S_j})}{m_j} + \frac{(1-Q_j)[4\lambda_j(1-\pi_{Sj}) + (1-\lambda_j)Q_j(3-Q_j)]}{4m_jQ_j}, \tag{5.10}$$

where $m_j = m_{j1} + m_{j2}$ and $\lambda_j = m_{j1}/m_j$.

The unbiased estimator of $\pi_S = \sum_{j=1}^{L} w_j\pi_{Sj}$ is given by

$$\hat{\pi}_S = \sum_{j=1}^{L} w_j\hat{\pi}_{mS_j} = \sum_{j=1}^{L} w_j\left\{\frac{m_{j_1}}{m_j}\hat{\pi}_{aj} + \frac{m_j - m_{j1}}{m_j}\hat{\pi}_{b_j}\right\} \tag{5.11}$$

where N is the number of units in the whole population, N_j is the total number of units in stratum j, and $w_j = N_j/N$ for $j = 1,2,\ldots,L$, so that $w = \sum_{j=1}^{L} w_j = 1$.

The variance of the estimator $\hat{\pi}_S$ is given by

$$V(\hat{\pi}_S) = \sum_{j=1}^{L}\frac{w_j^2}{m_j}\left\{\pi_{S_j}(1-\pi_{S_j}) + \frac{(1-Q_j)[4\lambda_j(1-\pi_{Sj}) + (1-\lambda_j)Q_j(3-Q_j)]}{4Q_j}\right\} \tag{5.12}$$

Here, the requirement of doing the optimal allocation of a sample size n, is to know $\lambda_j = m_{j1}/m_j$ and π_{Sj}. In practice it is difficult to have information on $\lambda_j = m_{j1}/m_j$ and π_{Sj}. However if prior information about $\lambda_j = m_{j1}/m_j$ and π_{Sj} is available from past experience, it assists to derive the following optimal allocation formula.

Theorem 5.1 *The optimal allocation of n to n_1, $n_{2,}\ldots$, n_{L-1} and n_L to derive the minimum variance of the $\hat{\pi}_S$ subject to $n=\sum_{j=1}^{L} m_j$ is approximately given by*

$$\frac{m_j}{n}=\frac{w_j\left\{\pi_{S_j}(1-\pi_{Sj})+\frac{(1-Q_j)[4\lambda_j(1-\pi_{Sj})+(1-\lambda_j)Q_j(3-Q_j)]}{4Q_j}\right\}^{1/2}}{\sum_{j=1}^{L} w_j\left\{\pi_{S_j}(1-\pi_{Sj})+\frac{(1-Q_j)[4\lambda_j(1-\pi_{Sj})+(1-\lambda_j)Q_j(3-Q_j)]}{4Q_j}\right\}^{1/2}}, \tag{5.13}$$

where $m_j=m_{j1}+m_{j2}$ and $\lambda_j=m_{j1}/m_j$.

Thus the minimal variance of the estimator $\hat{\pi}_S$ is given by

$$V(\hat{\pi}_S)=\frac{1}{n}\left\{\sum_{j=1}^{L} w_j\left[\pi_{S_j}(1-\pi_{S_j})+\ \frac{(1-Q_j)[4\lambda_j(1-\pi_{Sj})+(1-\lambda_j)Q_j(3-Q_j)]}{4Q_j}\right]^{1/2}\right\}^2, \tag{5.14}$$

where $n=\sum_{j=1}^{L} m_j$, $m_j=m_{j1}+m_{j2}$ and $\lambda_j=m_{j1}/m_j$.

5.2. Efficiency comparison

In this section we have made the comparison of proposed estimator $\hat{\pi}_S$ with the proposed mixed randomized estimator $\hat{\pi}_t$, Kim and Warde's (2005) mixed randomized response estimator $\hat{\pi}_m$ and Kim and Warde's (2005) stratified mixed randomized response estimator $\hat{\pi}_{mS}$. The comparisons are given in the form of following theorems.

Theorem 5.2 *Assume that there are two strata in the population (i.e. $L=2$) and $\lambda_j=m_{j1}/m_j$. The proposed estimator $\hat{\pi}_S$ of a stratified mixed RR is more efficient than the estimator $\hat{\pi}_t$ of a mixed model, where $P_1=Q_1=Q_2$ and $\lambda=\lambda_1=\lambda_2$.*

Proof. We denote by

$$a_1=\pi_{S1}(1-\pi_{S1}),\qquad a_2=\pi_{S2}(1-\pi_{S2}),$$

$$b_1=\frac{\lambda(1-P_1)(1-\pi_{S1})}{P_1},\quad b_2=\frac{\lambda(1-P_1)(1-\pi_{S2})}{P_1},\quad c=\frac{(1-\lambda)(1-P_1)(3-P_1)}{4}$$

Then for $L=2$, $P_1=Q_1=Q_2$, $\lambda=\lambda_1=\lambda_2$ and from (2.11) and (5.14) we have

$$V(\hat{\pi}_t)=\frac{1}{n}\left\{w_1\pi_{S1}+w_2\pi_{S2})(1-w_1\pi_{S1}-w_2\pi_{S2})+\frac{\lambda(1-P_1)(1-w_1\pi_{S1}-w_2\pi_{S2})}{P_1}+c\right\} \tag{5.15}$$

and

$$V(\hat{\pi}_S) = \frac{1}{n}\{w_1(a_1+b_1+c)^{1/2} + w_2(a_2+b_1+c)^{1/2}\}^2 \tag{5.16}$$

Now subtracting (5.16) from (5.15) we have

$$n[V(\hat{\pi}_t) - V(\hat{\pi}_S) = \ w_1 w_2 \left\{(\pi_{S1}-\pi_{S2})^2 + \left(\sqrt{(a_1+b_1+c)} - \sqrt{(a_2+b_2+c)}\right)^2\right\}$$

which is always positive.

Thus the proposed estimator $\hat{\pi}_S$ of stratified mixed RR is more efficient than the proposed estimator $\hat{\pi}_t$ (with $L=2$) of a mixed model.

This proves Theorem 5.2.

Theorem 5.3 *Suppose there are two strata in the population and $\lambda_j = m_{j1}/m_j$. The proposed estimator $\hat{\pi}_S$ of a stratified mixed RR is more efficient than Kim and Warde's (2005) estimator $\hat{\pi}_{kw}$ of a mixed model, where $P_1 = Q_1 = Q_2$ and $\lambda = \lambda_1 = \lambda_2$.*

Proof. For $L=2$, $\pi_S = w_1\pi_{S1} + w_2\pi_{S2}$, $P_1 = Q_1 = Q_2$, $\lambda = \lambda_1 = \lambda_2$ and from Kim and Warde (2005, Eq (2.10), p. 213) we have

$$V(\hat{\pi}_m) = \frac{1}{n}\left\{w_1\pi_{S1} + w_2\pi_{S2})(1 - w_1\pi_{S1} - w_2\pi_{S2}) + \frac{\lambda(1-P_1)(1-w_1\pi_{S1}-w_2\pi_{S2})}{P_1} + \frac{(1-\lambda)(1-P_1)}{P_1^2}\right\}. \tag{5.17}$$

From (5.16) and (5.17) we have

$$n[V(\hat{\pi}_m) - V(\hat{\pi}_S) = \left[w_1 w_2 \left\{(\pi_{S1}-\pi_{S2})^2 + \left(\sqrt{(a_1+b_1+c)} - \sqrt{(a_2+b_2+c)}\right)^2\right\} + \frac{(1-\lambda)(1-P_1)(4-3P_1^2+P_1^3)}{4P_1^2}\right] \tag{5.18}$$

which is always positive.

Thus the proposed estimator $\hat{\pi}_S$ of a stratified mixed RR is more efficient than Kim and Warde's estimator $\hat{\pi}_m$ of a mixed model.

This proves the theorem.

Theorem 5.4 *Assume that there are two strata in the population (i.e. $L=2$) and $\lambda_j = m_{j1}/m_j$. The proposed estimator $\hat{\pi}_S$ of a stratified mixed RR is more efficient than Kim and Warde's (2005) estimator $\hat{\pi}_{mS}$, $P_1 = Q_1 = Q_2$ and $\lambda = \lambda_1 = \lambda_2$.*

Proof. For $L=2$, $P_1=Q_1=Q_2$, $\lambda=\lambda_1=\lambda_2$ and from Kim and Warde (2005, Eq (4.12), p. 218) we have

$$V(\hat{\pi}_{mS})=\frac{1}{n}\left\{w_1(a_1+b_1+c_1)^{1/2}+w_2(a_2+b_2+c_1)^{1/2}\right\}, \tag{5.19}$$

where $c_1=(1-\lambda)(1-P_1)/P_1^2$.

From (5.16) and (5.19) we have

$$n[V(\hat{\pi}_{mS})-V(\hat{\pi}_S)=(c_1-c)\left[(w_1^2+w_2^2)+\frac{2w_1w_2(A_1+A_2^*)}{\sqrt{A_1A_2}+\sqrt{A_1^*A_2^*}}\right] \tag{5.20}$$

where

$$A_1=(a_1+b_1+c_1),\ A_2=(a_2+b_2+c_1),A_1^*=(a_1+b_1+c) \quad \text{and} \quad A_2^*=(a_2+b_2+c),$$

Since

$$(c_1-c)=\frac{(1-\lambda)(1-P_1)(4-3P_1^2+P_1^3)}{4P_1^2}>0,$$

therefore $n[V(\hat{\pi}_{mS})-V(\hat{\pi}_S)>0$.

It follows that the proposed estimator $\hat{\pi}_S$ of stratified mixed RR is more efficient than Kim and Warde's estimator $\hat{\pi}_{mS}$.

Thus the theorem 5.4 is proved.

If prior information on π_{S1}, π_{S2}, w_1, w_2, π_S and λ can be obtained and a researcher set Q_j, $j=1,2$ then we can compute the percent relative efficiency of the proposed estimator $\hat{\pi}_S$ with respect to Kim and Warde's estimator $\hat{\pi}_{mS}$ (for $L=2$, $\lambda_1=\lambda_2=\lambda$) by using the formula:

$$\begin{aligned}\mathrm{PRE}(\hat{\pi}_S,\hat{\pi}_{mS})&=\frac{V(\hat{\pi}_{mS})}{V(\hat{\pi}_S)}\times 100\\ &=\frac{\left(w_1\sqrt{B_1}+w_2\sqrt{B_2}\right)^2}{\left(w_1\sqrt{B_1^*}+w_2\sqrt{B_2^*}\right)^2}\times 100,\end{aligned}$$

where

$$B_1=\left[\pi_{S_1}(1-\pi_{S1})+\frac{(1-Q_1)[\lambda Q_1(1-\pi_{S1})+(1-\lambda)]}{Q_1^2}\right]$$

$$B_2 = \left[\pi_{S2}(1-\pi_{S2}) + \frac{(1-Q_2)[\lambda Q_2(1-\pi_{S2}) + (1-\lambda)]}{Q_2^2}\right]$$

$$B_1^* = \left[\pi_{S1}(1-\pi_{S_1}) + \frac{(1-Q_1)[4\lambda(1-\pi_{S1}) + (1-\lambda)Q_1(3-Q_1)]}{4Q_1}\right]$$

$$B_2^* = \left[\pi_{S2}(1-\pi_{S_2}) + \frac{(1-Q_2)[4\lambda(1-\pi_{S2}) + (1-\lambda)Q_2(3-Q_2)]}{4Q_2}\right]$$

We have computed $\mathrm{PRE}(\hat{\pi}_S, \hat{\pi}_{mS})$ for $n = 1000$, $\lambda = 0.2, 0.4, 0.6, 0.8$ and different values of w_1, w_2, Q_1, Q_2 , π_{S1} and π_{S2}. Findings are depicted in Table 3. Pictorial representation of $\mathrm{PRE}(\hat{\pi}_S, \hat{\pi}_{mS})$ is also given in Figure 3.

We have set eight different values of Q_j $(j = 1, 2)$ and four different values of λ to verify the percent relative efficiency of the suggested estimator $\hat{\pi}_S$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{mS}$. Table 3 and Figure 3 show that the value of percent relative efficiency $\mathrm{PRE}(\hat{\pi}_S, \hat{\pi}_{mS})$ decreases as the values of Q_j $(j = 1, 2)$ and λ increase.

The values of $\mathrm{PRE}(\hat{\pi}_S, \hat{\pi}_{mS})$ are greater than 100 for all values of π_{S1}, π_{S2}, w_1, w_2, Q_1, Q_2 and λ considered here. So we can say that the envisaged estimator $\hat{\pi}_S$ is more efficient than Kim and Warde's (2005) estimator $\hat{\pi}_{mS}$.

Figure 3 exhibits results from Tables 3 for $Q_1 = 0.1, 0.3, 0.5, 0.7$, $Q_2 = 0.2, 0.4, 0.6, 0.8$ and $\pi_S = 0.1, 0.2, 0.3, 0.4, 0.5$.

Remark 5.1. Proceeding as in Section 4 and the procedure adopted in Kim and Warde (2004) and Kim and Elam (2005, sec.4, p.4) the problem of "Less than completely truthful reporting" can be studied for the proposed mixed stratified RR model.

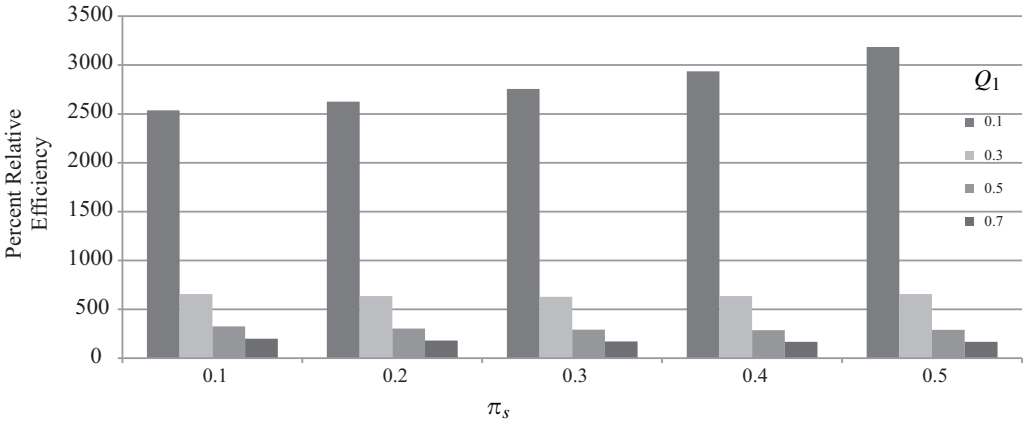


Figure 3: *Percent relative efficiency of the proposed estimator $\hat{\pi}_S$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{mS}$.*

Table 3: *Percent relative efficiency of the proposed estimator $\hat{\pi}_S$ with respect to Kim and Warde's (2005) estimator $\hat{\pi}_{mS}$.*

π_S	π_{S1}	π_{S2}	λ	$Q_1=0.1$	$Q_1=0.2$	$Q_1=0.3$	$Q_1=0.4$	$Q_1=0.5$	$Q_1=0.6$	$Q_1=0.7$	$Q_1=0.8$
				$Q_2=0.2$	$Q_2=0.3$	$Q_2=0.4$	$Q_2=0.5$	$Q_2=0.6$	$Q_2=0.7$	$Q_2=0.8$	$Q_2=0.9$
0.1	0.08	0.13	0.2	2535.64	1106.56	655.86	444.83	326.11	251.15	199.36	159.99
0.1	0.08	0.13	0.4	1239.07	623.89	411.76	304.55	240.19	197.23	166.05	141.21
0.1	0.08	0.13	0.6	652.74	367.72	266.11	212.85	179.69	156.76	139.55	125.36
0.1	0.08	0.13	0.8	317.85	208.66	169.21	148.15	134.74	125.25	117.95	111.77
0.2	0.18	0.23	0.2	2626.11	1102.85	635.71	421.70	303.50	230.24	180.81	144.90
0.2	0.18	0.23	0.4	1322.53	644.81	414.52	299.86	232.04	187.53	155.97	131.97
0.2	0.18	0.23	0.6	702.79	385.05	272.40	213.78	177.60	152.88	134.67	120.29
0.2	0.18	0.23	0.8	339.78	217.44	173.24	149.66	134.70	124.17	116.19	109.68
0.3	0.28	0.33	0.2	2755.76	1117.51	629.28	410.36	291.63	219.37	171.69	138.19
0.3	0.28	0.33	0.4	1430.06	675.34	423.49	300.33	228.80	182.78	150.98	127.75
0.3	0.28	0.33	0.6	766.92	408.18	282.12	217.28	177.82	151.33	132.32	117.94
0.3	0.28	0.33	0.8	367.94	228.95	178.86	152.29	135.59	124.00	115.41	108.70
0.4	0.38	0.43	0.2	2935.97	1152.12	635.57	408.54	287.39	214.75	167.61	135.20
0.4	0.38	0.43	0.4	1571.60	718.58	439.77	306.03	229.77	181.61	149.06	125.93
0.4	0.38	0.43	0.6	851.54	439.50	296.41	223.84	180.38	151.75	131.68	116.99
0.4	0.38	0.43	0.8	405.31	244.45	186.75	156.40	137.54	124.67	115.36	108.35
0.5	0.48	0.53	0.2	3185.50	1211.03	655.53	415.93	289.83	215.08	167.11	134.53
0.5	0.48	0.53	0.4	1763.77	779.78	465.61	317.90	235.14	183.73	149.56	125.75
0.5	0.48	0.53	0.6	967.70	483.13	317.29	234.52	185.76	154.22	132.55	117.07
0.5	0.48	0.53	0.8	457.16	266.14	198.09	162.66	140.93	116.03	116.03	108.49

6. Discussion

In this article, we have proposed an alternative to Kim and Warde's (2005) mixed randomized response model to estimate the proportion of a qualitative sensitive characteristic under the conditions presented in both the cases of completely truthful reporting and less than completely truthful reporting by the respondents. We have also developed the proposed model to stratified sampling. It has been shown that the proposed mixed randomized response model is more efficient than Kim and Warde's (2005) mixed randomized response model.

7. Acknowledgements

The authors are grateful to the Editor-in-Chief and to the learned referee for their valuable suggestions regarding improvement of the paper.

References

Chang, H. J. and Huang, K. C. (2001). Estimation of proportion and sensitivity of a qualitative character. *Metrika*, 53, 269–280.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. Marcel-Dekker, New York, USA.

Cochran, W. G. (1977). *Sampling Technique*, 3rd Edition. New York: John Wiley and Sons, USA.

Fox, J. A. and Tracy, P. E. (1986). *Randomized Response: A Method of Sensitive Surveys*. Newbury Park, CA: SEGE Publications.

Greenberg, B., Abul-Ela, A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response: Theoretical framework. *Journal of the American Statistical Association*, 64, 529–539.

Greenberg, B., Kubller, R. R., Abernathy, J. R. and Horvitz, D. G. (1971). Applications of the RR technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243–250.

Hong, K., Yum, J. and Lee, H. (1994). A stratified randomized response technique. *Korean Journal of Applied Statistics*, 7, 141–147.

Horvitz, D. G., Shah, B. V. and Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, 65–72.

Kim, J. M., Tebbs, J. M. and An, S. W. (2006). Extensions of Mangat's randomized response model. *Journal of Statistical Planning and Inference*, 136, 1554–1567.

Kim, J. M. and Warde, W. D. (2004). A stratified Warner randomized response model. *Journal of Statistical Planning and Inference*, 120, 155–165.

Kim, J. M. and Warde, W. D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211–221.

Kim, J. M. and Elam, M. E. (2005). A two-stage stratified Warner's randomized response model using optimal allocation. *Metrika*, 61, 1–7.

Kim, J. M. and Elam, M. E. (2007). A stratified unrelated randomized response model. *Statistical Papers*, 48, 215–233.

Lanke, J. (1976). On the degree of protection in randomized interview. *International Statistical Review* , 44, 80–83.

Mangat, N. S. and Singh, R. (1990). An alternative randomized procedure. *Biometrika*, 77, 439–442.

Mangat, N. S. (1994a). An improved randomized response strategy. *Journal of the Royal Statistical Society*, B, 56(1), 93–95.

Mangat, N. S. (1994b). An optional randomized response sampling technique. *Journal of the Indian Statistical Association*, 32(2), 71–75.

Mangat, N. S., Singh, R. and Singh, S. (1997). Violation of respondent's privacy in Moors model-its rectification through a random group strategy response model. *Communications in Statistics-Theory and Methods*, 26(3), 243–255.

Moors, J. A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, 66, 627–629.

Nazuk, A. and Shabbir, J. (2010). A new mixed randomized response model. *International Journal of Business and Social Science*, 1, 186–190.

Ryu, J. B., Hong, K. H. and Lee, G. S. (1993). *Randomized Response Model*. Freedom. Academy, Seoul, Korea.

Singh, H. P. and Tarray, T. A. (2012). A stratified unknown repeated trials in randomized response sampling. *Communication of the Korean Statistical Society*, 19, (6), 751–759.

Singh, H. P. and Tarray, T. A. (2013a). Two-stage stratified partial randomized response strategies. Accepted in *Communication in Statistics Theory-Methods*.

Singh, H. P. and Tarray, T. A. (2013b). An alternative to stratified Kim and Warde's randomized response model using optimal (Neyman) allocation. *Model Assisted Statistical Application-Special-Issue* (in press).

Singh, H. P. and Tarray, T. A. (2013c). An improved mixed randomized response model. *Model Assisted Statistical Application-Special-Issue* (in press).

Singh, R. and Mangat, N. S. (1996). *Elements of Survey Sampling*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Singh, S. (2003). *Advanced Sampling Theory with Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sing, S., Singh, R. and Mangat, N. S. (2000). Some alternative strategies to Moor's model in randomized response model. *Journal of Statistical Planning and Inference*, 83, 243–255.

Tracy, D. S. and Mangat, N. S. (1996). Some developments in randomized response sampling during the last decade-A follow up of review by Chaudhuri and Mukherjee. *Journal of Applied Statistical Science*, 4(2/3), 147–158.

Tracy, D. S. and Osahan, S. S. (1999). An improved randomized response technique. *Pakistan Journal of Statistics*, 15(1), 1–6.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.

SORT 37 (2) July-December 2013, 211-230

Double bounded Kumaraswamy-power series class of distributions

Hamid Bidram[1,*] and Vahid Nekoukhou[2]

Abstract

In this paper, we will introduce the new Kumaraswamy-power series class of distributions. This new class is obtained by compounding the Kumaraswamy distribution of Kumaraswamy (1980) and the family of power series distributions. The new class contains some new double bounded distributions such as the Kumaraswamy-geometric, -Poisson, -logarithmic and -binomial, which are used widely in hydrology and related areas. In addition, the corresponding hazard rate function of the new class can be increasing, decreasing, bathtub and upside-down bathtub. Some basic properties of this class of distributions such as the moment generating function, moments and order statistics are studied. Some special members of the class are also investigated in detail. The maximum likelihood method is used for estimating the unknown parameters of the members of the new class. Finally, an application of the proposed class is illustrated using a real data set.

MSC: 60E05, 62E10.

Keywords: Kumaraswamy distribution, Maximum likelihood estimation, Power series distributions, Uniform-power series distributions.

1. Introduction

Many times, the data are modelled by the finite range distributions. For many years, the beta distribution has been used as one of the most basic and useful distributions supported on finite range $(0,1)$ which has been utilized widely in both practical and theoretical aspects of Statistics. This distribution is very flexible to model data which are restricted to any finite interval in view of the fact that it can take an amazingly great

* Corresponding author: h.bidram@sci.ui.ac.ir

[1] Department of Statistics, University of Isfahan, Isfahan, 81746-73441, Iran.

[2] Department of Statistics, University of Isfahan, Khansar Unit, Isfahan, Iran. v.nekoukhou@gmail.com

Received: March 2013
Accepted: September 2013

variety of forms depending on the values of the index parameters (cf. Lemonte and Barreto-Souza, 2013). In econometrics, hydrological processes and related areas several types of data can be modelled by the beta distribution.

An alternative distribution like the beta distribution, which is easier to work with it, is the K distribution proposed by Kumaraswamy (1980). Unlike the beta distribution, the K distribution has a simple closed form of cumulative distribution function (cdf) given by

$$G(x) = 1 - (1 - x^a)^b; \quad 0 < x < 1, \tag{1}$$

where $a > 0$ and $b > 0$ are the shape parameters. The K distribution, similar to the beta distribution, can be unimodal, uniantimodal, increasing, decreasing or constant depending on the values of its parameters. In addition, one can easily show that the K distribution has the same basic shape properties of the beta distribution. But, because of the cdf of the K distribution, which has a simple closed form, it has received much attention in simulating hydrological data and related areas. For more detailed properties of the K distribution see Kumaraswamy (1980) and Jones (2009).

To model data with the finite range on $(0,1)$, we can only address a few distributions in the literature. Here, we attempt to introduce a new family of distributions in this connection. Indeed, to obtain some new double bounded distributions, we compound the K distribution with the family of power series distributions and construct the *Kumaraswamy-power series (KPS)* class of distributions. Compounding a continuous distribution with a discrete one is a known method to introduce new continuous distributions. In recent years, many authors have been interested using this method for constructing new models. For example, the four compound classes proposed by Chahkandi and Ganjali (2009), Morais and Barreto-Souza (2011), Mahmoudi and Jafari (2012) and Silva et al. (2013) are some researches in this regard.

The rest of the paper is organized as follows. In Section 2, we introduce the KPS class of distributions. The density, survival, hazard rate and moment generating functions as well as the moments, quantiles and order statistics are given in this section. In Section 3, we obtain some special distributions and study some of their distributional properties in detail. In addition, the stress-strength parameter is obtained for a special member of the family of KPS distributions in this section. Estimation of the parameters involved using the maximum likelihood method and some related inferences are discussed in Section 4. An application of the new class, using a real data set, is illustrated in Section 5. Finally, some concluding remarks are given in Section 6.

2. The KPS class of distributions

Given N, let $X_1, X_2, \ldots, X_N$ be independent and identically distributed (iid) random variables following a K distribution with cdf (1). Here, N is independent of X_i's and

it is a member of the family of power series distributions, truncated at zero, with the probability mass function

$$\pi_n = P(N=n) = \frac{a_n\theta^n}{C(\theta)}; \quad n=1,2,\ldots,$$

where $a_n \geq 0$ depends only on n, $C(\theta) = \sum_{n=1}^{\infty} a_n\theta^n$ and $\theta \in (0,s)$ (s can be $+\infty$). $C(\theta)$ is finite and $C'(.)$, $C''(.)$ and $C'''(.)$ denote its first, second and third derivatives, respectively. Useful quantities of some power series distributions, truncated at zero, such as geometric, Poisson, logarithmic and binomial (with m being the number of replicates) distributions are shown in Table 1. For more detailed properties of the power series class of distributions, see Noack (1950).

Table 1: *Useful quantities for some power series distributions.*

Model	a_n	$C(\theta)$	$C'(\theta)$	$C''(\theta)$	$C'''(\theta)$	s
Geometric	1	$\theta(1-\theta)^{-1}$	$(1-\theta)^{-2}$	$2(1-\theta)^{-3}$	$3(1-\theta)^{-4}$	1
Poisson	$n!^{-1}$	$e^{\theta}-1$	e^{θ}	e^{θ}	e^{θ}	$+\infty$
Logarithmic	n^{-1}	$-\log(1-\theta)$	$(1-\theta)^{-1}$	$(1-\theta)^{-2}$	$2(1-\theta)^{-3}$	1
Binomial	$\binom{m}{n}$	$(\theta+1)^m-1$	$\frac{m}{(\theta+1)^{1-m}}$	$\frac{m(m-1)}{(1+\theta)^{2-m}}$	$\frac{m(m-1)(m-2)}{(1+\theta)^{3-m}}$	1

Now, let $X_{(1)} = \min\{X_i\}_{i=1}^{N}$. Then, the conditional cdf of $X_{(1)}|N=n$ is given by

$$G_{X_{(1)}|N=n}(x) = 1-[\overline{G}(x)]^n = 1-(1-x^a)^{nb}; \quad 0<x<1,$$

where $\overline{G}(.)$ is the survival function of K distribution associated to cdf (1). As we see, $X_{(1)}|N=n$ follows a K distribution with parameters a and nb. The marginal cdf of $X_{(1)}$, that is,

$$\begin{aligned} F(x) = \sum_{n=1}^{n} \frac{a_n\theta^n}{C(\theta)}\{1-[\overline{G}(x)]^n\} &= 1-\frac{C(\theta\overline{G}(x))}{C(\theta)} \\ &= 1-\frac{C(\theta(1-x^a)^b)}{C(\theta)}; \quad 0<x<1, \end{aligned} \tag{2}$$

defines the cdf of the family of KPS distributions. We denote a random variable X following the KPS distribution with parameters a, b, and θ by $KPS(a,b,\theta)$.

2.1. Density, survival and hazard rate functions

The probability density function (pdf) of a random variable X following a $KPS(a,b,\theta)$ distribution is given by

$$f(x) = \theta abx^{a-1}(1-x^a)^{b-1}\frac{C'(\theta(1-x^a)^b)}{C(\theta)}; \quad 0<x<1. \tag{3}$$

Proposition 2.1 *The pdf of KPS distributions has at least a mode, for $a>1$ and $b>1$. It is increasing, for $a>1$ and $b<1$, and decreasing or bathtub elsewhere.*

Proof. See Appendix A.

Proposition 2.2 *The K distribution with parameters a and bc is a limiting distribution of the KPS distribution when $\theta \to 0^+$, where $c = \min\{n \in \mathbb{N} : a_n > 0\}$.*

Proof. See Appendix B.

Proposition 2.3 *The pdf of KPS distributions can be written as a mixture of the K distribution with parameters a and nb.*

Proof. Using a conditional argument on N, the proof is completed.

The survival and hazard rate functions of KPS distributions are given by

$$\overline{F}(x) = \frac{C(\theta(1-x^a)^b)}{C(\theta)} \tag{4}$$

and

$$h(x) = \theta abx^{a-1}(1-x^a)^{b-1}\frac{C'(\theta(1-x^a)^b)}{C(\theta(1-x^a)^b)}, \tag{5}$$

respectively. To see the density and hazard rate functions shapes of KPS distributions, let $C(\theta) = \theta + \theta^{20}$ (see also Mahmoudi and Jafari, 2012; Morais and Barreto-Souza, 2011). Then, for $\theta = 1$, we have $f(x) = \frac{ab}{2}x^{a-1}(1-x^a)^{b-1}[1+20(1-x^a)^{19b}]$ and $h(x) = abx^{a-1}(1-x^a)^{b-1}\frac{1+20(1-x^a)^{19b}}{(1-x^a)^b+(1-x^a)^{20b}}$. The plots of this density and the corresponding hazard rate function are given in Figure 1 for some selected values of parameters.

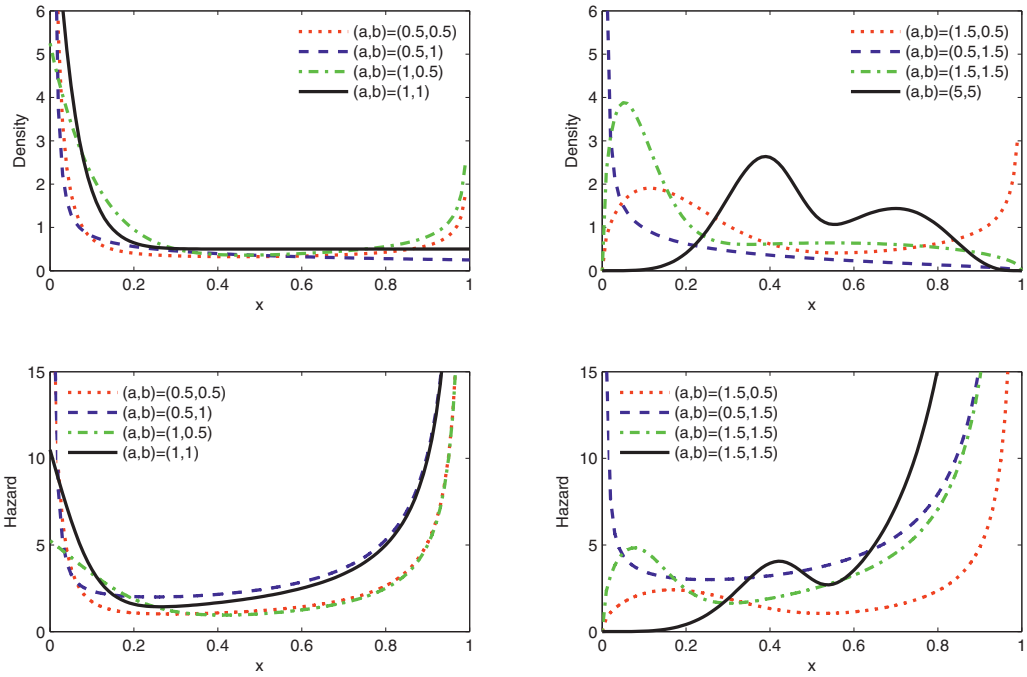


Figure 1: *Plots of density and hazard rate functions of KPS*$(a,b,1)$ *distribution with* $C(\theta)=\theta+\theta^{20}$.

2.2. Quantiles and median

The q-th quantile, say x_q, of the KPS distributions is given by

$$x_q=\{1-[\frac{1}{\theta}C^{-1}((1-q)C(\theta))]^{1/b}\}^{1/a},$$

where $C^{-1}(.)$ is the inverse function of $C(.)$. In particular, the median is immediately obtained by

$$m=\{1-[\frac{1}{\theta}C^{-1}(\frac{C(\theta)}{2})]^{1/b}\}^{1/a}.$$

2.3. Moment generating function and moments

Let Y be a random variable following the K distribution with parameters a and b. Lemonte and Barreto-Souza (2013) obtained the moment generating function (mgf) of the random variable Y as follows:

$$M_Y(t) = b\sum_{s=0}^{\infty}\frac{\Gamma(b)(-1)^s}{\Gamma(b-s)(s+1)!}{}_1F_1(a(s+1),a(s+1)+1;t), \tag{6}$$

where ${}_1F_1$ denotes the confluent hypergeometric function defined by

$${}_1F_1(a,b;t) = \sum_{m=0}^{\infty}\frac{a_{(m)}}{b_{(m)}m!}t^m$$

in which $a_{(m)} = \frac{\Gamma(a+m)}{\Gamma(a)} = a(a+1)...(a+m-1)$ is the ascending factorial. Combining Eq. (6) and Prop. 2.3 yields the mgf of the random variable $X \sim KPS(a,b,\theta)$ as follows:

$$M_X(t) = b\sum_{n=1}^{\infty}\sum_{s=0}^{\infty}\frac{n\Gamma(nb)(-1)^s}{\Gamma(nb-s)(s+1)!}{}_1F_1(a(s+1),a(s+1)+1;t)\pi_n. \tag{7}$$

The r-th moment of the K distribution is given by $bB(1+\frac{r}{a},b)$ (see Jones, 2009), where $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ denotes the beta function. Thus, the r-th moment of $X \sim KPS(a,b,\theta)$ is given by

$$E(X^r) = b\sum_{n=1}^{\infty} nB(1+\frac{r}{a},nb)\pi_n, \quad r = 1,2,\ldots. \tag{8}$$

2.4. Order statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample from a KPS distribution and $X_{i:n},\ i = 1,2,...,n$, denote its i-th order statistic. The pdf of $X_{i:n}$ is given by

$$f_{i:n}(x) = \frac{1}{B(i,n-i+1)}f(x)[F(x)]^{i-1}[1-F(x)]^{n-i}, \tag{9}$$

where F and f are the cdf and pdf of KPS distributions given by (2) and (3), respectively. Eq. (9) can be written as the following forms

$$f_{i:n}(x) = \frac{1}{B(i,n-i+1)}\sum_{k=0}^{n-i}\binom{n-i}{k}(-1)^k f(x)[F(x)]^{k+i-1} \tag{10}$$

or

$$f_{i:n}(x) = \frac{1}{B(i,n-i+1)}\sum_{k=0}^{i-1}\binom{i-1}{k}(-1)^k f(x)[1-F(x)]^{k+n-i}. \tag{11}$$

In view of the fact that

$$f(x)[F(x)]^{k+i-1} = \frac{1}{k+i}\frac{d}{dx}[F(x)]^{k+i},$$

the corresponding cdf of $f_{i:n}(x)$, denoted by $F_{i:n}(x)$, becomes

$$\begin{aligned} F_{i:n}(x) &= \frac{1}{B(i,n-i+1)}\sum_{k=0}^{n-i}\frac{\binom{n-i}{k}(-1)^k}{k+i}[F(x)]^{k+i} \\ &= \frac{1}{B(i,n-i+1)}\sum_{k=0}^{n-i}\frac{\binom{n-i}{k}(-1)^k}{k+i}[1-\frac{C(\theta(1-x^a)^b)}{C(\theta)}]^{k+i} \\ &= \frac{1}{B(i,n-i+1)}\sum_{k=0}^{n-i}\frac{\binom{n-i}{k}(-1)^k}{k+i}F_W(x;a,b,\theta,k+i), \end{aligned} \tag{12}$$

where W follows an *exponentiated* KPS (EKPS) distribution with parameters a, b, θ and $k+i$. For more details of *exponentiated F* distributions or, equivalently, *resilience parameter families*, see Marshall and Olkin (2007).

An alternative expression for $F_{i:n}(x)$, using Eq. (11), is

$$\begin{aligned} F_{i:n}(x) &= 1-\frac{1}{B(i,n-i+1)}\sum_{k=0}^{i-1}\frac{\binom{i-1}{k}(-1)^k}{k+n-i+1}[1-F(x)]^{k+n-i+1} \\ &= 1-\frac{1}{B(i,n-i+1)}\sum_{k=0}^{i-1}\frac{\binom{i-1}{k}(-1)^k}{k+n-i+1}[\frac{C(\theta(1-x^a)^b)}{C(\theta)}]^{k+n-i+1}. \end{aligned}$$

Expressions for moments of the i-th order statistic $X_{i:n},\ i=1,2,...,n$, with cdf (12), can be obtained using a result of Barakat and Abdelkader (2004) as follows:

$$\begin{aligned} E(X_{i:n}^r) &= r\sum_{k=n-i+1}^{n}(-1)^{k-n+i-1}\binom{k-1}{n-i}\binom{n}{k}\int_0^\infty x^{r-1}[\overline{F}(x)]^k dx \\ &= r\sum_{k=n-i+1}^{n}\frac{(-1)^{k-n+i-1}}{C(\theta)^k}\binom{k-1}{n-i}\binom{n}{k}\int_0^\infty x^{r-1}[C(\theta(1-x^a)^b)]^k dx, \end{aligned}$$

for $r=1,2,\ldots$ and $i=1,2,\ldots,n$, where $\overline{F}(x)$ is the survival function given by (4); see also Morais and Barreto-Souza (2011). An application of the first moments of order statistics can be considered in calculating the L-moments which are in fact the linear combinations of the expected order statistics. See Hosking (1990) for details.

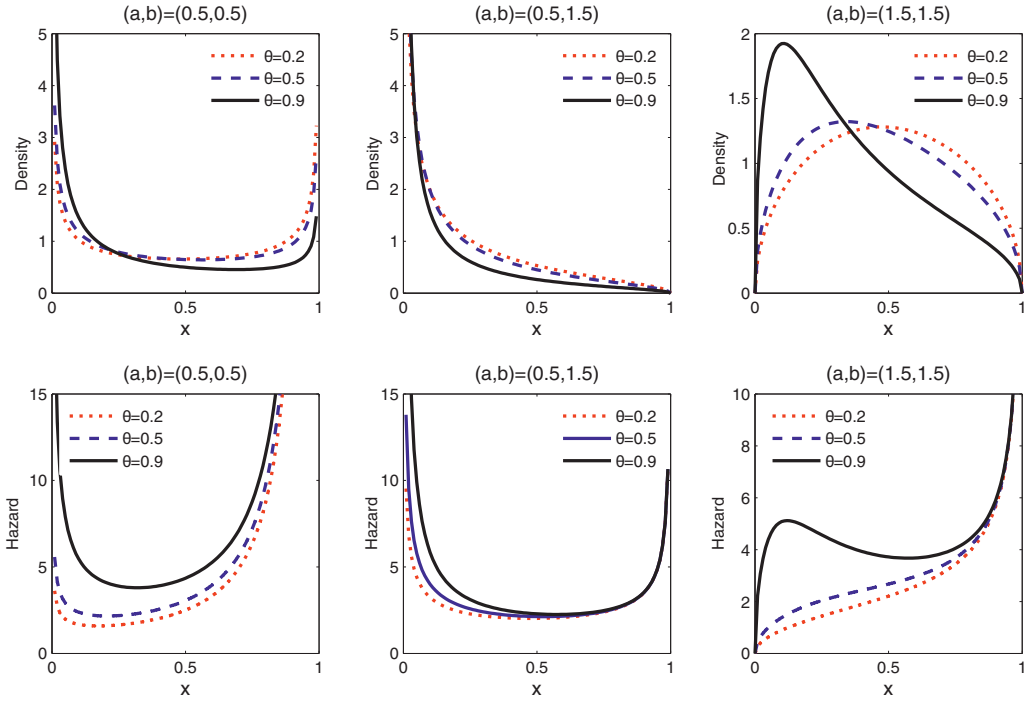


Figure 2: *Plots of $KG(a,b,\theta)$ density and hazard rate functions for some parameter values.*

3. Special cases of the KPS family

In this section, we study basic distributional properties of the Kumaraswamy-geometric (KG), Kumaraswamy-Poisson (KP), Kumaraswamy-logarithmic (KL) and Kumaraswamy-binomial (KB) distributions as special cases of KPS family. In addition, expressions for the pdf and moments of order statistics as well as the stress-strength parameter of the KG distribution are obtained. First, to illustrate the flexibility of the distributions, plots of the density and hazard rate functions are presented in Figures 2, 3, 4 and 5 for some selected values of the parameters.

3.1. Basic distributional properties

Using Table 1 and Eqs. (4-8) given in Section 2, basic distributional properties of the four special distributions of KPS family are immediately obtained. Table 2 contains the survival function, pdf, hazard rate function, mgf and the moments of KG, KP, KL and KB distributions.

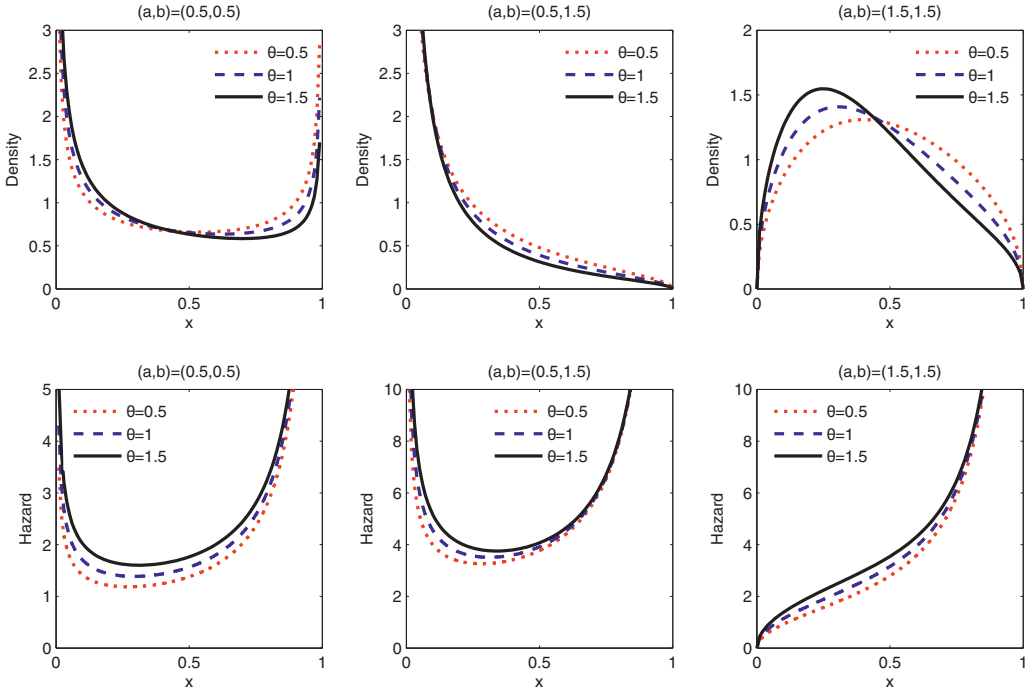


Figure 3: *Plots of $KP(a,b,\theta)$ density and hazard rate functions for some parameters values.*

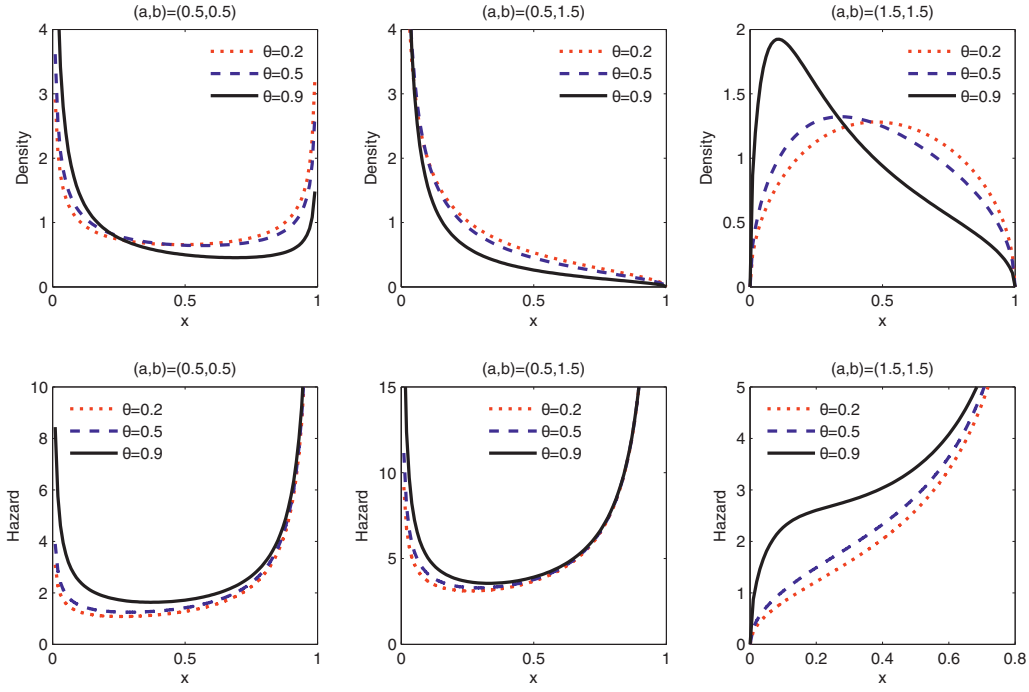


Figure 4: *Plots of $KL(a,b,\theta)$ density and hazard rate functions for some parameter values.*

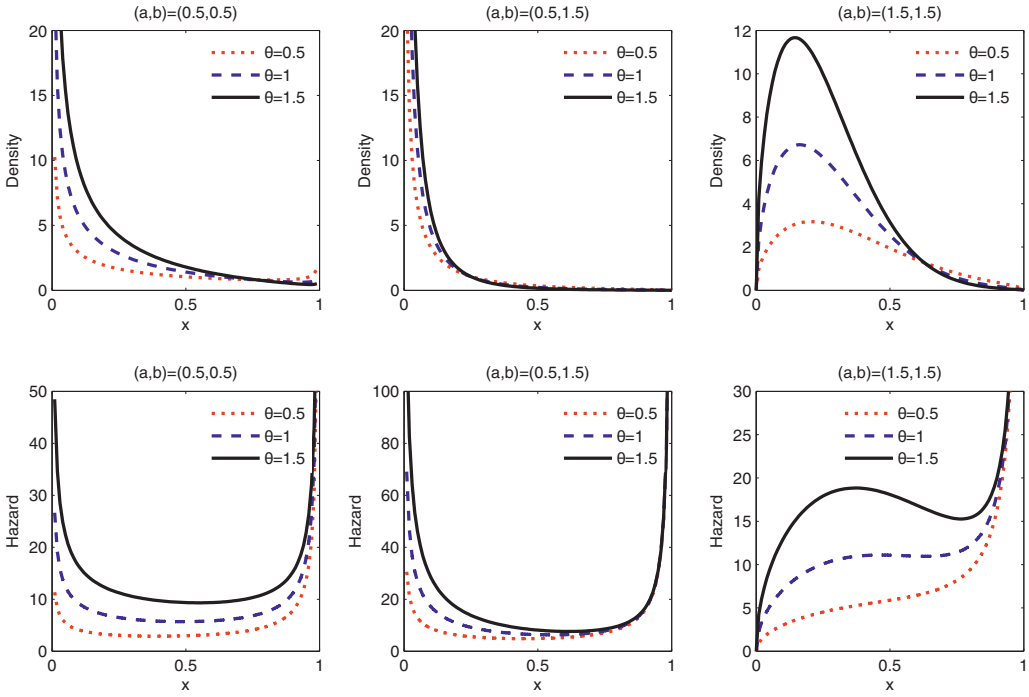


Figure 5: *Plots of $KB(a,b,\theta)$ density and hazard rate functions for some values of (a,b,θ) and $m=5$.*

Table 2: *Survival function, pdf, hazard rate function, mgf and moments of KG, KP, KL and KB dists.*

Model	survival function	pdf	hazard rate function
KG	$1-\frac{1-(1-x^a)^b}{1-\theta(1-x^a)^b}$	$\frac{(1-\theta)abx^{a-1}(1-x^a)^{b-1}}{\{1-\theta(1-x^a)^b\}^2}$	$\frac{abx^{a-1}(1-x^a)^{b-1}}{\{1-\theta(1-x^a)^b\}(1-x^a)}$
KP	$\frac{e^{\theta(1-x^a)^b}-1}{e^\theta-1}$	$\frac{\theta abx^{a-1}(1-x^a)^{b-1}e^{\theta(1-x^a)^b}}{e^\theta-1}$	$\frac{\theta abx^{a-1}(1-x^a)^{b-1}e^{\theta(1-x^a)^b}}{e^{\theta(1-x^a)^b}-1}$
KL	$\frac{\log(1-\theta(1-x^a)^b)}{\log(1-\theta)}$	$-\frac{\theta abx^{a-1}(1-x^a)^{b-1}}{\log(1-\theta)(1-\theta(1-x^a)^b)}$	$-\frac{\theta abx^{a-1}(1-x^a)^{b-1}}{\log(1-\theta(1-x^a)^b)(1-\theta(1-x^a)^b)}$
KB	$\frac{(\theta(1-x^a)^b+1)^m-1}{(\theta+1)^m-1}$	$\frac{m\theta abx^{a-1}(1-x^a)^{b-1}(\theta(1-x^a)^b+1)^{m-1}}{(\theta+1)^m-1}$	$\frac{m\theta abx^{a-1}(1-x^a)^{b-1}(\theta(1-x^a)^b+1)^{m-1}}{(\theta(1-x^a)^b+1)^m-1}$

mgf	moments
$b(1-\theta)\sum_{n=1}^{\infty}\sum_{s=0}^{\infty}\frac{n\Gamma(nb)(-1)^s\theta^{n-1}}{\Gamma(nb-s)(s+1)!}{}_1F_1(a(s+1),a(s+1)+1;t)$	$b(1-\theta)\sum_{n=1}^{\infty}nB(1+\frac{r}{a},nb)\theta^{n-1}$
$\frac{b}{e^\theta-1}\sum_{n=1}^{\infty}\sum_{s=0}^{\infty}\frac{\Gamma(nb)\theta^n(-1)^s}{\Gamma(nb-s)(s+1)!(n-1)!}{}_1F_1(a(s+1),a(s+1)+1;t)$	$\frac{b}{e^\theta-1}\sum_{n=1}^{\infty}B(1+\frac{r}{a},nb)\frac{\theta^n}{(n-1)!}$
$\frac{-b}{\log(1-\theta)}\sum_{n=1}^{\infty}\sum_{s=0}^{\infty}\frac{\Gamma(nb)\theta^n(-1)^s}{\Gamma(nb-s)(s+1)!}{}_1F_1(a(s+1),a(s+1)+1;t)$	$-\frac{b}{\log(1-\theta)}\sum_{n=1}^{\infty}B(1+\frac{r}{a},nb)\theta^n$
$\frac{b}{(\theta+1)^m-1}\sum_{n=1}^{\infty}\sum_{s=0}^{\infty}\frac{n\Gamma(nb)\binom{m}{n}\theta^n(-1)^s}{\Gamma(nb-s)(s+1)!}{}_1F_1(a(s+1),a(s+1)+1;t)$	$\frac{b}{(\theta+1)^m-1}\sum_{n=1}^{\infty}n\binom{m}{n}\theta^nB(1+\frac{r}{a},nb)$

3.2. Order statistics of the KG distribution

By inserting the pdf and cdf of KG distribution into Eq. (10), we obtain the pdf of the i-th order statistic of KG distribution as follows:

$$f_{i:n}(x) = \frac{1}{B(i,n-i+1)} \sum_{k=0}^{n-i} \frac{\binom{n-i}{k}(-1)^k(1-\theta)abx^{a-1}(1-x^a)^{b-1}\{1-(1-x^a)^b\}^{k+i-1}}{\{1-\theta(1-x^a)^b\}^{k+i+1}}.$$

Expanding the binomial term $\{1-\theta(1-x^a)^b\}^{k+i+1}$ by the series representation

$$(1-z)^{-k} = \sum_{i=0}^{\infty} \frac{\Gamma(k+i)}{\Gamma(k)i!} z^i; \quad k>0,\ |z|<1, \tag{13}$$

the pdf of the i-th order statistic can be rewritten as

$$f_{i:n}(x) = \frac{1-\theta}{B(i,n-i+1)} \sum_{j=0}^{\infty} \sum_{k=0}^{n-i} \frac{\binom{n-i}{k}(-1)^k\theta^j}{k+i} f_{BK}(x;k+i,j+1,a,b),$$

where

$$f_{BK}(x;\alpha,\beta,a,b) = \frac{1}{B(\alpha,\beta)} abx^{a-1}(1-x^a)^{b\beta-1}\{1-(1-x^a)^b\}^{\alpha-1} \tag{14}$$

is the density function of beta-Kumaraswamy (BK) distribution of Carrasco et al. (2012).

An alternative expression for the pdf of the i-th order statistic of KG distribution can be obtained by Eq. (11). Hence,

$$f_{i:n}(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{i-1} \frac{\binom{i-1}{k}(-1)^k\binom{k+n-i+j}{j}\theta^j(1-\theta)^{k+n-i+1}}{(k+n-i+1)B(i,n-i+1)} f_K(x;a,b(k+n-i+j+1)), \tag{15}$$

where f_K is the density function of K distribution. As we see, the pdf of order statistics of KG distribution can be expressed as a linear combination of the pdf of BK or K distributions. Therefore, some properties of the i-th order statistic, such as the mgf and moments, can be obtained directly from those of BK or K distributions. For example, from Eq. (15), the moments of the i-th order statistic of KG distribution are given by

$$E(X_{i:n}^r) = \sum_{j=0}^{\infty} \sum_{k=0}^{i-1} \frac{\binom{i-1}{k}(-1)^k\binom{k+n-i+j}{j}\theta^j(1-\theta)^{k+n-i+1}}{(k+n-i+1)B(i,n-i+1)}$$
$$\times b(k+n-i+j+1)B(1+r/a,b(k+n-i+j+1)), \quad r=1,2,\ldots.$$

3.3. Stress-strength parameter of the KG distribution

The stress-strength parameter $R = P(X > Y)$ is a measure of component reliability and its estimation problem when X and Y are independent and follow a specified distribution has been discussed widely in the literature. Let X be the random variable of the strength of a component which is subjected to a random stress Y. The component fails whenever $X < Y$ and there is no failure when $X > Y$. Here, we obtain an expression for the stress-strength parameter of the KG distribution.

Let $X \sim KG(a,b,\theta_1)$ and $Y \sim KG(a,b,\theta_2)$ be independent random variables. The stress-strength parameter is defined as

$$R = P(X > Y) = \int_0^1 f_X(x)F_Y(x)dx$$
$$= \int_0^1 \frac{(1-\theta_1)abx^{a-1}(1-x^a)^{b-1}\{1-(1-x^a)^b\}}{\{1-\theta_1(1-x^a)^b\}^2\{1-\theta_2(1-x^a)^b\}}dx.$$

Expanding the binomial terms $\{1-\theta_1(1-x^a)^b\}^2$ and $\{1-\theta_2(1-x^a)^b\}$ as in Eq. (13), we obtain

$$R = (1-\theta_1)\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{\theta_1^i\theta_2^j(i+1)}{(i+j+1)^2(i+j+2)}\int_0^1 f_{BK}(x;2,i+j+1,a,b)dx$$
$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{\theta_1^i\theta_2^j(i+1)}{(i+j+1)^2(i+j+2)},$$

where f_{BK} has been already defined by Eq. (14). It is clear that R can be estimated when the parameters θ_1 and θ_2 are estimated by the maximum likelihood method.

Remark 3.1 *If $a = 1$ $[b = 1]$ in a $KPS(a,b,\theta)$ distribution, then we obtain the beta-$PS(1,b,\theta)$ $[beta\text{-}PS(a,1,\theta)]$ distribution. In addition, $KPS(a,b,\theta)$ distribution reduces to a standard uniform-PS distribution, when $a = b = 1$. All properties of KPS distribution are valid for these special distributions.*

4. Estimation and inference

Let $x_1, x_2, \ldots, x_n$ be n observations of a random sample from a $KPS(a,b,\theta)$ distribution and $\boldsymbol{\theta} = (a,b,\theta)^{\mathsf{T}}$ be the unknown parameter vector in the rest of the paper. The log-likelihood function is given by

$$\ell_n = \ell_n(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n) = n\log\theta + n\log a + n\log b + (a-1)\sum_{i=1}^{n}\log x_i$$
$$+(b-1)\sum_{i=1}^{n}\log(1-x_i^a) + \sum_{i=1}^{n}\log C'(\theta(1-x_i^a)^b) - n\log(C(\theta)).$$

The associated score function is given by $U_n(\boldsymbol{\theta}) = (\partial\ell_n/\partial a, \partial\ell_n/\partial b, \partial\ell_n/\partial\theta)^T$, where

$$\frac{\partial\ell_n}{\partial a} = \frac{n}{a} + \sum_{i=1}^{n}\log x_i - (b-1)\sum_{i=1}^{n}\frac{x_i^a\log x_i}{1-x_i^a} - \sum_{i=1}^{n}\frac{\theta b x_i^a\log x_i(1-x_i^a)^{b-1}C''(\theta(1-x_i^a)^b)}{C'(\theta(1-x_i^a)^b)},$$

$$\frac{\partial\ell_n}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n}\log(1-x_i^a) + \sum_{i=1}^{n}\frac{\theta(1-x_i^a)^b\log(1-x_i^a)C''(\theta(1-x_i^a)^b)}{C'(\theta(1-x_i^a)^b)}$$

and

$$\frac{\partial\ell_n}{\partial\theta} = \frac{n}{\theta} + \sum_{i=1}^{n}\frac{(1-x_i^a)^bC''(\theta(1-x_i^a)^b)}{C'(\theta(1-x_i^a)^b)} - n\frac{C'(\theta)}{C(\theta)}.$$

The maximum likelihood estimation (MLE) of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, is obtained by solving the nonlinear system $U_n(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. The solution of this nonlinear system of equations can be found by using a numerical method. We need the Fisher information matrix for interval estimation and hypotheses testing on the model parameters. The 3×3 Fisher information matrix is given by

$$I_n(\boldsymbol{\theta}) = -\begin{bmatrix} I_{aa} & I_{ab} & I_{a\theta} \\ I_{ba} & I_{bb} & I_{b\theta} \\ I_{\theta a} & I_{\theta b} & I_{\theta\theta} \end{bmatrix},$$

whose elements are obtained by the relationship $I_{\theta_i\theta_j} = E[\frac{\partial^2\ell_n}{\partial\theta_i\partial\theta_j}];\ i,j = 1,2,3$ (see Appendix C). However, for usual large sample, the Fisher information matrix can be approximated by its observed matrix. That is,

$$I_n(\hat{\boldsymbol{\theta}}) \approx -[\frac{\partial^2\ell_n}{\partial\theta_i\theta_j}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}];\ i,j = 1,2,3,$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of θ. See, for example, Cox and Hinkley (1974) for more discussions of MLEs properties.

Table 3: *Phosphorus concentration in leaves data set.*

0.22	0.17	0.11	0.10	0.15	0.06	0.05	0.07	0.12	0.09	0.23	0.25	0.23	0.24	0.20	0.08
0.11	0.12	0.10	0.06	0.20	0.17	0.20	0.11	0.16	0.09	0.10	0.12	0.12	0.10	0.09	0.17
0.19	0.21	0.18	0.26	0.19	0.17	0.18	0.20	0.24	0.19	0.21	0.22	0.17	0.08	0.08	0.06
0.09	0.22	0.23	0.22	0.19	0.27	0.16	0.28	0.11	0.10	0.20	0.12	0.15	0.08	0.12	0.09
0.14	0.07	0.09	0.05	0.06	0.11	0.16	0.20	0.25	0.16	0.13	0.11	0.11	0.11	0.08	0.22
0.11	0.13	0.12	0.15	0.12	0.11	0.11	0.15	0.10	0.15	0.17	0.14	0.12	0.18	0.14	0.18
0.13	0.12	0.14	0.09	0.10	0.13	0.09	0.11	0.11	0.14	0.07	0.07	0.19	0.17	0.18	0.16
0.19	0.15	0.07	0.09	0.17	0.10	0.08	0.15	0.21	0.16	0.08	0.10	0.06	0.08	0.12	0.13

5. Application of the KPS distributions

Fonseca and Franca (2007) studied the soil fertility in influence and the characterization of the biologic fixation of N_2 for the *Dimorphandra wilsonii rizz growth*. For 128 plants, they made measures of the phosphorus concentration in the leaves. The data, which have also been analyzed by Silva et al. (2013), are listed in Table 3.

We fit the KG, KP, KL and K models to the data to show the capability and potentiality of the new class of distributions in data modelling. In addition, we fit the Weibull-geometric (WG) distribution of Barreto-Souza et al. (2011), which is also a member of the proposed class of Silva et el. (2013), and compare it with our models. We first estimate unknown parameters of the models by the maximum likelihood method and, then, we obtain the values of Akaike information criterion (AIC) and Bayesian information criterion (BIC) as well as Kolmogorov-Smirnov (K-S) statistic and their corresponding p-values. A summary of computations is given in Table 4.

Table 4: *MLE, maximized log-likelihood, AIC, BIC and K-S statistic (p-value) for fitted models.*

Model	MLEs of parameters	$\log L$	AIC	BIC	K-S (p-value)
KG	$(\hat{a},\hat{b},\hat{\theta}) = (3.5909, 318.2081, 0.7338)$	196.7994	-387.5989	-380.0127	0.0944 (0.1911)
KP	$(\hat{a},\hat{b},\hat{\theta}) = (3.1424, 73.3827, 5.1828)$	194.4806	-382.9613	-374.4052	0.1110 (0.0792)
KL	$(\hat{a},\hat{b},\hat{\theta}) = (2.6380, 130.8358, 0.0327)$	194.3899	-382.7797	-374.2236	0.0943 (0.1927)
K	$(\hat{a},\hat{b}) = (2.8104, 176.3491)$	194.8007	-385.6015	-379.8974	0.1181 (0.0517)
WG	$(\hat{\alpha},\hat{\gamma},\hat{\theta}) = (2.4471, 4.2041, 0.9995)$	192.2505	-378.5125	-370.0125	0.1208 (0.0461)

As we see from the results presented in Table 4, the KG model with the minimum values of AIC and BIC gives a better fit than the other rival models. However, the KG, KP and KL models (even K model with the two parameters) have better fits than the WG model of Silva et al. (2013). Further, Figures 6 and 7 also confirm these conclusions.

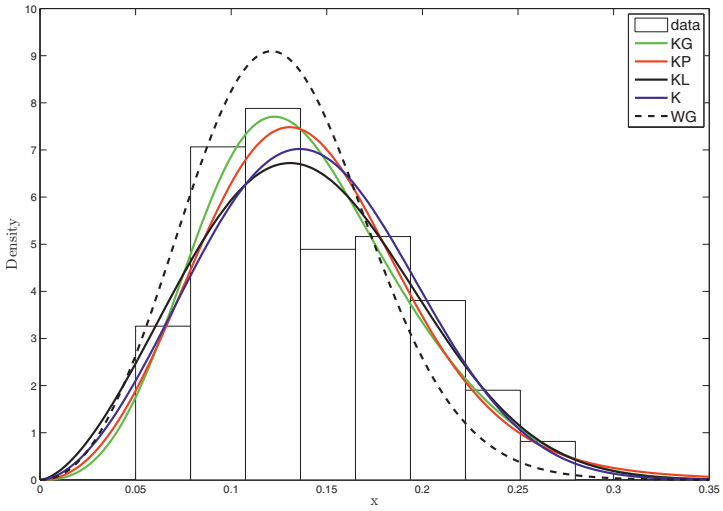


Figure 6: *Plots of the fitted KG, KP, KL, K and WG densities.*

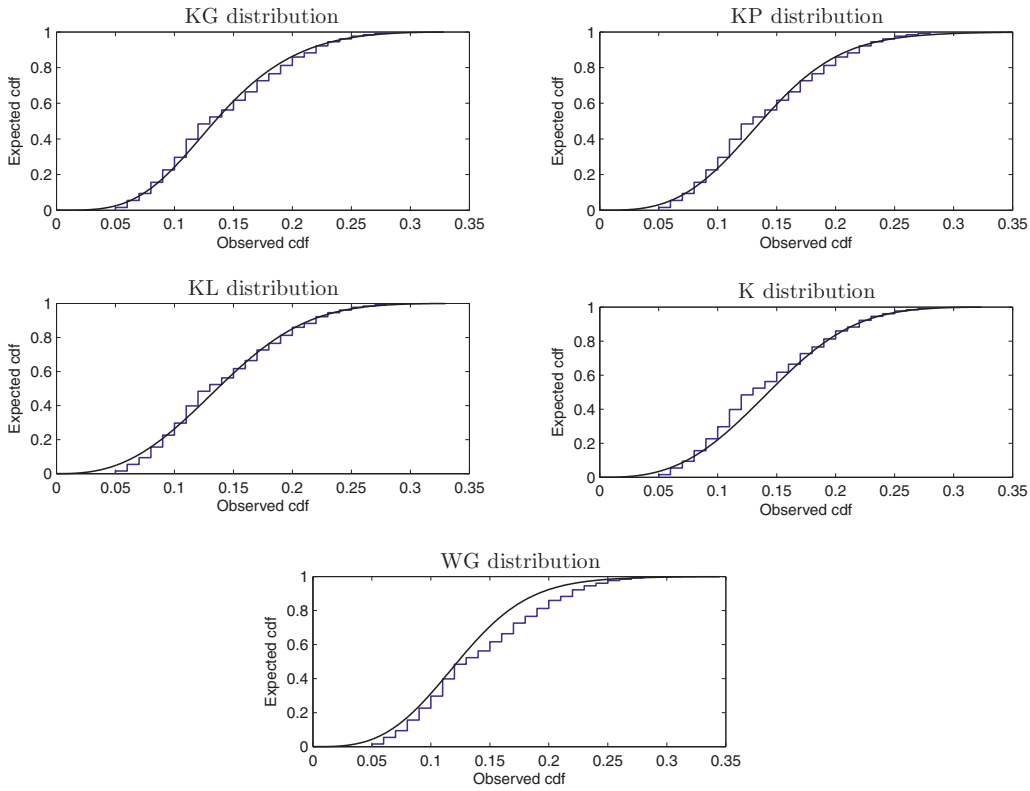


Figure 7: *Empirical cdf plots of the fitted KG, KP, KL, K and WG models.*

6. Concluding remarks

A new compound class of distributions with a finite range on $(0,1)$ is defined by the stochastic representation $X_{(1)} = \min\{X_i\}_{i=1}^{N}$, where X_i's have a Kumaraswamy distribution and N is a member of the family of the power series distributions, independent of X_i's. The new class, namely KPS, contains four new distributions with applications to hydrological areas. We had a comprehensive study on this class of distributions and investigated some their important distributional properties. In the application section, we fitted some special members of the KPS class to a real data set to indicate the potential of the new class in data modelling. As a new family of distributions in this connection, one can establish a new class by considering the stochastic representation $X_{(n)} = \max\{X_i\}_{i=1}^{N}$. In the context of reliability, the stochastic representations $X_{(1)}$ and $X_{(n)}$ have important roles in the series and parallel systems, respectively, which appear in many industrial applications and biological organisms.

Acknowledgments

The authors would like to sincerely thank the Associate Editor and the two anonymous referees for carefully reading the paper and whose useful suggestions led to the improvement of the paper.

Appendix A

Here, we examine the density shapes of KPS distributions. For this purpose, for all $b > 0$, we have

$$\lim_{x\to 0^+} f(x) = \begin{cases} \infty, & a < 1 \\ \theta b \dfrac{C'(\theta)}{C(\theta)}, & a = 1 \\ 0, & a > 1 \end{cases}$$

and, for all $a > 0$,

$$\lim_{x\to 1^-} f(x) = \begin{cases} \infty, & b < 1 \\ \theta a \dfrac{C'(\theta)}{C(\theta)}, & b = 1 \\ 0, & b > 1. \end{cases}$$

Therefore, as wee see, for $a > 1$ and $b > 1$, the pdf of KPS distributions has at least a mode and for $a > 1$ and $b < 1$, the pdf is increasing.

Appendix B

Below, we give a proof for Proposition 2.2:

$$\begin{aligned}\lim_{\theta\to 0^+} F(x) = \lim_{\theta\to 0^+}\{1-\frac{C(\theta\overline{G}(x))}{C(\theta)}\} &= 1-\lim_{\theta\to 0^+}\frac{\sum_{n=1}^{\infty}a_n\theta^n[\overline{G}(x)]^n}{\sum_{n=1}^{\infty}a_n\theta^n}\\ &= 1-\lim_{\theta\to 0^+}\frac{a_c[\overline{G}(x)]^c+\sum_{n=c+1}^{\infty}a_n\theta^{n-c}[\overline{G}(x)]^n}{a_c+\sum_{n=c+1}^{\infty}a_n\theta^{n-c}}\\ &= 1-[\overline{G}(x)]^c = 1-(1-x^a)^{bc}.\end{aligned}$$

Appendix C

Let $p_i=(1-x_i^a)$. Then, the elements of $I_n(\boldsymbol{\theta})$ are given by

$$I_{a,a}=\frac{\partial^2\ell_n}{\partial a^2}=-\frac{n}{a^2}-(b-1)\sum_{i=1}^{n}x_i^a[\frac{\log x_i}{p_i}]^2+\sum_{i=1}^{n}z''_{(aa)i},$$

$$I_{a,b}=I_{b,a}=\frac{\partial^2\ell_n}{\partial a\partial b}=-\sum_{i=1}^{n}\frac{x_i^a\log x_i}{p_i}+\sum_{i=1}^{n}z''_{(ab)i},$$

$$I_{a,\theta}=I_{\theta,a}=\frac{\partial^2\ell_n}{\partial a\partial\theta}=\sum_{i=1}^{n}z''_{(a\theta)i},\qquad I_{b,b}=\frac{\partial^2\ell_n}{\partial b^2}=-\frac{n}{b^2}+\sum_{i=1}^{n}z''_{(bb)i},$$

$$I_{b,\theta}=I_{\theta,b}=\frac{\partial^2\ell_n}{\partial b\partial\theta}=\sum_{i=1}^{n}z''_{(b\theta)i}$$

and

$$I_{\theta,\theta}=\frac{\partial^2\ell_n}{\partial\theta^2}=-\frac{n}{\theta^2}-n\frac{C''(\theta)C(\theta)-[C'(\theta)]^2}{[C(\theta)]^2}+\sum_{i=1}^{n}z''_{(\theta\theta)i},$$

where

$$\begin{aligned}z''_{(aa)i}&=\frac{\partial^2}{\partial a^2}\log C'(\theta p_i^b)\\ &=\frac{-b\theta(\log x_i)^2x_i^a}{[C'(\theta p_i^b)]^2}\{[p_i^{b-1}C''(\theta p_i^b)-(b-1)p_i^{b-2}C''(\theta p_i^b)-\theta bx_i^ap_i^{2b-2}C'''(\theta p_i^b)]C'(\theta p_i^b)\\ &+\theta bx_i^ap_i^{2b-2}[C''(\theta p_i^b)]^2\},\end{aligned}$$

$$z''_{(ab)i} = \frac{\partial^2}{\partial a \partial b} \log C'(\theta p_i^b)$$

$$= \frac{-\theta p_i^{b-1} x_i^a \log x_i \{[b \log p_i C''(\theta p_i^b) - C''(\theta p_i^b) - \theta C'''(p_i^b)] C'(\theta p_i^b) - \theta b p_i^b [C''(\theta p_i^b)]^2\}}{[C'(\theta p_i^b)]^2},$$

$$z''_{(a\theta)i} = -\frac{b x_i^a p_i^{b-1} \log x_i \{[C''(\theta p_i^b) + \theta p_i^b C'''(\theta p_i^b)] C'(\theta p_i^b) - \theta p_i^b [C'(\theta p_i^b)]^2\}}{[C'(\theta p_i^b)]^2},$$

$$z''_{(bb)i} = \frac{\theta \log p_i \{[p_i^b \log p_i C''(\theta p_i^b) + \theta p_i^{2b} \log p_i C'''(\theta p_i^b)] C'(\theta p_i^b)\}}{[C'(\theta p_i^b)]^2}$$

$$- \frac{\theta^2 p_i^{2b} (\log p_i)^2 [C''(\theta p_i)]^2}{[C'(\theta p_i)]^2},$$

$$z''_{(b\theta)i} = \frac{p_i^b \log p_i \{[C''(\theta p_i^b) + \theta p_i^b C'''(\theta p_i^b)] C'(\theta p_i^b) - \theta p_i^b [C''(\theta p_i^b)]^2\}}{[C'(\theta p_i^b)]^2},$$

$$z''_{(\theta\theta)i} = \frac{p_i^{2b} C'''(\theta p_i^b) C'(\theta p_i^b) - p_i^{2b} [C''(\theta p_i^b)]^2}{[C'(\theta p_i^b)]^2}.$$

References

Barakat, H. M. and Abdelkader, Y. H. (2004). Computing the moments of order statistics from nonidentical random variables. *Statistical Methods and Applications*, 13, 15–26.

Barreto-Souza, W., de Morais, A. L. and Cordeiro, G. M. (2011). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation*, 81, 645–657.

Carrasco, J. M. F., Ferrari, S. L. P. and Cordeiro, G. M. (2012). A new generalized Kumaraswamy distribution. *Submitted*, arXiv:1004.0911v1.

Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics and Data Analysis*, 53, 4433–4440.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Fonseca, M. B. and Franca, M. G. C. (2007). *A influoencia da fertilidade do solo e caracterizacao da fixacao biologica de N2 para o crescimento de Dimorphandra wilsonii rizz*. Master's thesis, Universidade Federal de Minas Gerais.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society-Series B*, 52, 105–124.

Jones, M. C. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6, 70–81.

Kumaraswamy, P. (1980). Generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46, 79–88.

Lemonte, A. J. and Barreto-Souza, W. (2013). The exponentiated Kumaraswamy distribution and its log-transform. *Brazilian Journal of Probability and Statistics*, 27, 31–53.

Mahmoudi, E. and Jafari, A. A. (2012). Generalized exponential-power series distributions. *Computational Statistics and Data Analysis*, 56, 4047–4066.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*. Springer Science+Business Media, LLC, New York.

Morais, A. L. and Barreto-Souza, W. (2011). A compound class of Weibull and power series distributions. *Computational Statistics and Data Analysis*, 55, 1410–1425.

Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, 21, 127–132.

Silva, R. B., Bourguignon, M., Dias, C. R. B. and Cordeiro, G. M. (2013). The compound class of extended Weibull power series distributions. *Computational Statistics and Data Analysis*, 58, 352–367.

SORT 37 (2) July-December 2013, 233-240

Analysis of inequality in fertility curves fitted by gamma distributions

Héctor M. Ramos, Antonio Peinado, Jorge Ollero and María G. Ramos*

Abstract

The aim of this paper is to analyse fertility curves from a novel viewpoint, that of inequality. Through sufficient conditions that can be easily verified, we compare inequality, in the Lorenz and Generalized Lorenz sense, in fertility curves fitted by gamma distributions, thus achieving a useful complementary instrument for demographic analysis. As a practical application, we examine inequality behaviour in the distributions of specific fertility curves in Spain from 1975 to 2009.

MSC: 60E15, 62P25.

Keywords: Inequality, Fertility curves, Lorenz order, Generalized Lorenz order.

1. Introduction

The basic concept of inequality arises in many and diverse fields, and so it is difficult to provide a brief definition that will command universal acceptance. More specific contexts give rise to different versions of the concept that can be defined indirectly if we assume certain comparative criteria. Roughly speaking, inequality is a particular aspect of variability when the variables considered are nonnegative and represent quantities that can be transferred from one unit to another. Champernowne and Cowell (1998) provide a convenient reference on this topic. Several studies have approached the problem of ranking distributions by seeking a dominance relationship between concentration curves. In this context, the Lorenz curve and the Generalized Lorenz curve have been used to compare two income distributions in terms of inequality.

The purpose of this paper is to show that the partial orderings of distributions induced by such curves can provide a useful instrument for the demographic analysis of fertility curves. Therefore, our particular interest lies in the concepts of inequality

* Department of Statistics and Operational Research. University of Cádiz. Duque de Nájera 8, 11002 Cádiz (Spain).
hector.ramos@uca.es, antonio.peinado@uca.es, jorge.ollero@uca.es, mramosalcala@gmail.es
Received: March 2012
Accepted: May 2013

underlying the comparison of probability distributions using the Lorenz curve and the Generalized Lorenz curve. The first of these, strictly speaking, is an order of inequality (concentration), whereas the second, when the variables being compared represent incomes, is considered a welfare ordering (see Arnold et al. (1987) and Lambert (2001)). Furthermore, Ramos and Sordo (2002) showed that the Generalized Lorenz order is equivalent to the increasing concave order (Stoyan, 1983). The main results for stochastic orders can be found in Shaked and Shanthikumar (2007).

The aim of this study is to consider and analyze fertility curves from a new viewpoint, that of inequality. Through sufficient conditions that can be easily verified, we compare inequality, in the Lorenz and Generalized Lorenz sense, in fertility curves fitted by gamma distributions. The age-specific fertility rates f_x^t for each maternal age (X) and each year (t) are conventionally defined as the ratio of the number of births to women (x) years of age and the population of women of the same age at the midpoint of year (t). For each year, the observed series of age-specific fertility rates can be fitted (Duchêne and Gillet de Stefano, 1974) using the following curve:

$$g(x) = \frac{a\beta^{-\alpha}(x-\theta)^{\alpha-1}\exp[-(x-\theta)/\beta]}{\Gamma(\alpha)} \tag{1}$$

where $\Gamma(\cdot)$ denotes the complete gamma function and where $a = SFI^{(t)} = \sum_{x=15}^{49} f_x^t$ is the Synthetic Fertility Index (SFI), in which the summation extends from 15 to 49 years, the bounds being the woman's fertile period, with the value of 15 assigned when the mother is aged 15 years or younger and 49 when aged 49 years or older. Then, $(x-\theta)$ is the class mark of the age interval considered less the minimum fertile age; that is, $x-\theta = x+0.5-15 = x-14.5$. Expression (1) enables us to compare series corresponding to different years and to analyze behaviour over a broader time span. Abad et al. (2006) used this approach to fit fertility curves in the Andalusia region in southern Spain.

In the present study, we are not interested in the values of age-specific fertility rates, rather in analyzing the inequality present in the corresponding vectors. Neither are we interested in numerically quantifying inequality, which could be done using various standard measures, such as those associated with the Lorenz curve. On the contrary, our interest lies in comparing in absolute terms (whatever the specific measure applied) the inequality corresponding to two different years within a given population and the inequality corresponding to two different populations in a single period of time. To fulfill this aim, we do not consider age-specific fertility rates, rather the quotients $g_x^t = f_x^t(SFI)^{-1}$. In this way, the corresponding fitting curve $g(x)$ is the density function of a Gamma distribution (α,β,θ), with $\theta = 14.5$:

$$g(x) = \frac{\beta^{-\alpha}(x-\theta)^{\alpha-1}\exp[-(x-\theta)/\beta]}{\Gamma(\alpha)}, \quad x>\theta,\ \alpha>0,\ \beta>0,\ \theta=14.5\,. \tag{2}$$

2. Results

The study of the distribution of fertility according to mothers' age is a scenario that can be easily transferred to the context of income distributions. Thus, it is only necessary to consider specific rates of fertility as the "income" contribution of a given age group of women to the "wealth" of the community in terms of the birth of new population members. This similarity enables us to approach the study of fertility curves from a new perspective. For the analysis and comparison of inequality, let us first employ the Lorenz order.

The Lorenz curve of any income distribution is the graph of the fraction of the total income owned by the lowest p–th fraction $(0 \leq p \leq 1)$ of the population as a function of p. If a nonnegative random variable X represents the income of a community, with distribution function $F_X(x)$ and finite expectation μ_X, then the Lorenz curve $L_X(p)$ is given by (Gastwirth, 1971):

$$L_X(p) = \mu^{-1} \int_0^p F_X^{-1}(t)dt, \quad 0 \leq p \leq 1,$$

where F_X^{-1} denotes the inverse of F_X:

$$F_X^{-1}(a) = \inf\{x : F_X(x) \geq a\},\ a \in [0,1].$$

The Lorenz curve can be used to define a partial ordering as:

$$X \leq_L Y \Leftrightarrow L_X(p) \geq L_Y(p) \text{ for every } 0 \leq p \leq 1.$$

In this case, we can say that X does not show more inequality than Y (in the Lorenz sense). While for any finite population there is no problem in evaluating Lorenz curves, for a continuous distribution, a simple closed form for these curves is rarely available. In our case, the analytical difficulties involved in comparing two gamma distributions by means of the Lorenz order are overcome by taking into account that this order is invariant to scale transformation (i.e., it does not depend on β) and by applying the following sufficient condition (Arnold et al., 1987):

$$\begin{gathered} \text{Let } X_1 \sim \text{gamma}\,(\theta, \alpha_1) \text{ and } X_1 \sim \text{gamma}\,(\theta, \alpha_2)\ (\theta \text{ fixed}). \text{ Then,} \\ \alpha_1 \leq \alpha_2 \implies X_2 \leq_L X_1. \end{gathered} \tag{3}$$

The definition of the Generalized Lorenz curve $GL_X(p)$ corresponding to the non-negative random variable X with distribution function F_X defined by Shorrocks (1983) is:

$$GL_X(p) = \int_0^p F_X^{-1}(t)dt, \quad p \in [0,1]. \tag{4}$$

Consequently, scaling up the Lorenz curves to form the Generalized Lorenz curves will often reveal a dominance relationship that is not apparent from an examination of the means and Lorenz curves on their own. The Generalized Lorenz curve can be used to define a partial ordering on the class of nonnegative random variables as:

$$X \leq_{GL} Y \Leftrightarrow GL_X(p) \geq GL_Y(p) \text{ for all } p \in [0,1]. \tag{5}$$

We then say that X exhibits less inequality than Y in the Shorrocks (or Generalized Lorenz) sense. Generalized Lorenz ordering reflects a desire for both greater equality and higher mean values. Kleiber and Krämer (2003) made a detailed study of the decomposition of the Generalized Lorenz order for both components. Some results on this ordering can be found in Ramos et al. (2000). Once again, the analytical difficulties arising from comparing two gamma distributions by means of the Generalized Lorenz order are overcome by using the following result (Ramos et al., 2000):

$$\text{Let } X_i \sim \text{gamma}(\alpha_i, \beta_i) \ \ (i:1,2). \text{ If } \alpha_1 \leq \alpha_2 \text{ and } \alpha_1\beta_1 \leq \alpha_2\beta_2, \text{ then } X_2 \leq_{GL} X_1. \tag{6}$$

These sufficient conditions enable us to readily compare, from the standpoint of inequality in the Generalized Lorenz sense, distributions of age-specific fertility rates, normalized and fitted by gamma distributions.

In the following section as a practical application, we study the behaviour of inequality in the distributions of normalized specific fertility rates in Spain from 1975 to 2009. The analysis of fertility curves from the standpoint of inequality provides a tool that usefully complements demographic analysis based solely on the behaviour of the SFI, as the latter sometimes fails to detect certain situations of interest, as shown in the Conclusions section.

3. Application to Spanish data

Using official data (INE, 2010) for age-specific fertility rates in Spain for each maternal age and year from 1975 until 2009, we fitted the corresponding normalized rates by gamma (α, β, θ) distributions, with $\theta = 14.5$ using the maximum likelihood method to estimate the parameters. As shown in Table 1 and Figure 1, the α parameter decreases up until 1980, increases from 1980 to 1996 and then decreases again after 1996. According to the sufficient condition (3) (Arnold et al., 1987), we show that inequality (in the Lorenz sense) correspondingly increases in the first and third periods and decreases in the second period.

Table 1: *SFI and estimated α, β parameters, 1975–2009.*

Year	SFI	α	β	$\alpha \cdot \beta$
1975	2.799	5.059	2.728	13.800
1976	2.799	4.850	2.786	13.512
1977	2.671	4.745	2.830	13.427
1978	2.550	4.658	2.865	13.349
1979	2.370	4.524	2.926	13.238
1980	2.213	4.457	2.962	13.202
1981	2.035	4.565	2.897	13.225
1982	1.940	4.643	2.870	13.323
1983	1.797	4.690	2.851	13.370
1984	1.726	4.742	2.831	13.424
1985	1.640	4.803	2.801	13.454
1986	1.556	5.019	2.695	13.528
1987	1.495	5.107	2.656	13.562
1988	1.449	5.191	2.615	13.574
1989	1.398	5.468	2.509	13.718
1990	1.361	5.715	2.425	13.857
1991	1.328	5.905	2.377	14.036
1992	1.316	6.178	2.306	14.247
1993	1.266	6.448	2.242	14.459
1994	1.202	6.646	2.215	14.722
1995	1.173	6.838	2.188	14.963
1996	1.160	7.020	2.163	15.186
1997	1.173	6.969	2.206	15.375
1998	1.153	6.895	2.254	15.541
1999	1.191	6.695	2.339	15.658
2000	1.231	6.570	2.394	15.725
2001	1.241	6.239	2.526	15.758
2002	1.259	6.115	2.582	15.792
2003	1.306	5.975	2.652	15.844
2004	1.325	5.890	2.694	15.871
2005	1.341	5.735	2.774	15.912
2006	1.377	5.549	2.865	15.896
2007	1.392	5.312	2.982	15.838
2008	1.459	5.252	3.015	15.833
2009	1.394	5.404	2.969	16.046

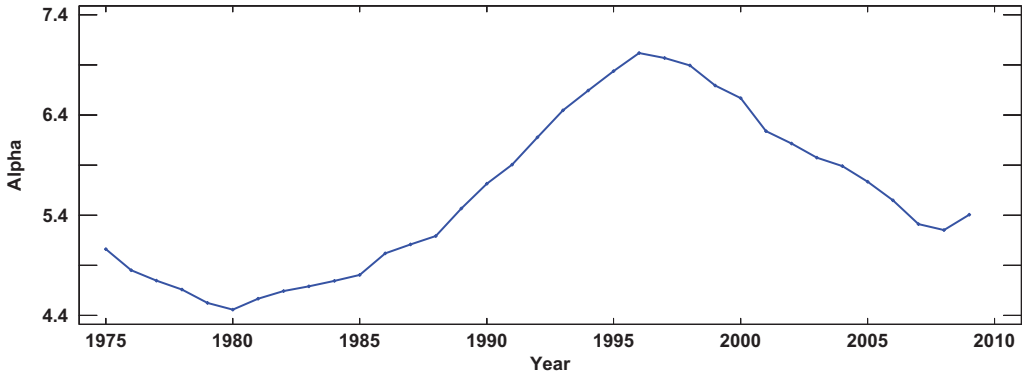


Figure 1: *Estimated α parameters, 1975–2009.*

We can account for the decrease in inequality from 1980 as a consequence of greater birth control and the increased entry of women into the workplace. This meant that the maternal age increased and therefore became less concentrated. This trend was interrupted after 1996. We could explain this with the impact of major immigration into Spain during the previous decade, a population movement that contributed a substantial number of young women, most of whom arrived from countries with cultural backgrounds tending to favour maternity at a younger age. This means that a greater contribution to

births was concentrated in younger age groups (in the case of migrants), and in older groups within the native population, thus increasing inequality in accordance with the concept implicit in the Lorenz order.

From 2002 onward, the Spanish Institute of Statistics (INE, 2010) provides data broken down into the Spanish population and the foreign population. Thus, we can obtain the corresponding values for the estimated parameters (Table 2). Note that, for each year, the value of the α parameter is clearly higher for the population of Spanish origin. According to the sufficient condition of Arnold et al. (1987), the inequality in the distribution of fertility by age groups can then be said to be higher among foreign women. Once again, we can account for this by the higher concentration of fertility among younger age groups within the foreign (mainly immigrant) population. However, in this population and by taking into account the SFI values, the fertility rate appears to be decreasing. We could interpret this as the gradual assimilation of the cultural references of the host country with respect to birth rates.

Table 2: *SFI and estimated values of α, β parameters for Spanish and foreign populations.*

Spanish

Year	SFI	α	β
2002	1.209	6.852	2.350
2003	1.256	6.883	2.356
2004	1.275	6.862	2.376
2005	1.296	6.824	2.402
2006	1.331	6.656	2.472
2007	1.328	6.501	2.540
2008	1.382	6.441	2.571
2009	1.331	6.528	2.563

Foreign

Year	SFI	α	β
2002	2.047	3.028	4.229
2003	1.901	2.949	4.363
2004	1.792	3.095	4.178
2005	1.703	2.997	4.338
2006	1.696	3.070	4.208
2007	1.750	3.103	4.160
2008	1.813	3.233	4.045
2009	1.671	3.422	3.930

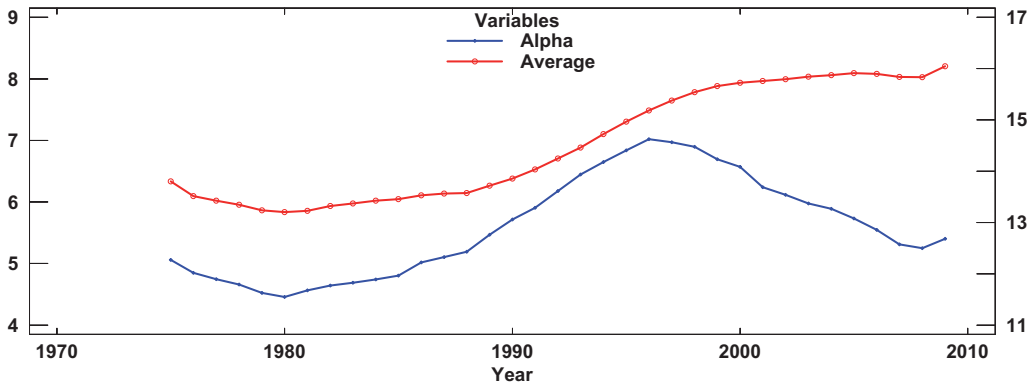


Figure 2: *Estimated α parameters and $\alpha\beta$ mean values, 1975–2009.*

Otherwise, our analysis of fertility curves could be approached from the standpoint of inequality in the sense of the Generalized Lorenz order. To do this, we employ sufficient condition (6) (Ramos et al., 2000). In this case, we must take into account not only the α values but also the $\alpha\beta$ mean values of the distributions being compared (Table 1). We can see that during the periods 1975–80 and 1980–96, the inequality behaviour was the same as the Lorenz sense. The abovementioned sufficient condition, however, was not attained during the period 1996–2005, during which the $\alpha\beta$ value increased while α decreased (Fig. 2).

In this case, we could resort to numerical procedures to determine whether any dominance exists between the corresponding Generalized Lorenz curves. As the expectations $\alpha\beta$ increased during this period, if there were a monotonic pattern of inequality in the Generalized Lorenz sense, then the inequality would necessarily be decreasing. This is shown immediately from (4) and (5) by taking account of the fact that when $p = 1$, expression (4) corresponds to the expression of the mean of the random variable X.

4. Conclusions

The analysis of fertility curves from the standpoint of inequality provides a tool that usefully complements demographic analysis based solely on the behaviour of the SFI, as the latter sometimes fails to detect certain situations of interest, as described below.

It appears reasonable to believe that a sustained rise in birth rates would arise naturally from higher fertility rates among younger women. In such a situation, the concentration of birth rates and, therefore, the degree of inequality must increase. Conversely, a sustained decline in birth rates over a given period would be associated with a decrease in inequality. However, it can be seen in Table 1 that a particular situation occurred in the period 1975–80 that could remain unnoticed if only the SFI values were considered. Although the latter index decreased during 1975–1996, inequality, according to both the Lorenz and the generalized Lorenz curves, did not fall during the period 1975–80, as would have been expected; on the contrary, it increased during this period. Thus, a detailed analysis of the age-specific fertility rates for the period 1975–80 (Table 3) reveals an anomalous behaviour pattern of the birth rate among younger women, in relation to the overall fertility rate. Indeed, SFI values indicate that the fertility rate decreased during this period while the age-specific fertility rates increased. This unexpected birth rate pattern among young women gave rise to a higher concentration of the birth rate and thus greater inequality in the distribution of fertility among age groups, despite the decrease in the SFI.

The sociological reasons for the above lie in the specificity of this particular historical period in Spain. It was a time of great social change, of transition from a dictatorship to a democracy. Society evolved from a situation of severe restrictions on individual freedoms affecting, among other aspects, sexual customs and behaviour, to a democratic context in which these freedoms were guaranteed. This, together with the fact that

birth control and family planning were less well established at the outset, provides a plausible explanation for the specificity of the period 1975–80 that was detected by simple inequality analysis.

Table 3: *Age-specific fertility rates, 1975–2009.*

	Year					
Age	1975	1976	1977	1978	1979	1980
15	3.323	4.059	4.205	4.644	5.182	5.151
16	8.168	9.633	9.643	10.21	10.912	11.259
17	17.574	20.783	21.23	21.64	22.29	22.175
18	33.238	37.494	37.955	38.887	38.182	36.832
19	50.679	56.712	59.153	59.226	59.573	55.133
20	76.447	85.061	85.313	83.838	80.556	78.252

References

Abad-Montes, F., Huete–Morales, M. D. and Vargas–Jiménez, M. (2006). Using the Gamma distribution to fit fecundity curves for application in Andalusia (Spain). Ch. 9 in *Distribution Modeling Theory.* World Scientific Publishing. Singapore.

Arnold, B. C., Robertson, C. A., Brockett, P. L. and Shu, B. Y. (1987). Generating ordered families of Lorenz curves by strongly unimodal distributions. *Journal of Business and Economic Statistics*, 5, 305–308.

Champernowne, D. G. and Cowell, F. A. (1998). *Economic Inequality and Income Distribution*. Cambridge University Press.

Duchêne, J. and Gillet de Stefano, S. (1974). Ajustement analytique des courbes de fécondité générale. *Population et Famille*, 32 (2), 53–94.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, 39, 1037–1039.

INE (2010). Indicadores Demográficos Básicos. Metodología. www.ine.es.

Kleiber, C. and Krämer, W. (2003). Efficiency, equity, and Generalized Lorenz dominance. *Estadística (Journal of the Inter–American Statistical Institute)*, 55 (Special Issue on Income Distribution, Inequality and Poverty), 173–186.

Lambert, P. (2001). *The Distribution and Redistribution of Income, 3rd ed.* Manchester University Press.

Ramos, H. M., Ollero, J. and Sordo, M. A. (2000). A sufficient condition for Generalized Lorenz order. *Journal of Economic Theory, 90, 286–292*.

Ramos, H. M. and Sordo, M. A. (2002). Characterizations of inequality orderings by means of dispersive orderings. *Qüestiió*, 26 (1–2), 15–28.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders.* Springer, New York.

Shorrocks A. F. (1983). Ranking income distributions. *Economica*, 50, 3–17.

Stoyan, D. (1983). *Comparison Methods for Queues and Other Stochastic Models*. Wiley, New York.