Statistics and Operations Research Transactions, vol. 39, n. 2 (2015)

Twenty years of P-splines (invited article)..... p. 149-186
Paul H.C. Eilers, Brian D. Marx and Maria Durbán

Likelihood-based inference for the power regression model..... p. 187-208
Guillermo Martínez-Flórez, Heleno Bolfarine and Héctor W. Gómez

On the bivariate Sarmanov distribution and copula. An application on insurance data using truncated marginal distributions..... p. 209-230
Zuhair Bahraoui, Catalina Bolancé, Elena Pelican and Raluca Vernic

On the interpretation of differences between groups for compositional data..... p. 231-252
Josep-Antoni Martín-Fernández, Josep Daunis-i-Estadella and Glòria Mateu-Figueras

Robust project management with the tilted beta distribution ..... p. 253-272
Eugene D. Hahn and María del Mar López Martín

A note on "Double bounded Kumaraswamy-power series class of distributions"..... p. 273-280
Tibor K. Pogány and Saralees Nadarajah

Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: a simulation study..... p. 281-308
Martí Casals, Klaus Langohr, Josep Lluís Carrasco and Lars Rönnegård

Multinomial logistic estimation in dual frame surveys..... p. 309-336
David Molina, Maria del Mar Rueda, Antonio Arcos and Maria Giovanna Ranalli

SORT 39 (2) July-December 2015, 149-186

Twenty years of P-splines

Paul H.C. Eilers[1], Brian D. Marx[2] and Maria Durbán[3]

Abstract

P-splines first appeared in the limelight twenty years ago. Since then they have become popular in applications and in theoretical work. The combination of a rich B-spline basis and a simple difference penalty lends itself well to a variety of generalizations, because it is based on regression. In effect, P-splines allow the building of a "backbone" for the "mixing and matching" of a variety of additive smooth structure components, while inviting all sorts of extensions: varying-coefficient effects, signal (functional) regressors, two-dimensional surfaces, non-normal responses, quantile (expectile) modelling, among others. Strong connections with mixed models and Bayesian analysis have been established. We give an overview of many of the central developments during the first two decades of P-splines.

MSC: 41A15, 41A63, 62G05, 62G07, 62J07, 62J12.

Keywords: B-splines, penalty, additive model, mixed model, multidimensional smoothing.

1. Introduction

Twenty years ago, *Statistical Science* published a discussion paper under the title "Flexible smoothing with B-splines and penalties" (Eilers and Marx, 1996). The authors were two statisticians with only a short track record, who finally got a manuscript published that had been rejected by three other journals. They had been trying since 1992 to sell their brainchild P-splines (Eilers and Marx, 1992). Apparently it did have some value, because two decades later the paper has been cited over a thousand times (according to the *Web of Science*, a conservative source), in both theoretical and applied work. By now, P-splines have become an active area of research, so it will be useful, and hopefully interesting, to look back and to sketch what might be ahead.

[1] Erasmus University Medical Centre, Rotterdam, the Netherlands, p.eilers@erasmusmc.nl

[2] Dept. of Experimental Statistics, Louisiana State University, USA, bmarx@lsu.edu

[3] Univ. Carlos III Madrid, Dept of Statistics, Leganés, Spain, mdurban@est-econ.uc3m.es
Received: October 2015

P-splines simplify the work of O'Sullivan (1986). He noticed that if we model a function as a sum of B-splines, the familiar measure of roughness, the integrated squared second derivative, can be expressed as a quadratic function of the coefficients. P-splines go one step further: they use equally-spaced B-splines and discard the derivative completely. Roughness is expressed as the sum of squares of differences of coefficients. Differences are extremely easy to compute and generalization to higher orders is straightforward.

The plan of the paper is as follows. In Section 2 we start with a description of basic P-splines, the combination of a B-spline basis and a penalty on differences of coefficients. The penalty is the essential part, and in Section 3 we present many penalty variations to enforce desired properties of fitted curves. The penalty is tuned by a smoothing parameter; it is attractive to have automatic and data-driven methods to set it. Section 4 presents model diagnostics that can be used for this purpose, emphasizing the important role of the effective model dimension. We present the basics of P-splines in the context of penalized least squares and errors with a normal distribution. For smoothing with non-normal distributions, it is straight-forward to adapt ideas from generalized linear models, as is done in Section 5. There we also lay connections to GAMLSS (generalized additive models for location, scale and shape), where not only the means of conditional distributions are modelled. We will see that P-splines are also attractive for quantile and expectile smoothing. The first step towards multiple dimensions is the generalized additive model (Section 6). Not only can smoothing be used to estimate trends in expected values (and other statistics), but it also can be used to find smooth estimates for regression coefficients that change with time or another additional variable. The prototypical case is the varying-coefficient model (VCM). We discuss the VCM in Section 7, along with other models like signal regression. In modern jargon these are examples of functional data analysis. In Section 8, we take the step to full multidimensional smoothing, using tensor products of B-splines and multiple penalties. In Section 9, we show how all the models from the previous sections can be added to each other and so combined into one structure. Here again the roots in regression pay off.

One can appreciate the penalty as just a powerful tool. Yet it is possible to give it a deeper meaning. In Section 10, P-splines are connected to mixed models. This leads to further insights, as well as to new algorithms for finding reasonable values for the penalty parameters. From the mixed model perspective, it is just a small step to a Bayesian approach, interpreting the penalty as (minus) the logarithm of the prior distribution of the B-spline coefficients. This is the subject of Section 11.

Asymptotics and boosting do not have a natural place in other sections, so we put them together in Section 12, while computational issues and availability of software are discussed in Section 13. We close the paper with a discussion

As far as we know, this is the first review on P-splines. Earlier work by Ruppert et al. (2009) took a broader perspective, on the first five years after appearance of their book (Ruppert et al., 2003). We do not try to be exhaustive. That would be impossible (and boring), given the large number of citations. With the availability of *Google Scholar*

and commercial citation databases such as *Scopus* and the *Web of Science*, anyone can follow the trail through history in detail.

We have done our best, in good faith, to give an overview of the field, but we do not claim that our choice of papers is free from subjectivity. The advent of P-splines has led to formidable developments in smoothing, and we have been actively shaping many of them. We hope that we will not offend any reader by serious omissions.

2. P-spline basics

The two components of P-splines are B-splines and discrete penalties. In this section we briefly review them, starting with the former. We do not go much into technical detail; see Eilers and Marx (2010) for that.

2.1. B-splines

Figure 1 shows four triangles of the same height and width, the middle ones overlapping with their two neighbours. These are linear B-splines, the non-zero parts consisting of two linear segments. Imagine that we scale the triangles by different amounts and add them all up. That would give us a piecewise-linear curve. We can generate many shapes by changing the coefficients, and we can get more or less detail by using more or fewer B-splines. If we indicate the triangles by $B_j(x)$ and if a_1 to a_n are the scaling coefficients, we have $\sum_{j=1}^{n} a_j B_j(x)$ as the formula for the function. This opens the door to fitting data pairs (x_i, y_i) for $1, \ldots, m$. We minimize the sum of squares

$$S = \sum_i (y_i - \sum_j a_j B_j(x_i))^2 = ||\boldsymbol{y} - \boldsymbol{Ba}||^2,$$

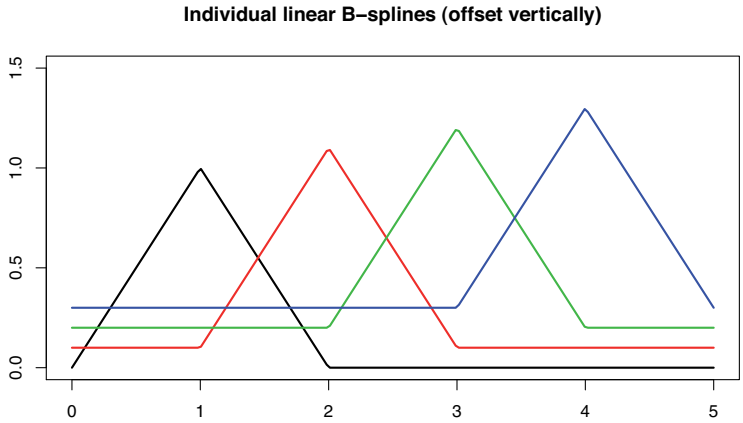


Figure 1: *Linear B-splines illustrated. The individual splines are offset for clarity. In reality the horizontal sections are zero.*

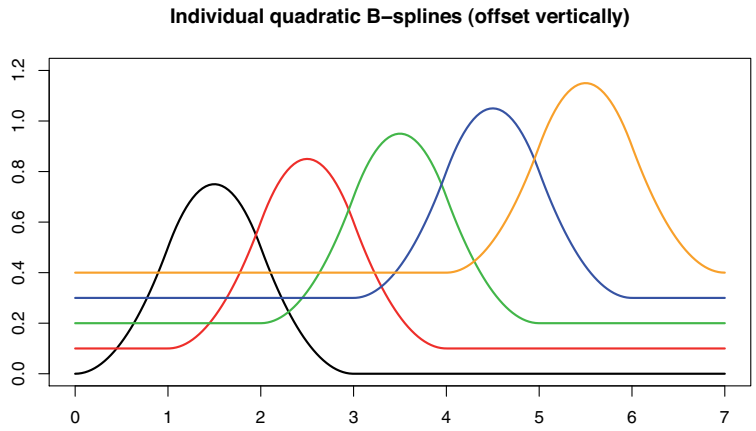


Figure 2: *Quadratic B-splines illustrated. The individual splines are offset for clarity. In reality the horizontal sections are zero.*

where $\boldsymbol{B} = [b_{ij}]$, the so-called basis matrix. This is a standard linear regression problem and the solution is well known: $\hat{\boldsymbol{a}} = (\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{y}$. The flexibility can be tuned by changing the width of the triangles (and hence their number).

A piecewise-linear fit to the data may not be pleasing to the eye, nor be suitable for computing derivatives (which would be piecewise-constant). Figure 2 shows quadratic B-splines, each formed by three quadratic segments. The segments join smoothly. In a similar way cubic B-splines can be formed from four cubic segments. The recipe for forming a curve and fitting the coefficients to data stays the same.

The positions at which the B-spline segments join are called the knots. In our illustrations the knots are equally-spaced and so all B-splines have identical shapes. This is not mandatory for general B-splines, but rather it is a deliberate choice for P-splines, as it makes the construction of penalties trivial.

One should take care when computing the B-splines. The upper panel of Figure 3 shows a basis using equally-spaced knots. Note the "incomplete" B-splines at both ends, of which not all segments fall within the domain of x. The lower panel shows a basis as computed by the R function `bs()`. It has so-called multiple knots at both ends and therefore is unsuitable for P-splines. To avoid this, one should specify an enlarged domain, and cut off the splines at both ends, by removing the corresponding columns in the basis matrix. Alternatively, one can use the code that is presented by Eilers and Marx (2010).

2.2. Discrete penalties

With the number of B-splines in the basis we can tune the smoothness of a curve to the data at hand. A smaller number of splines gives a smoother result. However, this

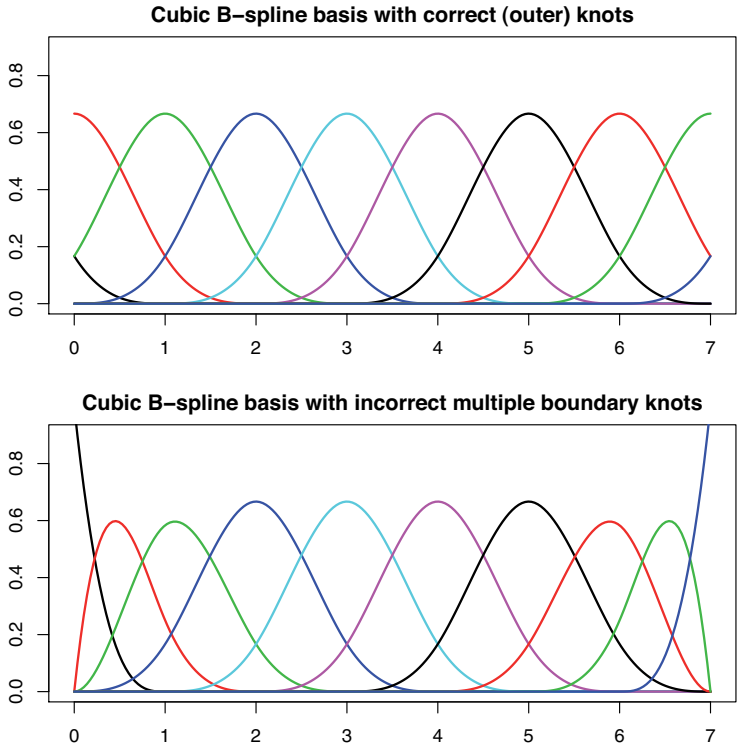


Figure 3: *B-splines bases with different choices of knots. Top: equally spaced knots, the proper basis for P-splines. Bottom: multiple knots at both ends of the domain, which is the result of the R function* `bs()` *and is unsuitable for P-splines.*

is not the only possibility. We can also use a large basis and additionally constrain the coefficients of the B-splines, to achieve as much smoothness as desired. A properly chosen penalty achieves this.

O'Sullivan (1986) had the brilliant idea to take a basis with many B-splines and to use a discrete penalty. The latter was derived from the integrated square of the second derivative of the curve. This was, and still is, an established way to measure roughness of a curve $f(x)$:

$$R = \int_l^u [f''(x)]^2 dx, \tag{1}$$

where l and u indicate the bounds of the domain of x. If $f(x) = \sum_j a_j B_j(x)$, we can derive a (banded) matrix $\boldsymbol{P}$ such that $\boldsymbol{R} = \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a}$. The elements of $\boldsymbol{P}$ are computed as integrals of products of second derivatives of neighbouring B-splines.

O'Sullivan proposed to minimize

$$\boldsymbol{Q} = \boldsymbol{S} + \lambda \boldsymbol{R} = S + \lambda \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a} = ||\boldsymbol{y} - \boldsymbol{a}||^2 + \lambda \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a}, \tag{2}$$

where λ is the parameter that sets the influence of the penalty. The larger the λ, the smoother the result. In the limit the second derivative is forced to be very close to zero and a straight line fit will result. Note that we only have to compute $\boldsymbol{P}$ once. The system to be solved is

$$(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{y}. \tag{3}$$

The computation of $\boldsymbol{P}$ is not trivial, and it becomes quite tedious when the third or fourth order derivative is used to measure roughness. Wand and Ormerod (2008) have extended O'Sullivan's idea to higher orders of the derivative. They used a computer algebra system to construct a table of formulas. P-splines circumvent the issue by dropping derivatives and integrals completely. Instead they use a discrete penalty matrix from the start. It is also simple to compute, as it is based on difference formulas. Let $\Delta a_j = a_j - a_{j-1}$, $\Delta^2 a_j = \Delta(\Delta a_j) = a_j - 2a_{j-1} + a_{j-2}$ and in general $\Delta^d a_j = \Delta(\Delta^{d-1} a_j)$. Let $\boldsymbol{D}_d$ be a matrix such that $\boldsymbol{D}_d\boldsymbol{a} = \Delta^d\boldsymbol{a}$. If we replace the penalty by $\lambda||\boldsymbol{D}_d\boldsymbol{a}||^2 = \lambda\boldsymbol{a}^{\mathsf{T}}\boldsymbol{D}_d^{\mathsf{T}}\boldsymbol{D}_d\boldsymbol{a} = \lambda\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{a}$, we get a similar construction as O'Sullivan's, but with a minimal amount of work. In modern languages like R and Matlab, $\boldsymbol{D}_d$ can be obtained mechanically as the dth order difference of the identity matrix.

It is surprising that nothing is lost by using a simplified penalty. Eilers and Marx (1996) showed how many many useful properties can be proved in a few lines of simple mathematics. Wand and Ormerod (2008) motivate their work by claiming that extrapolation by P-splines goes wrong. They recommended their "O-splines" as a better alternative; see also (Ruppert et al., 2009). In Appendix A we present a small study that lays severe doubt on their conclusion.

2.3. The power of the penalty

A fruitful way of looking at P-splines is to give the coefficients a central position as a skeleton, with the B-splines merely putting "the flesh on the bones." This is illustrated in Figure 4. A smoother sequence of coefficients leads to a smoother curve. The number of splines and coefficients is immaterial, as long as the latter are smooth. The role of the penalty is to make such happen.

The penalty makes interpolation easy (Currie et al., 2004; Eilers and Marx, 2010). At the positions where interpolated values are desired one introduces pseudo-observations with $y = 0$ (or any arbitrary number) and zero weights and solves the system. The true observations get weight 1. One solves

$$(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{W}$ is a diagonal matrix with the weights on the diagonal. Smooth interpolation takes place automatically. Extrapolation can be implemented in the same way, by introducing pseudo-observations outside the domain of the data.

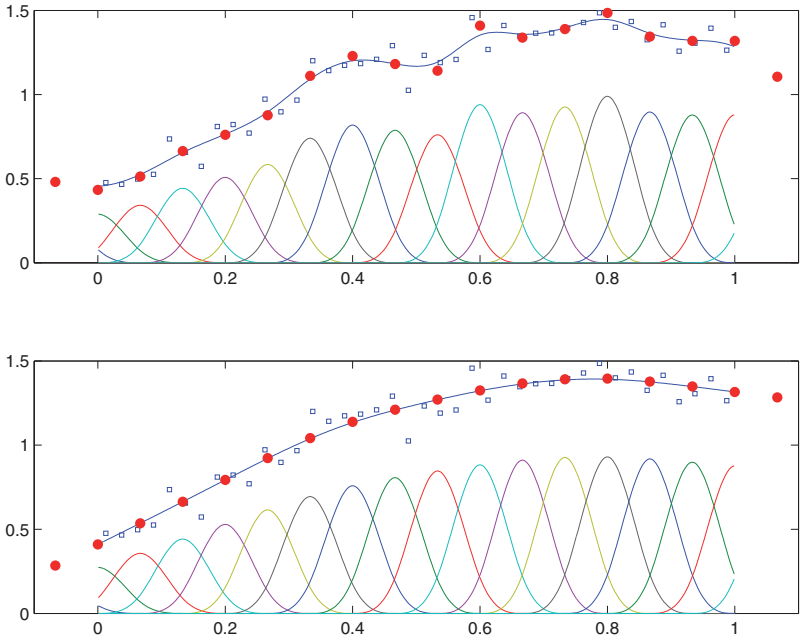


Figure 4: *Illustration of the role of the penalty. The number of B-splines is the same in both panels. In the upper panel the fit to the data (the squares) is more wiggly than in the lower panel, because the penalty is weaker there. The filled circles show the coefficients of the B-splines. Because of a stronger penalty they form a smooth sequence in the lower panel, resulting in a smoother curve fit.*

The number of B-splines can be (much) larger than the number of observations. The penalty makes the fitting procedure well-conditioned. This should be taken literally: even a thousand splines will fit ten observations without problems. Such is the power of the penalty. Figure 5 illustrates this for simulated data. There are 10 data points and 40 (+3) cubic B-splines. Unfortunately, this property of P-splines (and other types of penalized splines) is not generally appreciated. But one simply cannot have too many B-splines. A wise choice is to use 100 of them, unless computational constraints (in large models) come into sight.

We will return to this example in Section 4, after introducing the effective model dimension, and further address this issue of many splines in Appendix B.

2.4. Historical notes

The name P-splines was coined by Eilers and Marx (1996) to cover the combination of B-splines and a discrete difference penalty. It has not always been used with that specific meaning. Ruppert and Carroll (2000) published a paper on smoothing that also used the idea of a rich basis and a discrete penalty. Their basis consists of truncated power functions (TPF), the knots are quantiles of x, and the penalty is on the size of the

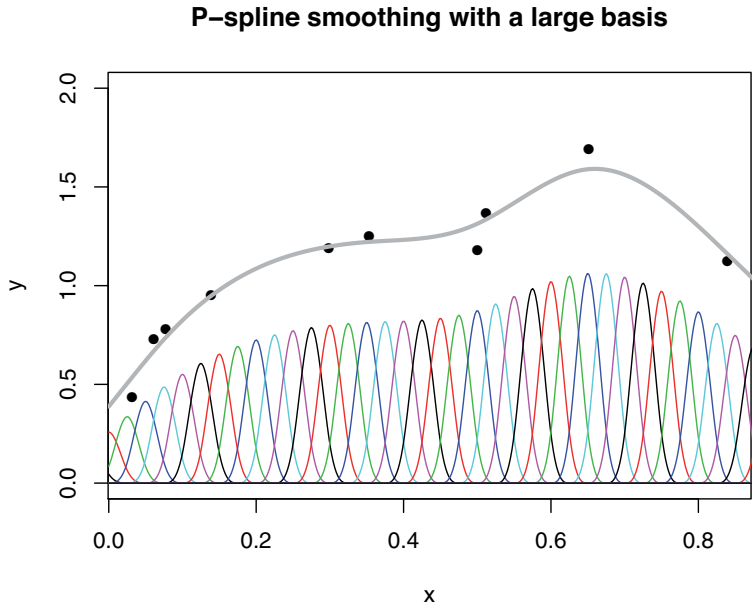


Figure 5: *P-spline smoothing of 10 (simulated) data points with 43 cubic B-splines.*

coefficients. This work has been extended in the book by Ruppert et al. (2003). Some people have called the TPF approach P-splines too. This is confusing and unfortunate because TPF are inferior to the original P-splines; Eilers and Marx (2010) documented their poor numerical condition.

B-splines and TPF are strongly related (Greven, 2008; Eilers and Marx, 2010). Actually B-splines can be computed as differences of TPF, but in the age of single precision floating point numbers it was avoided, for fear of large rounding errors. Eilers and Marx (2010) showed that this no longer holds. P-splines allow to select the degree of the B-splines and the order of the penalty independently. With TPF there is no choice: they imply a difference penalty the order of which is determined by the degree of the TPF.

3. Penalty variations

Standard P-splines use a penalty that is based on repeated differences. Many variations are possible. As stated, the B-spline coefficients form the skeleton of the fit, so if we can find other useful discrete penalties, then we can get curve fits with a variety of desired properties. Eilers and Marx (2010) called them "designer penalties" and they presented several examples. We give a summary here:

- A circular penalty connects the first and last elements of the coefficient vector using differences, making both ends connect smoothly. Combined with a circular

B-spline basis, this is the right tool for fitting periodic data or circular observations, like directions.

- With second order differences, $a_j - 2a_{j-1} + a_{j-2}$, in the penalty, the fit approaches a straight line when λ is increased. If we change the equation to $a_j - 2\phi a_{j-1} + a_{j-2}$, the limit is a (co)sine with period p such that $\phi = \cos(2\pi/p)$. The phase of the (co)sine is adjusted automatically to minimize the sum of squares of the residuals. For smoothing (and interpolation) of seasonal data (with known period) this harmonic penalty usually is more attractive than the standard one.
- Eilers and Goeman (2004) combined penalties of first and second order to eliminate negative side lobes of the impulse response (as would be the case with only a second order penalty). This guarantees that smoothing of positive data never can lead to negative fitted values.
- As described, the P-spline penalty is quadratic: it uses a sum of squares norm. This leads to a smooth result. Other norms have been used. The sum of absolute values (the L_1 norm) of first order differences allows jumps (Eilers and de Menezes, 2005) between neighbouring coefficients, making it suitable for piecewise constant smoothing. This norm is a natural choice when combined with an L_1 norm on the residuals; standard linear programming software can be used. See also Section 5 for quantile smoothing.
- The jumps that are obtained with the L_1 norm are not really "crisp," but slightly rounded. The reason is that the L_1 norm selects *and* shrinks. Much better results are obtained with the L_0 norm, the number of non-zero coefficients (Rippe et al., 2012b). Although a non-convex objective function results, in practice it can be optimized reliably and quickly by an iteratively updated quadratic penalty.

Other types of penalties can be used to enforce shape constraints. An example is a monotonously increasing curve fit (Bollaerts et al., 2006). A second, asymmetric, penalty $\kappa \sum_j v_j(\Delta a_j)$ is introduced, with $v_j = 1$ when $\Delta a_j < 0$ and $v_j = 0$ otherwise. The value of κ regulates the influence of the penalty. Iterative computation is needed, as one needs to know v to do the smoothing and then to know the solution to determine (update) v. In practice, starting from $v = 0$ works well.

Many variations are possible, to force sign constraints, to ensure (increasing or decreasing) monotonicity, or to require a convex or concave shape. One can also mix and match the asymmetric penalties to implement multiple shape constraints. Eilers (2005) used them for unimodal smoothing, while Eilers and Borgdorff (2007) used them to fit mixtures of log-concave non-parametric densities. This scheme has been extended to two dimensions by Rippe et al. (2012a) and applied to genotyping of SNPs (we discuss multidimensional smoothing in Section 8).

Pya and Wood (2015) took a different approach. They write $\boldsymbol{a} = \boldsymbol{\Sigma}\exp(\boldsymbol{\beta})$ and structure the matrix $\boldsymbol{\Sigma}$ in such a way that $\boldsymbol{a}$ has the desired shape, for any vector $\boldsymbol{\beta}$. For

example $\Sigma_{ij} = I(i \geq j)$, with the indicator function $I(\cdot)$, provides a monotonic increasing function. Patterns for combinations of constraints on first and second derivative are tabulated in their paper.

4. Diagnostics

In contrast to many other smoothers, like kernels, local likelihood, and wavelets, P-splines use a regression model with clearly defined coefficients. Hence we can borrow from regression theory to compute informative properties of the model. What we do not learn is the selection of a good value for the penalty parameter λ. Classical theory only considers the fit of a model to the data and as such is useless for this purpose. Instead we need to measure prediction performance. In this section we look at standard errors, cross-validation, effective dimension, and AIC.

The covariance matrix of the spline coefficients (for fixed λ) is given by

$$\boldsymbol{C}_a = \sigma^2(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}, \tag{5}$$

where σ is the variance of the observation noise ϵ in the model $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}$. The covariance of the fitted values follows as $\check{\boldsymbol{B}}\boldsymbol{C}_a\check{\boldsymbol{B}}^{\mathsf{T}}$, where $\check{\boldsymbol{B}}$ contains the B-spline basis evaluated at any chosen set of values of x.

As it stands, this C_a is not very useful, because we need to know σ. It could be estimated from the residuals, but for that we would need to choose the right value of λ, leading to the proper "degrees of freedom."

Leave-one-out cross-validation (CV) provides a mechanism to determine the predictive power of a P-spline model for any value of λ. Let one observation, y_i, be left out and let the predicted value be indicated by $\hat{y}_{-i}$. By doing this for each observation in turn we can compute the prediction error

$$\mathrm{CV} = \sqrt{\sum_i (y_i - \hat{y}_{-i})^2}. \tag{6}$$

As such, CV is a natural criterion to select λ, through its minimization. Using brute force, the computation of CV is expensive, certainly when the number of observations is large. Fortunately there is an exact shortcut. We have that

$$\hat{\boldsymbol{y}} = \boldsymbol{B}(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}. \tag{7}$$

Commonly $\boldsymbol{H}$ is called the "hat" matrix. One can prove that

$$y_i - \hat{y}_{i-1} = (y_i - \hat{y}_i)/(1 - h_{ii}), \tag{8}$$

and the diagonal of $\boldsymbol{H}$ can be computed quickly. A derivation can be found in Appendix B of Myers (1989). An informal proof goes as follows. Imagine that we change element i of $\boldsymbol{y}$ to get a new vector $\boldsymbol{y}^*$; then $\hat{\boldsymbol{y}}^* = \boldsymbol{H}\boldsymbol{y}^*$. Now it holds that if we set $y_i^* = \hat{y}_{-i}$, then $\hat{y}_i^* = \hat{y}_{-i}$. Hence we have that $\hat{y}_{-i} - \hat{y}_i = h_{ii}(y_{-i} - y_i)$, as $\Delta\hat{y}_i = h_{ii}\Delta y_i$. After adding $y_i - y_i$ to the right part of this equation and rearranging terms we arrive at (8).

The hat matrix also gives us the effective model dimension, if we follow Ye (1998), who proposed

$$\mathrm{ED} = \sum_i \partial\hat{y}_i/\partial y_i = \sum h_{ii}. \tag{9}$$

In fact we can compute the trace of $\boldsymbol{H}$ without actually computing its diagonal, using cyclic permutation:

$$\mathrm{ED} = \mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}[(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \lambda\boldsymbol{D}^\mathsf{T}\boldsymbol{D})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B}]. \tag{10}$$

A further simplification is possible by noting that

$$(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} = (\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P} - \boldsymbol{P}) = \boldsymbol{I} - (\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{P}, \tag{11}$$

where $\boldsymbol{P} = \lambda\boldsymbol{D}^\mathsf{T}\boldsymbol{D}$. Harville (1977) presented a similar result for mixed models. In the case of P-splines, the expression is very simple because there are no fixed effects.

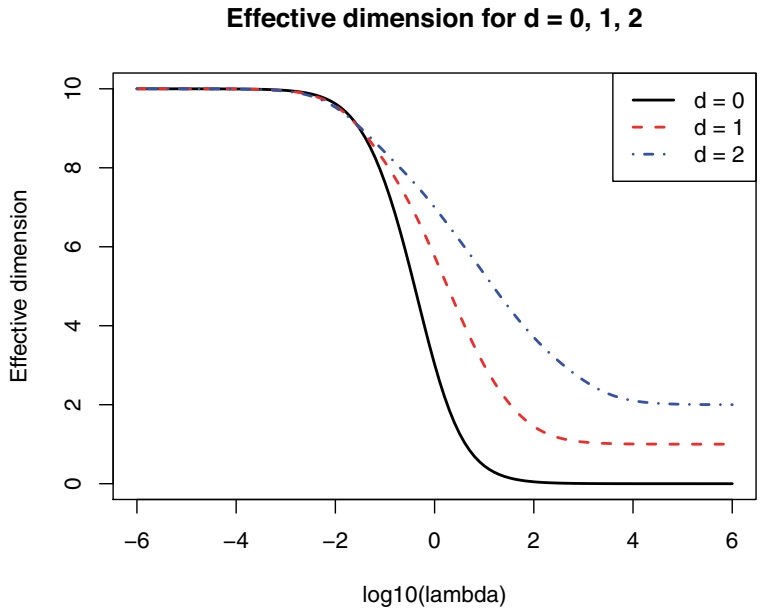


Figure 6: *Changes in the effective model dimension for P-spline smoothing of 10 (simulated) data points with 43 cubic B-splines, for different orders (d) of the differences in the penalty.*

The effective dimension is an excellent way to quantify the complexity of a P-spline model. It summarizes the combined influences of the size of the B-spline basis, the order of the penalty, and the value of the smoothing parameter. The last equation in (11) neatly shows that the effective dimension will always be smaller than n. Actually the effective dimension is always smaller than $\min(m,n)$. An illustration is presented by Figure 6, showing how ED changes with λ for the example with 10 observations and 43 B-splines in Figure 5. For small λ, ED approaches m, while for large λ it approaches d, the order of the differences.

The fact that ED $< n$ is obvious from the size of the system of penalized likelihood equations. A heuristic argument for ED $< m$ is that $\boldsymbol{B}(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B}+\lambda\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}$ is an m by m matrix. It is a hat matrix, having a trace smaller than m. A formal proof is given in Appendix B.

Additionally, the fact that ED $< m$ explains why smoothing with (many) more B-splines than observations works without a problem, for any value of λ. In our experience, many colleagues do not realize this fact. Maybe they fear singularities and stick to small numbers of basis functions.

To estimate σ^2, one divides the sum of squares of the residuals by their effective degrees of freedom, which is the number of observations minus the the effective model dimension: $\hat{\sigma^2} = \sum_i (y_i - \hat{y}_i)^2/(m - \text{ED})$.

Alternatively, one can use Akaike's Information Criterion to choose λ, where AIC $= -2\ell + 2\text{ED}$ and ℓ is the log-likelihood. The beauty of this formula is that it shows the balance between fidelity to the data and complexity of the model.

One should always be careful when using cross-validation or AIC to tune the smoothing parameter. An implicit assumption is that the observations are independent, conditional on their smooth expectations. If this is not the case, as for time series data, the serial correlation will be picked up as a part of the smooth component and severe undersmoothing can occur. One way to approach this problem is to explicitly model the correlation structure of the noise. We return to this subject in Section 10 on mixed models. A recent alternative strategy is the adaptation of the L-curve (Hansen, 1992). It was developed for ridge regression, but can be adapted to difference penalties. See Frasso and Eilers (2015) for examples and a variation, called the V-curve, which is easier to use.

In Section 10 the tuning parameter for the penalty will appear as a ratio of variances, and the effective dimension plays an essential role when estimating them.

5. Generalized linear smoothing and extensions

P-splines are based on linear regression, so it is routine to extend them for smoothing non-normal observations, by borrowing the framework of generalized linear models (GLM). Let $\boldsymbol{y}$ be observed and $\boldsymbol{\mu}$ the vector of expected values. Then the linear predictor $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{Ba}$ is modelled by B-splines, and a suitable distribution is chosen for y, given $\boldsymbol{\mu}$. The penalty is subtracted from the log-likelihood: $\ell^* = \ell - \lambda\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}/2$. The penal-

ized likelihood equations result in $\boldsymbol{B}^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{\mu}) = \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{a}$. This is a small change from the standard GLM, in which the right-hand side is zero (Dobson and Barnett, 2008).

The equations are non-linear, but penalized maximum likelihood leads to the iterative solution of

$$\hat{\boldsymbol{a}}_{t+1} = (\boldsymbol{B}^{\mathsf{T}}\hat{\boldsymbol{W}}_t\boldsymbol{B} + \lambda \boldsymbol{D}_d^{\mathsf{T}}\boldsymbol{D}_d)^{-1}\boldsymbol{B}^{\mathsf{T}}\hat{\boldsymbol{W}}_t\hat{\boldsymbol{z}}_t \quad \text{with} \quad \boldsymbol{z}_t = \hat{\boldsymbol{\eta}}_t + \hat{\boldsymbol{W}}_t^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_t), \tag{12}$$

where t denotes the iterate and $\hat{\boldsymbol{W}}_t$ and $\hat{\boldsymbol{\eta}}_t$ denote approximate solutions, while $\hat{\boldsymbol{z}}_t$ is the so-called working variable. The weights in the diagonal matrix $\hat{\boldsymbol{W}}$ depend on the link function and the chosen distribution. For example, the Poisson distribution, with $\eta = \log(\mu)$ has $\hat{w}_{ii} = \hat{\mu}_i$.

A powerful application of generalized linear smoothing with P-splines is density estimation (Eilers and Marx, 1996). A histogram with narrow bins is computed and the counts are smoothed, using the Poisson distribution and the logarithmic link function. There is no danger that bins are chosen too narrow: even if most of them contain only a few counts or zeros good results are obtained. The amount of smoothing is tuned by AIC.

It is essential (for any smoother) that enough bins with zero counts are included at the ends of the observed domain of the data, unless it is known to be bounded (as for intrinsically positive variables).

P-splines conserve moments of distributions up to order $d-1$, where d is the order of the differences in the penalty. This means that, if $d = 3$, the sum, the mean, and the variance of the smooth histogram are equal to those of the raw histogram, whatever the amount of smoothing (Eilers and Marx, 1996). In contrast, kernel smoothers increase variance.

Many variations on this theme have been published. We already mentioned one- and two-dimensional log-concave densities in Section 3.. Kauermann et al. (2013) explored flexible copula density estimation. They modelled the density directly as a sum of tensor products of linear B-splines (we discuss tensor products in Section 8). To reduce the number of coefficients, they used reduced splines, which are similar to nested B-splines (Lee et al., 2013).

Another variation is not to model the logarithm of the counts by a sum of B-splines, but rather the density itself, with constraints on the coefficients (Schellhase and Kauermann, 2012).

Mortality or morbity smoothing is equivalent to discrete density estimation with an offset for exposures. P-splines have found their way into this area, for both one- and two-dimensional tables (Currie et al., 2004; Camarda, 2012); both papers illustrate automatic extrapolation.

The palette of distributions that generalized linear smoothing can use is limited. A very general approach is offered by GAMLSS: generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005). An example is the normal distribution with smoothly varying mean and variance, combined with a (varying) Box-Cox trans-

form of the response variable. Many continuous and discrete distributions can be fitted by the GAMLSS algorithm, also in combination with mixtures, censoring and random components.

Instead of using a parametric distribution, one can estimate smooth conditional quantiles, minimizing an asymmetrically weighted sum of absolute values of the residuals. Bollaerts et al. (2006) combined it with shape constraints to force monotonicity. To avoid crossing of individually estimated smooth quantile curves, Schnabel and Eilers (2013) introduced the quantile sheet, a surface on the domain formed by the explanatory variable and the probability level.

Compared to the explicit solutions of (penalized, weighted) least squares problems, quantile smoothing is a bit less attractive for numerical work as it leads to linear programming or to quadratic programming if quadratic penalties are involved. In contrast, expectiles use asymmetrically weighted sums of squares and lead to simple iterative algorithms (Schnabel and Eilers, 2009). Sobotka and Kneib (2012) extended expectile smoothing to the spatial context, while Sobotka et al. (2013) provide confidence intervals. Schnabel and Eilers (2013) proposed a location-scale model for non-crossing expectile curves.

When analysing counts with a generalized linear model, often the Poisson distribution is assumed, with $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$ for the expected values. When counts are grouped or aggregated, the composite link model (CLM) of Thompson and Baker (1981) is more appropriate. It states that $\boldsymbol{\mu} = \boldsymbol{C}\exp(\boldsymbol{\eta})$, where the matrix $\boldsymbol{C}$ encodes the aggregation or mixing pattern. In the penalized CLM, a smooth structure for $\boldsymbol{\eta}$ is modelled with P-splines (Eilers, 2007). It is a powerful model for grouped counts (Lambert and Eilers, 2009; Lambert, 2011; Rizzi et al., 2015), but it has also found application in misclassification and digit preference (Camarda et al., 2008; Azmon et al., 2014). de Rooi et al. (2014) used it to remove artifacts in X-ray diffraction scans.

6. Generalized additive models

The generalized additive model (GAM) constructs the linear predictor as a sum of smooth terms, each based on a different covariate (Hastie and Tibshirani, 1990). The model is $\boldsymbol{\eta} = \sum_j f_j(\boldsymbol{x}_j)$; it can be interpreted as a multidimensional smoother without interactions.

The GAM with P-splines, or P-GAM, was proposed by Marx and Eilers (1998). We illustrate the main idea in two dimensions. Let

$$\boldsymbol{\eta} = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) = [\boldsymbol{B}_1|\boldsymbol{B}_2] \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} = \boldsymbol{B}\boldsymbol{a}. \tag{13}$$

By combining the two bases into one matrix and chaining the coefficients in one vector we are back in a standard regression setting.

The roughness penalties are $\lambda_1||\boldsymbol{D}_1\boldsymbol{a}_1||^2$ and $\lambda_2||\boldsymbol{D}_2\boldsymbol{a}_2||^2$ (where the indices here refer to the variables, not to the order of the differences), leading to two penalty matrices $\boldsymbol{P}_1 = \lambda_1\boldsymbol{D}_1^{\mathsf{T}}\boldsymbol{D}_1$ and $\boldsymbol{P}_2 = \lambda_2\boldsymbol{D}_2^{\mathsf{T}}\boldsymbol{D}_2$, which can be combined in the block-diagonal matrix P. The resulting penalized likelihood equations are $\boldsymbol{B}^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{a}$, which have exactly the same form as those for generalized linear smoothing. The weighted regression equations follow immediately. The same is true for the covariance matrix of the estimated coefficients, cross-validation, and the effective dimension.

Originally, backfitting was used for GAMs (Hastie and Tibshirani, 1990), updating each component function in turn, using any type of smoother. Convergence can be slow and diagnostics are hard or impossible to obtain. Direct fitting by P-splines does not have these disadvantages.

As presented the model is unidentifiable, because an arbitrary upward shift of $f_1(\boldsymbol{x}_1)$ can be compensated by an equal shift downward of $f_2(\boldsymbol{x}_2)$. A solution is to introduce an (unpenalized) intercept and to constrain each component to have a zero average.

The P-GAM has multiple smoothing parameters, so optimization of AIC, say, by a simple grid search involves much work. Heim et al. (2007) proposed a searching strategy that cycles over one-dimensional grid searches. As a more principled approach, Wood (2004) presented algorithms for numerical optimization in cross-validation. His book (Wood, 2006a) contains a wealth of material on GAMs. See also Section 13 for the `mgcv` software.

In Section 10 we will present Schall's algorithm for variance estimation. It is attractive for tuning multiple penalty parameters.

7. Smooth regression coefficients

In the preceding sections P-splines were used to model expected values of observations. There is another class of models in which the goal is to model regression coefficients as a curve or surface. In this section we discuss varying coefficient models (Hastie and Tibshirani, 1993), penalized signal regression (Marx and Eilers, 1999), and generalizations. In modern jargon these are all cases of functional data analysis (Ramsay and Silverman, 2003).

Varying coefficient models (VCM) were first introduced by Hastie and Tibshirani (1993). They allow regression coefficients to interact with another variable by varying smoothly. The simplest form is $E[y(t)] = \mu(t) = \beta(t)x(t)$, where $\boldsymbol{y}$ and $\boldsymbol{x}$ are observed and $\boldsymbol{\beta}$ is to be estimated and forced to change slowly with t. The model assumes that $\boldsymbol{y}$ is proportional to $\boldsymbol{x}$, with a varying slope of the regression line. If we introduce a B-spline basis B and write $\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{a}$, we get $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{a}$, where $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{x})$. With a difference penalty on $\boldsymbol{a}$ we have the familiar P-spline structure, with only a modified basis $\boldsymbol{X}\boldsymbol{B}$. A varying offset can be added: $E[y(t)] = \mu(t) = \beta(t)x(t) + \beta_0(t)$. This has the form of an additive model. Building $\boldsymbol{\beta}_0$ with P-splines we effectively get a P-GAM.

This simple VCM can be extended by adding more additive or varying-coefficient terms. For non-normal data we model the linear predictor and choose a proper response distribution.

VCM with P-splines were proposed by Eilers and Marx (2002). Lu et al. (2008) studied them too and presented a Newton-Raphson procedure to minimize the cross-validation error. Andriyana et al. (2014) brought quantile regression into VCMs using P-splines. Kauermann (2005b) and Kauermann and Khomski (2006) developed P-spline survival and hazard models, respectively, to accommodate varying-coefficients. Wang et al. (2014) used VCMs for longitudinal data (with errors in variables) with Bayesian P-splines. Heim et al. (2007) used a 3D VCM in brain imaging.

Modulation models for seasonal data are an interesting application of the VCM (Eilers et al., 2008; Marx et al., 2010). The amplitude of a sinusoidal (or more complex) waveform is made to vary slowly over time. This assumes that the period is known. If that is not the case, or when it is not constant, it is possible to estimate both varying amplitude and phase of a sine wave (Eilers, 2009).

In a VCM, $\boldsymbol{y}$ and $\boldsymbol{x}$ are parallel vectors given at the same sampling positions in time or space. In penalized signal regression (PSR) we have a set of $\boldsymbol{x}$ vectors and corresponding scalars in $\boldsymbol{y}$ and the goal is to predict the latter. If the $\boldsymbol{x}$ vectors form the rows of a matrix $\boldsymbol{X}$, we have linear regression $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. The problem is ill-posed, because $\boldsymbol{X}$ has many more columns then rows. Take for example optical spectra that have been measured with many hundreds of wavelengths. The elements of $\boldsymbol{y}$ are known concentrations of a substance. Because the columns of $\boldsymbol{X}$ are ordered, it makes sense to force $\boldsymbol{\beta}$ to be smooth, by putting a difference penalty on it, thereby making the problem well-posed (Marx and Eilers, 1999).

In principle there is no need to introduce P-splines, by writing $\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{a}$ and putting the penalty on $\boldsymbol{a}$, but it reduces the computational load when $\boldsymbol{X}$ has many columns. Effectively we get penalized regression on the basis $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{B}$. After this step the machinery for cross-validation, standard errors and effective dimension becomes available. Notice that $\boldsymbol{a}$ is forced to be smooth, but $\boldsymbol{\mu}$ does not have to be smooth at all. Also not the rows of $\boldsymbol{X}$ are smoothed, but the regression coefficients.

Li and Marx (2008) proposed signal sharpening to enhance external prediction by incorporating PLS weights.

An extensive review of functional regression was presented by Morris (2015).

The standard PSR model implicitly assumes the identity link function. However, it can be bent through $\boldsymbol{\mu} = f(\boldsymbol{X}\boldsymbol{\beta}) = f(\boldsymbol{U}\boldsymbol{a})$, where $f(\cdot)$ is unknown. We call this model single-index signal regression (SISR), which is closely related to projection pursuit (Eilers et al., 2009). To estimate f, a second B-spline basis and corresponding coefficients are introduced. The domain is that of $\boldsymbol{U}\boldsymbol{a}$, and $\boldsymbol{a}$ has to be standardized (e.g. mean zero and variance 1) to make the model identifiable. For given coefficients, the derivative of $f(\boldsymbol{U}\boldsymbol{a})$ can be computed and inserted in a Taylor expansion. Using that, $\boldsymbol{a}$ and the coefficients for f are updated in turn until convergence.

P-splines have been implemented in other types of single-index models, e.g. see Yu and Ruppert (2002) and Lu et al. (2006). Leitenstorfer and Tutz (2011) used boosting and Antoniadis et al. (2004) used a Bayesian approach.

In the next section, we review the tensor product fundamentals that enable PSR extensions into two-dimensions. For example, Eilers and Marx (2003) and Marx and Eilers (2005) extended PSR to allow interaction with a discrete variable and to the two-dimensional case where each $\boldsymbol{x}$ is not a vector but a matrix. In these models there is no notion of time. When each element of $\boldsymbol{y}$ is not a scalar but a time series, as is $\boldsymbol{x}$, the historical functional linear model (HFLM) assumes that in principle all previous $\boldsymbol{x}$ can influence the elements of $\boldsymbol{y}$ (Malfait and Ramsay, 2003). Harezlak et al. (2007) introduced P-spline technology for the HFLM.

A mirror image of the HFLM is the interest term structure, estimating the expected future course of interest rates; see Jarrow et al. (2004) and Krivobokova et al. (2006).

Additionally, Marx et al. (2011) extended SISR to two dimensions, whereas Marx (2015) presented a hybrid varying-coefficient single-index model. In SISR, a weighted sum of $x(t)$ is formed and transformed. McLean et al. (2014) went one step further: $E(y_i) = \mu_i = \int F(x_i(t),t)dt$. This can be interpreted as first transforming x (with a different function for each t) and then adding the results, or "transform and add" in contrast to "add and transform".

8. Multi-dimensional smoothing

A natural extension of P-splines to higher dimensions is to form tensor products of one-dimensional B-spline bases and to add difference penalties along each dimension (Eilers and Marx, 2003). Figure 7 illustrates this idea, showing one tensor product $T_{jk}(x,y) = B_j(x)\check{B}_k(y)$. Figure 8 illustrates a "thinned" section of a tensor product basis; for clarity not all tensor products are shown. A matrix of coefficients determines the height of each "mountain": $\boldsymbol{A} = [a_{kl}]$, $k = 1,\ldots,\ n$ and $l = 1,\ldots,\ \check{n}$. The situation is completely analogous to Figure 4, but extended to two dimensions. The roughness of the elements of $\boldsymbol{A}$ determines how smooth the surface will be. To tune roughness, each column and each row of $\boldsymbol{A}$ is penalized.

One can choose to use one penalty parameter for both directions, (isotropic smoothing), or separate ones (anisotropic smoothing). In the latter case optimizing the amount of smoothing generates much more work. Many useful properties of one-dimensional P-splines carry over to higher dimensions. Weighting of (missing) observations and interpolation and extrapolation work well. Effective model dimension and fast cross-validation are available. They can also be used as a building block in smooth structures (see the next section).

Technically, multidimensional P-splines are challenging. The main issue is that, to be able to estimate A with the usual matrix-vector operations, we need to write it as a

vector and to put the tensor products in a proper basis matrix. With careful organization of the computations this can be solved elegantly (Eilers et al., 2006).

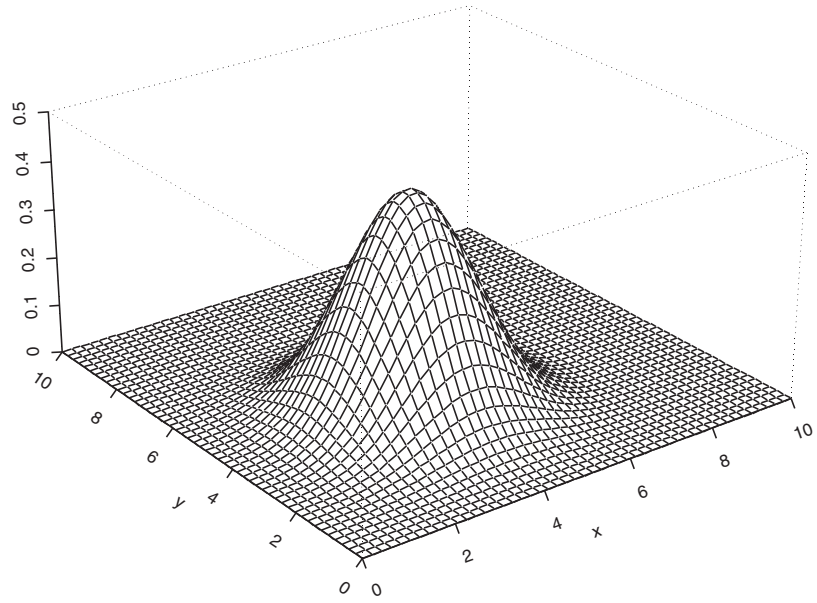


Figure 7: *The tensor product building block.*

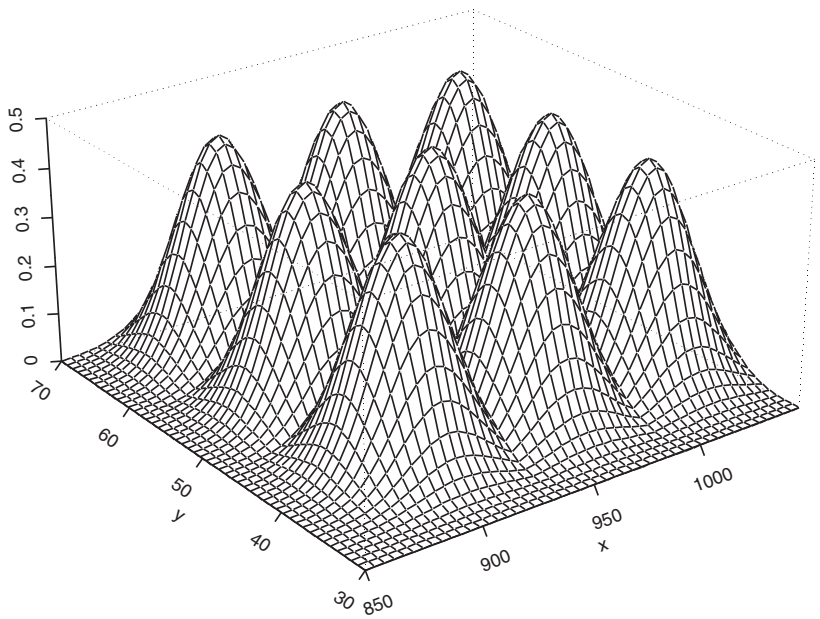


Figure 8: *Sparse portion of a complete tensor product B-spline basis.*

A natural application of multidimensional P-splines is the smoothing of data on a grid. For larger grids the demands on memory and computation time can become too large and special algorithms are needed. See Section 13 for details.

Multi-dimensional P-splines are numerically well-behaved, in contrast to truncated power functions. The poor numerical condition of the latter becomes almost insurmountable in higher dimensions. Proponents of TPF have avoided this issue by using radial basis functions (Ruppert et al., 2003; Kammann and Wand, 2003). This is, however, not an attractive scheme: a complicated algorithm is being used for placing the centres of the basis functions.

We emphasize that the set of tensor products does not have to be rectangular, although Figure 8 might give that impression. When dealing with, say, a ring-shaped data domain, we can remove all tensor products that do not overlap with the ring, and reduce the penalty matrix accordingly. This can save much computation, and in the case of a ring also is more realistic, because it prevents the penalty from working across the inner region.

As we showed for one-dimensional smoothing, the number of basis elements, here the tensor products, can be larger than the number of observations without problems, thanks to the penalties.

9. Additive smooth structures

As we have seen for the generalized additive and varying-coefficient model, the use of P-splines leads to a set of (modified) B-spline basis matrices which can be combined side-by-side into one large matrix. The penalties lead to a block-diagonal matrix. This idea extends to other model components like signal regression and tensor products. Standard linear regression and factor terms can be added too. This leads to additive smooth structures. Eilers and Marx (2002) proposed GLASS (generalized linear additive smooth structures), while Brezger and Lang (2006), referring to Fahrmeir et al. (2004), proposed STAR (structured additive regression). Belitz and Lang (2008) introduced simultaneous selection of variables and smoothing parameters in structured additive models.

The geoadditive model has received much attention; it is formed by the addition of one-dimensional smooth components and a two-dimensional spatial trend. Often the spatial component is modelled as a conditional autoregressive model. Brezger and Lang (2006) presented a Bayesian version of GLASS/STAR, also using 2D P-splines for modelling spatial effects in a multinomial logit model for forest health. Fahrmeir and Kneib (2009) further built on Bayesian STAR models by incorporating geoadditive features and Markov random fields, while addressing improper prior distributions. Also considering geoadditive structure, Kneib et al. (2011) expanded and unified Bayesian STAR models to further accommodate high-dimensional covariates.

Hierarchies of curves form a special type of additive smooth structures. For example, in growth data for children we can introduce an overall mean curve and two additional

curves that show the difference between boys and girls. Moreover, we can have a smooth curve per individual child. Durbán et al. (2005) gave an example (using truncated power functions), while Bugli and Lambert (2006) used proper P-splines in a Bayesian context.

10. P-splines as a mixed model

The connection between nonparametric regression and mixed models was first established over 25 years ago by Green (1987) and Speed (1991), but it was not until the late 1990s before it became a "hot" research topic (Wang, 1998; Zhang et al., 1998; Verbyla et al., 1999), partly due to the developments in mixed model software. These initial references were based on the use of smoothing splines. In the penalized spline context, several authors quickly extended the model formulation into a mixed model (Brumback et al., 1999; Coull et al., 2001; Wand, 2003). They used truncated power functions as the regression basis, since these have an intuitive connection with a mixed model. However, as previously mentioned, the numerical properties of TPFs are poor, compared to P-splines. In a short comment, that largely went unnoticed, Eilers (1999) showed how to interpret P-splines as a mixed model. Currie and Durbán (2002) used this approach and extended it to handle heteroscedastic or autocorrelated noise. Work on the general approach for a mixed model representation of smoothers with quadratic penalty was also presented in Fahrmeir et al. (2004).

With $\lambda = \sigma^2/\sigma_a^2$, the minimization problem in (2) is equivalent to:

$$Q^* = ||\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}||^2/\sigma^2 + \boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{a}/\sigma_a^2, \tag{14}$$

with σ_a^2 denoting the variance of the random effects a and σ^2 as the error variance. In fact, this is the minimization criterion in a random effects model of the form:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}, \quad \boldsymbol{a} \sim N(0, \sigma_a^2 \boldsymbol{P}^{-1}) \quad \epsilon \sim N(0, \sigma^2 \boldsymbol{I}). \tag{15}$$

As presented, difference penalties of order d do not penalize powers of x up to degree $d-1$. Therefore, $\boldsymbol{P}$ is singular (d eigenvalues are zero), and thus $\boldsymbol{a}$ has a degenerate distribution. One solution is to rewrite the model as $\boldsymbol{B}\boldsymbol{a} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, such that the d columns of $\boldsymbol{X}$ span the polynomial null space of $\boldsymbol{P}$ and the $(n-d)$ columns of $\boldsymbol{Z}$ span its complement. In this presentation, the random effects u have a non-degenerate distribution. This type of re-parametrization can be done in many ways. Eilers (1999) proposed $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}^{\mathsf{T}}(\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}})^{-1}$ (where $\boldsymbol{D}$ is the differencing matrix). A more principled approach (which can be used for any quadratic penalty) was introduced by Currie et al. (2006) and is based on the singular value decomposition of $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}$, yielding $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}$. In either case, the equivalent mixed model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{I}) \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}). \tag{16}$$

Instead of one smoothing parameter, we now have two variances, and we can profit from the stable and efficient algorithms and software that are available for mixed models. Especially in complex models with multiple smooth components, this approach can be more attractive than optimizing cross-validation or AIC. Yet, which approach (based on prediction error or maximum likelihood) is optimal for the selection of the smoothing parameter? Several papers on this subject have appeared along the years, but no unified opinion has been reached: Kauermann (2005a) showed that the ML estimate has the tendency to under-smooth, and prediction error methods give better performance than maximum likelihood based approaches, Gu (2002) also found that ML delivers rougher estimates than GCV, while Ruppert et al. (2003) found, through simulation studies, that REML will produce smoother fits than GCV (simular conclusion was also found in Kohn et al., 1991). Also, Wood (2011) concluded that REML/ML estimation is preferable to GCV for semiparametric GLMs due to its better resistance to over-fitting, less variability in the estimated smoothing parameters, and reduced tendency to having multiple minima. So, it is clear that there is no unique answer to this question, since different scenarios, will yield different conclusions. Moreover, behind the criteria used to select the smoothing parameter, there is, in our opinion, a deeper question: is it fair to use mixed models methodology for estimation and inference, when the mixed model representation of a P-spline could be considered just a "trick" to facilitate parameter estimation? This is a question for which we have no answer; researchers have different (and strong) opinions about the mixed model approach (even the authors of this paper do not always agree on this matter), but the truth is that it has become a revolution that has yielded incredible advances in a very short time. It certainly has helped to make penalized splines "salonfähig": nowadays they are acceptable and even attractive to a large part of the statistical community.

The estimation of the fixed and random effects is based on the maximization of the joint density of $(\boldsymbol{y}, \boldsymbol{u})$ for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ which results in the well-known Henderson's equations (Henderson, 1975):

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{Z}^\top \boldsymbol{y} \end{bmatrix}, \tag{17}$$

where $\lambda = \sigma^2/\sigma_u^2$. The solution of these equations yields $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. The variance components σ^2 and σ_u^2 are, in general, estimated by REML (Restricted Maximum Likelihood), see Patterson and Thompson (1971)), and the solutions are obtained by numerical optimization.

Other approaches can be used, and among them, it is worth mentioning the algorithm given by Schall (1991), which estimated random effects and dispersion parameters without the need to specify their distribution. The key is that each variance component is connected to an effective dimension. The sum of squares of the corresponding random

coefficients is equal to their variance times their respective effective dimension. This fact can be exploited in an iterative algorithm. After each cycle of smoothing, the sums of squares and effective dimensions are recomputed, which then are used to update the variances for the next round. See Marx (2010) and Rodriguez-Alvarez et al. (2015) for details and for extensions to multidimensional smoothing.

It is important to note a fact about prediction. Although the fitted model is the same regardless of parametrization (i.e. as mixed model or not), the standard errors for the predicted values are not invariant. This results because the variability of the random effects is taken into account in the mixed model case (and not in the other). The confidence interval obtained from the original parametrization is $\hat{f}(\boldsymbol{x}) \pm 2\hat{\sigma}\sqrt{(\boldsymbol{HH})_{ii}}$ (where $\boldsymbol{H}$ is such that $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$). This confidence interval covers $E[\hat{f}(\boldsymbol{x})]$ rather than $f(\boldsymbol{x})$, since $\hat{f}(\boldsymbol{x})$ is not an unbiased estimate of $f(\boldsymbol{x})$. Whereas in the mixed model framework, $\hat{f}(\boldsymbol{x})$ is unbiased due to the random $\boldsymbol{u}$, and the biased adjusted confidence interval is $\hat{f}(\boldsymbol{x}) \pm 2\hat{\sigma}\sqrt{(\boldsymbol{H})_{ii}}$ (Ruppert et al., 2003).

Of course, the interest of the mixed model representation of P-splines has been motivated by the possibility of including smoothing in a larger class of models. In fact, during the last 15 years, there has been an explosion of models: ranging from estimating subject-specific curves in longitudinal data (Durbán et al., 2005), to extending classical models in economics Basile et al. (2014), to playing a key role in the recent advances in functional data analysis (Scheipl et al., 2015; Brockhaus et al., 2015), among others.

10.1. P-splines and correlated errors

Although the mixed model approach has allowed the generalization of many existing models, there is an area in which it has played a key role: data with serial correlation. For many years the main difficulty when fitting a smooth model in the presence of correlation has been the joint estimation of the smoothing and correlation parameters. It is well known that the standard methods for smoothing parameter selection (based on minimization of the mean squared prediction error) generally under-smooth the data in the presence of positive correlation, since a smooth trend plus correlated noise can be seen as a less smooth trend plus white noise.

The solution is to take into account the correlation structure explicitly, i.e. $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ can depend on one or more correlation parameters. Durbán and Currie (2003) presented a strategy to select the smoothing parameter and estimate the correlation based on REML. Krivobokova and Kauermann (2007) showed that maximum likelihood estimation of the smoothing parameter is robust, even under moderately misspecified correlation. This method has allowed the inclusion of temporal non-linear trends and filtering of time series (Kauermann et al., 2011).

Recently, and motivated by the need to improve the speed and stability of forecasting models, Wood et al. (2015) have developed efficient methods for fitting additive models to large data sets with correlated errors.

Correlation also appears in more complex situations, for example in the case of spatial data. Lee and Durbán (2009) combined two-dimensional P-splines and random effects with a CAR (conditional auto-regressive) structure to estimate spatial trends when data are geographically distributed over locations on a map. Other authors have taken different approaches; they combined additive mixed models with spatial effects represented by Markov or Gaussian random fields (Kneib and Fahrmeir, 2006; Fahrmeir et al., 2010).

10.2. Multidimensional P-splines as mixed models

Multidimensional P-splines can be handled as a mixed model too. A first attempt was made by Ruppert et al. (2003) using radial basis functions. Currie et al. (2006) analysed tensor product P-splines as mixed models. Here, the singular value decomposition of the penalty matrix (as in the 1D case) is used to construct the mixed model matrices. This approach works for any sum of quadratic penalties (Wood, 2006a). However, when the penalty is expressed as the sum of Kronecker product of marginal bases (the Kronecker sum of penalties), the representation as a mixed model is based on the reparametrization of the marginal bases. An important by-product of this parametrization is that the transformed penalty matrix (i.e. the covariance matrix of the random effects), and the mixed model matrices lead to an interesting decomposition of the fitted values as the sum of main effects and interactions (Lee and Durbán, 2011):

$$E(\boldsymbol{Y}) = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + f_3(\boldsymbol{x}_1, \boldsymbol{x}_2).$$

This decomposition is strongly related to the work proposed by Gu (2002) on smoothing spline analysis of variance.

The model now has multiple smoothing parameters, which makes estimating them less efficient, if numerical optimization were to be used. Several steps have been taken to make computation efficient. Wood (2011) used a Laplace approximation to obtain an approximate REML suitable for efficient direct approximation. Lee et al. (2013) improved computational efficiency by using nested B-spline bases, and modified the penalty so that optimization could be carried out in standard statistical software. Wood and Scheipl (2013) proposed an intermediate low-rank smoother. Recently, Schall's algorithm has been extended (Rodriguez-Alvarez et al., 2015) to the case of multidimensional smoothing. This work also shows the fundamental role of the effective dimensions of the components of the model.

11. Bayesian P-splines

It is a small step to go from a latent distribution in a mixed model to a prior in a Bayesian interpretation. Bayesian P-splines were proposed by Lang and Brezger (2004), and they were made accessible by appropriate software (Brezger et al., 2005). Their approach is based on Markov chain Monte Carlo (MCMC) simulation. As for the mixed model, the penalty leads to a singular distribution. This is solved by simulation using a random walk of the same order as that of the differences.

It is also possible to start from a mixed model representation. Crainiceanu et al. (2007) did this in one dimension, using truncated power functions. They avoid tensor products of TPFs and switch to radial basis functions for spatial smoothing. These authors also allowed for varying (although isotropic) smoothness and for heteroscedasticity. Jullion and Lambert (2007) proposed a Bayesian model for adaptive smoothing.

As an alternative to MCMC, integrated nested Laplace approximation (INLA) is powerful and fast, and it is gaining in popularity (Rue et al., 2009). INLA avoids stochastic simulation for precision parameters and uses numerical integration instead. Basically INLA uses a parameter for each observation so a (B-spline) regression basis has to be implemented in an indirect way, as a matrix of constraints (Fraaije et al., 2015).

INLA is an attractive choice for anisotropic smoothing. By working with a sum of precision matrices it can handle the equivalent of a mixed model with overlapping penalties (Rodriguez-Alvarez et al., 2015).

12. Varia

In this section we discuss some subjects that do not find a natural home in one of the preceding sections. We take a look at asymptotic properties of P-splines and at boosting.

Several authors have studied the asymptotic behaviour of P-splines. See Li and Ruppert (2008); Claeskens et al. (2009); Kauermann et al. (2009); Wang et al. (2011). Although we admire the technical level of these contributions, we do not fully see their practical relevance. The problem is their very limited interpretation of increasing the number of observations: it is all about more observations on the same domain. In that case it is found that the number of knots should grow as a small power of the number of observations. Yet, the whole idea of P-splines is to use far too many knots and let the penalty do the work. Trying to optimize the number of knots, as Ruppert (2002) did, is not worthwhile. He reports some cases where more knots increase the estimation error, but the numbers are not dramatic. His analysis was based on truncated power functions, which he confusingly calls P-splines, with knots at quantiles of the observed x. It is not clear how this design influences the results. For a proper analysis equally-spaced knots have to be used.

Most asymptotic analyses of penalized splines use the framework of truncated power functions (TPF). There is an unpenalized part, a polynomial in x, of the same degree

as that of the TPF. The penalty on the TPF is on the size of the coefficients, not on differences thereof. This makes analytical work easier. In Section 10, we have presented two alternative representations of P-splines that have a unpenalized polynomial part and a size penalty on the other basis functions. We believe that such are more suitable than TPF, if only because of the decoupling of the degree of the splines and the order of the penalty. Boer (2015) recently presented a variant that keeps the basis sparse.

What is neglected in papers on asymptotic theory is that often we have to deal with observation in time or space, where more observations bring about a proportional increase in the size of the domain.

In Section 4, we have shown that there is no danger in using many splines even when fitting only a few data points. Hence one is always free to use many splines and not worry about optimization of their number. We therefore advise to use 100 B-splines, a safe choice.

Boosting for smoothing basically works as follows (Tutz and Binder, 2006; Schmid and Hothorn, 2008): (1) smooth with a very strong penalty and save the result, (2) smooth the residuals and add (a fraction of) this result to the previous result, (3) repeat step (2) many times. The result gets more flexible with each iteration. So one has to stop at some point, using AIC or another criterion. Boosting has many enthusiastic proponents, and its use has been extended to non-normal data and additive models and other smooth structures (Mayr et al., 2012). We find it difficult to see its advantages, especially when we compare it to Schall's algorithm for tuning multiple smoothing parameters, which we presented in Section 10. On the other hand, boosting allows to select relevant variables in a model and the use of non-standard objective functions.

13. Computation and software

For many applications standard P-splines do not pose computational challenges. The size of the B-spline basis will be moderate and many thousands of observations can be handled with ease. If the data are observed on an equidistant grid and only smoothed values on that grid are wanted, one can just as well use the identity matrix as a basis. This leads to the Whittaker smoother (Whittaker, 1923; Eilers, 2003). The number of coefficients will be equal to the number of observation, but in combination with sparse matrix algorithms a very fast smoother is obtained.

Sparse matrices also are attractive when long data series have to be smoothed with a large B-spline basis (de Rooi et al., 2014). Even though the basis matrix is sparse, one has to take care to avoid computing dense matrices along the way, as is the case when using truncated power functions. The key is to recognize that $\boldsymbol{B}_{j+k}(\boldsymbol{x}) = \boldsymbol{B}_j(\boldsymbol{x} - s\boldsymbol{k})$, where s is the distance between the knots. Each $\boldsymbol{x}_i$ is shifted by $-s\boldsymbol{k}_i$ to a chosen subdomain , and a basis of only four (cubic) B-splines is computed on that domain. In a last step a sparse matrix is constructed with the columns of row i shifted to the right by $\boldsymbol{k}_i$. An added advantage is that numerical roundoff is minimized (Eilers and Marx, 2010).

When the penalty parameter is large, forming $\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}$ explicitly to solve the penalized normal equations is not optimal and rounding problems can occur. It is better to use an augmented version of $\boldsymbol{B}$ yielding $\bar{\boldsymbol{B}}$, where $\bar{\boldsymbol{B}} = [\boldsymbol{B}^{\mathsf{T}}\ \sqrt{\lambda}\boldsymbol{D}^{\mathsf{T}}]^{\mathsf{T}}$, and an augmented $\boldsymbol{y}$ as $\bar{\boldsymbol{y}} = [\boldsymbol{y}^{\mathsf{T}}\ \boldsymbol{0}^{\mathsf{T}}]^{\mathsf{T}}$ and perform linear regression of $\bar{\boldsymbol{y}}$ on $\bar{\boldsymbol{B}}$ using the QR decomposition. Here $\boldsymbol{0}$ stands for a vector of zeros with length equal to the number of rows of $\boldsymbol{D}$. See Wood (2006a) for advice on stable computation.

The demands of additive models on computer memory and computation time often are modest. However, very large data sets need special treatment when they do not fit in the working memory. Wood et al. (2015) described such an application, forecasting electricity consumption in France. They developed a specialized algorithm, which is part of the the R package (`mgcv`) as the function `bam`.

In two-dimensional smoothing of large grids, using tensor products, one can run into yet another problem. The data and the one-dimensional bases may easily fit into memory, but the (inner products of) Kronecker products cannot be handled. The two-dimensional basis, $\breve{\boldsymbol{B}} \otimes \boldsymbol{B}$, has $m_1 m_2$ rows and $n_1 n_2$ columns. When smoothing a large 1000 by 1000 image using $n_1 n_2 \approx 1000$, the basis has one billion elements, taking 8 bytes each, and so will not fit into 8 Gb of main memory. Even if it would, the computation of the inner products will be extremely taxing. Note that in this case we have around 1000 coefficients, so it is not the size of the final system of penalized normal equations that is the problem. Fortunately, by rearranging the calculations, one can avoid the explicit Kronecker products and gain orders of magnitudes in computation speed and memory use (Currie et al., 2006; Eilers et al., 2006). This array algorithm, so-called GLAM, allows arbitrary weights and so is suitable for generalized linear smoothing.

When no weights are involved, even larger improvements are possible, by using the "sandwich smoother" (Xiao et al., 2013). The basic idea is that one can first apply one-dimensional smoothing to the rows of a matrix and then to the columns (or the other way around). The order is immaterial, as is easy to see from the explicit equation for $\boldsymbol{A} = (\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{Y}\breve{\boldsymbol{B}}(\breve{\boldsymbol{B}}^{\mathsf{T}}\breve{\boldsymbol{B}} + \breve{\lambda}\breve{\boldsymbol{D}}^{\mathsf{T}}\breve{\boldsymbol{D}})^{-1}$, the matrix of coefficients. A similar approach was followed by Eilers and Goeman (2004), using a modified Whittaker smoother.

Nowadays it is quite common to publish computer code on a website, or as supplementary material, to accompany statistical papers. This is certainly true for the literature on P-splines. We do not try to describe these individual efforts. Instead we point to some packages for R (R Core Team, 2015) with a rather wide scope.

Originally designed for fitting generalized additive model, and accompanying Wood (2006b), `mgcv` has grown into the Swiss army knife of smoothing. It offers a diversity of basis functions and their tensor products for multidimensional smoothing. Furthermore, it can fit varying-coefficient models and signal regression, and it can mix and match components in an additive way. It offers a diversity of distributions to handle (over-dispersed) non-normal data.

We described the GAMLSS model in Section 5. A very extensive package is available. Its core is aptly called `gamlss`; it can be extended with a suite of add-ons for censored data, mixed models, and a variety of continuous and discrete distributions.

The package `MortalitySmooth` focuses on smoothing of counts in one and two dimensions (Camarda, 2012). It also is a nice source for mortality data from several countries.

BayesX (Brezger et al., 2005) is a stand-alone program for Windows and Linux. It covers all the models that fit in the generalized linear additive smooth structure (or structured additive regression) framework. The Bayesian algorithms are based on Markov chain Monte Carlo. It also offers mixed model based algorithms. There are R packages to install BayesX and to communicate with it.

It is also possible to use the R-INLA (Rue et al., 2009) package for fitting additive models with P-splines. See Fraaije et al. (2015) and the accompanying software.

The package `mboost` offers boosting for a variety of models, including P-splines and generalized additive models (Hofner et al., 2014). With the extension `gamboostLSS`, one can apply boosting to models for location, scale and shape, similar to GAMLSS.

To estimate smooth expectile curves or surfaces, the package `expectreg` is available.

14. Discussion

The paper by Eilers and Marx (1996) that started it all contained a "consumer score card", comparing various smoothing algorithms. P-splines received the best marks and their inventors concluded that they should be the smoother of choice. Two decades later, it is gratifying to see that this opinion is being shared by many statisticians and other scientists. Once prominent tools like kernel smoothers and local likelihood are gradually fading into obscurity.

In twenty years, P-spline methodology has been extended in many directions. The analogy with mixed models is being exploited to the fullest, as is the Bayesian approach, leading to new interpretations of penalties and powerful recipes for optimizing the amount of smoothing. Multidimensional smoothing with tensor products has become practical and fast, thanks to array algorithms. Regression on (multidimensional) signals has also become practical. Smooth additive structures allow the combination of various models. The key is the combination of rich B-spline regression and a simple roughness penalty. Actually the penalties are the core and many variations have been developed, while the B-spline basis did not change. We expect to see exciting developments in the near future. For a start, we sketch some aspects that we hope will get much attention.

We wrote that the penalty forms the skeleton and that the B-splines put flesh on the bones. That means that new ideas for penalties have to be developed. One promising avenue is the application to differential equations. One can write the solution as a sum of B-splines (the collocation method) and use the differential equation (DE) as the penalty (Ramsay et al., 2007). In this light the usual penalty for smoothing splines is equivalent to a differential equation that says that the second derivative of the solution is zero everywhere. O'Sullivan (1986) took the step from a continuous penalty to a discrete one. This can also be done with a DE-based penalty. However, if the coefficients of the DE are

not fixed (e.g. estimated from the data), then this generates a significant computational load. It will be useful to study (almost) equivalent discrete penalties, based on difference equations.

It is remarkable that in one-dimensional smoothing, kriging is almost absent. Altman (2000) compared splines and kriging and found that serial correlation is a key issue. If it is present and ignored, splines do not perform well. There are ways to handle correlation, as discussed in Section 10.. In spatial data analysis, kriging is still dominant. We believe that for many applications, tensor product P-splines would be a much better choice, especially if one is more interested in estimating a trend rather than doing spatial interpolation. It may appear that attempting to estimate a covariance structure from the data is a worthwhile effort, but in practice it often leads to unstable procedures. Handling non-normal data with kriging is cumbersome. In contrast, P-splines impose a relatively simple covariance structure, and in practice do the job in a very stable way, as our experiences with the analysis of agricultural field trials has shown. Smoothing of data on large grids is problematic for kriging, but P-splines and array algorithms handle such data with ease. In some cases it might even be attractive to summarize the data (as counts and sums) on a grid before analysis. Combined with the PS-ANOVA approach (Lee et al., 2013), which avoids detailed modelling of higher-order interactions, powerful tools for large data sets can be developed.

In some applications extrapolation is very important. Mortality data are a prime example. The order of the differences in the penalty determines the result: for first order differences it is a constant, for second order a straight line and a weighted combination of both gives an exponential curve. A challenge is to determine which penalty to use and to set its tuning parameter(s) for optimal extrapolation. In one dimension extrapolation does not influence the fit to the data. This is not true in two dimensions, for example with life tables. The penalties for the age and the time dimension interact and the size of the extrapolation region also has an influence. More research is needed to better understand these issues.

In several places in this paper, we have encountered the effective dimension of (components) of a model. It is an important parameter when optimizing penalties. Yet it deserves more attention on its own right. The definition, by Ye (1998), in (10) is very powerful. The contribution to $\mathrm{ED} = \sum_i \partial \hat{y}_i / \partial y_i$ by a component of an additive model can be determined clearly by following a change in y_i down the model to the coefficients, and from there back up again to the corresponding change in $\hat{y}_i$. Partial effective dimensions can be calculated this way; they are important summaries of the contributions of the model components.

In this paper, we have tried to give a glimpse of the many landmarks created in the last 20 years. It has been a collective achievement, the result of the work of many researchers who believed in the power of P-splines. We see a great and exciting future ahead, as there are many problems to solve, new complex data to model, and especially a new generation of bright statisticians who are already showing that P-splines have much more to contribute to this century, the century of data.

Appendix A. O-splines or P-splines?

Wand and Ormerod (2008) introduced O'Sullivan splines, or O-splines for short. They were not entirely pleased with the pure discrete penalty of P-splines and returned to the integral of the squared second (or higher) derivative of the fitted function. This can be attractive, especially when the knots of the B-spline basis are not evenly spaced. There are cases when this can be very valuable. As an example, Whitehorn et al. (2013) presented an example of high-dimensional smoothing with tensor products in high-energy physics to model the response of a detector. In this case, more detail was needed in the centre than near the boundaries. However, this was not the motivation of Wand and Ormerod. They rather favour the use of quantiles of x for the knots.

The paper claims that P-splines do not extrapolate well, when compared to the smoothing spline. Hence O-splines should be preferred. This claim was repeated by Ruppert et al. (2009). The paper by Wand and Ormerod (2008) has been cited more than 50 times, so apparently the message did not get lost.

We were concerned about this analysis because a basis with multiple knots at the domain boundaries had been used for the O-splines. If multiple knots had also been used for the P-splines (similar to the one in the bottom panel of Figure 3), artifacts could have occurred. So we decided to take the example data from their paper (dataset `fossil` in the R package `SemiPar`) and re-analyse them. We downloaded the file `WandOrmerod08.Rs` from Matt Wand's personal web page. For fitting, we extracted the section that invokes the function `lme` in the package `nlme` for estimating an equivalent mixed model. This program was adapted for P-splines by changing the basis and the penalty matrix. For comparison we use the function `smooth.spline` that is a standard part of R. It tunes the amount of smoothing automatically to the data, using cross-validation.

Figure 9 shows the fits of P-splines and a smoothing spline. We used 40 B-splines on the domain from 85 to 130. There is strong correspondence between the two splines. This is also true for the estimated derivative, which was approximated by taking differences.

Surprisingly, the O-splines do not work as well as P-splines, as Figure 10 shows. This especially can be seen in the derivatives. It appears that the O-spline fit struggles near $x = 100$. The reason is that the knot density is low there, because the low local data density. What is more worrying is that derivatives of the extrapolated part is not constant, as they should be for a linear result.

We believe that the anomalous behaviour of the O-splines is caused by the choice of basis. Multiple knots do not go together with a discrete penalty on the spline coefficients. The root of all evil is the choice to use quantiles of x for the knots; there is absolutely no need for it.

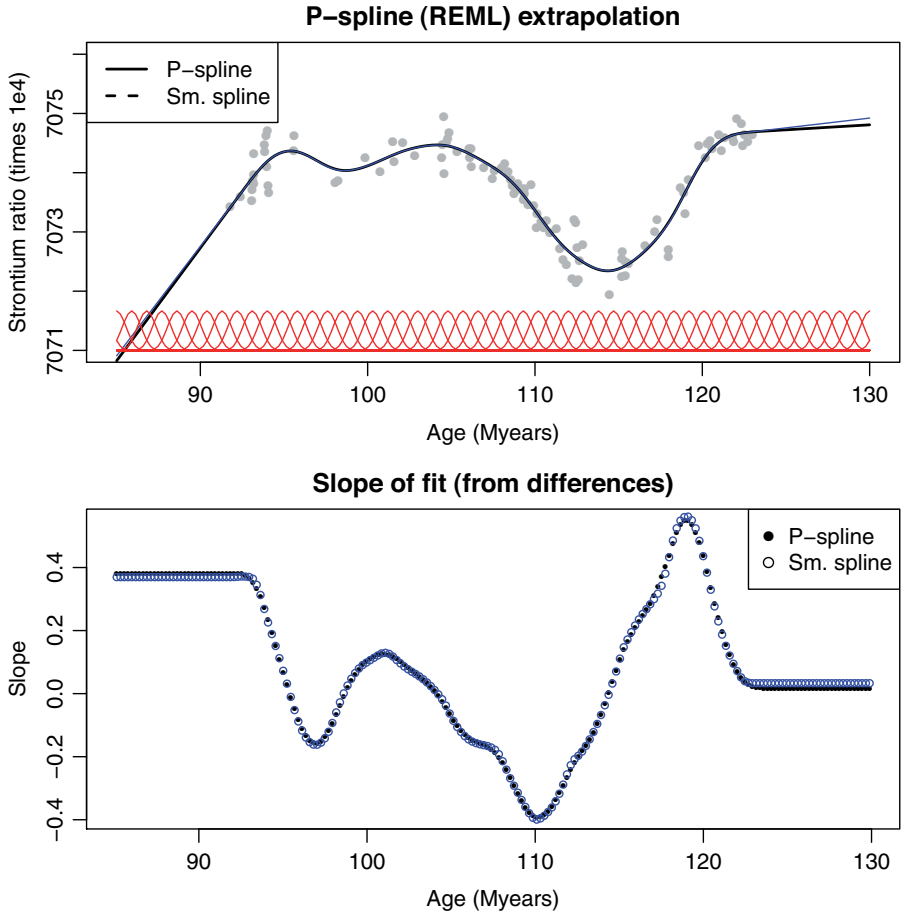


Figure 9: *Upper panel: P-spline and smoothing spline fit to the fossil data. On both sides the fit is extrapolated automatically. Lower panel: derivatives of both splines, as computed from first differences.*

Appendix B. Proof that the effective dimension is smaller than m

A formal proof starts from a simplified case, with $d = 0$ and a general basis $\boldsymbol{Z}$, where the system of equations to solve is

$$(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}. \tag{18}$$

The singular value decomposition of $\boldsymbol{Z}$ gives: $Z = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$, with $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}_m$ and $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}_n$. Through substitution:

$$(\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\boldsymbol{u} = (\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\hat{\boldsymbol{a}} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \tag{19}$$

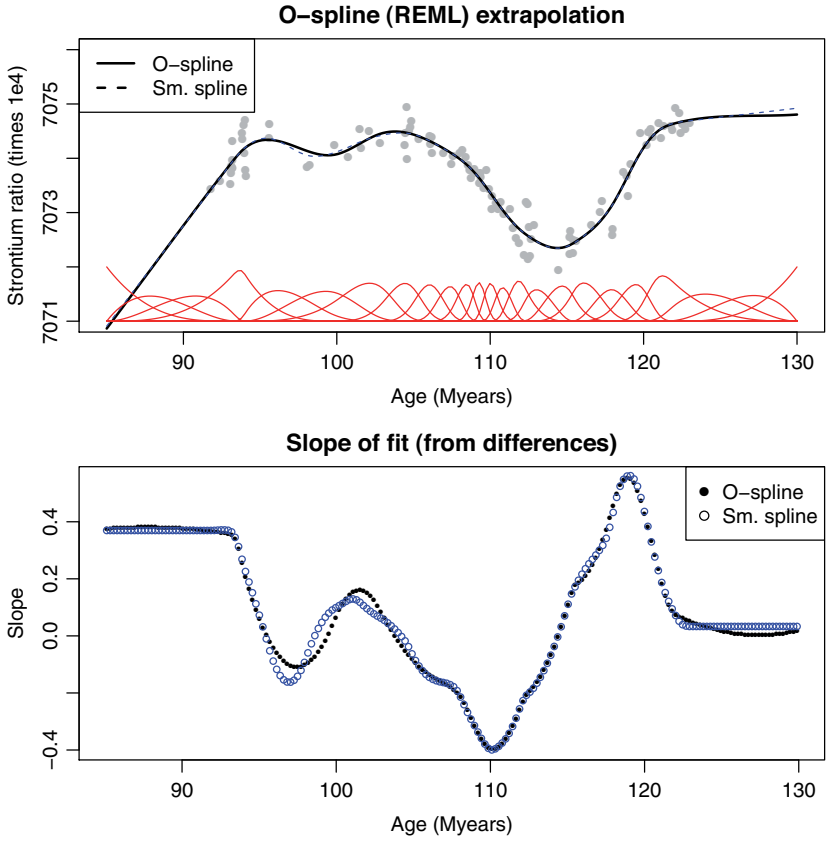


Figure 10: *Upper panel: O-spline and smoothing spline fit to the fossil data. On both sides the fit is extrapolated automatically. Lower panel: derivatives of both splines, as computed from first differences.*

Now assume that $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{\gamma}$. Fill in and multiply by $\boldsymbol{V}^{\mathsf{T}}$:

$$\boldsymbol{V}^{\mathsf{T}}(\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\boldsymbol{V}\boldsymbol{\gamma} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}. \tag{20}$$

Hence

$$(\boldsymbol{S}^2 + \lambda \boldsymbol{I}_m)\boldsymbol{\gamma} = \boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}. \tag{21}$$

This is a system with m equations in m unknowns. The system matrix is diagonal and non-singular.

The penalty is a special case here, but in Section 10 it was shown that P-splines can be transformed into a mixed model, specifically with $\boldsymbol{B}\boldsymbol{a} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$ and with a ridge penalty on $\boldsymbol{u}$. The Henderson equations (17) contain a part for $\boldsymbol{\beta}$. We now have

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}, \tag{22}$$

or

$$(\boldsymbol{S}^2 + \lambda \boldsymbol{I}_m)\boldsymbol{u} = \boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}). \tag{23}$$

Note that the value of $\boldsymbol{\beta}$ is immaterial, as it does not change the properties of the system matrix and only the right-hand side of the equation changes.

Acknowledgment

A part of the work of the first author was supported by a Chair of Excellence Grant from Carlos III University in Madrid, Spain.

References

Altman, N. (2000). Theory & methods: Krige, smooth, both or neither? *Australian & New Zealand Journal of Statistics*, 42, 441–461.

Andriyana, Y., Gijbels, I. and Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficient models. *Test*, 23, 153–194.

Antoniadis, A., Gregoire, G. and McKeague, I. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, 14, 1147–1164.

Azmon, A., Faes, C. and Hens, N. (2014). On the estimation of the reproduction number based on misreported epidemic data. *Statistics In Medicine*, 33, 1176–1192.

Basile, R., Durbán, M., Minguez, R., Montero, J. and Mur, J. (2014). Modeling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic and Dynamics Control*, 48, 229–245.

Belitz, C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics & Data Analysis*, 53, 61–81.

Bollaerts, K., Eilers, P. H. C. and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the l_1-norm. *Statistical Modelling*, 6, 189–207.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50, 967–991.

Brezger, A., Kneib, T. and Lang, S. (2005). BayesX: analysing Bayesian structured additive regression models. *Journal of Statistical Software*, 14.

Brockhaus, S., Scheipl, F., Torsten, H. and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, 15, 279–300.

Brumback, B., Ruppert, D. and Wand, M. (1999). Comment on: variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94, 794–797.

Bugli, C. and Lambert, P. (2006). Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Statistics in Medicine*, 25, 3718–3739.

Camarda, C. G. (2012). MortalitySmooth: an R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, 50, 1–24.

Camarda, C. G., Eilers, P. H. C. and Gampe, J. (2008). Modelling general patterns of digit preference. *Statistical Modelling*, 8, 385–401.

Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96, 529–544.

Coull, B., Schwartz, J. and Wand, M. (2001). Respiratory health and air pollution: additive mixed model analyses. *Biostatistics*, 2, 337–349.

Crainiceanu, C., Ruppert, D. and Carroll, R. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 17, 265–288.

Currie, I., Durbán, M. and Eilers, P. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4, 279–298.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, 2, 333–349.

Currie, I. D., Durbán, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 68, 259–280.

de Rooi, J. J., van der Pers, N. M., Hendrikx, R. W. A., Delhez, R., Bottger, A. J. and Eilers, P. H. C. (2014). Smoothing of X-ray diffraction data and K alpha(2) elimination using penalized likelihood and the composite link model. *Journal of Applied Crystallography*, 47, 852–860.

Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models, 3d ed.* CRC Press.

Durbán, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, 18, 251–262.

Durbán, M., Harezlak, J., Wand, M. and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24, 1153–1167.

Eilers, P. (1999). Discussion on: the analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, 307–308.

Eilers, P. (2005). Unimodal smoothing. *Journal Of Chemometrics*, 19, 317–328.

Eilers, P. (2009). The smooth complex logarithm and quasi-periodic models. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures*. Springer.

Eilers, P. and de Menezes, R. (2005). Quantile smoothing of array CGH data. *Bioinformatics*, 21, 1146–1153.

Eilers, P., Currie, I. and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50, 61–76.

Eilers, P. H. C. (2003). A perfect smoother. *Analytical Chemistry*, 75, 3631–3636.

Eilers, P. H. C. (2007). III-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7, 239–254.

Eilers, P. H. C. and Borgdorff, M. W. (2007). Non-parametric log-concave mixtures. *Computational Statistics & Data Analysis*, 51, 5444–5451.

Eilers, P. H. C. and Goeman, J. J. (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20, 623–628.

Eilers, P. H. C. and Marx, B. D. (1992). Generalized linear models with P-splines. In *Proceedings of GLIM 92 and the 7th International Workshop on Statistical Modelling*.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–102.

Eilers, P. H. C. and Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, 11, 758–783.

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66, 159–174.

Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653.

Eilers, P. H. C., Gampe, J., Marx, B. D. and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics In Medicine*, 27, 3430–3441.

Eilers, P. H. C., Li, B. and Marx, B. D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 96, 196–202.

Fahrmeir, L. and Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference*, 139, 843–859.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14, 731–761.

Fahrmeir, L., Kneib, T. and Konrath, S. (2010). Bayesian regularization in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20, 203–219.

Fraaije, R. G. A., ter Braak, C. J. F., Verduyn, B., Breeman, L. B. S., Verhoeven, J. T. A. and Soons, M. B. (2015). Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. *Functional Ecology*, 29, 971–980.

Frasso and Eilers (2015). L- and V-curves for optimal smoothing. *Statistical Modelling*, 15, 91–111.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Greven, S. (2008). *Non-Standard Problems in Inference for Additive and Linear Mixed Models*. Cuvillier Verlag.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34, 561–580.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R. and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational Statistics & Data Analysis*, 51, 4911–4925.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320–338.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, 55, 757–796.

Heim, S., Fahrmeir, L., Eilers, P. H. C. and Marx, B. D. (2007). 3D space-varying coefficient models with application to diffusion tensor imaging. *Computational Statistics & Data Analysis*, 51, 6212–6228.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423–447.

Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics*, 29, 3–35.

Jarrow, R., Ruppert, D. and Yu, Y. (2004). Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the Americal Statistical Association*, 99, 57–66.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis*, 51, 2542–2558.

Kammann, E. and Wand, M. (2003). Geoadditive models. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 52, 1–18.

Kauermann, G. (2005a). A note on smoothing parameter selection for penalized spline smoothing. *Journal of statistical planning and inference*, 127, 53–69.

Kauermann, G. (2005b). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49, 169–186.

Kauermann, G. and Khomski, P. (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics & Data Analysis*, 51, 1944–1956.

Kauermann, G., Krivobokova, T. and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 71, 487–503.

Kauermann, G., Krivobokova, T. and Semmler, W. (2011). Filtering time series with penalized splines. *Studies in Nonlinear Dynamic & Econometrics*, 15, 1–26.

Kauermann, G., Schellhase, C. and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical B-splines. *Scandinavian Journal of Statistics*, 40, 685–705.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, 62, 109–118.

Kneib, T., Konrath, S. and Fahrmeir, L. (2011). High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 60, 51–70.

Kohn, R., Ansley, C. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, 86, 1042–1050.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102, 1328–1337.

Krivobokova, T., Kauermann, G. and Archontakis, T. (2006). Estimating the term structure of interest rates using penalized splines. *Statistical Papers*, 47, 443–459.

Lambert, P. (2011). Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics & Data Analysis*, 55, 429–445.

Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53, 1388–1399.

Lang, S. and Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics & Data Analysis*, 53, 2968–2979.

Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11, 49–69.

Lee, D.-J., Durbán, M. and Eilers, P. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis*, 61, 22–37.

Leitenstorfer, F. and Tutz, G. (2011). Estimation of single-index models based on boosting techniques. *Statistical Modelling*, 11, 203–217.

Li, B. and Marx, B. D. (2008). Sharpening P-spline signal regression. *Statistical Modelling*, 8, 367–383.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95, 415–436.

Lu, X., Chen, G., Singh, R. and Song, P. (2006). A class of partially linear single-index survival models. *Canadian Journal of Statistics*, 34, 97–112.

Lu, Y., Zhang, R. and Zhu, L. (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics-Theory and Methods*, 37, 2249–2261.

Malfait, N. and Ramsay, J. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31, 185–201.

Marx, B. D. (2010). P-spline varying coefficient models for complex data. In G. Tutz and T. Kneib, editors, *Statistical Modelling and Regression Structures*, 19–43. Springer.

Marx, B. D. (2015). Varying-coefficient single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 143, 111–121.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28, 193–209.

Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, 41, 1–13.

Marx, B. D. and Eilers, P. H. C. (2005). Multidimensional penalized signal regression. *Technometrics*, 47, 13–22.

Marx, B. D., Eilers, P. H. C., Gampe, J. and Rau, R. (2010). Bilinear modulation models for seasonal tables of counts. *Statistics and Computing*, 20, 191–202.

Marx, B. D., Eilers, P. H. C. and Li, B. (2011). Multidimensional single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 109, 120–130.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data. A flexible approach based on boosting. *Journal of the Royal Statistical Society Series C*, 61, 403–427.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014). Functional Generalized Additive Models. *Journal of Computational and Graphical Statistics*, 23, 249–269.

Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Application*, 2, 321–359.

Myers, R. (1989). *Classic and Modern Regression with Applications*. PWS-KENT.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, 1, 505–527.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.

Pya, N. and Wood, S. (2015). Shape constrained additive models. *Statistics and Computing*, 25, 543–559.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. and Silverman, B. (2003). *Functional Data Analysis, 2nd ed.* Springer.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 741–796.

Rigby, R. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Computational Statistics & Applied Statistics*, 54, 507–544.

Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. (2012a). Reliable single chip genotyping with semi-parametric log-concave mixtures. *Plos One*, 7.

Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. (2012b). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS ONE*, 7.

Rizzi, S., Gampe, J. and Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, 182, 138–147.

Rodriguez-Alvarez, M., Lee, D., Kneib, T., Durbán, M. and Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, 25, 941–957.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal Of Statistics*, 42, 205–223.

Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193–1256.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719–727.

Scheipl, F., Staicu, A. and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24, 477–501.

Schellhase, C. and Kauermann, G. (2012). Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, 27, 757–777.

Schmid, M. and Hothorn, T. (2008). Boosting additive models using component-wise P-Splines. *Computational Statistics & Data Analysis*, 53, 298–311.

Schnabel, S. K. and Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53, 4168–4177.

Schnabel, S. K. and Eilers, P. H. C. (2013). Simultaneous estimation of quantile curves using quantile sheets. *ASTA-Advances In Statistical Analysis*, 97, 77–87.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56, 755–767.

Sobotka, F., Kauermann, G., Waltrup, L. S. and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, 23, 135–148.

Speed, T. (1991). Comment on: that BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15–51.

Thompson, R. and Baker, R. (1981). Composite link functions in generalized linear models. *Applied Statistics*, 30, 125–131.

Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62, 961–971.

Verbyla, A., Cullis, B., Kenward, M. and Welham, S. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, 269–312.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223–249.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal Of Statistics*, 50, 179–198.

Wang, X., Shen, J. and Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, 5, 1–17.

Wang, X.-F., Hu, B., Wang, B. and Fang, K. (2014). Bayesian generalized varying coefficient models for longitudinal proportional data with errors-in-covariates. *Journal of Applied Statistics*, 41, 1342–1357.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, 60, 159–174.

Whitehorn, N., van Santen, J. and Lafebre, S. (2013). Penalized Splines for smooth representation of high-dimensional Monte Carlo datasets. *Computer Physics Communications*, 184, 2214–2220.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673–686.

Wood, S. (2006a). *Generalized Additive Models: An Introduction with R*. CRC Press.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, 73, 3–36.

Wood, S., Scheipl, F. and Faraway, J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23, 341–360.

Wood, S. N. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, 48, 445–464.

Wood, S. N., Goude, Y. and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 64, 139–155.

Xiao, L., Li, Y. and Ruppert, D. (2013). Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 75, 577–599.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection . *Journal of the American Statistical Association*, 93, 120–131.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the Americal Statistical Association*, 97, 1042–1054.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the Americal Statistical Association*, 93, 710–719.

SORT 39 (2) July-December 2015, 187-208

Likelihood-based inference for the power regression model

Guillermo Martínez-Flórez[1], Heleno Bolfarine[2] and Héctor W. Gómez[3]

Abstract

In this paper we investigate an extension of the power-normal model, called the alpha-power model and specialize it to linear and nonlinear regression models, with and without correlated errors. Maximum likelihood estimation is considered with explicit derivation of the observed and expected Fisher information matrices. Applications are considered for the Australian athletes data set and also to a data set studied in Xie et al. (2009). The main conclusion is that the proposed model can be a viable alternative in situations were the normal distribution is not the most adequate model.

MSC: 60E05.

Keywords: Correlation, maximum likelihood, power-normal distribution, regression.

1. Introduction

Linear and nonlinear regression models are statistical techniques typically used for modeling and studying relationships between variables in several areas of human knowledge such as biomedical and agricultural sciences, engineering, and many others, being extremely useful for data analysis. One important step in regression analysis is parameter estimation, usually made under the assumption of normality. However, there are situations were the normal assumption is not realistic and several distributions have been suggested as alternatives to the normal model. Among such models we have the Student-t, logistic and exponential power distributions (Cordeiro et al., 2000 and Galea et al., 2005), whereas for the asymmetric nonlinear model we have only the work of Cancho

[1] Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad de Córdoba, Córdoba, Colombia, gmartinez@correo.unicordoba.edu.co

[2] Departamento de Estatística, IME, Universidad de Sao Paulo, Sao Paulo, Brasil, hbolfar@ime.usp.br

[3] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile, hector.gomez@uantof.cl

Received: February 2013

Accepted: April 2015

et al. (2008). In this paper we suggest an alternative asymmetric model, the alpha-power model, for fitting linear and nonlinear regression models. The maximum likelihood approach is used for parameter estimation and the normality assumption can be tested using the likelihood ratio statistics since large sample properties are satisfied for the maximum likelihood estimator (Pewsey et al., 2012). Real data applications reveal that the model considered can be a viable alternative to existing asymmetric models in the literature.

The paper is organized as follows. In Section 2 asymmetric models are reviewed and some of their main properties discussed. Emphasis is placed on the alpha-power model, a special case of which is the power-normal model (Gupta and Gupta, 2008). In Section 3 a general definition of asymmetric regression models is presented and previous works on linear and nonlinear versions are listed. Section 4 is devoted to the study of the linear multiple regression model with power-normal errors. Inference via maximum likelihood for this model is also considered. The nonlinear power-normal model is considered in Section 5. Estimation is considered via maximum likelihood. The autoregressive model is studied in Section 6, with inference via maximum likelihood. A score type statistic is developed for testing null correlation. A small-scale Monte Carlo study is conducted in Section 7, including a study on model robustness. The main conclusion is that estimators under the regression model studied are fairly robust against data contamination. Results of two real data applications are reported illustrating the usefulness of the models considered in Section 8. In Section 9 (Appendix), we present the elements for the observed information matrices for the models considered in the previous Sections.

2. Skew distributions

Lehmann (1953) studied the family of distributions with a general distribution function given by

$$\mathscr{F}_F(z;\alpha) = \{F(z)\}^\alpha, \quad z \in \mathbb{R}, \tag{1}$$

where F is a distribution function and α is a rational number.

Durrans (1992), in a hydrological context, extended Lehmann's model by considering α real (and positive) for the special case $F = \Phi$, the distribution function of the normal distribution. We consider in this paper an extension of Lehmann's model, which we call the *alpha-power model*, with density function given by

$$\varphi_f(z;\alpha) = \alpha f(z)\{F(z)\}^{\alpha-1}, \qquad z \in \mathbb{R}, \ \ \alpha \in \mathbb{R}^+, \tag{2}$$

where F is an absolutely continuous distribution function with density function $f = dF$. Properties for a particular case of this distribution (with $F = \Phi$, the distribution function

of the normal distribution), were studied in Gupta and Gupta (2008). We use the notation $Z \sim P_F(\alpha)$. We refer to this model as the standard alpha-power distribution (see also Pewsey et al., 2012). This is an alternative to asymmetric models with higher amounts of asymmetry and kurtosis as is the case with the skew-normal distribution (Azzalini, 1985), see also Mudholkar and Hutson (2000) for some special cases. Parameter α is a shape parameter that controls the amount of asymmetry in the distribution. Extensions of the power-normal model are also considered in Rego et al. (2012).

In the particular case that $F = \Phi$, the distribution function of the normal distribution, Z is said to follow a power-normal distribution (denoted PN(α)) with density function given by

$$\varphi(z;\alpha) = \alpha\phi(z)\{\Phi(z)\}^{\alpha-1}, \qquad z \in \mathbb{R}. \tag{3}$$

If Z is a random variable from a standard $P_F(\alpha)$ distribution then the location-scale extension of Z, $X = \xi + \eta Z$, where $\xi \in \mathbb{R}$ and $\eta \in \mathbb{R}^+$, has probability density function given by

$$\varphi_F(x;\xi,\eta,\alpha) = \frac{\alpha}{\eta} f\left(\frac{x-\xi}{\eta}\right)\left\{F\left(\frac{x-\xi}{\eta}\right)\right\}^{\alpha-1}. \tag{4}$$

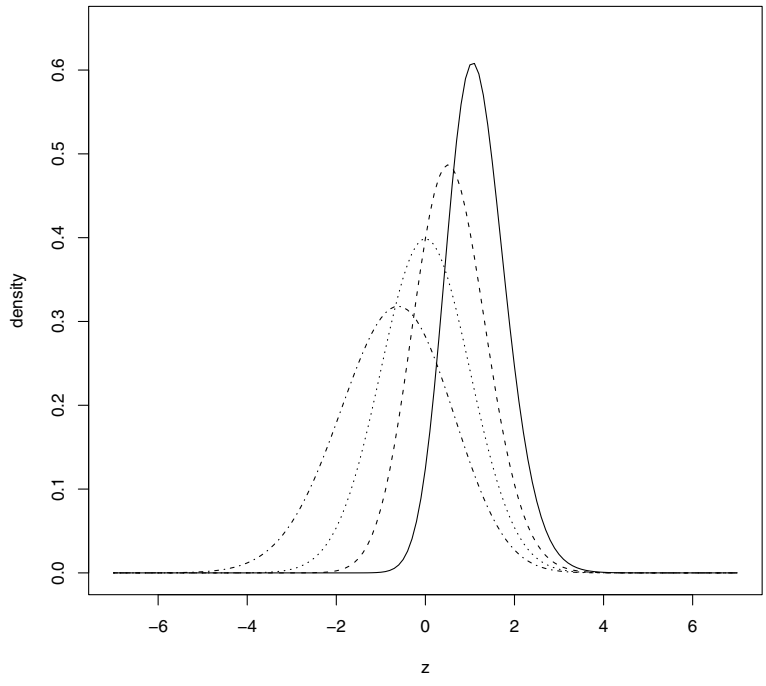


Figure 1: *Density $\varphi_Z(z;\alpha)$ for α equals to* 5 *(solid line),* 2 *(dashed line),* 1 *(dotted line) and* 0.5 *(dashed and dotted line).*

We will denote this extension by using the notation $X \sim P_F(\xi, \eta, \alpha)$. Notice that this model can be further extended by considering $\xi_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ replacing ξ, where $\boldsymbol{\beta}$ is an unknown vector of regression coefficients and x_i a vector of known regressors possibly correlated with the response vector.

As can be deduced from Figure 1, parameter α controls also the distribution kurtosis. Moreover, it can be noticed that for $\alpha > 1$, the kurtosis is greater than that of the normal distribution and, for $0 < \alpha < 1$, the opposite is observed.

Pewsey et al. (2012) derived the Fisher information matrix for the location-scale version of the power-normal model and have shown that it is not singular for $\alpha = 1$. We recall that the Fisher information matrix for the skew-normal distribution (Azzalini, 1985) is singular under the symmetry hypothesis. Hence, with the power-normal model, normality can be tested using ordinary large sample properties of the likelihood ratio statistics. They also found the asymmetry and kurtosis ranges to be $[-0.6115, 0.9007]$ and $[1.7170, 4.3556]$, respectively. This illustrates the fact that the model is more flexible, respective to kurtosis, than the model skew-normal (Azzalini, 1985), for which the kurtosis range is given by $[3, 3.8692)$.

A generalization for the PN(α) model is given in Eugene et al. (2002), by introducing the beta-normal distribution, denoted BN(α, β), with BN($\alpha, 1$)=PN(α). Therefore, model BN(α, β) is more flexible than model PN(α). However, model BN(α, β) contains two parameters to be estimated and the asymmetry and kurtosis ranges for both models are the same, namely $[-0.6115, 0.9007]$ and $[1.7170, 4.3556]$, respectively. General properties of the model BN where studied by Gupta and Nadarajah (2004) and Rego et al. (2012).

3. The asymmetric regression model

The multiple regression model is typically represented by

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{5}$$

where $\boldsymbol{\beta}$ is a vector of unknown constants and x_i are values of known explanatory variables. The error terms ϵ_i are independent random variables with N$(0, \sigma^2)$ distribution. It may occur that the symmetry assumption is not an adequate assumption for the error term so that an asymmetric model may present a better fit for the data set under study. As seen in the literature, some asymmetric distributions that can be considered are the epsilon-skew-normal (ESN, Mudholkar and Hutson, 2000) distribution, the skew-exponential power (SEP, see Azzalini, 1986) distribution and the Beta-Normal (BN) distribution, among others. Hutson (2004) replaces in (5) the normal distribution by the ESN distribution, DiCiccio and Monti (2004) consider that the error terms follow model SEP while Razzaghi (2009) consider the BN distribution for fitting a quadratic

dose-response modeling. Asymmetric nonlinear regression is studied in Cancho et al. (2008) by considering that the error terms follow a skew-normal model distribution. Xie et al. (2009) studied the case where the error term follows the skew-t-normal model (see Gómez et al., 2007).

4. The multiple regression model with PN errors

In this section, we assume under the ordinary multiple regression model that the error term follows a PN (denoted PNR) distribution with parameters 0, η_e and α, that is,

$$\epsilon_i \sim PN(0,\eta_e,\alpha) \quad \text{for} \quad i=1,2,\ldots,n.$$

Hence, it follows that the density function of ϵ_i is given by

$$\varphi(\epsilon_i;\boldsymbol{\beta},\eta_e,\alpha)=\frac{\alpha}{\eta_e}\phi\left(\frac{y_i-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\eta_e}\right)\left\{\Phi\left(\frac{y_i-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\eta_e}\right)\right\}^{\alpha-1}, \quad i=1,2,\ldots,n, \tag{6}$$

Therefore, it follows that y_i given $\mathbf{x}_i$, $(y_i|\mathbf{x}_i)$, also follows a PN distribution, that is,

$$y_i|\mathbf{x}_i \sim PN(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta},\eta_e,\alpha), \quad i=1,2,\ldots,n, \tag{7}$$

with location parameter $\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$, $i=1,2,\ldots,n$, scale parameter η_e and shape parameter α. Under the PN model,

$$E(\epsilon_i)=\alpha\eta_e\int_0^1\Phi^{-1}(z)z^{\alpha-1}dz\neq 0$$

so that the expected value of the error term is not null as is the case under normality. Therefore, $E(y_i)\neq\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ and we have to make the following correction to obtain the regression line as the expected value of the response variable: $\beta_0^*=\beta_0+\mu_\epsilon$, where $\mu_\epsilon=E(\epsilon_i)$. Thus,

$$E(y_i)=\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}^* \text{ where } \boldsymbol{\beta}^*=(\beta_0^*,\beta_1,\ldots,\beta_p)^{\mathsf{T}}.$$

The next section discusses maximum likelihood estimation for the corrected model.

4.1. Inference for the multiple PNR model

We discuss in the following maximum likelihood estimation for the multiple power-normal regression model. A detailed derivation of the Fisher information matrix is considered, resulting that it is nonsingular at the vicinity of symmetry.

4.2. Likelihood and score functions

Considering now a matrix notation where $\mathbf{y}$ denotes the vector with entries y_i and dimension $(n\times 1)$ and $\mathbf{X}$ the $(n\times(p+1))$-matrix with rows $\mathbf{x}_i^\top$, the likelihood function for $\boldsymbol{\theta}=(\boldsymbol{\beta}^\top,\eta_e,\alpha)^\top$ given a random sample of size n, $\mathbf{y}=(y_1,y_2,\ldots,y_n)^\top$, can be written as

$$\ell(\boldsymbol{\theta};\mathbf{y})=n\log\left(\frac{\alpha}{\eta_e}\right)-\frac{1}{2\eta_e^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+(\alpha-1)\sum_{i=1}^{n}\log\left\{\Phi\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)\right\},$$

with score function:

$$U(\boldsymbol{\beta})=\frac{1}{\eta_e^2}\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\frac{\alpha-1}{\eta_e}\mathbf{X}^\top\Lambda_1,\qquad U(\alpha)=n\left(\frac{1}{\alpha}+\overline{u}\right),\tag{8}$$

$$U(\eta_e)=-\frac{n}{\eta_e}+\frac{1}{\eta_e^3}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\frac{\alpha-1}{\eta_e^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\Lambda_1\tag{9}$$

where

$$\Lambda_1=(w_1,\ldots,w_n)^\top\qquad\text{and}\qquad u_i=\log\left\{\Phi\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)\right\},$$

with $w_i=\phi\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)/\Phi\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)$, for $i=1,2,\ldots,n$. After some algebraic manipulations, maximum likelihood estimating equations are given by

$$\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{MQ}-(\alpha-1)\eta(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Lambda_1,\qquad\alpha=-\frac{1}{\overline{u}},\tag{10}$$

$$\eta=\frac{(1-\alpha)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\Lambda_1}{2n}+\frac{\sqrt{(1-\alpha)^2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\Lambda_1\Lambda_1^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+4n(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}}{2n},\tag{11}$$

where $\hat{\boldsymbol{\beta}}_{MQ}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

Hence, the maximum likelihood estimator for the parameter vector $\boldsymbol{\beta}$ is equal to the least squares estimator for $\boldsymbol{\beta}$ plus the symmetry correcting term. No analytical solutions are available for the likelihood equations and hence they have to be solved numerically.

For the simple linear regression model, namely $p=1$, the following system of equations results

$$\beta_1=\eta_e(\alpha-1)\frac{S_{xw}}{S_x^2}+\frac{S_{xy}}{S_x^2},\qquad\beta_0=-\eta_e(\alpha-1)\overline{w}+\overline{y}-\beta_1\overline{x},\qquad\alpha=-\frac{1}{\overline{u}},$$

and

$$\eta_e = \frac{(1-\alpha)(\overline{wy}-\beta_0\overline{w}-\beta_1\overline{wx})}{2n} + \frac{\sqrt{(1-\alpha)^2(\overline{wy}-\beta_0\overline{w}-\beta_1\overline{wx})^2+4n\sum_{i=1}^n(y_i-\beta_0-\beta_1x_i)^2}}{2n}$$

with

$$S_x^2=\sum_{i=1}^n(x_i-\overline{x})^2/n,\ \ S_{xy}=\sum_{i=1}^n(x_i-\overline{x})(y_i-\overline{y})/n \ \text{ and } \ S_{wx}=\sum_{i=1}^n(w_i-\overline{w})(x_i-\overline{x})/n,$$

where $\overline{w}=\sum_{i=1}^n w_i/n,\ \overline{u}=\sum_{i=1}^n u_i/n,\ \overline{x}=\sum_{i=1}^n x_i/n,\ \overline{x^2}=\sum_{i=1}^n x_i^2/n,\ \overline{y}=\sum_{i=1}^n y_i/n,\ \overline{xy}=\sum_{i=1}^n x_iy_i/n,\ \overline{wx}=\sum_{i=1}^n w_ix_i/n$ and $\overline{wy}=\sum_{i=1}^n w_iy_i/n$.

For $\alpha=1$, the model with normal error terms follow and the estimators reduce to the well known $\hat{\beta}_1=\frac{S_{xy}}{S_x^2},\ \hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$ and $\hat{\eta}_e=\sqrt{\frac{1}{n}\sum_{i=1}^n(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}$. To initialize the likelihood approach, we can take as initial values the vector $\hat{\boldsymbol{\beta}}$ and for parameter η_e the ones obtained by the least squares approach. They can be computed as follows: for $\epsilon_i^*=\epsilon_i-\mu_\epsilon$, we have that $E(\epsilon^*)=0$ and $Var(\epsilon^*)=\eta_e^2\Phi_2(\alpha)$, where Φ_2 is the variance of the random variable PN$(0,1,\alpha)$.

Hence, after minimizing the error sum of squares, namely,

$$\sum_{i=1}^n\epsilon_i^{*2}=\sum_{i=1}^n\left(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}^*\right)^2$$

we obtain the least squares estimators of $\boldsymbol{\beta}^*$ and η_e, which are given by:

$$\hat{\boldsymbol{\beta}}^*=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \ \text{ and } \ \hat{\eta}_e^2=\frac{\Phi_2^{-1}(\hat{\alpha})}{n-2}\sum_{i=1}^n\left(y_i-\hat{\beta}_0^*-\hat{\beta}_1x_i\right)^2.$$

On the other hand, an initial value for α can be obtained by fitting the PN model for the errors obtained or by using the elemental percentile approach of Castillo and Hadi (1995), assuming β and η_e known (usually computed using the least-squares approach).

The elements needed to compute the observed information matrix are given in Appendix 10.1. The expected (Fisher) information matrix follows then by taking expectations of those components (multiplied by n^{-1}).

Approximation $N_{p+3}(\boldsymbol{\theta},\Sigma_{\hat{\boldsymbol{\theta}}})$ can be used to construct confidence intervals for θ_r, which are given by $\hat{\boldsymbol{\theta}}_r\mp z_{1-\delta/2}\sqrt{\hat{\sigma}(\hat{\boldsymbol{\theta}}_r)}$, where $\hat{\sigma}(\cdot)$ corresponds to the r-th diagonal element of the matrix $\Sigma_{\hat{\boldsymbol{\theta}}}$ and $z_{1-\delta/2}$ denotes $100(\delta/2)$-quantile of the standard normal distribution.

For the simple linear regression model, that is, $p = 1$, denoting the elements of the observed information matrix by

$$i_{\beta_0\beta_0},\ i_{\beta_1\beta_0},\ i_{\eta_e\beta_0},\ i_{\alpha\beta_0},\ldots,i_{\eta_e\eta_e},\ i_{\alpha\alpha},$$

and making $a_{jk} = E\left(W^j Y^k\right)$ for $k = 0,1,2,3$ and $j = 0,1,2$, we obtain the expected information matrix, the elements of which are given in the appendix.

5. The alpha-power nonlinear regression model

A more general model can be defined replacing the linearity assumption by a nonlinear one. Therefore, we define the nonlinear alpha-power model as

$$y_i = f(\boldsymbol{\beta},\mathbf{x}_i) + \epsilon_i, \quad i = 1,2,\ldots,n,$$

where y_i is the response variable, f is an injective continuous and twice differentiable function with respect to the parameter $\boldsymbol{\beta}$, $\mathbf{x}_i$ is an explanatory variable vector and ϵ_i are independent and identically distributed $P_F(0,\eta,\alpha)$ random variables with

$$\mu_\epsilon = \alpha\eta\int_0^1 z^{\alpha-1}F^{-1}(z)dz.$$

As in the linear case, $E(Y_i) = f(\boldsymbol{\beta},\mathbf{x}_i) + \mu_\epsilon$, so that corrections are required so that the error term is unbiased for zero, that is,

$$y_i \sim P_F(f(\boldsymbol{\beta},\mathbf{x}_i),\eta,\alpha).$$

In the PN situation we have the density function

$$\varphi(y_i;\boldsymbol{\beta},\alpha) = \frac{\alpha}{\eta}\phi\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)\left\{\Phi\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)\right\}^{\alpha-1}. \tag{12}$$

which we denote by $y_i|\mathbf{x}_i \sim \text{PN}(f(\boldsymbol{\beta},\mathbf{x}_i),\eta,\alpha)$. The log-likelihood function (disregarding constants) for the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top,\eta,\alpha)^\top$ for a random sample of size n from y_i with distribution $\text{PN}(f(\boldsymbol{\beta},\mathbf{x}_i),\eta,\alpha)$, is given by

$$\ell(\boldsymbol{\theta};\mathbf{X},\mathbf{y}) = n\log\left(\frac{\alpha}{\eta}\right) - \frac{1}{2\eta^2}\sum_{i=1}^n (y_i - f(\boldsymbol{\beta},\mathbf{x}_i))^2 + (\alpha-1)\sum_{i=1}^n \log\left\{\Phi\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)\right\}.$$

The score function $U(\boldsymbol{\theta}) = (U(\boldsymbol{\beta}), U(\eta), U(\alpha))^{\mathsf{T}}$ is given by

$$U(\beta_i) = \frac{1}{\eta^2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{\beta}, x_i))\frac{\partial f(\boldsymbol{\beta}, x_i)}{\partial \beta_i} - \frac{\alpha - 1}{\eta}\sum_{i=1}^{n} w_i \frac{\partial f(\boldsymbol{\beta}, x_i)}{\partial \beta_i}, \quad U(\alpha) = n\left(\frac{1}{\alpha} + \overline{u}\right),$$

$$U(\eta) = -\frac{n}{\eta} + \frac{1}{\eta^3}\sum_{i=1}^{n}(y_i - f(\boldsymbol{\beta}, x_i))^2 - \frac{\alpha - 1}{\eta^2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{\beta}, x_i))w_i,$$

with $u_i = \Phi\left(\frac{y_i - f(\boldsymbol{\beta}, x_i)}{\eta}\right)$ and $w_i = \frac{\phi\left(\frac{y_i - f(\boldsymbol{\beta}, x_i)}{\eta}\right)}{\Phi\left(\frac{y_i - f(\boldsymbol{\beta}, x_i)}{\eta}\right)}$.

Differentiating the scores above, we arrive at the observed information matrix, see appendix. Hence, the maximum likelihood estimator for θ, can be obtained by implementing the following Newton-Raphson type iterative procedure:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}(k) + [J(\hat{\boldsymbol{\theta}}^{(k)})]^{-1} U(\hat{\boldsymbol{\theta}}^{(k)}), \tag{13}$$

where $J(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}}$.

6. Nonlinear autoregressive alpha-power-normal model

We consider now the extension of the nonlinear-normal model with autoregressive errors to the PN distribution. Hence, the stochastic representation for the nonlinear PN model with autoregressive errors is given by

$$y_i = f(\boldsymbol{\beta}, \mathbf{x}_i) + \epsilon_i, \quad \text{with} \quad \epsilon_i = \rho\epsilon_{i-1} + a_i, \quad i = 1, 2, \cdots, n, \tag{14}$$

where y_i, $i = 1, \ldots, n$ are the observed responses, the $\mathbf{x}_i$, $i = 1, \ldots n$ are known covariate vectors with ρ as the autoregressive coefficient satisfying $|\rho| < 1$; $\boldsymbol{\beta}$ is an unknown p-dimensional vector of real parameters, f is a known continuous and twice differentiable function with respect to $\boldsymbol{\beta}$, a_i are independent random variables with $a_i \sim \text{PN}(0, \eta^2, \alpha)$ and $\epsilon_0 = 0$.

It then follows that the expectation of the random response is

$$E(Y_i) = f(\boldsymbol{\beta}, \mathbf{x}_i) + E(a_i)\sum_{k=0}^{i-1}\rho^k, \tag{15}$$

$i = 1,\ldots,n$, where $E(a_i)$ is the expectation of a random variable with PN$(0,\eta^2,\alpha)$ distribution.

6.1. Maximum likelihood estimation

Given a random sample of size n from the above model, the log-likelihood function for parameter vector $\boldsymbol{\theta} = (\rho, \boldsymbol{\beta}^\top, \eta^2, \alpha)^\top$, can be written as

$$\ell_n(\boldsymbol{\theta};\mathbf{y}) = n\left\{\log(\alpha) - \log(\eta) - \frac{1}{2}\log(2\pi)\right\} - \sum_{i=1}^{n}\frac{(\epsilon_i - \rho\epsilon_{i-1})^2}{2\eta^2} + (\alpha-1)\sum_{i=1}^{n}\log\{\Phi(z_i)\},$$

with $z_i = \frac{\epsilon_i - \rho\epsilon_{i-1}}{\eta}$. Therefore, for $w_i = \frac{\phi(z_i)}{\Phi(z_i)}$, $D_i = -\frac{\partial f(\boldsymbol{\beta},x_i)}{\partial \boldsymbol{\beta}} + \rho\frac{\partial f(\boldsymbol{\beta},x_{i-1})}{\partial \boldsymbol{\beta}}$ and $Q_i = -w_i(z_i + w_i)$, $i = 1,2,\ldots,n$, the score function $U_{\boldsymbol{\theta}} = (U_\rho, U_{\boldsymbol{\beta}}^\top, U_{\eta^2}, U_\alpha)^\top$ has elements:

$$U(\rho) = \sum_{i=1}^{n}\left[\frac{a_i}{\eta^2} - \frac{\alpha-1}{\eta}w_i\right]\epsilon_{i-1}, \qquad U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[\frac{a_i}{\eta^2} - \frac{\alpha-1}{\eta}w_i\right]D_i,$$

$$U(\eta^2) = \sum_{i=1}^{n}\left[-\frac{1}{2\eta^2} + \frac{a_i^4}{2\eta^4} - \frac{\alpha-1}{2\eta^3}a_i w_i\right], \qquad U(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n}\log\{\Phi(z_i)\},$$

where $a_i = \epsilon_i - \rho\epsilon_{i-1}$ and $\epsilon_i = y_i - f(\boldsymbol{\beta},x_i)$. Hence, taking $G_i = -\frac{\partial^2 f(\boldsymbol{\beta},x_i)}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top} + \rho\frac{\partial^2 f(\boldsymbol{\beta},x_{i-1})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}$, we obtain the Hessian matrix, see Appendix, from which the expected information can be obtained.

Therefore, the maximum likelihood estimators can be obtained by iteratively solving the equation:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} + [J(\hat{\boldsymbol{\theta}}^{(k)})]^{-1}U(\hat{\boldsymbol{\theta}}^{(k)}), \tag{16}$$

where $J(\boldsymbol{\theta}) = -\dfrac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}$.

6.2. Score statistics for testing ρ

In the particular case where $\rho = 0$, the autoregressive model (14) reduces to the nonlinear PN regression model. Hence, it is important to verify whether this is the case or not. Considering $\boldsymbol{\beta}$, η^2 and α as nuisance parameters, we want to test the hypotheses

$$H_0 : \rho = 0 \ \text{ versus } \ H_1 : \rho \neq 0.$$

It can be shown that the score statistics (Cox and Hinkley, 1974) for testing H_0 is given by :

$$\mathrm{SC}_1 = [U_\rho^2 J^{\rho\rho}(\theta)]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0}, \tag{17}$$

where $J^{\rho\rho}$ is the block of J^{-1} corresponding to ρ and $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. Under H_0, statistics (17) follows, asymptotically the chi-square distribution (χ_1^2) with one degree of freedom.

7. Simulation study

We report next results of a simulation study designed at investigating the performance of the maximum likelihood estimators for parameters β_0, β_1 and η_e. We simulated 1000 samples of sizes $n = 50$, 75 and 100. Without loss of generality we took $\eta_e = 1$. Values for X were generated from the $U(0,1)$, the uniform distribution on the $(0,1)$ interval and $p = 1$, with $\beta_0 = 1.5$ and $\beta_1 = -2.5$. Moreover, we took $\epsilon_i \sim \mathrm{PN}(0, \eta_e, \alpha)$. Estimators performance were evaluated by computing the relative empirical bias (RB = empirical bias/parameter value) and the square root of the empirical mean squared error ($\sqrt{\mathrm{MSE}}$) and the covering probability of the 95% large sample intervals (discussed above) or, equivalently, the rejection rate for testing $\beta_1 = 0$ at the 5% significance level. This study was implemented using software R.

Table 1: *Empirical RB and $\sqrt{MSE}$ for the simple PNR model.*

		$\alpha = 0.75$			$\alpha = 1.50$			$\alpha = 2.25$		
n	$\hat{\theta}$	RB(%)	$\sqrt{\mathrm{MSE}}$	$1-\delta$	RB(%)	$\sqrt{\mathrm{MSE}}$	$1-\delta$	RB(%)	$\sqrt{\mathrm{MSE}}$	$1-\delta$
	$\hat{\beta}_0$	7.96	1.25	0.66	4.06	1.26	0.70	5.93	1.32	0.74
	$\hat{\beta}_1$	0.21	0.16	0.77	0.13	0.13	0.79	0.11	0.11	0.83
50	$\hat{\eta}$	15.58	0.52	0.85	12.08	0.46	0.83	12.05	0.46	0.84
	$\hat{\alpha}$	58.81	1.36	0.62	84.57	3.22	0.67	101.80	5.51	0.70
	$\hat{\beta}_0$	5.07	1.11	0.66	2.19	1.12	0.76	3.89	1.12	0.82
	$\hat{\beta}_1$	0.16	0.13	0.83	0.12	0.11	0.87	0.10	0.09	0.90
75	$\hat{\eta}$	10.85	0.44	0.82	5.53	0.38	0.83	4.29	0.36	0.88
	$\hat{\alpha}$	46.53	1.17	0.65	76.44	2.86	0.72	92.73	4.84	0.79
	$\hat{\beta}_0$	1.10	0.47	0.80	1.01	0.51	0.92	2.64	0.53	0.93
	$\hat{\beta}_1$	0.06	0.04	0.94	0.02	0.04	0.94	0.04	0.03	0.95
500	$\hat{\eta}$	0.56	0.16	0.80	0.46	0.16	0.92	0.47	0.15	0.93
	$\hat{\alpha}$	12.08	0.47	0.78	16.09	1.13	0.88	21.51	1.88	0.88

Table 2: *Empirical RB and $\sqrt{MSE}$ for simple regression model with contaminated model.*

		$\alpha = 0.75$			$\alpha = 1.50$			$\alpha = 2.25$		
n	$\hat{\theta}$	RB(%)	$\sqrt{\text{MSE}}$	$1-\delta$	RB(%)	$\sqrt{\text{MSE}}$	$1-\delta$	RB(%)	$\sqrt{\text{MSE}}$	$1-\delta$
	$\hat{\beta}_0$	6.61	1.22	0.63	8.74	1.28	0.68	5.98	1.28	0.71
	$\hat{\beta}_1$	0.24	0.16	0.75	0.21	0.13	0.80	0.15	0.12	0.83
50	$\hat{\eta}$	15.12	0.51	0.85	13.69	0.48	0.81	11.06	0.44	0.84
	$\hat{\alpha}$	56.72	1.32	0.59	75.31	3.12	0.64	92.29	5.29	0.67
	$\hat{\beta}_0$	5.82	1.12	0.68	7.11	1.14	0.73	3.77	1.13	0.76
	$\hat{\beta}_1$	0.16	0.13	0.82	0.16	0.10	0.86	0.18	0.08	0.85
75	$\hat{\eta}$	10.94	0.44	0.83	9.47	0.40	0.82	7.49	0.37	0.85
	$\hat{\alpha}$	47.76	1.22	0.65	60.14	2.72	0.70	70.91	4.39	0.72
	$\hat{\beta}_0$	1.66	0.49	0.80	2.20	0.53	0.92	3.68	0.54	0.93
	$\hat{\beta}_1$	0.06	0.04	0.93	0.06	0.04	0.94	0.04	0.03	0.94
500	$\hat{\eta}$	0.46	0.16	0.80	0.13	0.15	0.92	0.54	0.15	0.94
	$\hat{\alpha}$	13.70	0.50	0.78	19.67	1.22	0.89	24.79	2.01	0.89

Results in Table 1 show that the relative bias and $\sqrt{\text{MSE}}$ for the maximum likelihood estimators for parameters η_e, β_0 and β_1 decreases as the sample sizes increase which is expected. It can also be noted that the relative bias can be large in small and moderate sample sizes situations. As parameter α increases, relative bias also increases for parameters η_e and $\hat{\beta}_0$ which is also expected. Relative bias for $\hat{\beta}_1$ is below 1.5%. To reduce bias for β_0 procedures such as bootstrap and jacknife could be implemented.

We also developed a simulation study designed at evaluating the robustness of the estimation procedure under the PN regression model obtained by contaminating the error terms with a skew-normal random variable. It was considered that the first observation was generated according to the distribution $\text{SN}(0,1,-1) + \text{PN}(0,1,\alpha)$. Maximum likelihood estimators were than computed for each generated sample, as described above and Table 1 presents the results. It can be deduced from the table that empirical RB and $\sqrt{\text{MSE}}$ does not seem affected by changes in the model generating the data.

8. Numerical illustrations

8.1. Linear model

The following illustration is based on the Australian athletes data set available for downloading at the directory http://azzalini.stat.unipd.it/SN/. The linear model considered is

$$Bfat_i = \beta_0 + \beta_1 Wt_i + \beta_2 sex_i + \epsilon_i, \;\; i = 1,2,\ldots,202,$$

where $Bfat_i$ is the body fat percentage for the i-th athlete, and covariates Wt_i and sex_i the weight and sex, respectively, for the i-th athlete; variable sex is dichotomized with 1 for male and zero for female. A residual analysis has indicated that symmetric models may not be the most adequate ones and that an asymmetric model can present a better fit, see Table 3, where quantities $\sqrt{b_1}$ and b_2 indicate sample asymmetry and kurtosis coefficients.

Table 3: *Summary statistics for estimated residuals under normality.*

n	Mean	Variance	$\sqrt{b_1}$	b_2
202	0.0050	11.8431	0.6030	3.9321

We fitted linear regression models under the assumption that model errors follow an asymmetric distribution, namely the the skew-normal (SNR), the skew-t$_\nu$ (StR), the student con ν degrees of freedom and power-normal (PN) distributions. For estimating under skew-normal and skew-Student-t R Development Core Team (2014) package is used, which uses the centred parametrization (CP), namely $E(Y) = \mathbf{x}^\mathsf{T}\boldsymbol{\beta}$ and $\mathrm{Var}(Y) = \eta^2$ (see Chiogna (2005) and Pewsey (2000)), whereas for model PN we use the optim program in the R package.

We use the AIC (Akaike, 1974), written as $\mathrm{AIC} = -2\hat{\ell}(\cdot) + 2k$ and BIC, written as $\mathrm{BIC} = -2\hat{\ell}(\cdot) + (\log(n))k$, where k is the number of unknown parameters, for comparing the normal and power-normal which are nested models. The best model is the one with the smallest AIC or BIC.

Moreover, the results in Table 4 present estimates for model parameters. It also reveals that, according to the PN regression model, % of body fat depends on weight and sex of the athlete. Estimating β_0^* in the PN regression model leads to $\hat{\beta}_0^* = 0.39$.

Table 4: *Estimates (standard error) for normal and PN linear models.*

Parameters	Normal model	SNC model	St$_{14}$ model	PN model
β_0	1.62 (1.43)	2.91 (1.34)	−0.52 (1.35)	−5.97(2.00)
β_1	0.24 (0.02)	0.21 (0.02)	0.21 (0.02)	0.24 (0.02)
β_2	−12.25 (0.57)	−11.10 (0.71)	−11.09 (0.68)	−11.25 (0.60)
η	3.43 (0.17)	3.43 (0.18)	4.47 (0.75)	5.29 (0.48)
α		0.57 (0.14)	2.07 (0.50)	5.38 (1.83)

The model $\epsilon_i \sim \mathrm{PN}(0, 5.29, 5.38)$ seems to present a good fit for the data set under study. A more emphatic justification for using a PN type model comes from testing the normality assumption, that is, the hypotheses

$$H_0 : \alpha = 1 \ \text{ versus } \ H_1 : \alpha \neq 1,$$

by using the likelihood ratio statistics,

$$\Lambda = \frac{\ell_{\mathrm{N}}(\widehat{\boldsymbol{\theta}})}{\ell_{\mathrm{PN}}(\widehat{\boldsymbol{\theta}})},$$

which, for the data set under study, leads to $-2\log(\Lambda) = 4.97$, so that p-value $=$ Prob$(\chi^2_1 > 4.97) < 0.05$. with strong indication against the null hypothesis.

Computing AIC and BIC for normal and PN regression models lead to AIC $=$ 1079.54 and BIC $=$ 1092.77 and AIC $=$ 1076.56 and BIC $=$ 1093.10, respectively. According to the values obtained for AIC and BIC, the power-normal (PN) linear regression model presents the better fit when compared with normal linear model.

We use Voung (1989) approach (generalized LR statistic) for comparing the skew-normal (SNR), skew-Student-t (StR) and power-normal (PNR) linear non-nested models fitted to the data. A description of the procedure is described next. Being F_θ and G_ζ two non-nested models and $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ and $g(y_i|\mathbf{x}_i,\boldsymbol{\zeta})$ the corresponding densities, the likelihood ratio statistics to compare both models is given by

$$\mathrm{LR}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\zeta}}) = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{f(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\zeta}})} \right\},$$

which does not follow a chi-square distribution. To overcome this problem, Vuong (1989) proposed an alternative approach based on the Kullback-Liebler divergence criterion. Based on the divergence between each model and the true process generating the data, namely the model $h^0(y|x)$, one arrives at the statistics

$$T_{\mathrm{LR,NN}} = \frac{1}{\sqrt{n}} \frac{\mathrm{LR}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\zeta}})}{\widehat{w}}, \tag{18}$$

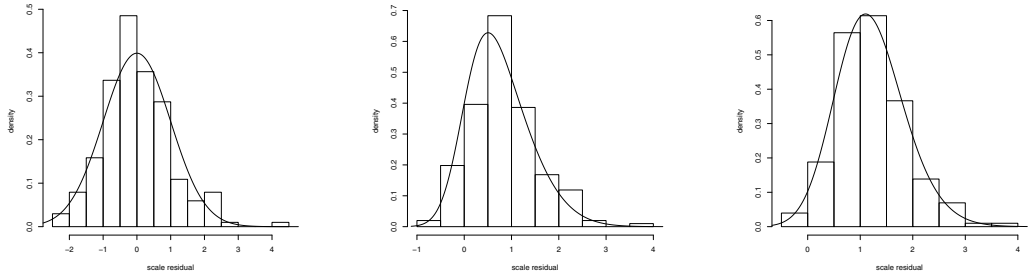
where

$$\widehat{w}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\zeta}})} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i,\widehat{\boldsymbol{\zeta}})} \right)^2.$$

For strictly non-nested models, it can be shown that the statistic $T_{\mathrm{LR,NN}}$ converges in distribution to a standard normal distribution under the null hypothesis. Thus, the null hypothesis is not rejected if $|T_{\mathrm{LR,NN}}| \leq z_{p/2}$. On the other hand, we reject at significance level p the null hypothesis of equivalence of the models in favor of model F_θ being better (or worse) than model G_ζ if $T_{\mathrm{LR,NN}} > z_p$ (or $T_{\mathrm{LR,NN}} < -z_p$).

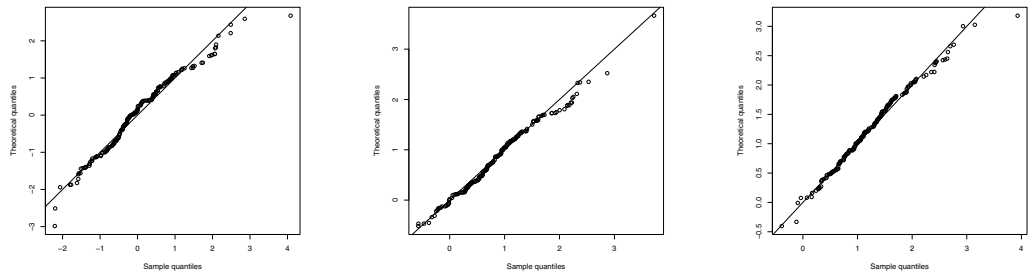
For testing PNR versus SNR, we obtain $T_{\mathrm{LR,NN}} = 22.59$ (p-value < 0.05) and for the PNR versus RSt_{14} model, $T_{\mathrm{LR,NN}} = 0.61$ (p-value > 0.05). Therefore, the PNR model is significantly better than the SNR model according to the generalized LR statistic. In a similar fashion it can be concludes that there is no significant difference between models PNR and RSt_{14}. However, favouring model PNR we have the fact that it involves one less parameter. Authors Lange et al. (1989), Berkane et al. (1994), Fernández and Steel (1999), Taylor and Verbyla (2004) and Leiva et al. (2008), all reported dificulties in estimating the degrees of freedom parameter.

We also computed the scaled residuals $e_i = (y_i - \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}})/\hat{\eta}$ to investigate model fit. Figures 2-(a), (b) and (c) and 3-(a), (b) and (c) depict the histograms and Q-Q plots for the scaled residuals under normal, SNR and PNR models, which also indicate a good fit for the PNR model.



(a) (b) (c)

Figure 2: *Graphs for residuals, of the fitted models. (a) Normal, (b) SN and (c) PN.*



(a) (b) (c)

Figure 3: *Q-qplots for the scaled residuals Z, from the fitted models. (a) Normal, (b) SN and (c) PN.*

8.2. Nonlinear model with correlated errors

In the following we present an application of the PN model fitting to the palm oil data set presented in Foong (1999) and studied in Xie et al. (2009) using a skew-normal nonlinear model. This data set was previously analysed in Azme et al. (2005), were parameter estimates are obtained under nonlinear growth curve models using Marquardat's iterative procedure. They found that the best fit is presented by the logistic growth curve model (see, Ratkowsky, 1983), followed by the Gompertz model, which was followed by the Morgan-Mercer-Flodin, Chapman-Richard model. Cancho et al. (2008) also analysed the model using a nonlinear skew-normal model with logistic growth. We focus now on analyzing the data set under a PN nonlinear regression model with logistic growth. Therefore, the model considered can be written as

$$y_i = \frac{\beta_1}{1+\beta_2 \exp(-\beta_3 x_i)} + \epsilon_i \tag{19}$$

with $\epsilon_i = \rho\epsilon_{i-1} + a_i$, $a_i \sim PN(0,\eta^2,\alpha)$, $i = 1,\ldots,n$.

We are now implementing the correlated nonlinear normal model with normally distributed errors (NLCM) and the correlated nonlinear model with errors PN distributed (NLCPN). As Table 5 reveals, according to both criteria (AIC and BIC), the nonlinear PN model with correlated errors fits the data better.

Table 5: *AIC and BIC for the oil palm data.*

Statistics	Log-likelihood	AIC	BIC
Normal	−41.2656	92.5312	97.2534
PN	−39.1004	90.2008	95.8674

Table 6: *Parameter estimates (standard errors) for the following models: NLCN, NLCPN and NLPN.*

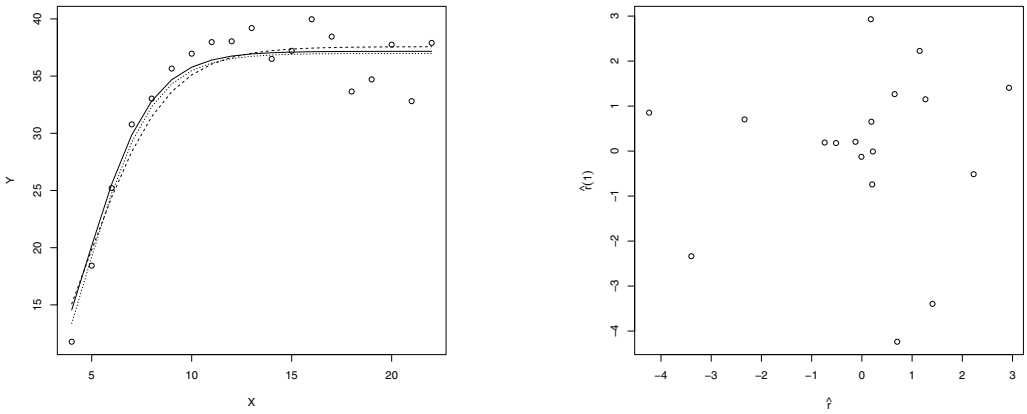
	NLCN	NLCPN	NLPN
Parameter	estimate	estimate	estimate
ρ	0.3222(0.2757)	0.2574(0.2114)	—
β_1	37.5699(0.3038)	37.9163(0.4041)	38.8798(0.2485)
β_2	11.4310(0.8327)	17.5880(1.2504)	17.5833(1.7888)
β_3	0.5092(0.0227)	0.6140(0.0135)	0.6079(0.0172)
η^2	5.5559(0.7392)	2.6815(0.3658)	1.2010(0.1550)
α	—	0.7010(0.1564)	0.2547(0.0589)

We consider now testing the hypotheses $H_0 : \alpha = 1$ versus $H_1 : \alpha \neq 1$, that is, a nonlinear normal model with correlated errors against a nonlinear PN model with correlated errors. The likelihood ratio statistics for testing the above hypotheses, namely, $\Lambda = \frac{\ell_{\text{NLCN}}(\hat{\boldsymbol{\theta}})}{\ell_{\text{NLCPN}}(\hat{\boldsymbol{\theta}})}$, leads to $-2\log(\Lambda) = 4.3304$, a value greater than the corresponding 5%

chi-square critical values which is $\chi^2_{1,5\%} = 3.8414$. Hence there is strong evidence that the nonlinear PN model with correlated errors fit the oil palm data set far better than the corresponding normal one.

Parameter estimates and standard errors for models NLPN, NLCN and NLCPN are presented in Table 6.

Figure 4(a), presents the nonlinear fitted models graphs and (b), and the fitted residuals for model PN, $\hat{r}_i$ against $\hat{r}_{i-1} = \hat{r}(1)$, under the assumption that $\rho = 0$; which does not reveal presence of correlation. Therefore, we implement a nonlinear model with errors PN$(0, \eta, \alpha)$, (NLPN) for which parameter estimates are given in Table 6.



(a) (b)

Figure 4: *(a) Graphs for fitted models, NLCN (dashed line), NLCPN (solid line) and NLPN (mixed (dashed-dotted) line); (b) graph for $\hat{r}_i$ against $\hat{r}_{i-1}$.*

9. Final discussion

In this paper we extended the power models in Pewsey et al. (2012) for the case of regression models. Linear models were considered as well as a non-linear extension. Emphasis was placed on the PN regression model situation. Estimation was performed by implementing the maximum likelihood approach. Large sample point and interval estimates were obtained by using the observed information (minus the inverse of the Hessian matrix evaluated at the maximum likelihood estimates). The exact Fisher information matrix is also derived and shown to be non-singular, so that large sample distribution for the alternative likelihood ratio statistics is central chisquare. For some comparisons, models are not nested so that an appropriate statistics with limiting normal distribution is considered.

The methodology implemented presented satisfactory results when applied to real data sets. Results of a small scale simulation indicate that the estimation approach leads

to good parameter recovery and that for large sample sizes bias and mean square error are significantly reduced. One of the applications is to a linear model applied to the Australian athletes data set (available for downloading from the internet) previously analysed by several other authors. It was seen that data present moderate to large skewness so that the PN regression model can be a viable alternative. The second data set that was analysed is the palm oil data set previously analysed by several authors. It turned out that the non-linear model with PN errors fitted the data better than the ordinary normal model.

10. Appendix

In this section we present in closed form the elements of the observed and expected (Fisher) information matrices for the PNR type models considered in this. Their derivation (requiring extensive algebraic manipulations) extends results in Pewsey et al. (2012). The relevance of the results rely on the fact one can conclude they are nonsingular so that large sample properties of the maximum likelihood estimators hold for such models. A similar discussion for skew-normal type models is considered in Azzalini (2013).

10.1. Observed information matrix for the PNR model

In this section we present the observed information matrix for the general PNR model.

$$j_{\boldsymbol{\beta}^\top\boldsymbol{\beta}} = \frac{1}{\eta_e^2}\mathbf{X}^\top\mathbf{X} + \frac{\alpha-1}{\eta_e^2}\mathbf{X}^\top\Lambda_2\mathbf{X}, \qquad j_{\eta_e\boldsymbol{\beta}} = \frac{2}{\eta_e^3}\mathbf{X}^\top(y-\mathbf{X}\boldsymbol{\beta}) + \frac{\alpha-1}{\eta_e^2}\mathbf{X}^\top\Lambda_3,$$

$$j_{\eta_e\eta_e} = -\frac{n}{\eta_e^2} + \frac{3}{\eta_e^2}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)^2 - \frac{2(\alpha-1)}{\eta_e^2}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)w_i$$
$$+ \frac{\alpha-1}{\eta_e^2}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)^3 w_i + \frac{\alpha-1}{\eta_e^2}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)^2 w_i^2$$

$$j_{\alpha\boldsymbol{\beta}} = \frac{1}{\eta_e}\mathbf{X}^\top\Lambda_1, \qquad j_{\alpha\eta_e} = \frac{1}{\eta_e^2}(y-\mathbf{X}\boldsymbol{\beta})^\top\Lambda_1, \qquad j_{\alpha\alpha} = n/\alpha^2,$$

where

$$\Lambda_2 = \operatorname{diag}\left\{\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\eta_e}\right)w_i + w_i^2\right\}_{i=1,2,\ldots,n}$$

and $\Lambda_3 = (a_1, a_2, \ldots, a_n)^\top$ with

$$a_i = \left\{ \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\eta_e} \right)^2 w_i + \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\eta_e} \right) w_i^2 - w_i \right\}_{i=1,2,\ldots,n}.$$

10.2. Information matrix for the simple PNR model

The elements of the FIM for the case $p = 1$ are given by

$$i_{\beta_0\beta_0} = \left\{ 1 + \frac{\alpha - 1}{\eta_e} [a_{11} - a_{10}(\beta_0 + \beta_1 \overline{x})] + (\alpha - 1) a_{20} \right\} / \eta_e^2,$$

$$i_{\beta_1\beta_0} = \left\{ \overline{x} + \frac{\alpha - 1}{\eta_e} \left[\overline{x}(a_{11} - \beta_0 a_{10}) - \beta_1 a_{10} \overline{x^2} \right] + (\alpha - 1) a_{20} \overline{x} \right\} / \eta_e^2,$$

$$\begin{aligned} i_{\eta_e\beta_0} = {} & \frac{1 - \alpha}{\eta_e^2} a_{10} + \frac{1}{\eta_e^3} [2a_{01} + (\alpha - 1)a_{21} - (2 + (\alpha - 1)a_{20})(\beta_0 + \beta_1 \overline{x})] + \\ & + \frac{\alpha - 1}{\eta_e^4} \left\{ a_{12} + a_{10}(\beta_0^2 + \beta_1^2 \overline{x^2} + 2\beta_0\beta_1 \overline{x}) - 2a_{11}(\beta_0 + \beta_1 \overline{x}) \right\}, \end{aligned}$$

$$i_{\beta_1\beta_1} = \left\{ \overline{x^2}(1 + (\alpha - 1)a_{20}) + \frac{\alpha - 1}{\eta_e} \left[a_{11} \overline{x^2} - a_{10}(\beta_0 \overline{x^2} + \beta_1 \overline{x^3}) \right] \right\} / \eta_e^2,$$

$$\begin{aligned} i_{\eta_e\beta_1} = {} & \frac{1 - \alpha}{\eta_e^2} a_{10} \overline{x} + \frac{1}{\eta_e^2} \left[\overline{x}(2(a_{01} - \beta_0) + (\alpha - 1)(a_{21} - \beta_0 a_{20})) - \beta_1(2 + (\alpha - 1)a_{20}) \overline{x^2} \right] \\ & + \frac{\alpha - 1}{\eta_e^4} \left[a_{12} \overline{x} + a_{10}(\beta_0^2 \overline{x} + 2\beta_0\beta_1 \overline{x^2} + \beta_1^2 \overline{x^3}) - 2a_{11}(\beta_0 \overline{x} + \beta_1 \overline{x^2}) \right], \end{aligned}$$

$$\begin{aligned} i_{\eta_e\eta_e} = {} & -\frac{1}{\eta_e^2} + \frac{1}{\eta_e^4} [3a_{02} + (\alpha - 1)a_{22} - 2(\beta_0 + \beta_1 \overline{x})(3a_{01} + (\alpha - 1)a_{21})] \\ & + \frac{1}{\eta_e^4}(3 + (\alpha - 1)a_{20})(\beta_0^2 + 2\beta_0\beta_1 \overline{x} + \beta_1^2 \overline{x^2}) - 2\frac{\alpha - 1}{\eta_e^3}(a_{11} - a_{10}(\beta_0 + \beta_1 \overline{x})) \\ & + \frac{\alpha - 1}{\eta_e^5} \left[a_{13} - 3a_{12}(\beta_0 + \beta_1 \overline{x}) + 3a_{11}(\beta_0^2 + 2\beta_0\beta_1 \overline{x} + \beta_1^2 \overline{x^2}) \right] \\ & - \frac{\alpha - 1}{\eta_e^5} a_{10}(\beta_0^3 + \beta_1^3 \overline{x^3} + 3\beta_0\beta_1 \overline{x^2} + 3\beta_0^2\beta_1 \overline{x}), \end{aligned}$$

$$i_{\alpha\beta_0} = a_{10}/\eta_e,\ \ i_{\alpha\beta_1} = a_{10}\overline{x}/\eta_e,\ \ i_{\alpha\eta_e} = [a_{11} - a_{10}(\beta_0+\beta_1\overline{x})]/\eta_e^2 \ \text{ and } \ i_{\alpha\alpha} = 1/\alpha^2.$$

The above expressions can be computed numerically.

10.3. Observed information matrix for the nonlinear PNR model

The elements of the observed information matrix for the nonlinear PNR model are given by

$$j_{\beta_k\beta_j} = \frac{1}{\eta^2}\sum_{i=1}^{n}[(y_i - f(\boldsymbol{\beta},\mathbf{x}_i))w_i + w_i^2 + \alpha - 1]\frac{\partial f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_k}\frac{\partial f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_k} +$$

$$\frac{1}{\eta^2}\sum_{i=1}^{n}[-(y_i - f(\boldsymbol{\beta},\mathbf{x}_i)) + \eta(\alpha-1)w_i]\frac{\partial^2 f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_k\partial\beta_j},$$

$$j_{\eta\beta_j} = \frac{\alpha-1}{\eta^2}\sum_{i=1}^{n}\left\{-w_i + \frac{(y_i - f(\boldsymbol{\beta},\mathbf{x}_i))}{\eta}[(y_i - f(\boldsymbol{\beta},\mathbf{x}_i))w_i + w_i^2]\frac{\partial f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_j}\right\} +$$

$$\frac{2}{\eta^3}\sum_{i=1}^{n}(y_i - f(\boldsymbol{\beta},\mathbf{x}_i))\frac{\partial f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_j},$$

$$j_{\alpha\beta_j} = \frac{1}{\eta}\sum_{i=1}^{n}w_i\frac{\partial f(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\beta_j}, \qquad j_{\alpha\eta} = \frac{1}{\eta}\sum_{i=1}^{n}\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)w_i, \qquad j_{\alpha\alpha} = n/\alpha^2,$$

$$j_{\eta\eta} = -\frac{n}{\eta^2} + \frac{3}{\eta^2}\sum_{i=1}^{n}\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)^2 - \frac{2(\alpha-1)}{\eta^2}\sum_{i=1}^{n}\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)w_i$$

$$+\frac{\alpha-1}{\eta^2}\sum_{i=1}^{n}\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)^3 w_i + \frac{\alpha-1}{\eta^2}\sum_{i=1}^{n}\left(\frac{y_i - f(\boldsymbol{\beta},\mathbf{x}_i)}{\eta}\right)^2 w_i^2.$$

10.4. Hessian matrix for the nonlinear PNR model with correlated errors

For the case of the nonlinear model with correlated errors, we have the following elements for the Hessian matrix:

$$\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\rho^2} = \frac{1}{\eta^2}\sum_{i=1}^{n}[-1 + (\alpha-1)Q_i]\,\epsilon_{i-1}^2,$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{\mathsf{T}} \partial \rho} = \sum_{i=1}^{n} \left[[1-(\alpha-1)Q_i] \frac{\epsilon_{i-1}}{\eta^2} D_i^{\mathsf{T}} - \left[\frac{a_i}{\eta^2} - \frac{1}{\eta} w_i \right] \frac{\partial f(\boldsymbol{\beta}, \mathbf{x}_{i-1})}{\partial \boldsymbol{\beta}^{\mathsf{T}}} \right],$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \eta^2 \partial \rho} = \sum_{i=1}^{n} \left[-\frac{a_i}{\eta^4} + \frac{\alpha-1}{2\eta^2} \left[\frac{a_i Q_i}{\eta^2} + \frac{w_i}{\eta} \right] \right] \epsilon_{i-1}, \qquad \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \alpha \partial \rho} = -\frac{1}{\eta} \sum_{i=1}^{n} w_i \epsilon_{i-1},$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^{\mathsf{T}}} = \sum_{i=1}^{n} \left[\frac{1}{\eta^2} [-1+(\alpha-1)Q_i] D_i D_i^{\mathsf{T}} - \frac{a_i}{\eta^2} G_i + \frac{\alpha-1}{\eta} w_i G_i \right],$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \eta^2 \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{a_i}{\eta^4} D_i - \frac{\alpha-1}{2\eta^2} \left[\frac{a_i}{\eta^2} Q_i D_i + \frac{1}{\eta} w_i D_i \right] \right],$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \alpha \partial \boldsymbol{\beta}} = \frac{1}{\eta} \sum_{i=1}^{n} w_i D_i, \qquad \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \alpha \partial \eta^2} = -\frac{1}{2\eta^3} \sum_{i=1}^{n} a_i w_i, \qquad \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \alpha^2} = -\frac{n}{\alpha^2}.$$

Acknowledgments

We thank two referees for comments and suggestions that substantially improved the presentation. The research of H. Bolfarine was supported by CNPq-Brasil. The research of H. W. Gómez was supported by FONDECYT-Chile 1130495.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Azme, K., Ismail, Z., Haron, K. and Ahmad, T.M. (2005). Nonlinear growth models for modeling oil palm yield growth. *Journal of Mathematical Statistics*, I(3), 225–233.

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

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46, 199–208.

Azzalini, A. (2013). *Skew Normal and Related Families*. Cambridge University Press.

Berkane, M., Kano, Y. and Bentle, P. M. (1994). Pseudo maximum likelihood estimation in elliptical theory: effects of misspecification. *Computational Statistics and Data Analysis*, 18, 255–267.

Cancho, V. G., Lachos, V. H. and Ortega, E. M. (2008). A nonlinear regression model with skew-normal errors. *Statistical Papers*, 51, 547–558.

Castillo, E. and Hadi, A. S. (1995). A method for estimating parameters and quantiles of distributions of continuous random variables. *Computational Statistics and Data Analysis*, 20, 421–439.

Chiogna, M. (2005). Notes on estimation problems with scalar skew-normal distributions. *Statistical Methods and Applications*, 14, 331–341.

Cordeiro, G. M., Ferrari, S. L. P., Uribe-Opazo, M. A. and Vasconcellos, K. L. P. (2000). Corrected maximum likelihood estimation in a class of symmetric nonlinear regression models. *Statistics and Probability Letters*, 46, 317–328.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall, London.

DiCiccio, T. J. and Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association*, 99, 439–450.

Durrans, S. R. (1992). Distributions of fractional order statistics in hydrology. *Water Resources Research*, 28, 1649–1655.

Eugene, N., Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods*, 31, 497–512.

Fernández, C. and Steel, M. (1999). Multivariate Student-t regression models: pitfalls and inference. *Biometrika*, 86, 153–167.

Foong, F. S. (1999). Impact of mixture on potential evapotranspiration, growth and yield of palm oil. Pro 1999, *PORIM International Palm Oil Congress (Agric.)*. 265–287.

Galea, M., Paula, G. A. and Cysneiros, J. A. (2005). On diagnostic in symmetrical nonlinear models. *Statistics & Probability Letters*, 73, 459–467.

Gómez, H. W., Venegas, O. and Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics*, 18, 395–407.

Gupta, A. K. and Nadarajah, S. (2004). On the moments of the beta normal distribution. *Communications in Statistics-Theory and Methods*, 33, 1–13.

Gupta, D. and Gupta, R. C. (2008). Analyzing skewed data by power normal model. *Test*, 17, 197–210.

Hutson, A. D. (2004). Utilizing the flexibility of the epsilon-skew-normal distribution for common regression problems. *Journal of Applied Statistics*, 31, 673–683.

Lange, K. L., Little, J. A. and Taylor, M. G. J. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881–896.

Lehmann, E. L. (1953). The power of rank tests. *Annals of Mathematical Statistics*, 24, 23–43.

Leiva, V., Riquelme, M., Balakrishnan, N. and Sanhueza, A. (2008). Lifetime analysis based on the generalized Birnbaum-Saunders distribution. *Computational Statistics and Data Analysis*, 52, 2079–2097.

Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, 83, 291–309.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics*, 27, 859–870.

Pewsey, A., Gómez, H. W. and Bolfarine, H. (2012). Likelihood-based inference for power distributions. *Test*, 21, 775–789.

R Development Core Team (2014). A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Razzaghi, M. (2009). Beta-normal distribution in dose-response modeling and risk assessment for quantitative responses. *Enviromental and Ecological Statistics*, 16, 25–36.

Ratkowsky, D. A. (1983). *Nonlinear Regression Models*. Marcel Dekker, New York.

Rego, L. C., Cintra, R. J. and Cordeiro, G. M. (2012). On some properties of the beta normal distribution. *Communications in Statistics-Theory and Methods*, 41, 3722–3738.

Taylor, J. and Verbyla, A. (2004). Joint modeling of location and scale parameters of t distribution. *Statistical Modelling*, 4, 91–112.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.

Xie, F. C., Lin, J. G. and Wei, B.C. (2009). Diagnostics for skew nonlinear regression models with AR(1) errors. *Computational Statistics and Data Analysis*, 53, 4403–4416.

SORT 39 (2) July-December 2015, 209-230

On the bivariate Sarmanov distribution and copula. An application on insurance data using truncated marginal distributions

Zuhair Bahraoui[1], Catalina Bolancé[1], Elena Pelican[2] and Raluca Vernic[2]

Abstract

The Sarmanov family of distributions can provide a good model for bivariate random variables and it is used to model dependency in a multivariate setting with given marginals. In this paper, we focus our attention on the bivariate Sarmanov distribution and copula with different truncated extreme value marginal distributions. We compare a global estimation method based on maximizing the full log-likelihood function with the estimation based on maximizing the pseudo-log-likelihood function for copula (or partial estimation). Our aim is to estimate two statistics that can be used to evaluate the risk of the sum exceeding a given value. Numerical results using a real data set from the motor insurance sector are presented.

MSC: MSC62-07 MSC62E15 and MSC62F10

Keywords: Bivariate Sarmanov distribution, truncated marginal distributions, copula representation, risk measures.

1. Introduction

Modelling bivariate variables that represent economic losses is not a straightforward task. To analyse such data, the usual approach involves fitting univariate distributions to both marginals and then considering the corresponding theoretical bivariate distribution for the entire data set. However, this procedure might not be successful if the marginals present different distribution types or if the dependency structure of the theoretical bivariate distribution is inappropriate for the real data. Furthermore, given the shape of the likelihood function or moments, estimating the parameters can be challenging.

[1] Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain.

[2] Faculty of Mathematics and Computer Science, Ovidius University of Constanta, 124 Mamaia Blvd, 900527 Constanta, Romania.

Received: December 2014
Accepted: April 2015

On the other hand, when the marginals follow extreme value distributions, in some cases we have infinite moments. In an economic context, this means that the loss amount is unlimited. However, in practice, this is an unrealistic scenario.

In this paper, we limit ourselves to the Sarmanov family of distributions, originally introduced in its bivariate form by Sarmanov (1966) to join given marginals. This distribution has also been proposed in a more general form in the field of physics by cohen (1984), its multivariate version was suggested by Lee (1996) and generalised further by Bairamov et al. (2001) and Bairamov et al. (2011). Recently, the Sarmanov distribution has attracted interest in other fields (see, for example, Danaher, 2007; Gómez-Déniz and Pérez-Rodríguez, 2015), including finance and insurance. Thus, Hernández-Bastida et al. (2009) and Hernández-Bastida and Fernández-Sánchez (2013) used the bivariate Sarmanov distribution for evaluating premiums in insurance compound models, while further applications related to the theory of ruin were presented by Yang and Hashorva (2013). Furthermore, Hashorva and Ratovomirija (2015) have analysed the Sarmanov distribution with mixed Erlang marginal distributions and have used it for capital allocation. In general, this family of distributions is useful for analysing multivariate loss data, whose marginal distributions may be of the extreme value type or may present very different behaviours. We propose a global estimation (GE) method for the parameters of the Sarmanov distribution with right truncated extreme value marginal distributions.

The bivariate Sarmanov copula is derived from the bivariate Sarmanov distribution and can be a good, quite simple alternative for representing dependency. A copula is a function that relates a bivariate distribution function to its univariate marginal distribution functions, thus allowing the structure of dependence between variables to be fitted separately from the marginal distributions. Specifically, we focus our attention on the bivariate Sarmanov distribution and copula with different log-types of truncated marginal distributions: truncated log-normal, mixture of truncated log-normals and truncated log-logistic. The proposed models may be useful for measuring the risk of loss.

When analising data that represent univariate losses, the univariate distribution that generates the observations is often an extreme value distribution and, therefore, the mean or variance (first or second moment) of the corresponding random variable can be infinite. In finance and insurance, for quantifying the risk it is useful to assume a finite value for the first two moments of the distribution, leading to the right truncation of the distribution of the random variable analised, which was the procedure adopted in this paper. Furthermore, we use a bivariate Sarmanov distribution that requires marginal distributions with finite first moment.

Using a real data set from the motor insurance sector, we compare the estimated risk of loss evaluated for the bivariate Sarmanov distribution with truncated extreme value marginal distributions whose parameters result by the GE method, with the estimation of the same risk obtained after Monte Carlo simulation from the corresponding copula (as examples of fitting alternative copulas and marginals on this data set see, Bolancé et al., 2014; Bahraoui et al., 2014).

The paper is structured as follows: in Section 2 we present two truncated log-normal-type univariate distributions, plus the heavier-tailed truncated log-logistic (Champernowne) distribution, for which we also obtained the first and second moments. In Section 3 we introduce the bivariate Sarmanov distribution and its copula representation, and discuss the parameters estimation. Some comments on the evaluation of two statistics that are used to quantify the risk of loss (Value at Risk - VaR and Tail Value at Risk - TVaR) are presented in Section 4. Finally, in Section 5 we present the results of the proposed fits and risk estimations. Section 6 concludes.

2. Some univariate truncated distributions

We begin by introducing some notations and some univariate truncated distributions to be used as marginals for the bivariate Sarmanov distribution and copula in Section 3.

Let $\mathbf{X} = (X_1, X_2)$ be a bivariate random vector that represents two dependent losses. The random variable (r.v.) $S = X_1 + X_2$ is the total loss and we are interested in measuring the risk associated with the distribution of S; for this, we need to consider both the joint distribution of $\mathbf{X}$ and the marginal distributions of X_1, X_2.

In this section, we analise the probability distribution function (pdf), the cumulative distribution function (cdf) and the first two moments of three distributions that can be useful to model losses: the truncated log-normal, the mixture of two truncated log-normals and the truncated log-logistic, also known as the Champernowne distribution; we let m and M be the truncation points[1] on the left and right side, respectively.

2.1. Truncated log-normal distributions

Let $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cdf, respectively, of the standard normal $N(0,1)$ distribution. To denote the pdf of the general normal $N\left(\mu, \sigma^2\right), \mu \in \mathbb{R}, \sigma > 0$ distribution, we use the same symbol φ emphasizing the parameters, i.e. $\varphi\left(x; \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$. The truncated normal distribution $TN\left(\mu, \sigma^2; m, M\right), \mu \in \mathbb{R}, \sigma > 0$, with truncation points $m < M$, has the pdf

$$f_{TN}(x) = \frac{\varphi\left(x; \mu, \sigma^2\right)}{\Phi(A) - \Phi(a)} = \frac{1}{(\Phi(A) - \Phi(a))\,\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, m \leq x \leq M,$$

where $A = \frac{M-\mu}{\sigma}, a = \frac{m-\mu}{\sigma}$. Its expected value and variance are given respectively, by (see, for example, Kotz et al.,2000)

1. In our numerical application we assume $m = 0$.

$$\xi = \mu + \frac{\varphi(a) - \varphi(A)}{\Phi(A) - \Phi(a)}\sigma,\ \delta^2 = \left(1 + \frac{a\varphi(a) - A\varphi(A)}{\Phi(A) - \Phi(a)} - \left(\frac{\varphi(a) - \varphi(A)}{\Phi(A) - \Phi(a)}\right)^2\right)\sigma^2.$$

We recall that a random variable (r.v.) X follows a log-normal distribution $LN\left(\mu, \sigma^2\right)$ if $\ln X \sim N\left(\mu, \sigma^2\right)$, having hence the pdf $f_X(x) = \frac{1}{x}\varphi\left(\ln x; \mu, \sigma^2\right)$ and cdf $f_X(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), x > 0$. Moreover, we say that X follows a truncated log-normal distribution $TLN\left(\mu, \sigma^2; m, M\right)$ with truncation points $0 < m < M$, if $\ln X \sim TN\left(\mu, \sigma^2; \ln m, \ln M\right)$; hence, its pdf is $f_X(x) = \frac{1}{x}\frac{\varphi\left(\ln x; \mu, \sigma^2\right)}{\Phi(B) - \Phi(b)}$, where $B = \frac{\ln M - \mu}{\sigma}, b = \frac{\ln m - \mu}{\sigma}$.

Proposition 1 *If $X \sim TLN\left(\mu, \sigma^2; m, M\right), 0 < m < M$, its first two moments are given by*

$$E[X] = e^{\mu + \frac{\sigma^2}{2}}\frac{\Phi(C) - \Phi(c)}{\Phi(B) - \Phi(b)},\ E\left[X^2\right] = e^{2\left(\mu + \sigma^2\right)}\frac{\Phi(D) - \Phi(d)}{\Phi(B) - \Phi(b)},$$

where $C = B - \sigma, c = b - \sigma, D = B - 2\sigma, d = b - 2\sigma$.

Proof Changing variable $y = \ln x$, we obtain

$$E[X] = \int_m^M \frac{x}{x}\frac{\varphi\left(\ln x; \mu, \sigma^2\right)}{\Phi(B) - \Phi(b)}dx = \int_{\ln m}^{\ln M} \frac{\varphi\left(y; \mu, \sigma^2\right)e^y}{\Phi(B) - \Phi(b)}dy$$

$$= \frac{e^{\mu + \frac{\sigma^2}{2}}}{\Phi(B) - \Phi(b)}\int_{\ln m}^{\ln M} \varphi\left(y; \mu + \sigma^2, \sigma^2\right)dy,$$

which immediately yields the stated formula of $E[X]$. The formula of $E\left[X^2\right]$ results in a similar way. ∎

2.2. Mixtures of two truncated log-normal distributions

Consider two truncated normal distributions $TN(\mu_i, \sigma_i^2; m, M), \mu_i \in \mathbb{R}, \sigma_i > 0, i = 1, 2$, having the same truncation points $m < M$. Then, denoting their mixture by $TN_{mixt}\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, r; m, M\right), r \in (0, 1)$, its pdf has the form

$$f_{TN_{mixt}}(x) = r\frac{\varphi\left(x; \mu_1, \sigma_1^2\right)}{\Phi(A_1) - \Phi(a_1)} + (1 - r)\frac{\varphi\left(x; \mu_2, \sigma_2^2\right)}{\Phi(A_2) - \Phi(a_2)}, m \le x \le M,$$

where $A_i = \frac{M - \mu_i}{\sigma_i}, a_i = \frac{m - \mu_i}{\sigma_i}, i = 1, 2$, and r is the mixing parameter.

Similarly, we say that the r.v. X follows a mixture of two truncated log-normal distributions $TLN_{mixt}\left(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,r;m,M\right),m>0$, if its pdf is

$$f_{TLN_{mixt}}(x)=r\frac{\varphi\left(\ln x;\mu_1,\sigma_1^2\right)}{x\left(\Phi\left(B_1\right)-\Phi\left(b_1\right)\right)}+(1-r)\frac{\varphi\left(\ln x;\mu_2,\sigma_2^2\right)}{x\left(\Phi\left(B_2\right)-\Phi\left(b_2\right)\right)},m\leq x\leq M,$$

with $B_i=\frac{\ln M-\mu_i}{\sigma_i},b_i=\frac{\ln m-\mu_i}{\sigma_i},i=1,2$. In this case, $\ln X\sim TN_{mixt}\left(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,r;m',M'\right)$, where $m'=\ln m,M'=\ln M$.

To obtain the moments of the above mixtures of truncated distributions, we note that the pdf of such a mixture of distributions is of the form $f(x)=rf_1(x)+(1-r)f_2(x)$, where f_1 and f_2 are themselves pdf's. If we denote by X_i a r.v. having pdf f_i and by X a r.v. with pdf f, then the first two moments of the mixed distribution results as

$$E[X]=rE[X_1]+(1-r)E[X_2],\ E\left[X^2\right]=rE\left[X_1^2\right]+(1-r)E\left[X_2^2\right],$$

from where a straightforward calculation yields the variance

$$Var[X]=rVar[X_1]+(1-r)Var[X_1]+r(1-r)\left(E[X_1]-E[X_2]\right)^2.$$

Using these formulas, the first moments of the TN_{mixt} and TLN_{mixt} distributions are immediate.

Moreover, we also note that fitting a truncated log-normal distribution or a mixture of two truncated log-normal distributions to a data set, is the same as fitting a truncated normal distribution or, correspondingly, a mixture of two truncated normal distributions to the log-data set.

2.3. Champernowne (log-logistic) distribution

Introduced by Champernowne in 1952 (see, Champernowne, 1952), the log-logistic distribution is the distribution of a r.v. whose logarithm follows a logistic distribution. In economics, where it is also known as the Fisk distribution (see, Fisk, 1961), it is used to model the distribution of wealth or income. Its shape is similar to the log-normal distribution, but it has heavier tails; moreover, as an asymptotic behaviour, it converges towards a Pareto distribution in the tail (see, Buch-Larsen et al., 2005). Denoted by $Ch(\alpha,H),\alpha,H>0$, its pdf is defined by

$$f_{Ch}(x)=\frac{\alpha H^{\alpha}x^{\alpha-1}}{\left(x^{\alpha}+H^{\alpha}\right)^2},x\geq 0,$$

having cdf $F_{Ch}(x)=\frac{x^{\alpha}}{x^{\alpha}+H^{\alpha}}, x\geq 0$, expected value $\frac{\pi H}{\alpha}\left(\sin\frac{\pi}{\alpha}\right)^{-1}$, for $\alpha>1$, and variance $\frac{\pi H^2}{\alpha^2}\left(\sin\frac{\pi}{\alpha}\right)^{-1}\left(\left(\cos\frac{\pi}{\alpha}\right)^{-1}-\pi\left(\sin\frac{\pi}{\alpha}\right)^{-1}\right)$, for $\alpha>2$. Note that H is a scale parameter and the median of the distribution, while α is a shape parameter.

We also consider the truncated form $TCh(\alpha,H;M),\alpha,H,M>0$, having pdf[2]

$$f_{TCh}(x)=\alpha\left(M^{\alpha}+H^{\alpha}\right)\left(\frac{H}{M}\right)^{\alpha}\frac{x^{\alpha-1}}{\left(x^{\alpha}+H^{\alpha}\right)^2}, 0\leq x\leq M. \tag{1}$$

Its moments do not have a closed form, but they can be expressed in terms of the hypergeometric function ${}_2F_1$ defined for $|z|<1$ by the following integral or power series

$${}_2F_1(a,b;c;z)=\frac{1}{B(b,c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\,dt=\sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!},$$

where

$$(q)_k=\begin{cases}1, & k=0\\ q(q+1)\cdots(q+k-1), & k>0\end{cases}, c\notin\{0,-1,-2,...\},$$

and $B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ is the Beta function.

Proposition 2 *Letting $X\sim TCh(\alpha,H;M)$, its first two moments are given by*

$$E[X]=\frac{\alpha M}{\alpha+1}\left(1+\left(\frac{M}{H}\right)^{\alpha}\right){}_2F_1\left(2,1+\frac{1}{\alpha};2+\frac{1}{\alpha};-\left(\frac{M}{H}\right)^{\alpha}\right), \tag{2}$$

$$E\left[X^2\right]=\frac{\alpha M^2}{\alpha+2}\left(1+\left(\frac{M}{H}\right)^{\alpha}\right){}_2F_1\left(2,1+\frac{2}{\alpha};2+\frac{2}{\alpha};-\left(\frac{M}{H}\right)^{\alpha}\right). \tag{3}$$

Proof We evaluate the expected value of X by changing variable $x=My^{1/\alpha}$ in

$$\begin{aligned}E[X]&=\alpha\left(M^{\alpha}+H^{\alpha}\right)\left(\frac{H}{M}\right)^{\alpha}\int_0^M\frac{x^{\alpha}}{\left(x^{\alpha}+H^{\alpha}\right)^2}dx\\ &=\alpha\left(M^{\alpha}+H^{\alpha}\right)\left(\frac{H}{M}\right)^{\alpha}\int_0^1\frac{\left(My^{1/\alpha}\right)^{\alpha}}{\left(\left(My^{1/\alpha}\right)^{\alpha}+H^{\alpha}\right)^2}\frac{My^{1/\alpha-1}}{\alpha}dy\end{aligned}$$

2. Since in our application we assume $m=0$, for the sake of simplicity, we only present the properties for $M>0$.

$$= M\left(M^{\alpha}+H^{\alpha}\right)\left(\frac{H}{M}\right)^{\alpha}\left(\frac{M}{H^{2}}\right)^{\alpha}\int_{0}^{1}\frac{y^{1/\alpha+1-1}}{\left(1+\left(\frac{M}{H}\right)^{\alpha}y\right)^{2}}dy$$

$$= M\left(1+\left(\frac{M}{H}\right)^{\alpha}\right)B\left(1+\frac{1}{\alpha},1\right)\ {}_{2}F_{1}\left(2,1+\frac{1}{\alpha};2+\frac{1}{\alpha};-\left(\frac{M}{H}\right)^{\alpha}\right),$$

with the last relation resulting from the definition of the function ${}_2F_1$. Note that

$$B\left(1+\frac{1}{\alpha},1\right)=\frac{\Gamma\left(1+\frac{1}{\alpha}\right)\Gamma(1)}{\Gamma\left(2+\frac{1}{\alpha}\right)}=\frac{1}{1+\frac{1}{\alpha}}=\frac{\alpha}{1+\alpha},$$

where $\Gamma(a)=\int_0^{\infty}x^{a-1}e^{-x}dx$ denotes the Gamma function. Inserting this result into the last expression of $E[X]$ immediately yields formula (2). Formula (3) results in a similar way. ■

3. Bivariate Sarmanov distribution

3.1. The general distribution

We say that the random vector $\mathbf{X}=(X_1,X_2)$ follows a bivariate Sarmanov's distribution if its joint pdf is given by (see, Kotz et al., 2000).

$$f_{\mathbf{X}}(x_1,x_2)=f_1(x_1)\,f_2(x_2)\,(1+\omega\phi_1(x_1)\,\phi_2(x_2)), \tag{4}$$

where $(f_i)_{i=1,2}$ are the corresponding marginal pdf's, $(\phi_i)_{i=1,2}$ are bounded non-constant kernel functions and ω is a real number such that

$$\int_{-\infty}^{\infty}\phi_i(x_i)\,f_i(x_i)\,dx_i=0,\ i=1,2,\ \text{and} \tag{5}$$

$$1+\omega\phi_1(x_1)\,\phi_2(x_2)\geq 0,\ \text{for all}\ (x_1,x_2)\in\mathbb{R}^2. \tag{6}$$

If we denote $\nu_i=\int_{-\infty}^{\infty}x\phi_i(x)\,f_i(x)\,dx,\ i=1,2,$ then the covariance and correlation coefficient are given, respectively, by

$$\operatorname{cov}(X_1,X_2)=\omega\,\nu_1\,\nu_2,$$

$$\operatorname{corr}(X_1,X_2)=\frac{\omega\,\nu_1\,\nu_2}{\sqrt{Var[X_1]\,Var[X_2]}}. \tag{7}$$

Note that when ω=0, X_1 and X_2 are independent. As to the choice of the kernel functions ϕ_i, some particular cases satisfying (5) have already been discussed in the literature (see, Lee, 1996), from which we recall:

- $\phi_i = 1 - 2F_i$, where F_i is the cdf of X_i. In this case, the Sarmanov distribution is known as the Farlie-Gumbel-Morgenstern distribution (see Farlie, 1960), verifying the restrictive condition that the correlation coefficient $\text{corr}(X_1, X_2)$ cannot exceed 1/3 in absolute value. However, in general, the Sarmanov distribution is not restricted by such a condition (see, for example, Shubina and Lee, 2004).
- $\phi_i(x) = e^{-\alpha x} - E\left[e^{-\alpha X_i}\right]$ (we say no more about this form as it did not provide a good fit to our data).
- $\phi_i(x) = x^\alpha - E[X_i^\alpha]$, assuming that $E[X_i^\alpha] < \infty$. In this case, $\nu_i = E\left[X_i^{\alpha+1}\right] - E[X_i^\alpha]E[X_i]$, if it is finite.

Given its simplicity and better fit for our data, in our study we consider $\phi_i(x) = x^\alpha - E[X_i^\alpha]$ with $\alpha = 1$, yielding from (7) the correlation

$$\text{corr}(X_1, X_2) = \omega\sqrt{Var[X_1]\,Var[X_2]}. \tag{8}$$

Therefore, assuming that $E[X_i] < \infty$, in the following we limit ourselves to the pdf form

$$f_{\mathbf{X}}(x_1, x_2) = f_1(x_1) f_2(x_2)(1 + \omega(x_1 - E[X_1])(x_2 - E[X_2])) \tag{9}$$

that requires the existence of a finite first moment for both marginals. In this case, condition (6) obviously restricts the domain of $f_{\mathbf{X}}$. For simplicity, we preferred to work with truncated marginals, which meant imposing restrictions on the coefficient ω. More precisely, if the support of X_i is $[m_i, M_i], i = 1,2$, then condition (6) yields $l \le \omega \le u$, where

$$l = \max\left\{\frac{-1}{(M_1 - E[X_1])(M_2 - E[X_2])}, \frac{-1}{(m_1 - E[X_1])(m_2 - E[X_2])}\right\}, \tag{10}$$

$$u = \min\left\{\frac{-1}{(M_1 - E[X_1])(m_2 - E[X_2])}, \frac{-1}{(m_1 - E[X_1])(M_2 - E[X_2])}\right\}. \tag{11}$$

Because of the restriction imposed by condition (6), we used marginal distributions with bounded support. Therefore, we considered the truncated distributions presented in Section 2, their choice being driven by the real data to be studied in Section 5.

3.2. Copula representation and simulation

A copula can be defined as a multivariate cdf with standard uniform $[0,1]$ marginals. Then the cdf of a random vector $\mathbf{X} = (X_1, \ldots, X_m)$ can be written in terms of its marginal cdf's using a copula function $C : [0,1]^m \to [0,1]$, as follows $F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \ldots, F_m(x_m))$; for details on copulas see Nelsen (2006).

Since the Sarmanov bivariate distribution is defined directly from its marginal distributions, its cdf can be immediately expressed as $F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), F_2(x_2))$ using the following copula function

$$C(u_1,u_2) = u_1u_2 + \omega \int_0^{u_1} \phi_1(F_1^{-1}(t))dt \int_0^{u_2} \phi_2(F_2^{-1}(s))ds, 0 \le u_1, u_2 \le 1, \qquad (12)$$

assuming that F_1^{-1}, F_2^{-1} exist; the corresponding density is

$$c(u_1,u_2) = 1 + \omega\phi_1(F_1^{-1}(u_1))\phi_2(F_2^{-1}(u_2)). \qquad (13)$$

Working with the copula representation of the Sarmanov family of distributions has some advantages. The copula representation is straightforward and its estimation procedure is simple. Furthermore, this representation enables us to generate pseudo-random samples from the Sarmanov bivariate distribution. To do this, we first generate values from the Sarmanov copula (12) using the procedure described in Nelsen (2006), which is based on the conditional distribution of a random vector (U_1, U_2) having uniform $[0,1]$ marginals and cdf C, i.e., on $C_{u_1}(u_2) = \Pr(U_2 \le u_2 | U_1 = u_1)$. Note that

$$C_{u_1}(u_2) = \lim_{\Delta u_1 \to 0^+} \frac{C(u_1 + \Delta u_1, u_2) - C(u_1, u_2)}{\Delta u_1} = \frac{\partial C(u_1,u_2)}{\partial u_1}.$$

The corresponding algorithm is implemented as follows:

1. Generate two independent random values u_1 and z from the uniform $U(0,1)$ distribution.
2. Set $u_2 = C_{u_1}^{(-1)}(z)$, where $C_{u_1}^{(-1)}$ denotes a quasi-inverse of C_{u_1}. The desired pair from the Sarmanov copula is (u_1, u_2).
3. Solving now $F_1(x_1) = u_1$ for x_1 and $F_2(x_2) = u_2$ for x_2 yields the pseudo-random pair (x_1, x_2) simulated from the corresponding bivariate Sarmanov's distribution.

In our case, assuming that the inverses F_1^{-1}, F_2^{-1} exist, the partial derivative of (12) is

$$C_{u_1}(u_2) = u_2 + \omega\phi_1(F_1^{-1}(u_1)) \int_0^{u_2} \phi_2(F_2^{-1}(s))ds.$$

If, in particular, we take the kernel functions $\phi_i(z) = z - E[X_i], i = 1,2$, this gives

$$C_{u_1}(u_2) = u_2 + \omega(F_1^{-1}(u_1) - E[X_1]) \int_0^{u_2} (F_2^{-1}(s) - E[X_2])ds.$$

3.3. Estimation of parameters

Let Θ denote the parameters set of the Sarmanov distribution. First, we estimate the parameters using the maximum likelihood (ML) method, that we named global estimation (GE), based on the random data sample $\{(x_{1j}, x_{2j})\}_{j=1}^n$ consisting of n couples of observations. For estimating the Sarmanov copula, we use the maximum pseudo-likelihood method that we named partial estimation (PE).

3.3.1. Global estimation (GE) method

From density (4), the log-likelihood function to be maximized is

$$\ln L\left(\{(x_{1j}, x_{2j})\}_{j=1}^n; \Theta\right) = \sum_{j=1}^n \left(\ln f_1(x_{1j}) + \ln f_2(x_{2j}) + \ln\left(1 + \omega \phi_1(x_{1j})\phi_2(x_{2j})\right)\right). \tag{14}$$

The parameters to be estimated are ω, the parameters of f_i, and, eventually, the parameters of ϕ_i. Let θ denote a generic parameter of f_i. The corresponding ML system is

$$\begin{cases} 0 = \dfrac{\partial \ln L}{\partial \theta} = \displaystyle\sum_{j=1}^n \left(\frac{\partial \ln f_1(x_{1j})}{\partial \theta} + \frac{\partial \ln f_2(x_{2j})}{\partial \theta}\right) + \omega \sum_{j=1}^n \frac{1}{1 + \omega \phi_1(x_{1j})\phi_2(x_{2j})} \\ \qquad \times \left(\phi_1(x_{1j}) \dfrac{\partial \phi_2(x_{2j})}{\partial \theta} + \phi_2(x_{2j}) \dfrac{\partial \phi_1(x_{1j})}{\partial \theta}\right), \quad \theta \in \Theta \\ 0 = \dfrac{\partial \ln L}{\partial \omega} = \displaystyle\sum_{j=1}^n \frac{\phi_1(x_{1j})\phi_2(x_{2j})}{1 + \omega \phi_1(x_{1j})\phi_2(x_{2j})}. \end{cases} \tag{15}$$

This system can become quite complex and, therefore, it must be solved using numerical methods that require starting values for the unknown parameters. Such starting values readily result from the method of moments (MM); for example, a value for ω can be obtained from the empirical correlation coefficient, ρ. For more details on this procedure see Pelican and Vernic (2013).

Alternatively, instead of solving the ML system, numerical methods can be used to find the maximum of the log-likelihood function directly. Such an optimization problem

can be solved using, for example, a variable neighborhood search (VNS) algorithm (see, Mladenovic and Hansen, 1997).

3.3.2. Partial estimation (PE) method

As discussed above, the GE method can result in cumbersome calculations. For this reason, we suggest comparing it with the alternative method based on maximizing the pseudo-log-likelihood corresponding to the copula representation of the Sarmanov distribution (see, for example, Joe, 1997):

- Using the ML method, we estimate the parameters of the univariate marginal distributions of X_1 and X_2, starting from the corresponding data samples $(x_{1j})_{j=1}^n$ and $(x_{2j})_{j=1}^n$, respectively.
- To obtain the parameter ω of the copula, we use again the ML method on (14), after setting the marginal parameters at the values obtained in the previous step. Note, that it is enough to maximize only the last part of (14), i.e., $\sum_{j=1}^n \ln\left(1+\omega\phi_1(x_{1j})\phi_2(x_{2j})\right)$, since the rest does not depend on ω; in fact, this is reduced to applying the ML method to the copula density (13).

4. Evaluating the total risk of loss

Evaluating risk measures for aggregate losses is a challenging task. Let S denote an insurance risk, that is, a non-negative random variable whose cdf is denoted by F_S. A risk measure is generally formulated as a functional from the space of insurance risks to $[0,\infty]$, and its purpose is to provide a single value for the degree of risk associated with the corresponding risk. Among the common risk measures, the Value-at-Risk (VaR) is probably the most frequently adopted. To define it, let $q \in (0,1)$ denote the confidence level required by regulations; then

$$VaR_q[S] := \inf\{x : F_S(x) \geq q\}.$$

The Solvency II Accord drawn up by the EU Commission sets $q = 0.995$ over a one year time horizon.

When heavy tails occur in risk management (see recent episodes of financial instability), a risk measure providing information above a given threshold is recommended. In this respect, the Tail Value-at-Risk (TVaR, also known as the expected shortfall or conditional tail expectation) measure is defined, for $q \in (0,1)$, as

$$TVaR_q[S] := E[S|S > VaR_q[S]].$$

TVaR is considered a coherent risk measure, see Artzner et al. (1999). In some countries, TVaR has already replaced VaR in the regulatory requirements; the current practice is $q = 0.99$ over a one year time horizon.

Let now $S = X_1 + X_2$ be the sum of two possibly dependent insurance risks X_1 and X_2. In this section, our goal is to show how to calculate VaR and TVaR for the risk S when $\mathbf{X} = (X_1, X_2)$ follows the bivariate Sarmanov distribution. Vernic (2014) has analised a closed form for the TVaR of the sum of random variables Sarmanov distributed with exponential marginals. We approach this task in two ways: by direct evaluation and by simulation based on the Sarmanov copula.

4.1. Direct evaluation

To obtain VaR, we must evaluate the cdf of S and then invert it. Letting f_S denote the pdf of S, its cdf results from

$$F_S(s) = \int_0^s f_S(x)\,dx = \int_0^s \int_0^x f_{\mathbf{X}}(x-y, y)\,dy dx$$
$$= \int_0^s \int_0^1 x f_{\mathbf{X}}(x(1-t), xt)\,dt dx.$$

Similarly, for TVaR we need

$$E[S|S > s_q] = \frac{1}{1-F_S(s_q)} \int_{s_q}^{\infty} x f_S(x)\,dx$$
$$= \frac{1}{1-F_S(s_q)} \int_{s_q}^{\infty} \int_0^1 x^2 f_{\mathbf{X}}(x(1-t), xt)\,dt dx,$$

where $s_q = VaR_q[S]$. As there are no closed formulas for these integrals, they have to be calculated using mathematical software. To do this, we wrote Matlab procedures based on Simpson's composite rule for double integrals (see, for example, Bourden and Faires, 2001), paying special attention to the integrals limits since the marginals are truncated.

4.2. Simulation of the Sarmanov copula

Using the Monte Carlo method, the procedure is as follows:

1. We apply the PE method to the data sample $\{(x_{1j}, x_{2j})\}_{j=1}^{n}$ from which we obtain the estimations of the marginals cdf's, denoted $\hat{F}_i, i = 1, 2$, and the estimated parameter of the Sarmanov copula, $\hat{\omega}$.

2. Using the algorithm described in Section 3.2, we generate the pseudo-random sample $\{(\hat{x}_{1j},\hat{x}_{2j})\}_{j=1}^{r}$ from the bivariate Sarmanov distribution with marginals $\hat{F}_1$ and $\hat{F}_2$, where the sample volume r is large (we used $r = 10000$).
3. We calculate $\hat{s}_j = \hat{x}_{1j} + \hat{x}_{2j}, j = 1,\ldots,r$, and we estimate $VaR_q[S]$ and $TVaR_q[S]$ empirically from the generated pseudo-sample $(\hat{s}_j)_{j=1}^{r}$.

5. Numerical study

We used the bivariate Sarmanov distribution and copula to model a random sample of motor insurance claims consisting of the costs of property damage and medical expenses, kindly provided by a major insurer in Spain for the year 2000. Since the data were collected two years later, in 2002, all the claims included in our sample had been settled. The sample size is $n = 518$ and for each claim, X_1 represents the cost of property damage (including third-part liability), while X_2 represents the cost of medical expenses (i.e., treatments and hospitalization as a result of the accident).

Previously, several bivariate distributions were fitted to these data, the best global fit being provided by the bivariate log-skew-normal distribution with a log-likelihood value of -7323.50 and AIC $= 14663.00$ (see, Bolancé et al., 2008). In an attempt to find a better model, in the numerical part of this paper we fitted the bivariate Sarmanov distribution with different normal-type marginals to the bivariate log-data set. Note that if we fit a bivariate Sarmanov distribution with pdf $f_{\mathbf{Y}}$ to the log-data, then the distribution corresponding to the original data is the bivariate log-Sarmanov with pdf

$$f_{\mathbf{X}}(x_1,x_2) = \frac{1}{x_1x_2} f_{\mathbf{Y}}(\ln x_1, \ln x_2), x_1,x_2 > 0.$$

This implies that the marginal distributions of the original data are the log-distributions of the corresponding marginals of $\mathbf{Y}$ (in our case, they become of log-normal and log-logistic types).

In the first attempt, we assumed that $\ln X_1$ follows a truncated normal (TN) distribution and we varied the distribution of $\ln X_2$, but since the best fit was provided by the mixture of two truncated normal distributions TN_{mixt} for $\ln X_2$, we decided not to provide details of the other distributions and we concentrated only on the best fit. This choice was also motivated by the fact that when studying separately the marginal distributions of our data set, we noticed that the normal distribution provided a good fit for $\ln X_1$, but unfortunately, this was not the case with $\ln X_2$, which has a less regular histogram; hence, we made use of the property of the Sarmanov distribution of joining different marginals. Alternatively, we also fitted the bivariate Sarmanov distribution with the heavier-tailed Champernowne marginal distributions to the original data.

In Table 1 we show the descriptive statistics for the original data and for the log-data.

Table 1: *Descriptive statistics.*

	Mean	Std.Dev.	Kurtosis	Skewness	Min	Max	Median
	Original data						
X_1	1827.60	6867.81	297.10	15.65	13.00	137936.00	677.00
X_2	283.92	863.17	82.02	8.04	1.00	11855.00	88.00
	Correlation between X_1 and X_2 is 0.73						
	Log-data						
$\ln X_1$	6.44	1.33	0.57	0.21	2.56	11.83	6.52
$\ln X_2$	4.38	1.52	0.45	0.12	0.00	9.38	4.48
	Correlation between $\ln X_1$ and $\ln X_2$ is 0.59						

Table 2: *AIC obtained for different estimated models and methods.*

Method	Marginals	max	max$\times$10	max$\times$100
GE	$X_1 \sim TLN, X_2 \sim TLN_{mixt}$	14839.58	14863.04	14869.42
	$X_1 \sim TCh, X_2 \sim TCh$	14849.26	14878.42	14883.74
PE	$X_1 \sim TLN, X_2 \sim TLN_{mixt}$	14854.79	14868.25	14873.40
	$X_1 \sim TCh, X_2 \sim TCh$	14880.99	14884.31	14884.52

Since we decided to work with truncated distributions (as discussed in Section 3.1), a key issue was the choice of the upper truncation limits, the lower ones being fixed at $m_1 = m_2 = 0$. We started by taking the upper limits as being equal to the maximum observed values, i.e., $M_i = \max\limits_{j=1,\ldots,n} x_{ij}, i = 1,2$. However, this choice most probably underestimates the real risk since it implies the assumption that the probability of a loss greater than the maximum observed is zero, which is not true in practice. Hence, we assumed that the upper truncation limits increase progressively, being equal to 10, 100 and 1000 times the maximum observed values (denoted in the following by max$\times$10, max$\times$100 and max$\times$1000, respectively). We found the results for the truncation limits of max$\times$100 and max$\times$1000 to be similar, hence, we present here only the former, i.e., max$\times$100, which is equivalent to almost eliminating the effect of truncation.

To estimate the parameters using the methods described in Section 3.3, we took the main empirical characteristics as starting values. Then, to compare the different fits, we calculated the corresponding log-likelihood and the Akaike information criterion (AIC) values. AIC is defined by $\text{AIC} = 2(s - \ln L)$, where s is the number of estimated parameters and L is the likelihood function. This criterion penalizes an increased number of parameters, so that the preferred model is the one with the lowest AIC value. In Table 2 we show the AIC obtained for each estimation, while the estimated parameters and their standard errors are shown in Tables 3-6 in the Appendix. It seems that GE yields a slightly better fit than PE, although we observe that the difference between the AICs for GE and PE is small. This is expected since the GE method maximizes the full likelihood,

while the PE method maximizes separately the partial likelihoods corresponding to the copula and the marginal distributions. Considering both methods (GE and PE) for all upper truncation limits, it results that the best model is the Sarmanov distribution with a truncated log-normal distribution for X_1 and a mixture of two truncated log-normal distributions for X_2.

In Figures 1 and 2 we plot the VaR and TVaR curves as functions of the confidence level q for $q \geq 0.98$, for all the distributions estimated. In Tables 7-10 in Appendix we also displayed the VaR and TVaR values obtained for the same distributions and for some confidence levels q, compared with the empirical values resulting from data. These values and plots clearly show that for $q \geq 0.95$, the Sarmanov distributions with log-normal-type marginals underestimate the empirical values. Although closer to the empirical curve, this is also the case of the Sarmanov distribution with TCh marginals and an upper truncation limit equal to max, while the other two distributions (i.e., max$\times$10 and max$\times$100) overestimate the empirical values. Therefore, from the point of view of the insurer, only these two last distributions would be of interest.

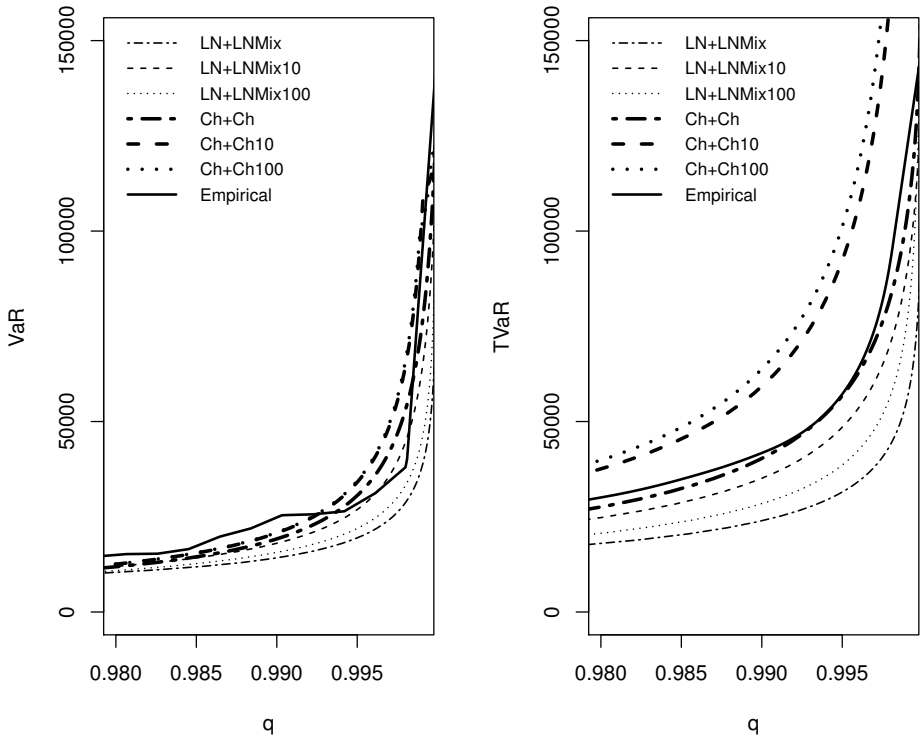


Figure 1: *Estimated VaR and TVaR with GE.*

Note that, the curves resulting from GE and PE methods look similar, although, from Tables 7-10 in Appendix it seems that, in general, PE leads to higher values of VaR and TVaR than those provided by GE.

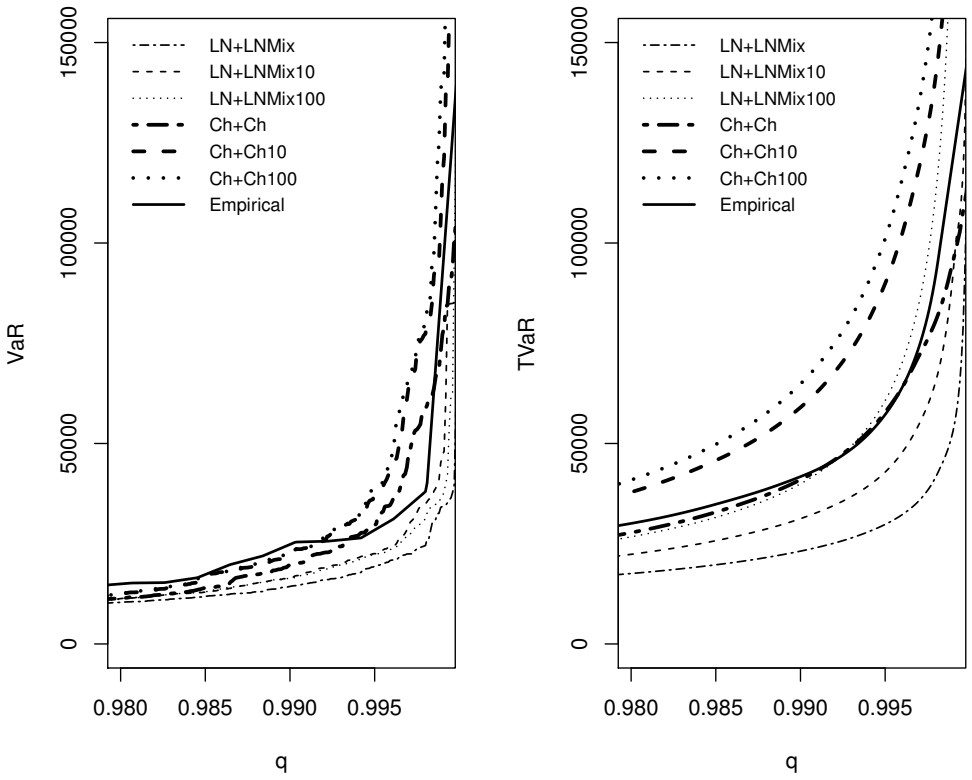


Figure 2: *Estimated VaR and TVaR with PE.*

On the other hand, note that the best globally fitted distribution (in our case, according to AIC, the Sarmanov distributions with LTN and LTN_{mixt} marginals) does not necessarily provide the best model for the risk measures VaR and TVaR, which are defined on the distribution tail – this is also the case with the previously fitted bivariate log-skew-normal distribution, which strongly overestimates the empirical TVaR curve (see Bolancé et al., 2008). For our data set the heavier-tailed Champernowne distribution provides a better model for Sarmanov's marginals when evaluating VaR and TVaR.

6. Conclusions

In this paper, we have proposed the Sarmanov bivariate distribution as a model for bivariate insurance losses and we have illustrated its applicability using a real data set from the motor insurance sector. The choice of this distribution was motivated by its flexible structure that allowed us to join given marginals. From the numerical study, we conclude that the distribution could be a good model for such bivariate insurance data, but special attention should be paid to the choice of the marginal distributions. More specifically, these distributions must fulfill the condition of a real pdf, see (5)-(6), so

that truncated marginal distributions can be selected. Moreover, the upper truncation limits have to be carefully fixed so that the real risk values (like VaR or TVaR) should not be underestimated, but also not overestimated to an exaggerated degree.

It should also be noted that a better global fit does not necessarily mean a better fit regarding the evaluation of some tail related risk measures.

As for the choice between GE and PE methods, it seems that GE yields a somewhat better fit than PE, although the differences are very small. However, the application of the GE method might be more time-consuming given the random search involved in the ML solution. Clearly, the complexity of the calculation should be taken into consideration when selecting the most suitable estimation method.

Appendix

Table 3: *GE for* $\ln X_1 \sim TN$ *and* $\ln X_2 \sim TN_{mixt}$ *and different upper truncation limits (standard errors between parentheses).*

	max	max×10	max×100
μ_1	6.4237 (0.0585)	6.4163 (0.0596)	6.4089 (0.0594)
μ_{21}	4.3661 (0.0836)	4.3758 (0.0713)	4.2860 (0.1199)
μ_{22}	4.3771 (0.5458)	4.0157 (0.4906)	4.4288 (0.2702)
σ_1	1.3310 (0.0412)	1.3560 (0.0431)	1.3517 (0.0428)
σ_{21}	1.2420 (0.0833)	1.2938 (0.0569)	1.1653 (0.1140)
σ_{22}	2.9064 (0.8128)	3.0008 (0.3984)	2.0079 (0.2070)
r	0.8079 (0.0889)	0.8456 (0.0383)	0.6733 (0.1348)
ω	0.0404 (0.0210)	0.0214 (0.0188)	0.0162 (0.0180)
$\ln L$	−7411.79	−7423.52	−7426.71
AIC	14839.58	14863.04	14869.42

Table 4: *PE for* $\ln X_1 \sim TN$ *and* $\ln X_2 \sim TN_{mixt}$ *and different upper truncation limits (standard errors between parentheses).*

	max	max×10	max×100
μ_1	6.4439 (0.0587)	6.4437 (0.0553)	6.4437 (0.0587)
μ_{21}	4.3115 (0.1274)	4.1975 (0.0801)	4.2743 (0.1560)
μ_{22}	4.4105 (0.2229)	5.0547 (0.2746)	4.4769 (0.2594)
σ_1	1.3351 (0.0416)	1.3350 (0.0415)	1.3350 (0.0415)
σ_{21}	1.1476 (0.1184)	1.3346 (0.0671)	1.2315 (0.1508)
σ_{22}	1.9488 (0.2144)	1.9550 (0.1994)	2.0372 (0.5330)
r	0.5899 (0.1580)	0.7770 (0.0587)	0.6396 (0.3122)
ω	0.0309 (0.0095)	0.0212 (0.0086)	0.0161 (0.0082)
$\ln L$	−7419.39	−7426.13	−7428.70
AIC	14854.79	14868.25	14873.40

Table 5: *GE for $X_1 \sim TCh$ and $X_2 \sim TCh$ and different upper truncation limits (standard errors between parentheses).*

	max	max×10	max×100
α_1	1.3344 (0.0950)	1.3420 (0.0489)	1.3423 (0.0492)
α_2	1.1767 (0.0444)	1.1771 (0.0431)	1.1706 (0.0427)
H_1	631.1100 (36.2700)	623.2490 (35.6333)	619.3690 (35.1012)
H_2	76.8340 (4.9617)	77.7100 (5.0423)	78.2220 (5.1128)
ω	3.0290×10^{-8} (1.4497×10^{-8})	2.3070×10^{-9} (2.6310×10^{-9})	1.9540×10^{-10} (8.773163×10^{-10})
$\ln L$	−7419.63	−7434.21	−7436.87
AIC	14849.26	14878.42	14883.74

Table 6: *PE for $X_1 \sim TCh$ and $X_2 \sim TCh$ and different upper truncation limits (standard errors between parentheses).*

	max	max×10	max×100
α_1	1.3362 (0.0497)	1.3409 (0.0492)	1.3407 (0.0492)
α_2	1.1564 (0.0755)	1.1693 (0.0768)	1.1706 (0.0769)
H_1	624.1119 (35.6350)	623.3819 (35.4695)	623.5835 (35.4896)
H_2	78.9157 (6.2698)	78.3094 (6.2351)	78.2899 (6.2332)
ω	9.4918×10^{-9} (9.6420×10^{-9})	9.7283×10^{-10} (3.2613×10^{-9})	9.7508×10^{-11} (2.4705×10^{-9})
$\ln L$	−7435.49	−7437.15	−7437.26
AIC	14880.99	14884.31	14884.52

Table 7: *VaR values for several truncated Sarmanov distributions and different confidence levels using GE.*

Distribution	Confidence level q			
$\phi_i = x - E[X_i]$	0.95	0.99	0.995	0.999
$Log-Sarmanov$ $TN+TNmixt(\max)$	3484.592	11221.492	16469.954	34770.477
$Log-Sarmanov$ $TN+TNmixt(\max\times10)$	6703.136	18043.612	26888.181	62829.676
$Log-Sarmanov$ $TN+TNmixt(\max\times100)$	6363.582	15658.461	21929.474	44858.422
$Sarmanov$ $TCh+TCh(\max)$	3307.784	16192.401	27588.607	71445.821
$Sarmanov$ $TCh+TCh(\max\times10)$	6399.348	20755.411	34073.251	113319.114
$Sarmanov$ $TCh+TCh(\max\times100)$	6405.983	20868.242	34416.052	106865.442
Empirical values	7905.600	24821.140	28420.870	92112.930

Table 8: *TVaR values for several truncated Sarmanov distributions and different confidence levels using GE.*

Distribution	Confidence level q			
$\phi_i = x - E[X_i]$	0.95	0.99	0.995	0.999
$Log-Sarmanov$ $TN+TNmixt(\max)$	10722.435	23966.182	31500.887	55354.994
$Log-Sarmanov$ $TN+TNmixt(\max\times 10)$	15198.865	35093.589	48500.570	88741.558
$Log-Sarmanov$ $TN+TNmixt(\max\times 100)$	13184.039	28457.449	38574.933	75462.460
$Sarmanov$ $TCh+TCh(\max)$	14236.400	40314.549	56765.029	103080.945
$Sarmanov$ $TCh+TCh(\max\times 10)$	20317.585	59231.604	92295.169	244553.953
$Sarmanov$ $TCh+TCh(\max\times 100)$	21255.717	63750.746	101169.957	284614.349
Empirical values	20836.960	49453.170	73078.330	149791.000

Table 9: *VaR values for several truncated Sarmanov distributions and different confidence levels using PE.*

Distribution	Confidence level q			
$\phi_i = x - E[X_i]$	0.95	0.99	0.995	0.999
$Log-Sarmanov$ $TN+TNmixt(\max)$	6146.651	15182.363	20345.663	36692.056
$Log-Sarmanov$ $TN+TNmixt(\max\times 10)$	6499.495	16546.371	22518.986	46979.657
$Log-Sarmanov$ $TN+TNmixt(\max\times 100)$	6485.068	16408.995	22269.078	36658.346
$Sarmanov$ $TCh+TCh(\max)$	5943.208	19699.385	29599.685	77228.707
$Sarmanov$ $TCh+TCh(\max\times 10)$	6229.109	23097.386	38009.139	116412.203
$Sarmanov$ $TCh+TCh(\max\times 100)$	6237.787	23074.898	38462.701	141907.139
Empirical values	7905.600	24821.140	28420.870	92112.930

Table 10: *TVaR values for several truncated Sarmanov distributions and different confidence levels using PE.*

Distribution	Confidence level q			
$\phi_i = x - E[X_i]$	0.95	0.99	0.995	0.999
$Log-Sarmanov$ $TN+TNmixt(\max)$	12100.52	24887.46	32197.80	57251.34
$Log-Sarmanov$ $TN+TNmixt(\max\times 10)$	14002.47	31249.10	42864.60	91912.35
$Log-Sarmanov$ $TN+TNmixt(\max\times 100)$	15727.76	39941.33	60676.94	190957.41
$Sarmanov$ $TCh+TCh(\max)$	16015.06	40962.05	57834.45	95298.44
$Sarmanov$ $TCh+TCh(\max\times 10)$	20355.52	58995.50	89891.05	191242.66
$Sarmanov$ $TCh+TCh(\max\times 100)$	21601.87	64854.65	101058.90	222460.19
Empirical values	20836.960	49453.170	73078.330	149791.000

Acknowledgements

Zuhair Bahraoui and Catalina Bolancé acknowledge the support received from Spanish Ministry and FEDER grant ECO2013-48326-C2-1-P. We also thank the members and affiliates of Riskcenter at the University of Barcelona. The authors gratefully acknowledge the two anonymous referees for very helpful comments and recommendations.

References

Artzner, P., Delbaen, F., Eber, J. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9, 203–228.

Bahraoui, Z., Bolancé, C. and Pérez-Marín, M. (2014). Testing extreme value copulas to estimate the quantile. *SORT-Statistics and Operations Research Transactions*, 38, 89–102.

Bairamov, I., Altinsoy, B. and Kerns, G. (2011). On generalized Sarmanov bivariate distributions. *TWMS Journal of Applied and Engineering Mathematics*, 1, 86–97.

Bairamov, I., Kotz, S. and Gebizlioglu, O. L. (2001). The Sarmanov family and its generalization: theory and methods. *South African Statistical Journal*, 35, 205–224.

Bolancé, C., Bahraoui, Z. and Artís, M. (2014). Quantifying the risk using copulae with nonparametric marginals. *Insurance: Mathematics and Economics*, 58, 64–56.

Bolancé, C., Guillén, M., Pelican, E. and Vernic, R. (2008). Skewed bivariate models and non-parametric estimation for CTE risk measure. *Insurance: Mathematics and Economics*, 43, 386–393.

Bourden, R. and Faires, J. (2001). *Numerical Analysis*. 7th Edition. Brooks/Cole, Pacific Grove, CA.

Buch-Larsen, T., Guillén, M., Nielsen, J. and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39, 503–518.

Champernowne, D. (1952). The graduation of income distribution. *Econometrica*, 20, 591–615.

Cohen, L. (1984). Probability distributions with given multivariate marginals. *Journal of Mathematical Physics*, 25, 2402–2403.

Danaher, P. (2007). Modeling page views across multiple websites with an application to internet reach and frequency prediction. *Marketing Science*, 26, 422–437.

Farlie, D. (1960). The performance of some correlation coeficients for a general bivariate distribution. *Biometrika*, 47, 307–323.

Fisk, P. (1961). The graduation of income distribution. *Econometrica*, 29, 171–185.

Gómez-Déniz, E. and Pérez-Rodríguez, J. (2015). Closed form solution for a bivariate distribution in stochastic frontier models with dependent errors. *Journal of Productivity Analysis*, 43, 215–223.

Hashorva, E. and Ratovomirija, G. (2015). On Sarmanov mixed Erlang risks in insurance applications. *ASTIN Bulletin*, 45, 175–205.

Hernández-Bastida, A. and Fernández-Sánchez, M. (2013). A Sarmanov family with beta and gamma marginal distributions: an application to the Bayes premium in a collective risk model. *Statistical Methods & Applications*, 21, 391–409.

Hernández-Bastida, A., Fernández-Sánchez, M. and Gómez-Déniz, E. (2009). The net Bayes premium with dependence between the risk profiles. *Insurance: Mathematics and Economics*, 45, 247–254.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, New York.

Kotz, S., Balakrishnan, N. and Johnson, N. (2000). *Continuous Multivariate Distributions*. Vol.1: Models and Applications. John Wiley & Sons, Inc., New York.

Lee, M. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods*, 25, 1207–1222.

Mladenovic, N. and Hansen, P. (1997). Variable neighborhood search. *Computers and Operations Research*, 24, 1097–1100.

Nelsen, R. (2006). *An Introduction to Copulas*. 2nd ed. Springer, New York.

Pelican, E. and Vernic, R. (2013). Maximum-likelihood estimation for the multivariate Sarmanov distribution: simulation study. *International Journal of Computer Mathematics*, 90, 1958–1970.

Sarmanov, O. (1966). Generalized normal correlation and two-dimensional Frechet classes. *Doclady (Soviet Mathematics)*, 168, 596–599.

Shubina, M. and Lee, T. (2004). On maximum attainable correlation and other measures of de- pendence for the Sarmanov family of bivariate distributions. *Communications in Statistics: Theory and Methods*, 33, 1031–1052.

Vernic, R. (2014). On the distribution of a sum of Sarmanov distributed random variables. *Journal of Theoretical Probability*, Forthcoming.

Yang, Y. and Hashorva, E. (2013). Extremes and products of multivariate AC-product risks. *Insurance: Mathematics and Economics*, 52, 312–319.

SORT 39 (2) July-December 2015, 231-252

On the interpretation of differences between groups for compositional data

Josep-Antoni Martín-Fernández[*,1], Josep Daunis-i-Estadella[1]
and Glòria Mateu-Figueras[1]

Abstract

Social polices are designed using information collected in surveys; such as the Catalan Time Use survey. Accurate comparisons of time use data among population groups are commonly analysed using statistical methods. The total daily time expended on different activities by a single person is equal to 24 hours. Because this type of data are compositional, its sample space has particular properties that statistical methods should respect. The critical points required to interpret differences between groups are provided and described in terms of log-ratio methods. These techniques facilitate the interpretation of the relative differences detected in multivariate and univariate analysis.

MSC: 62F40, 62H15, 62H99, 62J10, 62J15, 62P25

Keywords: Log-ratio transformations, MANOVA, perturbation, simplex, subcomposition.

1. Introduction

Statistical offices around the world (e.g., Eurostat) state that "a time use survey measures the amount of time people spend doing various activities, such as paid work, household and family care, personal care, voluntary work, social life, travel, and leisure activities". This type of survey offers exhaustive information concerning the social habits and the everyday life of the population. The time use data are compiled by factors such as, among others, sex, age group, household composition, level of education, professional status, and day of the week. In consequence, the analysis of time use data across the groups defined by these factors is of crucial importance because it supports the de-

[*] *Address for correspondence:* Dept. Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona, Spain, josepantoni.martin@udg.edu

[1] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Spain.

Received: October 2014
Accepted: April 2015

velopment of family and gender equality policies. When one has a preliminary look at time use data, one states that they are *closed* data (Aitchison, 1986). That is, the total daily time expended on different activities by anyone one person is always equal to 24 hours. In addition, the units (hours or minutes) are not relevant when one describes the time spent on one activity. The interest then, is on the proportion of time, that is, the part of the day that people do an activity. According to Aitchison (1986), time use data is one example of *compositional data*.

Compositional data (CoDa) are quantitative descriptions of the parts or components of a whole, conveying exclusively relative information. Typical examples of compositions appear in geochemistry, environmetrics, chemometrics, budget expenses and data from time use surveys. In this latter case, the compositions are *closed*. On the other hand, if one were to analyse the solid waste composition in a household the CoDa would not be closed because the kilograms of waste vary between the families. In such cases, for convenience, compositions are commonly expressed in terms of proportions, percentages or parts per million (ppm) and to do this the *closure* operation is applied (Aitchison, 1986).

When an analyst decides to analyse a data set $\mathbf{X}$ $(n \times D)$ using compositional methods, he or she is assuming that the information collected is relative rather than absolute. In this sense, it holds that the information collected in any observation $\mathbf{x}$ is the same as in $\alpha\mathbf{x}$, for any scalar $\alpha > 0$, property known as *scale invariance* (Aitchison, 1986). However, in some cases the closure operation may be useful when the analyst is interested in the interpretation of some univariate statistics, such as percentiles. As the ratios rather than the absolute values are of interest, any function used to measure the difference between two compositions should be expressed in terms of ratios between variables. Indeed, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two compositions, the vector of ratios $(\frac{x_{11}}{x_{21}}, \ldots, \frac{x_{1D}}{x_{2D}})$ should play an important role when one interprets the difference between $\mathbf{x}_1$ and $\mathbf{x}_2$ (Aitchison and Ng, 2005). These ideas for dealing with CoDa were introduced in the early 1980s, when the use of logratios was proposed by Aitchison (1986). At the beginning of the current century, the use of orthonormal log-ratio coordinates was introduced in Egozcue et al. (2003). The critical concept of these approaches is that compositions have a natural geometry, known as the Aitchison geometry (Egozcue and Pawlowsky-Glahn, 2006), which is coherent with the relative scale of compositions.

Our interest is to compare the use of time between groups of people defined by factors such as professional status, level of education or municipality size. When one compares groups of data, descriptive techniques, models and the corresponding inferential techniques are commonly used. All of these elements have to be appropriate for the type of data collected. For example, some models consider that random multivariate observations $\mathbf{X}$ from a group are generated by *adding* a random variation or noise $\boldsymbol{\epsilon}$ around a fixed centre $\boldsymbol{\mu}$. Although the typical model $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ can be appropriate for interval scale data, this model would not be useful for ratio scale data such as CoDa. For example, in the case of data from a time use survey, when the centre $\boldsymbol{\mu}$ takes a small value for some activity, the resulting composition $\mathbf{x}$ may take negative values. Moreover, most common parametric and non-parametric methods for analysing differences

between groups deal with *variability* matrices: total, between and within. The critical idea is how to compare the variability *inside* the groups with the variability *between* the groups. For interval scale data these variabilities are measured using the typical *sum of squares* matrices. However, for CoDa these elements should be appropriately defined in terms of ratios (Aitchison, 1986). Multivariate analysis of variance (MANOVA) is the conventional name for the contrast of the equality of means in several groups (Wilks, 1932; Smith et al., 1962). MANOVA and its related parametric methods include inferential techniques based on the multivariate normal distribution. The approach known as the *principle of working on log-ratio coordinates* (Mateu-Figueras et al., 2011) suggested the definition of the normal distribution for CoDa (Mateu-Figueras et al., 2013). With these elements at hand, the MANOVA contrast can be coherently defined to the particular geometry of CoDa.

The main objective of this article is to provide the critical points required to interpret differences between groups for CoDa. In Section 2, some descriptive statistics and techniques for CoDa are presented. Section 3 provides the complete proposal of a compositional MANOVA contrast. The interpretation of multiple comparisons and related techniques are also described. The example that motivated this article is presented in Section 4, where all the elements introduced for interpreting differences are applied. Finally, in Section 5, some concluding remarks are provided. The programming of the data analyses discussed in this article was carried out using the open source R statistical programming language and software (R development core team, 2014) and the freeware CoDaPack (Comas-Cufí and Thió-Henestrosa, 2011). The computer routines implementing the methods can be obtained from the website *http://www.compositionaldata.com* and the R-package "zCompositions" (Palarea-Albaladejo and Martín-Fernández, 2014).

2. Compositional descriptive techniques

2.1. Logratio coordinates

According to the ratio scale nature of CoDa, any function $f(\cdot)$ applied to a composition $\mathbf{x}$ must verify that $f(\mathbf{x}) = f(\alpha\mathbf{x})$, for any $\alpha > 0$. In particular,

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_D) = f\left(x_k \left(\frac{x_1}{x_k}, \ldots, \frac{x_{(k-1)}}{x_k}, 1, \frac{x_{(k+1)}}{x_k}, \ldots, \frac{x_D}{x_k}\right)\right) =$$
$$= f\left(\frac{x_1}{x_k}, \ldots, \frac{x_{(k-1)}}{x_k}, \frac{x_{(k+1)}}{x_k}, \ldots, \frac{x_D}{x_k}\right), \quad \text{for } k = 1, \ldots, D.$$

In other words, any function should be expressed in terms of ratios between variables. Note that any ratio x_j/x_k is not symmetric and takes values in $(0, +\infty)$. However, a logratio $\ln(x_j/x_k)$ takes values in the full real space and is symmetric with respect to

the zero origin. Following Aitchison (1986), the general expression of a logratio is a log-contrast

$$a_1 \ln(x_1) + \ldots + a_D \ln(x_D) = \ln\left(\prod_{j=1}^{D} x_j^{a_j}\right), \tag{1}$$

where $\sum a_j = 0$, so as to verify the scale invariance property. A log-contrast is, in essence, a logratio of parts because for $a_j > 0$ the corresponding part x_j appears in the numerator, but if $a_j < 0$ it appears in the denominator, and for those parts that do no contribute to the logratio, then $a_j = 0$ holds. Importantly, log-contrast (1) has the same role as linear combinations of variables in classic statistics. Note that ratios and logratios can not be calculated when one of the parts takes the value zero. The treatment of this difficulty, also known as the *zero problem*, has recently been described in numerous articles. A reader interested in this topic will find a general description in Palarea-Albaladejo et al. (2014).

Using a log-contrast one can define new variables (e.g., latent variables or principal components) where the information collected in the original variables is combined. One example of the very useful new variables is the *centred log-ratio* (clr) defined in Aitchison (1986) by $\mathrm{clr}(\mathbf{x})_k = \ln \frac{x_k}{\left(\prod x_j\right)^{1/D}} = \ln x_k - \overline{\ln \mathbf{x}}, \quad k = 1, \ldots, D$. The log-contrast expression (1) of a clr-variable verifies that $a_{kj} = -1/D$ for $j \neq k$ and $a_{kk} = 1 - 1/D$. The clr variables, also known as clr coordinates, have another interesting interpretation: they are the log-coordinates centred by rows. Therefore, it holds that $\sum_{k=1}^{D} \mathrm{clr}(\mathbf{x})_k = 0$, indicating that the dimension of the clr coordinates' space is $D-1$. The critical element of the Aitchison geometry is the scalar product defined via the log-ratio coordinates. Indeed, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two compositions, then $< \mathbf{x}_1, \mathbf{x}_2 >_a = < \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) >_e$. Here the subscripts a and e represents respectively Aitchison and Euclidean metric elements. As usual, one can derive a distance and norm from the scalar product and finally obtain $d_a(\mathbf{x}_1, \mathbf{x}_2) = d_e(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2))$, and $||\mathbf{x}_1||_a = ||\mathrm{clr}(\mathbf{x}_1)||_e$. Remarkably, the Aitchison distance verifies that $d_a(\mathbf{x}_1, \mathbf{x}_2) = ||(\frac{x_{11}}{x_{21}}, \ldots, \frac{x_{1D}}{x_{2D}})||_a$, providing information about the relative difference between two compositions.

These metric elements are used to construct orthonormal basis and calculate the corresponding orthonormal log-ratio coordinates (Egozcue et al., 2003). The expression of these coordinates, known as isometric log-ratio coordinates (ilr), depends on the basis selected. Following Egozcue and Pawlowsky-Glahn (2005) one can define a particular ilr coordinates created through a sequential binary partition (SPB). According to equation (1), to make any logratio consists of selecting which parts contribute to the logratio and decide if these will appear in the numerator or in the denominator. In the first step of an SBP, when the first ilr coordinate is created, the complete composition $\mathbf{x} = (x_1, \ldots, x_D)$ is split into two groups of parts: one for the numerator and the other for the denominator. In the following steps, to create the following ilr coordinates, each group is in turn split into two groups. That is, in step k when the $\mathrm{ilr}(\mathbf{x})_k$ coordinate

is created, the r parts $(x_{n1},\ldots,x_{nr})$ in the first group are coded as +1 and placed in the numerator, and the s parts $(x_{d1},\ldots,x_{ds})$ in the second group will appear in the denominator and coded as -1. As a result, the $\mathrm{ilr}(\mathbf{x})_k$ is

$$\mathrm{ilr}(\mathbf{x})_k = \sqrt{\frac{r_k\, s_k}{r_k + s_k}} \ln \frac{(x_{n1}\cdots x_{nr})^{1/r_k}}{(x_{d1}\cdots x_{ds})^{1/s_k}},\ k = 1,\ldots,D-1. \tag{2}$$

where $\sqrt{\frac{r_k s_k}{r_k+s_k}}$ is the factor for normalizing the coordinate. The log-contrast expression (1) of a ilr-variable verifies that $a_{kj} = \sqrt{\frac{s_k}{r_k(r_k+s_k)}}$ if the part x_j is placed in the numerator, and $a_{kj} = -\sqrt{\frac{r_k}{s_k(r_k+s_k)}}$ for parts that appear in the denominator.

The metric elements can be also expressed in terms of ilr coordinates (e.g., $d_a(\mathbf{x}_1,\mathbf{x}_2) = d_e(\mathrm{ilr}(\mathbf{x}_1),\mathrm{ilr}(\mathbf{x}_2))$) as these coordinates are equal to the clr coordinates expressed on an orthonormal basis (Egozcue et al., 2003). The most important point here is that, once an orthonormal basis has been chosen, all standard statistical methods can be applied to the log-ratio coordinates and transferred to the simplex preserving their properties (Mateu-Figueras et al., 2011). The log-ratio approach proposed by Aitchison (1986) and the proposal to work on log-ratio coordinates do not differ substantially. In fact, the only distinction is the recommended use of orthonormal (or ilr) coordinates in the latter approach instead of the use of clr transformed vectors (see Mateu-Figueras et al. (2011) for an in-depth discussion). Note that, when a statistical method is applied to the ilr coordinates, one must analyse whether the results depend on the particular orthonormal basis selected. In other words, one must assure oneself that the interpretations are invariant under changes of basis. In this scenario, the advantage of the ilr coordinates created by an SBP is the interpretation of results and the corresponding CoDa-dendrogram, described in the following section.

2.2. Descriptive statistics and plots

Most of the multivariate methods for dealing with groups are based on location and spread (shape) descriptive statistics. In this paper we focus on the common centre and variability elements, accordingly modified to take into account the Aitchison geometry. Let $\mathbf{X}$ be a random composition. In practical terms, Pawlowsky-Glahn and Egozcue (2001) stated that the centre $\boldsymbol{\mu}$ is the geometric mean of $\mathbf{X}$, whose ilr coordinates $\mathrm{ilr}(\boldsymbol{\mu})$ are, respectively, equal to the arithmetic mean of $\mathrm{ilr}(\mathbf{X})$. The covariance of $\mathbf{X}$ is $\mathrm{cov}_\mathrm{a}[\mathbf{X}] = \boldsymbol{\Sigma} = \mathrm{cov}[\mathrm{ilr}(\mathbf{X})]$. In consequence, a comparison of g groups with respect to its location will be based on the comparison between the corresponding g centres $\boldsymbol{\mu}_j,\ j = 1,\ldots,g$. Indeed, for a data set with n compositions distributed in g groups, we can use the matrices total ($\mathbf{T}$), between-groups ($\mathbf{B}$), and within-groups sums of squares matrices ($\mathbf{W}$). These matrices verify the variability decomposition property: $\mathbf{T} = \mathbf{B} + \mathbf{W}$. An *important* contribution of matrix $\mathbf{B}$ in this equality suggests that there are relevant

differences between the groups with respect to its location. The approach used for evaluating this *importance* means techniques differ from one another. In this article we focus on the MANOVA contrast that evaluates this contribution using measures based on the trace, determinant and eigenvalues of these matrices. To compare the spread of the groups, for example to evaluate the homoscedasticity, one will compare the within-groups sum of squares matrices. Besides the location case, there are few techniques for evaluating differences with respect to the variability.

For example, we considered that a first group is the data set *statistician's time budget* (Aitchison, 1986) formed by 20 compositions with six parts (T,C,A,R,O,S), corresponding to time spent on daily activities: Teaching, Consulting, Administration, Research, Other, and Sleep. Next, we generated a second group perturbing the 20 samples multiplying them component-wise by the vector $(1.2,1,1,1,1,1)$, that is, increasing by 20% the first activity ratio against the other activities. Finally, we created a third group perturbing the initial 20 samples by the vector $(1,1,1,1.3,1,1)$ to increase by 30% the fourth component ratio against the other components. Hereinafter, we refer to the whole CoDa set as the *ST3* data set. Note that the three groups have the same covariance matrix because the second and third groups were created by perturbing the first group (Aitchison, 1986). Table 1 shows the unitary representative of the centre ($\boldsymbol{\mu}$) of the whole ST3 data set and the centres ($\boldsymbol{\mu}_j,\ j=1,2,3$) of the three groups. As expected, the larger differences occur in parts T and R.

Table 1: *Centres in ST3: for the whole data set ($\boldsymbol{\mu}$) and for the three groups ($\boldsymbol{\mu}_j$).*

Centre	T	C	A	R	O	S
$\boldsymbol{\mu}$	3.69	2.47	2.93	2.90	5.45	6.56
$\boldsymbol{\mu}_1$	3.54	2.52	2.98	2.71	5.56	6.69
$\boldsymbol{\mu}_2$	4.13	2.44	2.90	2.63	5.40	6.50
$\boldsymbol{\mu}_3$	3.42	2.43	2.89	3.41	5.38	6.47

The geometric mean barplot (Figure 1) is an option for describing differences between groups. Given a CoDa set, for each group, the logratio between the whole geometric mean and the geometric mean of the group is calculated. Finally, each component is represented in a barplot using a logarithmic scale. If the centre of the group is equal to the whole centre, the ratio of each component is one and the corresponding logarithm is zero. If one part of the centre of the group is greater or smaller than the corresponding part of the whole centre, then the ratio is different than one and the corresponding logarithm is respectively positive or negative. Indeed, large bars (positive or negative) will indicate large differences in the means. Figure 1 shows that in the parts C, A, O, and S, the differences between the groups and the whole data set are not relevant. The samples from the second group have large values in part T, whereas they take small values in the rest of the parts. A similar situation occurs in the third group

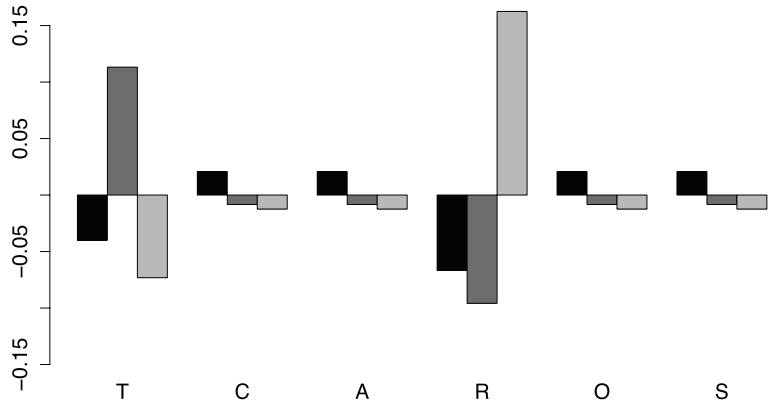


Figure 1: *Geometric mean barplot for ST3 data set: first group (black), second group (dark gray), third group (light gray).*

with respect to part R. Note that, for example, when a bar in one part is larger than 0.15, one can interpret that, on average, the samples of this group are in this part 16.18% ($\exp(0.15) = 1.1618$) larger than the whole centre.

To complete the basic description of a grouped CoDa set, one can represent the data using three specific plots: a biplot, a canonical variates plot and a coda-dendrogram. Aitchison and Greenacre (2002) adapted the typical biplot for CoDa and in doing so introduced the clr-biplot, that is, the biplot of clr coordinates. In other words, a clr-biplot draws on the same plot a projection of scores in the first two clr principal components together with the centred clr variables. Daunis-i-Estadella et al. (2011) described the interpretation of clr-biplots and introduced an extension for including supplementary elements. However, the statistical technique that underlies a biplot is not specially designed for highlighting differences between groups. In some cases, despite the groups being different, they appear mixed in the biplot. Figure 2(up) shows the clr-biplot of the ST3 data set. This representation is of a medium quality because the two first axes retain 61% of the variability. The samples of the first group are represented by circles. The compositions of the second group are shown by the triangles and shifted slightly to the positive part of the clr variable associated to the part T. On the other hand, as expected, the squares representing the third group are shifted to the positive direction of the clr-transformed part R. However, the samples from the different groups appear mixed, suggesting that there are no relevant differences between the groups.

As an alternative to the biplot, one can consider the canonical variates plot. Broadly speaking, a canonical variate is a new variable obtained as a linear combination of the original variables, where the linear combination attemps to highlight any differences between the g groups. For CoDa, we will use log-contrasts to create these new variables. Indeed, using ilr coordinates, a canonical variate y is equal to $\mathrm{y} = \mathbf{v}^{\mathsf{T}} \operatorname{ilr}(\mathbf{x}) = \sum_{j=k}^{D-1} v_k \operatorname{ilr}(\mathbf{x})_k$, that is, y is also a log-contrast. According to the general procedure, the first canonical variate y_1 is defined by the vector $\mathbf{v}_1$ which maximizes the F statistic

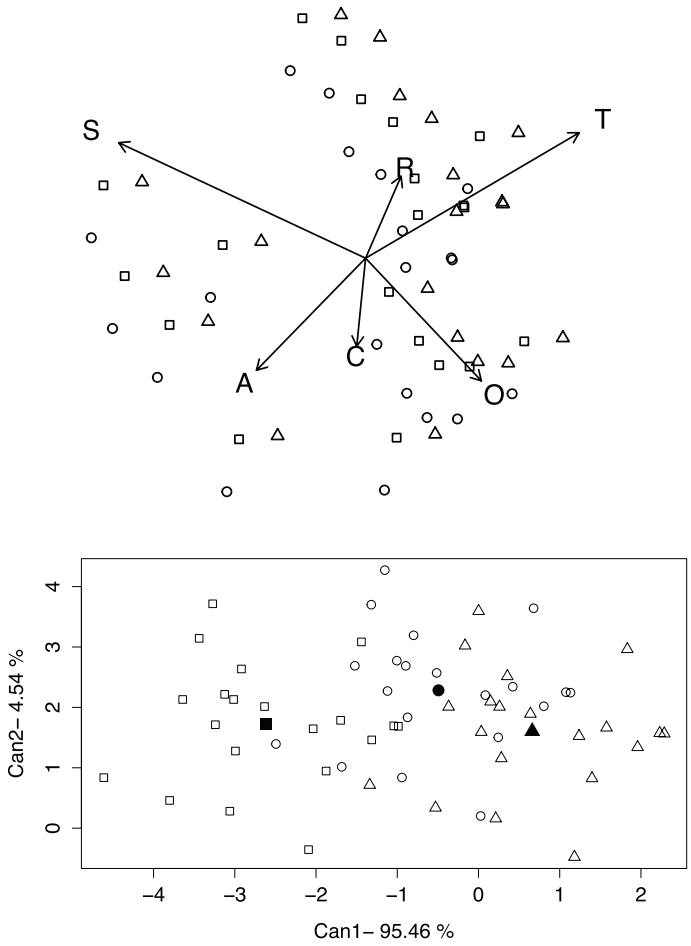


Figure 2: *Compositional plots for ST3 data set: clr-biplot (up) and canonical variates plot of ilr coordinates (down). Samples of the three groups are respectively represented by circles, triangles and squares.*

associated with the ANOVA test: $H_0 : \mathbf{v}_1^\mathsf{T} \boldsymbol{\mu}_1 = \ldots = \mathbf{v}_1^\mathsf{T} \boldsymbol{\mu}_g$ It could be proved that the vector $\mathbf{v}_1$ is the eigenvector of matrix $\mathbf{W}^{-1}\mathbf{B}$ associated to its maximum eigenvalue. Following this procedure iteratively, we can obtain the ordered $D-1$ eigenvectors that define the corresponding canonical variates. Importantly, if a change of basis is applied and the new ilr coordinates are $\mathbf{A}\,\mathrm{ilr}(\mathbf{x})$, with $\mathbf{A}$ an unitary matrix ($\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$), then taking $\mathbf{A}\,\mathbf{v}$, the same canonical covariate is obtained. In other words, the invariance under change of basis is guaranteed. Figure 2(down) shows the two first canonical variates plot for the ST3 data set. In addition, the centres of each group are represented by a filled symbol. The samples from different groups appear well separated, suggesting that there are relevant differences between the groups. In this case, the first eigenvector of matrix $\mathbf{W}^{-1}\mathbf{B}$ is $\mathbf{v}_1 = (3.75, -0.37, 10.31, 3.57, 0.31)^\mathsf{T}$ which, combined with the coefficients of log-contrast in equation (2), produces the first canonical variate

log-contrast with coefficients $\mathbf{a} = (6.33, 1.03, 4.13, -8.08, -3.13, -0.28)^{\mathsf{T}}$. To perturb only the first part of the samples of the first group by 1.2 is equivalent to adding $6.33\ln(1.2) = 1.15$ to the first canonical variate. The perturbation of the fourth part by 1.3 is equivalent to adding $-8.08\ln(1.3) = -2.12$ to the scores in the first canonical variate.

A CoDa-dendrogram is a descriptive plot for visualizing some univariate statistics of particular ilr coordinates created through a SPB (Pawlowsky-Glahn and Egozcue, 2011). Table 2 shows the complete SBP for the ST3 data set.

Table 2: *Sequential Binary Partition for ST3 CoDa set.*

$\mathrm{ilr}(\mathbf{x})_k$	T	C	A	R	O	S
$\mathrm{ilr}(\mathbf{x})_1$	1	1	1	1	−1	−1
$\mathrm{ilr}(\mathbf{x})_2$	1	−1	−1	1	0	0
$\mathrm{ilr}(\mathbf{x})_3$	1	0	0	−1	0	0
$\mathrm{ilr}(\mathbf{x})_4$	0	−1	1	0	0	0
$\mathrm{ilr}(\mathbf{x})_5$	0	0	0	0	1	−1

The SBP is represented by the dendrogram-type links between parts (Figure 3). The variability of each ilr coordinate is represented by the length of the vertical bars. Therefore, a short vertical bar, as in the case of $\mathrm{ilr}(\mathbf{x})_3$ and $\mathrm{ilr}(\mathbf{x})_4$ means that the ilr coordinate has a small variance. Conversely, when the ilr coordinate has a large variance its vertical bar is longer, as in $\mathrm{ilr}(\mathbf{x})_2$ which involves the parts T and R. The location of the mean of an ilr coordinate is determined by the intersection of the horizontal segment with the vertical segment (variance). When these intersections are not in the middle, this indicates a major contribution of one of the groups of parts. This is the case of $\mathrm{ilr}(\mathbf{x})_1$, where the intersection is close to parts O and S, according to its larger values (Table 1) with respect to the values in the other parts. In addition, the box-plot of the ilr coordinates is provided. Note that for the coordinate $\mathrm{ilr}(\mathbf{x})_3$, the box-plots are ordered according to the perturbation applied to create the corresponding group. One can analyse its symmetry and compare the median with the mean to interpret the symmetry of the corresponding univariate distribution. Figure 3 shows these statistics of the ilr coordinates for the three groups in ST3 data set. Note that there are no differences between the variances between the groups but there are differences between the means in the coordinates $\mathrm{ilr}(\mathbf{x})_1$, $\mathrm{ilr}(\mathbf{x})_2$, and $\mathrm{ilr}(\mathbf{x})_3$. For the coordinate $\mathrm{ilr}(\mathbf{x})_3$ there is a difference between the mean of first group (black colour) and the mean of second group (dark gray colour) and third group (light gray colour). This fact is consistent because $\mathrm{ilr}(\mathbf{x})_3$ evaluates the ratio between the parts T and R. The second group mean is close to part T and the mean of third group is close to R. This interpretation agrees with the construction of these groups by perturbing the first group.

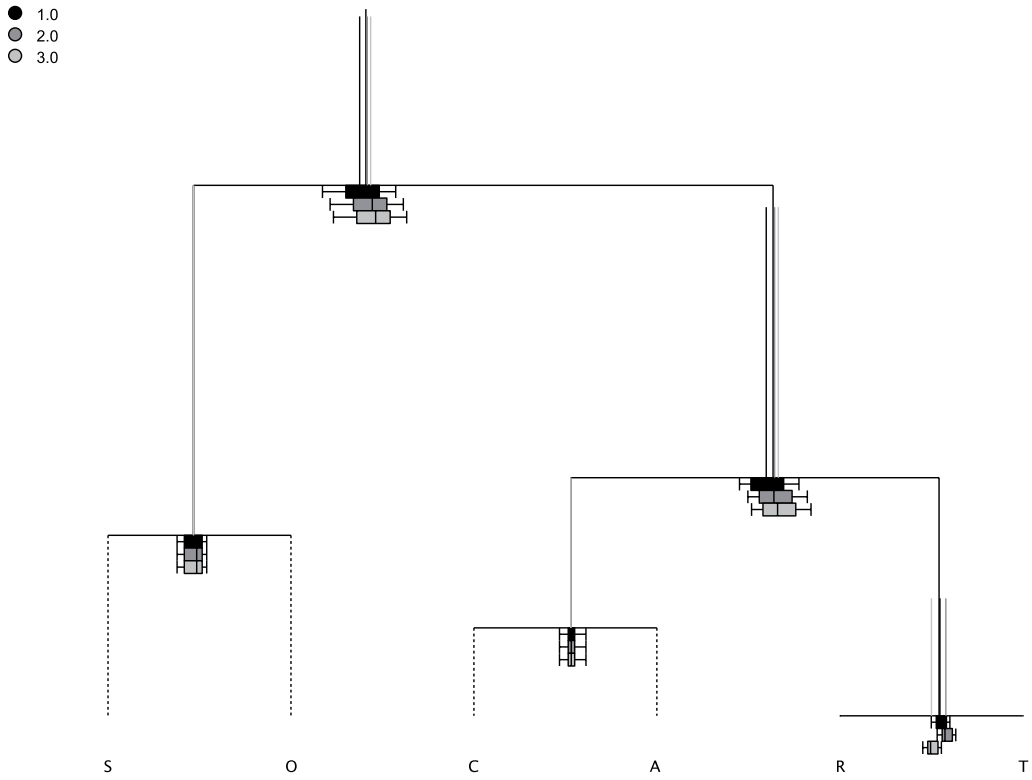


Figure 3: *CoDa-dendrogram for ST3 data set using the SBP from Table 2. Elements of three groups 1, 2, and 3 are distinguish by the colours black, dark gray and light gray, respectively.*

3. Inferential techniques to compare CoDa groups

3.1. MANOVA contrast for CoDa

In the statistical literature there are many inferential methods for comparing groups of data in the real space. Following Aitchison and Ng (2005), in this article we focus on the most basic methods to show how to proceed when one wants to be coherent with the Aitchison geometry. Other more sophisticated methods could be adapted accordingly to the compositional geometry by using an analogous procedure.

Let $\mathbf{X}_k$ be a random composition corresponding to the group k for $k = 1,\ldots,g$. The more basic model assumes that $\mathbf{X}_k$ is generated adding a random variability $\boldsymbol{\epsilon}_k$ around a centre $\boldsymbol{\mu}_k$ in a multiplicative part-wise way: $\mathbf{X}_k = \boldsymbol{\mu}_k \odot \boldsymbol{\epsilon}_k$. In this case, the expected value of variability $\boldsymbol{\epsilon}_k$ is the unit vector $\mathbf{1}$. Following Egozcue et al. (2003) and according to the principle of working on coordinates (Mateu-Figueras et al., 2011), this model is equivalent to $\mathrm{ilr}(\mathbf{X}_k) = \mathrm{ilr}(\boldsymbol{\mu}_k) + \mathrm{ilr}(\boldsymbol{\epsilon}_k)$, where $\mathrm{ilr}(\boldsymbol{\epsilon}_k)$ is centred at the origin of coordinates ($\mathbf{0}$). In other words, working on coordinates, one can assume the same model

as for interval scale data in the real space. From the different approaches for dealing with this type of model, in this article we focus on the MANOVA contrast (Wilks, 1932; Smith et al., 1962) and related techniques. The technical details of these methods are provided by the majority of books devoted to multivariate statistical techniques (e.g., Seber, 1984).

The critical assumption of the MANOVA contrast is that the ilr random variability $\mathrm{ilr}(\boldsymbol{\epsilon}_k)$ is homocedastic and normally distributed, that is, $\mathrm{ilr}(\boldsymbol{\epsilon}_k) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}); k = 1, \ldots, g$. Following Mateu-Figueras et al. (2013), this assumption is equivalent to assuming log-ratio normality and homocedasticity for the compositional term $\boldsymbol{\epsilon}_k$. In addition, the original null hypothesis $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$ is equivalent to the null hypothesis $H_0 : \mathrm{ilr}(\boldsymbol{\mu}_1) = \ldots = \mathrm{ilr}(\boldsymbol{\mu}_g)$. Therefore, the statistics of contrast will be based on the sum of square matrices $\mathbf{T}$, $\mathbf{B}$, and $\mathbf{W}$ calculated on ilr coordinates. The most common contrast statistics are: Wilks' Λ ($det(\mathbf{W})/det(\mathbf{T})$), Pillai's trace ($trace(\mathbf{B}\mathbf{T}^{-1})$)), Lawley-Hotelling trace ($trace(\mathbf{W}^{-1}\mathbf{B})$), and Roy's largest root of matrix $\mathbf{W}^{-1}\mathbf{B}$. Nowadays, the discussion over the merits of each statistic continues and the common software routines allow the four statistics to be calculated. In the case of two groups, the four statistics are equivalent and the MANOVA contrast reduces to Hotelling's T-square test. Importantly, the MANOVA contrast is invariant under a change of log-ratio basis because the four statistics are invariant functions of the eigenvalues of matrix $\mathbf{W}^{-1}\mathbf{B}$. This fact facilitates the use of the contrast because one can work with the ilr coordinates obtained from an SBP.

When the assumptions are accomplished the four statistics are associated to a value in an F probability distribution, permitting the calculation of a p-value for the MANOVA contrast. The first assumption, the homogeneity of variances and covariances ($H_0 : \boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g$), can be tested using the Box M test (Seber, 1984, p. 449). This test has been severely criticized because it is very sensitive to lack of normality, so that a significant value could indicate either unequal covariance matrices or non-normality or both. The general recommendation is to take a significant level less than 0.005. Nevertheless, if the number of subjects in each of the groups are approximately equal, the robustness of the MANOVA test is guaranteed and the impact, if the assumption of equal covariances is violated, is minimal (Johnson and Wichern, 2007). Noticeably, because Box M statistic is a function of the covariance matrices determinant, it could be proved that the results of the Box M test are invariant under changes of basis. Moreover, the equality of covariances can be descriptively checked using the CoDa-dendrogram. Because the vertical lines in the plot (Figure 3) represent the variance of each ilr coordinate, then we can evaluate if there are differences in the variances of each group. Figure 3 suggests that the variances of each ilr coordinate are equal because the three lines are of similar length. However, a CoDa-dendrogram only allows the variances of ilr coordinates to be compared, that is the diagonal of matrices $\boldsymbol{\Sigma}_k$. Despite being unusual, it could be that the random compositions $\mathbf{X}_k$ have equal variances but different covariances. That is, the matrices $\boldsymbol{\Sigma}_k$ have equal diagonals but the rest of elements are different. To investigate this case we propose to previously sphericize the data to plot the CoDa-dendrogram. A spherification, similar to the standardization in the univariate

case, consists of multiplying the residuals $\text{ilr}(\boldsymbol{\epsilon}_k)$ of each group by the squared root of $\mathbf{W}^{-1}$, the inverse of covariance matrix. After this transformation, if the homocedasticity is verified, the resulting covariances matrices in each group should be the identity matrix. In consequence, in the CoDa-dendrogram of spherized data, the vertical lines of all the groups for all the ilr coordinates should be equal.

The second assumption for the MANOVA contrast is the normality of the residuals $\boldsymbol{\epsilon}_k$. One can apply to the their ilr coordinates any of the multivariate normality tests that exist in the literature. For example, we can use the goodness-of-fit test suggested in Aitchison (1986) for compositional data. This test is based on the idea that, under the assumption of normality, the *radii* (or squared Mahalanobis distances) of the residuals are approximately distributed as a chi-squared distribution. We can use some empirical distribution function statistics, for instance Anderson-Darling or Cramer-von Mises, to test significant departures from the chi-squared distribution. Importantly, because Mahalanobis distances are invariant under change of basis, this normality test can be calculated using any ilr coordinates obtained with an SBP. To complete the analysis of this assumption, the normality can be explored using a typical Q-Q plot of the radii against the theoretical quantiles of a chi-squared distribution. Finally, according to Johnson and Wichern (2007), the assumption of normality in a MANOVA contrast can be relaxed when the sample sizes are large due to the multivariate version of the central limit theorem.

3.2. Analysing differences between groups

When MANOVA contrast or another test suggest rejecting the null hypothesis of equality of means, two questions immediately arise: (a) Which groups differ from the rest and (b) which variables are responsible of these differences? One common way to investigate the answer of these questions is by making the corresponding $g(g-1)/2$ comparisons between pairs of groups. Following the MANOVA approach, these comparisons can be analysed through the Hotelling's T-squared test, which is the multivariate generalization of a typical t-test. In this procedure, there is a general recommendation to avoid an artificial increase of the Type I Error rate: to adjust the alpha level of each test by making some kind of correction. Although there is no general agreement about the way to make this correction, a common technique is the Bonferroni correction for simultaneous tests (Seber, 1984). Following this technique one should modify the critical level α to $\alpha/(g(g-1)/2)$. However, this procedure tends to be conservative, especially when the number of comparisons is large. There are other more sophisticated procedures, such as the Scheffe's, the Tukey's or the Student-Newman-Keuls tests or another different approach provided by the FDR False Discovery Rate controlling method (Benjamini and Hochberg, 1995; Benjamini, 2010). The techniques for dealing with multiple comparisons are currently an open field whose development is beyond the scope of this article.

Regardless of this analysis, note that when using Hotteling's T-squared test on ilr coordinates the invariance under change of basis is guaranteed. In addition, these differences can be explored using the canonical variates plot, where one may also draw the corresponding confidence region for the mean and the predictive region for each group. With this plot one has a complete picture of the differences between groups, analysing if the corresponding regions overlap or not.

Once differences between two particular groups are detected, interest focuses on discovering if there are ilr coordinates responsible for these differences. That is, if the differences stated using multivariate techniques may be attributed to any separate variable. Following the previous approach, these univariate comparisons will be done through $D-1$ simultaneous t-tests. Again, in this case it is accordingly necessary to adjust the critical level or use a more complex technique. Remarkably, the results of these comparisons strongly depend on the ilr coordinates selected because one is making univariate t-test for a particular orthonormal basis. As a consequence, the choice of an interpretable SBP turns out to be crucial. For this analysis, the geometric mean barplot may be very useful for completing the interpretation of the univariate log-ratio differences because this plot allows all the parts to be compared directly. In addition, following Hesterberg et al. (2012), one can add the uncertainty associated to the geometric mean barplot using a bootstrap technique for comparing two populations. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two groups with n_1 and n_2 compositions, respectively:

1. Draw a resample of size n_1 with a replacement from the first group and a separate resample of size n_2 from the second group. Compute the centre of each group and calculate its log-ratio part-wise vector.
2. Repeat this resampling process B times (common B is 1000).
3. Construct the bootstrap confidence interval for each part of the log-ratio vector.

Note that the critical level of confidence intervals should be appropriately corrected. There are four common types of bootstrap confidence intervals: t, percentile, bias-corrected, and tilting. The description of their properties is provided by Hesterberg et al. (2012). For example, to calculate the $\alpha\%$ bootstrap percentile confidence interval, one should calculate the interval between the $\alpha/2$th and $(1-\alpha/2)$th percentiles of the bootstrap distribution of the corresponding part in the log-ratio vector. Regardless of which type of interval is calculated, if the interval for a part includes the value zero, it indicates that there is no difference between these groups with respect to this part. Only the parts with positive or negative intervals may be considered as responsible for the difference between the groups.

For the ST3 CoDa set we obtained p-values below 0.001 for the four statistics in the MANOVA contrast, indicating to us to reject the null hypothesis of equality of means. The radii normality tests based on the Anderson-Darling, on the Cramer von-Mises, and on the Watson statistic show p-values above 0.15, suggesting that the normal

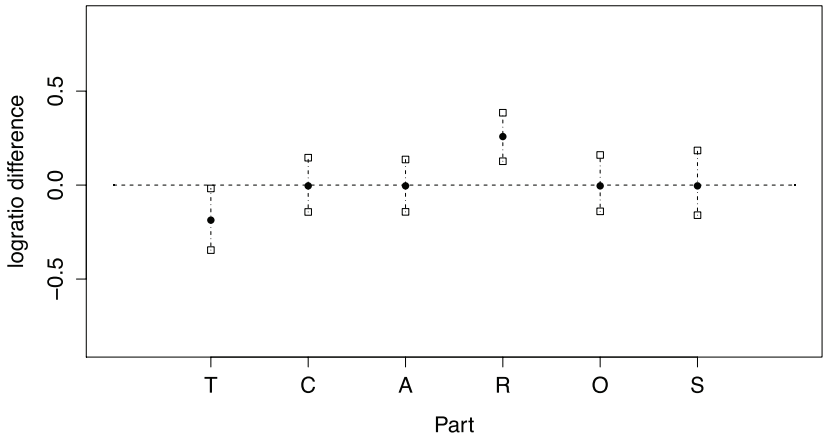


Figure 4: *Bootstrap percentile confidence intervals for log-ratio difference between centres of second and third group in CoDa set ST3. Filled circles are the log-ratio difference for the centres in ST3. Vertical dashed lines are the percentile intervals.*

probability distribution fits the data well. In this case, the Box M test is not necessary because by construction the three groups have the same covariance matrix. Figure 2(down) suggests that there are differences between the three groups, later confirmed by the three Hotelling T-squared simultaneous tests (p-value $< 0.05/((3\times 2)/2) = 0.0167$, Bonferroni correction). According to the construction of groups, the SBP in Table 2 suggests that the groups are equal in $\text{ilr}(\mathbf{x})_4$ and $\text{ilr}(\mathbf{x})_5$. They could only have relevant differences in the three first ilr coordinates. When the univariate ilr coordinates are analysed through simultaneous t-test, only $\text{ilr}(\mathbf{x})_3$ confirms the differences between the three groups because the larger p-value, obtained when comparing 1 and 2, was lower than the corrected alpha level (p-value $= 0.0034 < 0.0167/4 = 0.0042$, Bonferroni correction). These differences are associated to large values in activities T or R, or both. For example, we detected that the difference is on both parts when groups 2 and 3 were compared using the bootstrap percentile confidence intervals ($B = 1000$). Figure 4 shows these intervals for the corresponding corrected critical value. The filled circles represent the difference between the centres in the CoDa set ST3. The dashed vertical lines are the bootstrap percentile confidence intervals, whose extremes are the corresponding percentiles. Only the T and R intervals do not have the value zero (horizontal line), indicating that the univariate differences are confirmed.

4. Example

We used a data set kindly provided by the Statistical Institute of Catalonia (Idescat) in Catalonia, Spain. It consists of information collected using a face to face survey of 6471 persons aged 10 and over. Household participants were randomly selected according to

a double stratified sampling design to guarantee that the selected sample is a reflection of the general population. The survey provides information on the main activity the individual does during each of the 144 10-minute slots, which make up a day. On a primary level, the survey considers a list of 34 different possible activities. However, to better interpret the results, Idescat's official reports aggregate these activities into a minimum set of 5 main activities: personal care and sleep (CS), paid work and study (WS), household and family care (HF), social activities (SA), and commuting and others (CO). Moreover, the survey collected additional information related to many other aspects such as geographical area, municipality size, day of the week, household composition, sex, age, nationality or professional status. In consequence, any appropriate statistical analysis of the whole data set requires more general and complex methods. Because this type of analysis is beyond the scope of this article, we focused on a simple comparison between groups of data. In particular, to illustrate the CoDa techniques we attempted to solve the question: As regards to municipality size, are there any differences between time use composition of working people in a usual working day? Despite the fact that there are many other similar questions that might be analysed, this one is the most interesting to design regional polices. Note that, any other similar question may be analysed using an analogous procedure.

A preliminary inspection of data shows that the participant #1606 has a zero in the activity personal care and sleep. This participant is considered as an anomalous composition and it is removed accordingly from the data set. To obtain more realistic results we also removed those participants who's sampled day was an unusual day. That is, 668 participants who, for some unforeseen reason (illness, accident, or public holiday), did not carry out their usual activities and so were not included in the analysis. According to Idescat's reports, we considered that working days were from Monday to Thursday and working people, the participants that self-declared being in paid employment or studying. After the aforementioned steps, the sample size of the data set included in the analysis was reduced to 1051 participants. Only 253 from these compositions contain at least one zero, which represents an overall 5.2% of the values in the data set. The parts CS and WS have no zeros. The parts HF, SA and CO have respectively 18.55%, 4.28%, and 3.14% of their values equal to zero. According to the nature of these three parts, these zeros are considered as censored values consequence of the sampling design. Because of the data correspond to the main activity during a 10-minute slot, we assumed a threshold equal to 10 minutes for the censored values. These values were imputed using the log-ratio robust method based on a modified Expectation-Maximization algorithm (Palarea-Albaladejo et al., 2014; Palarea-Albaladejo and Martín-Fernández, 2014).

As regards to the size of the municipality, Idescat classified the participants into three groups: small, medium and large. Table 3 shows the number of inhabitants that define these sizes and the number of participants of each group. Remarkably, the three groups have comparable sample sizes.

Table 3: *Time Use data set: groups defined by municipality size.*

Size	Group numbering	Inhabitants limit	Participants
Small	1	<20000	369
Medium	2	[20000,100000]	311
Large	3	>100000	371

Figure 5 shows the ilr canonical variates, where the first one retains 77.93% of the variability. The participants who live in a small municipality (Group 1) are represented by circles. The triangles and squares represent participants of Groups 2 and 3, respectively. Participants appear mixed and no evidence of large differences between groups is detected. We used numbered circles to show the position of the geometric centre of each group. Apparently, Groups 1 and 2 have similar average values and the centre of Group 3 appears slightly separated.

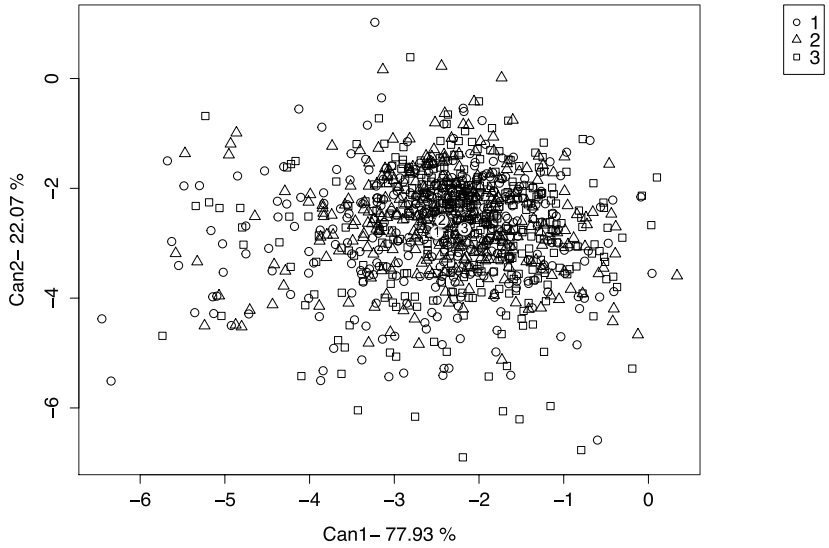


Figure 5: *Canonical variates plot of ilr coordinates for Time Use data set. Samples of the small (1), medium (2) and large (3) municipalities are respectively represented by circles, triangles and squares. The geometric centres of each group are accordingly represented by numbered circles.*

Using equation (2) the log-contrast coefficients of the first canonical variate are $\mathbf{a} = (-0.50, -0.83, -0.28, 0.34, 1.27)$. We can interpret the slight differences between Group 3 and the other groups in terms of an opposition between the three first parts (CS, WS, HF) and the parts (SA, CO). The largest weights correspond to WS and CO parts. This interpretation is coherent with the values shown in Table 4. The largest values the SA and CO parts are taken from the group from the large municipalities. To the contrary, Groups 1 and 2 take largest values in WS part. In summary, people from large

Table 4: *Centres of groups in Time Use data set: personal care and sleep (CS), paid work and study (WS), household and family care (HF), social activities (SA), and commuting and others (CO).*

Group	CS	WS	HF	SA	CO
1	10.97	8.72	1.15	1.89	1.28
2	10.66	8.74	1.19	2.11	1.31
3	10.80	8.50	1.02	2.13	1.55

municipalities expend more time on SA and CO, this time is subtracted to parts WS and HF.

These differences are illustrated in Figure 6. When compared with the whole geometric centre, the largest differences appear in HF, SA and CO. On the other hand, the barplot suggests that the values in CS and WS are very similar.

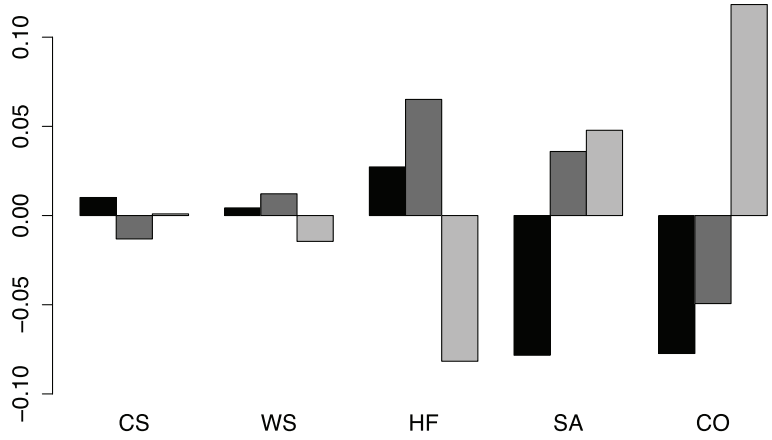


Figure 6: *Geometric mean barplot for Time Use data set: Group 1 (black), Group 2 (dark gray), and Group 3 (light gray).*

The MANOVA contrast confirms these differences because all the p-values provided by the common contrast statistics are lower than 0.05, where the largest p-value was 0.0001969, the value for the Wilks' Λ. When the groups were compared by pairs, the behaviour suggested in Figure 5 was confirmed. For Groups 1 and 2, the p-value was equal to 0.107. On the other hand, for Groups 1 and 3 was 0.000255 and for Groups 2 and 3, 0.003079, both lower than the Bonferroni correction level $0.05/3 = 0.0166$. For these cases, we investigated which log-ratio coordinate was contributing to these significant differences between groups. The alpha level was provided by the corresponding Bonferroni correction $0.05/(3 \times 4) = 0.0042$. After applying the t-test to the four coordinates for the data from Groups 1 and 3, we only obtained significant differences for the fourth coordinate (p-value $= 0.0006$). This behaviour was repeated when the data involved were from Groups 2 and 3 (p-value $= 0.0004$). According to the equation (2), the fourth coordinate provides information about the ratio of part CO over

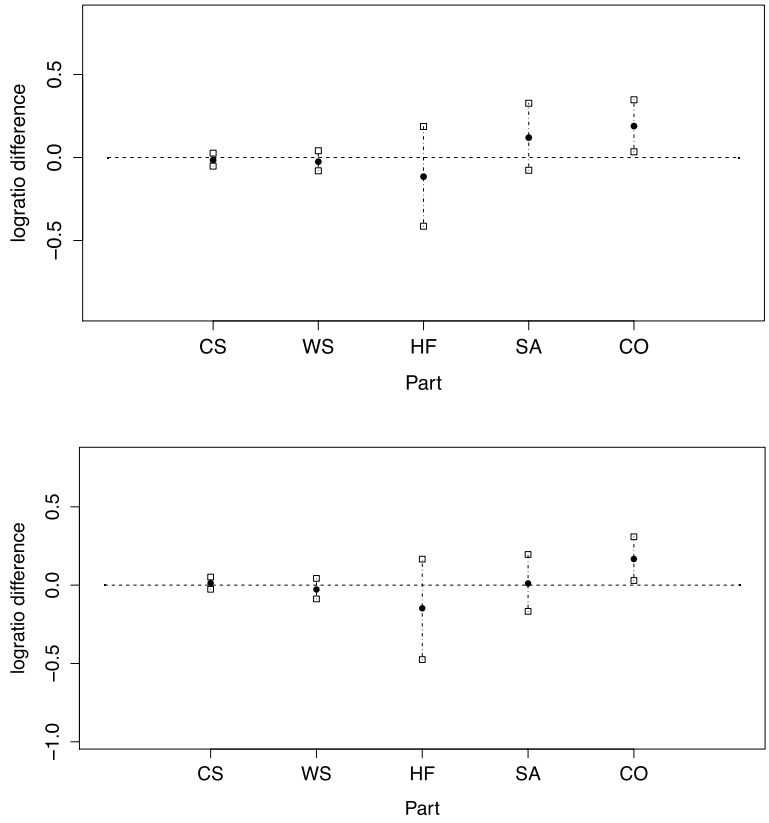


Figure 7: *Time Use data set. Bootstrap percentile confidence intervals for log-ratio difference between centres of: first and third group (up); second and third group (down). Filled circles are the log-ratio difference for the corresponding centres. Vertical dashed lines are the percentile intervals.*

the geometric mean of the other parts. Following this result, we investigated if the CO part was responsible of these differences. Figure 7(up) shows the bootstrap percentile confidence intervals when first and third groups are compared. The alpha level was provided by the corresponding Bonferroni correction $0.05/(3\times 5) = 0.0033$. The picture for the comparison between Groups 2 and 3 is shown in Figure 7(down). Both figures suggest the same behaviour, that is, the only significant difference appears in part CO. The percentile interval in both cases appears above the reference line. Because the log-ratio comparison uses the data from the third group in the numerator, this position means that participants in the third group take greater values than in the other two groups. In other words, people from large municipalities expend significantly more time on the Commuting and Others activities.

When the homogeneity of log-ratio variances and covariances was checked using the Box M test, we obtained a p-value equal to 0.8244. That is, we assumed that covariance matrices were not significantly different. On the other hand, when we applied the *radii* test to check the log-ratio normality of residuals, the p-value was lower than 0.01.

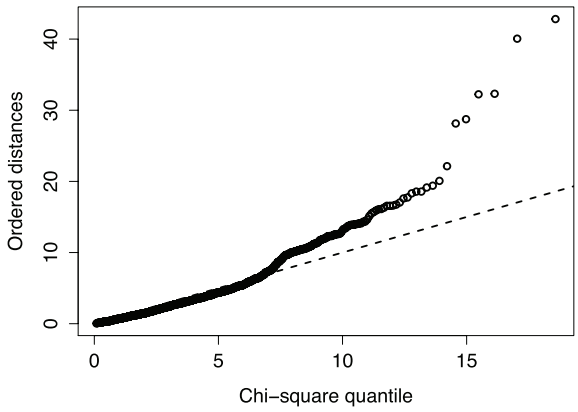


Figure 8: *Time Use data set. Chi-square plot of Manova contrast residuals.*

Figure 8 suggests that this lack of normality is caused by samples taking extreme values. This typical pattern is also evident in the histogram of the residuals shown in Figure 9. Despite the Gaussian shape of these histograms, the presence of extreme values may cause the lack of normality. In this case, due to our large sample size, the multivariate version of the central limit theorem guarantees the robustness of MANOVA results even the lack of normality. As indicated in Section 3.1, the non-normality could also affect the Box M test and a lower p-value could be obtained. However, this effect was not appreciated in our case because a large Box M test p-value was obtained.

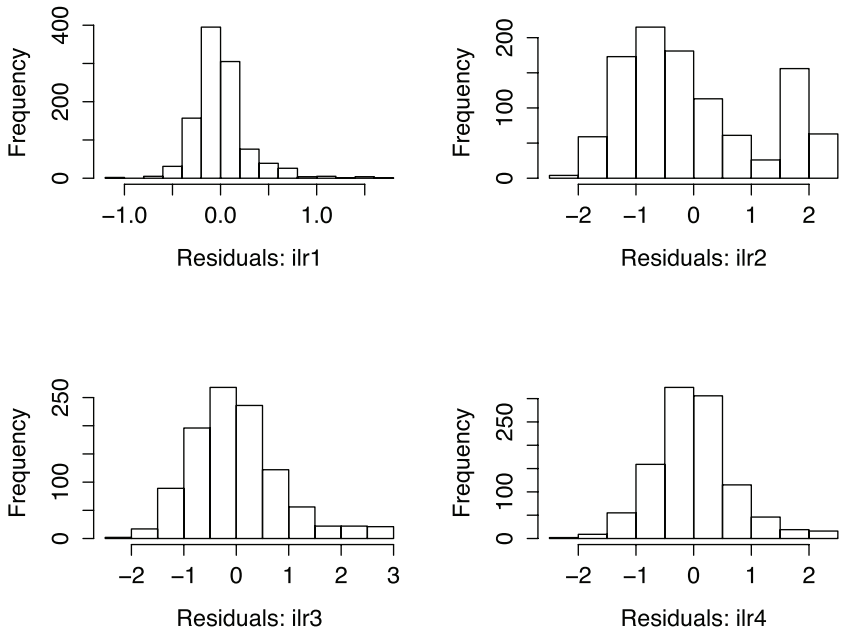


Figure 9: *Time Use data set. Histograms of Manova contrast residuals.*

5. Final remarks

Because time use data are compositional, any statistical analysis has to take into account their relative nature. This article fills the gap for basic methods for comparing groups of CoDa. We introduced descriptive techniques (log-ratio canonical variates and geometric mean barplots) for an initial exploration into the differences between groups. These differences can be confirmed by the typical inferential tools (MANOVA contrast). We introduced the bootstrap log-ratio percentiles to improve the interpretation of univariate differences and to complete the analysis of the log-ratio coordinates. Because most of these techniques are based on the principle of working on log-ratio coordinates, a detailed discussion of its invariance under change of basis was provided. The methods described assume normality and homocedasticity. When these assumptions are violated, another family of techniques should be explored, such as robust methods or distance based methods. These techniques should be applied accordingly to log-ratio coordinates to assure an appropriate analysis of the relative information collected in CoDa.

The Time Use data set provided by Idescat, is a complex data set that requires more sophisticated and general methods. However, we realised that no literature deals with these type of data using recent advances in CoDa analysis. The log-contrast approach provided in this article will be helpful to develop more complex methods, such as structural equation modelling. In addition, any general models for time use data have to include the presence of essential or structural zeros. These types of zeros represent absolute zeros, that is, it makes no sense to replace them by small values because they are not a consequence of the sampling design. The analysts should use their prior knowledge to decide what type of zero is present in a part. For example, survey participants that do not work or study have an essential zero in this part. On the other hand, in our example, after an appropriate amalgamation the zeros were considered as a consequence of sampling design. Because the greater the number of different activities are considered, the more zeros are collected, the appropriate amalgamation of parts is recommended. The development of these types of models is one of the more interesting challenges in current compositional data analysis.

Acknowledgments

This research was supported by the Ministerio de Economía y Competividad under the project “METRICS” Ref. MTM2012-33236; and by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2014SGR551. Finally, we would like to thank the Statistical Institute of Catalonia (Idescat) for kindly providing the Time Use data set.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of The Royal Statistical Society Series B*, 44, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London 416 pp. Reprinted in 2003 by Blackburn Press.

Aitchison, J. (2001). Simplicial inference. In: M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of Contemporary Mathematics Series, pp. 1–22. American Mathematical Society, Providence, Rhode Island (USA), 340 p.

Aitchison, J. and Greenacre, M. (2002). Biplots for compositional data. *Journal of The Royal Statistical Society Series C (Applied Statistics)*, 51, 375–392.

Aitchison, J. and Ng, K.W. (2005). The role of perturbation in compositional data analysis. *Statistical Modelling*, 5, 173–185.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society Series B*, 57, 289–300.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of The Royal Statistical Society Series B*, 72, 405–416.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96, 1205–1214.

Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds.) (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice.* Geological Society, London, Special Publication 264.

Comas-Cufí, M. and Thió-Henestrosa, S. (2011). CoDaPack 2.0: a stand-alone, multi-platform compositional software. In: Egozcue, J. J., Tolosana-Delgado, R., Ortego, M. I., eds. CoDaWork'11: *4th International Workshop on Compositional Data Analysis*. Sant Feliu de Guíxols, Spain.

Daunis-i-Estadella, J., Thió-Henestrosa, S. and Mateu-Figueras, G. (2011). Including supplementary elements in a compositional biplot. *Computers and Geosciences*, 37, 696–701.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37, 795–828.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, 264, 145–159.

Johnson, R. A. and Wichern, D. W. (2007) Applied Multivariate Statistical Analysis (6th Edition). Pearson Book, Prentice-Hall.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., Epstein, R., Craig, B. A. and McCabe, G.P. (2012). Bootstrap Methods and Permutation Tests. *Chapter 16 of Introduction to the Practice of Statistics*, 7th edition, W. H. Freeman, New York, 657 p.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology*, 35, 253–278.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. *Compositional Data Analysis: Theory and Applications*, Pawlowsky-Glahn, V. and Buccianti, A. eds., John Wiley & Sons, Chichester, 31–42.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2013). The normal distribution in some constrained simple spaces. *Statistics and Operations Research Transactions (SORT)*, 37, 29–56.

Palarea-Albaladejo, J., Martín-Fernández, J. A. and Olea, R. A. (2014). Bootstrap estimation of distributional statistics from compositional data with nondetects: a case study on coal ashes. *Journal of Chemometrics*, 28, 585–599.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2014). zCompositions: Imputation of zeros and nondetects in compositional data sets. *R package version 1.0.2*.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15, 384–398.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). Exploring compositional data with the CoDa-dendrogram, *Austrian Journal of Statistics*, 40, 103–113.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs, *Proceedings of the Royal Society of London*, 60, 489–502.

R development core team (2014). R: A language and environment for statistical computing: Vienna, http://www.r-project.org.

Seber, G. A. F. (1984). *Multivariate Observations*. Wiley, New York 685 pp. Reprinted in 2004 by Wiley.

Smith, H. , Gnanadesikan, R. and Hughes, J. B. (1962). Multivariate analysis of variance (MANOVA). *Biometrics*, 18, 22–41.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471–494.

SORT 39 (2) July-December 2015, 253-272

Robust project management with the tilted beta distribution

Eugene D. Hahn[1] and María del Mar López Martín[2]

Abstract

Recent years have seen an increase in the development of robust approaches for stochastic project management methodologies such as PERT (Program Evaluation and Review Technique). These robust approaches allow for elevated likelihoods of outlying events, thereby widening interval estimates of project completion times. However, little attention has been paid to the fact that outlying events and/or expert judgments may be asymmetric. We propose the tilted beta distribution which permits both elevated likelihoods of outlying events as well as an asymmetric representation of these events. We examine the use of the tilted beta distribution in PERT with respect to other project management distributions.

MSC: 62E15, 90B99, 62P30

Keywords: Activity times, finite mixture, PERT, tilted beta distribution, robust project management, sensitivity analysis.

1. Introduction

In project management it is important to be able to assess the total time for a project's completion. Since projects can be very complex, methodologies such as the Program Evaluation and Review Technique (PERT) (Malcolm et al., 1959) have been developed to assist in these assessments. PERT has been used for many decades but in recent years academics, managers, and policy makers have increasingly realized that conventional modeling approaches and tools may not be well-equipped to deal with extreme events. For example, few lenders would have predicted that the rise of lending to the sub-prime

[1] Department of Information and Decision Sciences, Salisbury University, Salisbury, MD 21801 USA. edhahn@salisbury.edu

[2] Department of Didactics of Mathematics. Campus Universitario Cartuja, s/n. 18071. University of Granada, Spain. mariadelmarlopez@ugr.es

Received: March 2015
Accepted: June 2015

market in the United States would cost them their own jobs, and fewer still would have predicted that it would have global repercussions. Hence there is a growing appreciation of the need for more robust models that assign greater probability to more extreme events.

Recent research by authors such as Hahn (2008) and López Martín et al. (2012) has described some ways for increasing the amount of distributional uncertainty in the context of project management tools such as PERT. The goal of this research stream has been to extend the PERT framework to accommodate greater likelihood to extreme tail-area outcomes. This has led to the ability to provide wider confidence intervals for activity and project duration times and hence more conservative results, while still retaining the classic PERT results as an important special case. The ability to increase distributional uncertainty is an important first step towards robust project management estimation; however, one consideration that has been underexplored is that one extreme may be more likely or more important than another. For example, as documented below, project managers tend to provide positively biased time estimates. Accordingly, project management tools which depend on these biased estimates are likely to underpredict the overall project time. In the current paper, we describe a new distribution that can be used by an independent agent (such as a risk manager) to differentially weight high versus low extremes. This can be used to help counteract some biased estimates.

This paper is structured as follows. Firstly, we review the literature about alternative distributions used in the area of PERT methodology. In Section 3 we present the tilting distribution, as a particular case of generalized Topp and Leone distribution; the tilted beta distribution, as a mixture between the tilting and beta distributions, and some stochastic characteristics. The elicitation for the distribution is presented in Section 4. The results are illustrated with an example in Section 5. Finally, Section 6 summarizes the main conclusions.

2. Literature

Projects often fail to meet various financial and scheduling targets despite management's best efforts to ensure success. For example, Bevilacqua et al. (2009) report on budget overruns and non-completion of tasks in projects undertaken with the use of PERT methodologies in the energy sector. Hence, there have been numerous studies which have tried to understand the sources of the persistent problem of project management overestimates or underestimates. Boulding et al. (1997) find that senior level executive subjects tended to ignore negative information or distort the information to fit preconceived notions and decisions. Hill et al. (2000) find that expert project managers sometimes overestimate and sometimes underestimate project durations, but that the underestimates were greater in magnitude leading on average to underestimation. Keil et al. (2007) conducted a laboratory experiment which revealed that failure to recognize problems early also leads to over-optimistic assessments regarding information technology

projects. Snow et al. (2007) describe an in-depth research program on the assessment of biases among software development project managers. The most commonly reported reason for giving optimistic judgments was to avoid being the bearer of bad news. Of the 56 surveyed project managers, 22% mentioned providing optimistic judgments because of a belief that senior management "shoots the messenger" while another 22% indicated that optimistic judgments were provided so as to make the project manager look good. Project managers were also two times more likely to be optimistically biased than they were to be pessimistically biased. Snow et al. (2007) conclude "optimistic bias leads to status reports that are very different from reality, while pessimistically biased status reports tend to be accurate because bias offsets error". Iacovou et al. (2009) also found that optimistically-biased reports were more prevalent than pessimistically biased ones in a sample of 390 information systems project managers, a finding consistent with work by Smith et al. (2001) and Gillard (2005). In a related vein, project managers who are able to accurately assess the risks of a troubled project are more likely to discontinue the project (Keil et al., 2000). Sengupta et al. (2008) conducted research on several hundred project managers and found that managers seem to strongly anchor on the initial risk assessment, and find it difficult to update their opinion with new information that should have prompted a re-assessment. One of the mitigation strategies identified by Sengupta et al. (2008) was better calibration of forecasting tools to project particulars.

Given the large volume of research which indicates project managers may tend toward having an optimistic bias, one possible solution is to provide a system whereby a third party (such as a risk manager) can provide an outside independent review to help remove bias in estimates. Öztaş and Ökmen (2005) describe a project management methodology called the judgmental risk analysis process that is explicitly pessimistic in nature. This is implemented by assessing a separate risk factor for each activity and assigning a probability distribution to the risk factor. The minimum and maximum activity times are then modified by including additive and subtractive offsets based on the activity risk factor to these activity times. Here we observe that if managers or other experts tend to overemphasize optimistic information, then counteracting this is a matter of de-emphasizing or downweighting optimistic information. In the current paper, we provide a probabilistic approach that permits a risk manager to introduce a negative (or positive) weighting across the range that nonetheless retains the usual PERT framework as a special case. This is accomplished by the introduction of a new distribution called the tilted beta distribution. In the following section, we present this distribution and explore some of its main properties which are relevant for project management. We then present an application of this distribution and conclude with discussion.

3. Distributions for project activity times

Malcolm et al. (1959) were the first to use the beta distribution to describe project management activity times. The beta distribution is the most prevalent distribution used in

stochastic project management due to its useful properties and appearance in the seminal work of Malcolm et al. (1959). Other widely used continuous probability models within the PERT methodology are the triangular distribution (Clark, 1962; MacCrimmon and Ryavec, 1964; Moder and Rodgeres, 1968; Vãduva, 1971; Megill, 1984; Williams, 1992; Keefer and Verdini, 1993; Johnson, 1997), the trapezoidal distribution (Pouliquen, 1970; Herrerías and Calvete, 1987; Herrerías, 1989; Powell and Wilson, 1997; Garvey, 2000), the doubly truncated normal distribution (Kotiah and Wallace, 1973), the uniform distribution (Suárez, 1980; Romero, 1991), the generalized biparabolic distribution (García et al., 2010) and the Parkinson distribution (Trietsch et al., 2012).

More recently, the literature on distributions for project management activity times has emphasized the importance of accounting for heavy tails and assigning more probability density to extreme values (Mohan et al., 2007). In addition, the research emphasis has moved away from using 'off the shelf' statistical distributions and instead has sought to engineer new distributions that are tailored to satisfy PERT desiderata. For example Hahn (2008) proposed the beta rectangular distribution which is a bounded distribution like the beta but assigns greater density to extreme values and can accommodate very heavy tails. Similarly, García et al. (2010) presented the generalized biparabolic distribution and demonstrated its capacity to have larger variances than the beta distribution. Our motivation in writing this paper is to engineer a distribution to address the documented optimistic bias discussed above while also addressing the need for heavy tails and large variances which has been identified previously.

3.1. Beta distribution

We begin with a brief review of the beta distribution given its importance in both the current work and in the project management literature. The standard beta distribution defined on $[0,1]$ has the following probability density function (pdf)

$$p(x|\alpha,\beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

It is necessary that both $\alpha > 0$ and $\beta > 0$ for (1) to be a valid pdf. The mean of (1) is $\alpha/(\alpha+\beta)$ while the variance is $(\alpha\beta)/\left((\alpha+\beta)^2(\alpha+\beta+1)\right)$.

The beta distribution is capable of a variety of shapes (see distributions having dotted lines in Figure 3). Unfortunately the beta distribution does not provide a great deal of flexibility when it is of interest to preserve the typically-preferred unimodal shapes but assign higher probability to extremal (or 'tail area') events. This observation led Hahn (2008) to propose the beta rectangular distribution, which is a mixture of the beta and rectangular distributions, for project management activity times and it is defined by

$$p(x|\alpha,\beta,\theta) = \begin{cases} \theta \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} + (1-\theta) & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where θ is a mixing parameter such that $0 \le \theta \le 1$.

Under the PERT conditions (to be discussed in Section 4), the beta rectangular distribution permits larger variances than the beta distribution and allows for elevated tail-area density (see Figure 1. The beta rectangular also has the beta distribution as a special case; hence, the classic PERT activity time parameters can be easily obtained as a particular case.

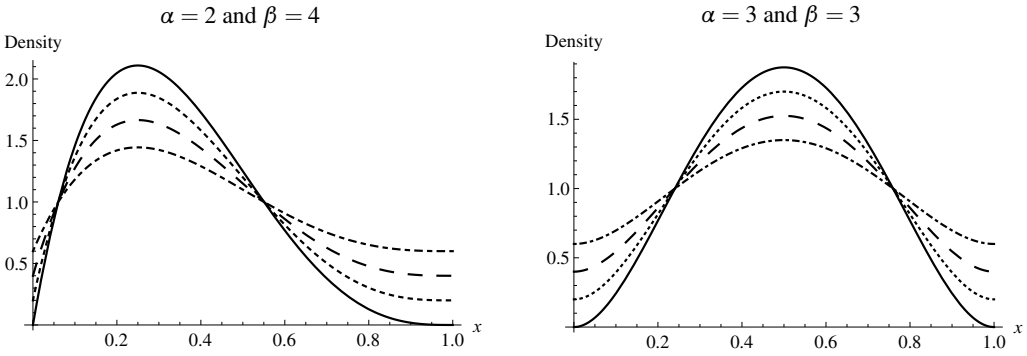


Figure 1: *Examples of beta rectangular distribution for $\theta = 1$ (solid), $\theta = 0.8$ (dotted), $\theta = 0.6$ (dashed) and $\theta = 0.4$ (dash-dotted).*

However, the previous discussion of Section 2 indicates that project managers may have an optimistic bias and the beta rectangular does not provide a way for addressing this issue. The remainder of this section is dedicated to formulating a distribution that addresses this issue. Accordingly next we describe the tilting distribution which allows for a straightforward way of expressing an optimistic (or pessimistic) bias. This will in turn allow us to construct the tilted beta distribution whereupon we will study in depth the characteristics of this distribution.

3.2. Tilting distribution

Topp and Leone (1955) present a distribution with probability density function (pdf) defined by $f(x,\beta) = \beta(2-2x)(2x-x^2)^{\beta-1}$, where $x \in [0,1]$ and $\beta > 0$. Depending on the values of β the distribution either has a J-shaped form ($0 < \beta < 1$); is unimodal ($\beta > 1$); or is left-triangular ($\beta = 1$). Kotz and van Dorp (2004) present a generalization of the Topp and Leone distribution by considering a slightly more general generating pdf, whose expression is

$$g(x|\alpha) = \alpha - 2(\alpha - 1)x, \tag{3}$$

where α defined on the interval $[0,2]$. The authors define the slope distributions as the distributions with pdf of the form (3).

Taking the reparametrization $\alpha = 2v$, we introduce a distribution, called the tilting distribution, which has the following density function:

$$p(x|v) = \begin{cases} 2v - 2(2v-1)x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

As $0 \le \alpha \le 2$, the parameter v is defined on the interval $[0,1]$. The reparameterization considered here leads to a parameter range consistency for the tilted beta, as shown later.

The cumulative density function (CDF) of (4) is

$$F(x|v) = \begin{cases} 0 & \text{if } x < 0, \\ 2vx - (2v-1)x^2 & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x > 1. \end{cases} \tag{5}$$

Graphical examples of density and cumulative density function of the tilting distribution are shown in Figure 2. Figure 2 reveals that when $v = 1/2$ the uniform distribution is obtained and when $v = 0$ or $v = 1$, a triangular distribution with mode v is obtained.

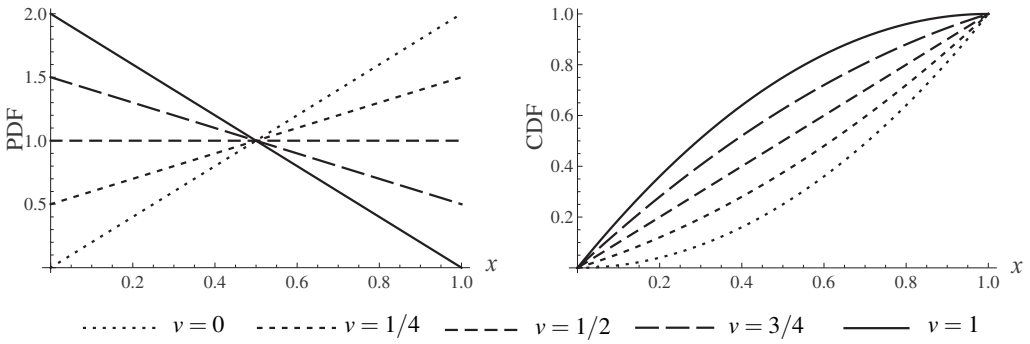


Figure 2: *Examples of PDF and CDF of tilting distributions.*

The mean, variance and coefficient of skewness of the tilting distribution are respectively

$$E(X) = \frac{2-v}{3}, \quad \operatorname{var}(X) = \frac{2v(1-v)+1}{18}, \quad \beta_1 = \frac{2\sqrt{2}\left(1+3v-15v^2+10v^3\right)}{5\left(1+2v-2v^2\right)^{3/2}}. \tag{6}$$

Taking the first derivative of $\text{var}(X)$ with respect to v the variance of the distribution is maximized for $v = 1/2$. When $v = 0$ or $v = 1$ the variance of the distribution is minimum whilst the coefficient of skewness is maximum.

In some contexts it can be necessary to work with a variable defined over the more general support $[a,b]$ instead of $[0,1]$. For cases where the variable may take on an arbitrary location and scale, we describe the variable $Y = a + (b-a)X$ with $b > a$. The inverse function is $X = \frac{y-a}{b-a}$ with Jacobian $\partial X/\partial y = \frac{1}{b-a}$. Then, the density function of (4) with to support $[a,b]$ is

$$p(y|w,a,b) = \frac{2}{b-a}\begin{cases} \dfrac{w-a}{b-a} - \left(2\dfrac{w-a}{b-a} - 1\right)\left(\dfrac{y-a}{b-a}\right) & \text{if } a \le y \le b, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $w = a + (b-a)v$. The quantile function of Y is

$$P^{-1}(q|w) = \begin{cases} \dfrac{a(2b-w) - bw + (b-a)\sqrt{a^2(1-q) + b^2 q - 2aw(1-q) - 2bwq + w^2}}{a - 2w + b} & \text{if } w \neq \frac{a+b}{2}, \\ a + (b-a)q & \text{if } w = \frac{a+b}{2}, \end{cases} \tag{8}$$

with $0 < q < 1$.

Although the introduction of additional parameters is associated with an increased complexity for the distributional expressions, the increase in flexibility makes it worthwhile to briefly summarize a few key expressions. In this case, the mean and variance of the tilting distribution are

$$E[Y] = \frac{2a - w + 2b}{3}, \quad \text{var}[Y] = \frac{a^2 + 2(a+b)w + b^2}{6} - \frac{(a + w + b)^2}{9}. \tag{9}$$

3.3. Tilted beta distribution

Having presented a few key properties of (4), we now introduce the tilted beta distribution. The density function of a random variable X having the tilted beta distribution with $\alpha > 0$, $\beta > 0$, $v \in [0,1]$, and $\theta \in [0,1]$ is

$$p(x|v,\alpha,\beta,\theta) = \begin{cases} (1-\theta)\left[2v - 2(2v-1)x\right] + \theta\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\right] & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

This distribution can be seen as a mixture of the tilting and beta distribution. The parameter θ indicates the relative proportionality of the tilting distribution to the beta distribution and v can be interpreted as the relative tilt proportionality. When $\theta = 1$

the beta distribution is obtained, for $\theta = 0$ we obtain the tilting distribution of (4) and $\theta = 1/2$ indicates a balance between the two distributions. With respect to the parameter v, $v = 0$ indicates the maximum downward tilt, $v = 1$ indicates maximum upward tilt, and $v = 1/2$ indicates a balance of upward and downward tilt. Figure 3 shows that the beta distribution, the uniform distribution, and the beta rectangular (Hahn, 2008) are all special cases of the distribution (10). Indeed, the density of the resulting tilted beta keeps the property of smoothness possessed by the beta distribution. This property can be contrasted with discontinuous or 'sharp' distributions that have been proposed for PERT such as the triangular (Johnson, 1997) and its extensions in the two-sided power distribution family (García Pérez et al., 2005; Herrerías et al., 2009).

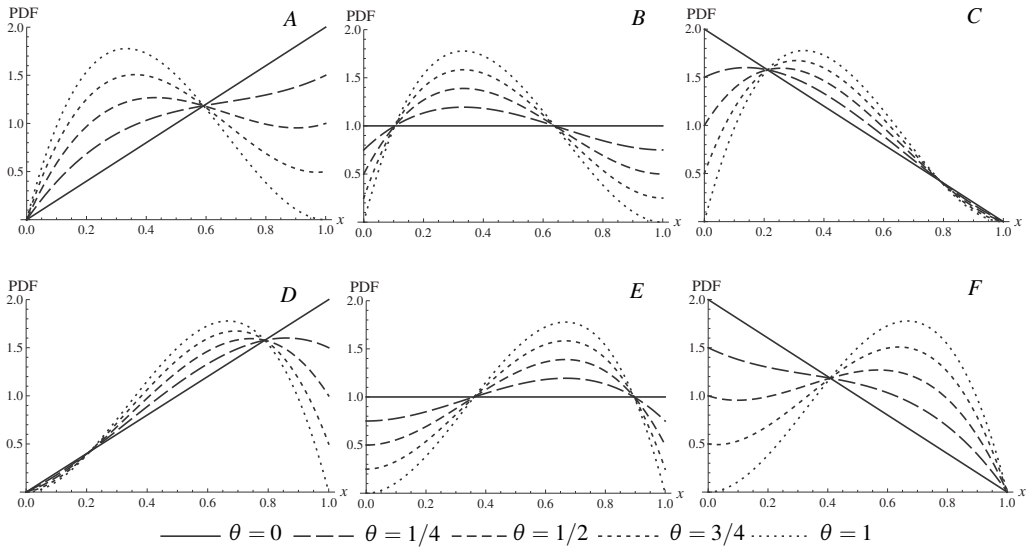


Figure 3: *Examples of tilted beta distributions for:* $\alpha = 2, \beta = 3, v = 0$ *(A);* $\alpha = 2, \beta = 3, v = 0.5$ *(B),* $\alpha = 2, \beta = 3, v = 1$ *(C),* $\alpha = 3, \beta = 2, v = 0$ *(D);* $\alpha = 3, \beta = 2, v = 0.5$ *(E),* $\alpha = 3, \beta = 2, v = 1$ *(F).*

The moment generating function of (10) is defined by

$$M_x(t) = 2\frac{e^t\,(t-1+(2-t)v)+1-(2+t)v}{t^2}(1-\theta) + {}_1F_1[\alpha, \alpha+\beta, t]\theta \qquad (11)$$

where ${}_1F_1$ indicates the Kummer confluent hypergeometric function. From (11) one can obtain the mean and the second moment of the tilted beta distribution

$$E(X) = (1-\theta)\frac{2-v}{3} + \theta\frac{\alpha}{\alpha+\beta}, \qquad (12)$$

$$E(X^2) = (1-\theta)\frac{3-2v}{6} + \theta\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}. \qquad (13)$$

We can consider the density of the pdf (10) at the endpoints making the usual assumption that α and β are both greater than or equal to 1. When $x = 0$, it can be shown that $p(x|v,\alpha,\beta,\theta) = (1-\theta)2v$. When $x = 1$, $p(x|v,\alpha,\beta,\theta) = (1-\theta)(2-2v)$. Observe that the density at the endpoints will be different as long as $v \neq 0.5$ and $\theta \neq 1$. As will be mentioned later, very few distributions applied in the PERT methodology have this property.

In the existing literature on this issue, few distributions exhibit a shape similar to that of the tilted beta. One such distribution is the elevated two-sided power distribution introducing by García et al. (2011). However, here we respectfully note that the elevated two-sided power distribution requires an additional parameter and possesses more complex expressions (cf. the simplicity of (13) above versus equation (23) of García et al. (2011)). In contrast, the current distribution is a mixture of two tractable distributions and hence it is straightforward to implement in any environment where the beta distribution of PERT has been previously applied. The reflected generalized Topp and Leone distribution (Van Dorp and Kotz, 2006) can achieve somewhat similar shapes but it does not seem possible for this distribution to have appreciable density at both extremes simultaneously. Thus, this distribution is not well-suited for circumstances where tail-area events have appreciable likelihood at both high and low extremes. Moreover, Van Dorp and Kotz (2006) indicate that this distribution does not have closed-form moment expressions except for special cases (cf. with (13) above), again adding computational cost for applications-oriented Monte Carlo simulation. Finally, Pham-Gia and Turkkan (1993) presented an explicit expression for the distribution of the difference of two beta distributions. This distribution can also take on many flexible shapes (Nadarajah and Gupta, 2004, see pp. 71–84). However, it also has a complex specification and, for example, its moments can only be analytically approximated.

Note that the procedure to raise any bounded continuous distribution by the tilting distribution is equivalent to the procedure of raising the density of the distribution linearly, and then re-normalizing.

4. Elicitation of the tilted beta distribution's parameters

In many project management applications, it is necessary to consider parameter elicitation for distributions. Direct elicitation of the beta distribution's α and β is always an option (e.g., Chaloner and Duncan, 1983). Historically however project management applications have used the classic PERT parameters: a (lower bound), m (most likely) and b (upper bound). The classic PERT formulas are then

$$E(Y) = \frac{a+4m+b}{6}, \tag{14}$$

$$\mathrm{V}(Y) = \frac{(b-a)^2}{36}. \tag{15}$$

A wide literature has been dedicated to examining the necessary conditions linking (14) and (15) to the parameters of the beta distribution (Malcolm et al., 1959; Clark, 1962; Grubbs, 1962; Sasieni, 1986; Gallagher, 1987; Littlefield and Randolph, 1987; Kamburowski, 1997). To summarize, (14) holds exactly when $k = \alpha + \beta = 6$ and $\alpha \neq \beta$. We may call this the Type I beta condition. Further, (14) and (15) simultaneously hold when: $\alpha = \beta = 4$; $\alpha = 3 + \sqrt{2}$, $\beta = 3 - \sqrt{2}$; and $\alpha = 3 - \sqrt{2}$, $\beta = 3 + \sqrt{2}$ (Grubbs, 1962). We may call this the Type II beta condition. Clearly the Type II condition is more restrictive than the Type I condition. In this case, all that is required is to select whether a symmetric, positively skewed, or negatively skewed distribution is required. Then the values of α and β are given as above. For the Type I condition, note that the mean and mode of the beta distribution are α/k and $(\alpha-1)/(k-2)$, respectively. Hence, solving simultaneous equations for the mean and the mode we find in the case of a standardized beta ($a = 0$ and $b = 1$) that the values of α and β under the Type I condition are $\alpha = 4m + 1$ and $\beta = 5 - 4m$.

Having addressed the elicitation of α and β, we turn to the elicitation of the remaining parameters of (10). The elicitation of the mixture parameter θ has been considered by Hahn (2008) and López Martín et al. (2012) using the parameter λ. Hence it remains to discuss v. Eliciting v can be accomplished by the following procedure. We assume the expert believes there is a linear increase or decrease in the probability density across time in accordance with the shape of the tilting distribution. Let A_j represent the event that a particular activity is completed on day j. Then we ask the expert to provide the probability of the event of activity completion in day j. This is denoted by $p(A_j)$. Next we ask her to give the probability of the event of activity completion in day $j+1$, which is denoted by $p(A_{j+1})$. Suppose a discrete approximation to the tilting distribution is used. The slope is (see Figure 4)

$$\frac{p(A_{j+1}) - p(A_j)}{(j+1) - j} \tag{16}$$

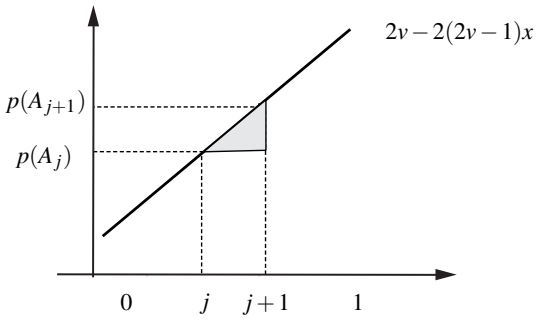


Figure 4: *Cumulative distribution function of a discrete variable with support* $[0,1]$.

by the definition of the slope as $\frac{y_2-y_1}{x_2-x_1}$. Since there are $b-a$ conceivable activity completion days, we may normalize the cumulative activity time until A_{j+1} as $x_2 = \frac{j+1}{b-a}$. Similarly we may normalize the cumulative activity time until A_j as $x_1 = \frac{j}{b-a}$. Since A_{j+1} and A_j differ by one day out of the $b-a$ total activity days, we substitute into the slope formula to define the rate of change as

$$r = \frac{p(A_{j+1}) - p(A_j)}{1/(b-a)}. \tag{17}$$

Note that other time units besides days may be alternatively used. Once we have obtained r, we can solve for the value of v by making r equal to the slope of the density function $-2(2v-1)$ and solving in terms of v, yielding

$$v = \frac{2-r}{4}. \tag{18}$$

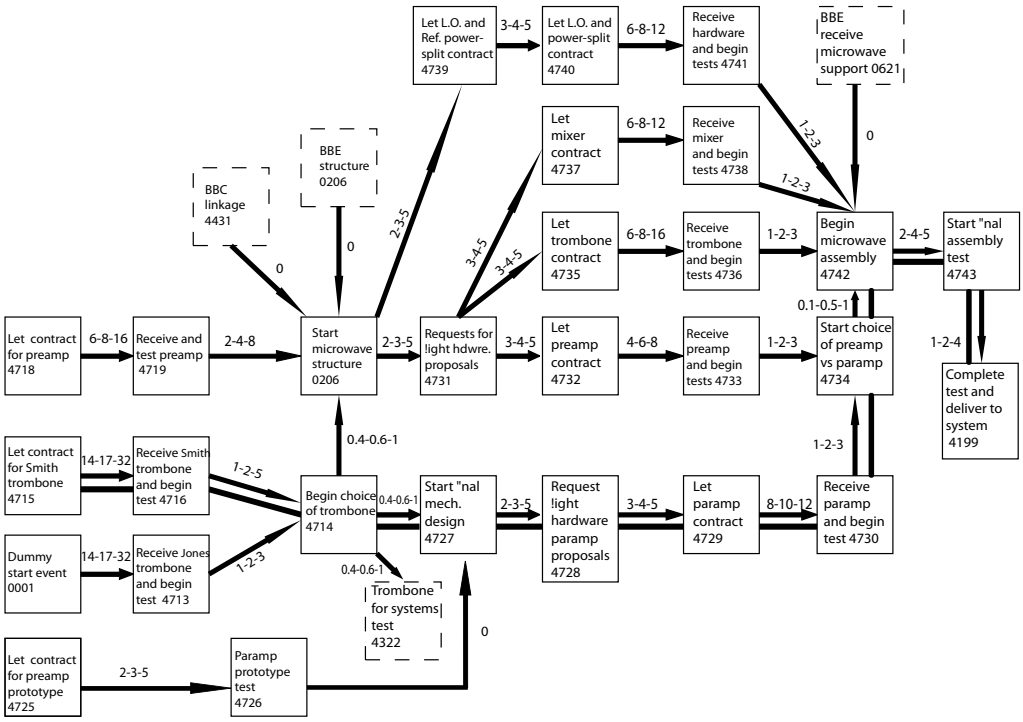
If $v \notin [0,1]$, then a re-examination will be required. Discussion with the expert can be undertaken to reveal whether, for example, the judgment task can be made easier by considering months instead of days. Alternatively, it may be that a linearly-sloped distribution does not correspond to the expert beliefs and if so the process would need to be terminated. Assuming a valid value of v, conversion to w is given by $w = a + (b-a)v$.

Alternatively, we ask the expert the probability of the event of activity completion in day $j+1, j+2, \ldots, j+k$, where the period $j+1$ is the following day of the first day after the start of the project and $j+k$ is the day before the end of the project. For each probability, and using the expression (18), we elicit the parameter v for each different day. For example we obtain $v_1, v_2, \ldots, v_k$. We can then find v as the arithmetic mean $v = \frac{1}{k}\sum_{i=1}^{k} v_i$.

Another more informal approach to elicitation of v can be contemplated by analogy with the elicitation of θ. Note in the beta rectangular distribution that $\theta = 0$ corresponds to the case of no (or 0%) additional uncertainty above and beyond that of the beta. Further, $\theta = 1$ corresponds to the case of complete (or 100%) uncertainty. So $v = 1$ would correspond to a 100% linear pessimistic belief or worst-case linear belief about the project activity completion time. In contrast, $v = 0$ would correspond to a 0% linear pessimistic belief (100% linear optimistic belief) or a best-case linear belief about the project activity completion time. More moderate values of v would represent various compromises between these extremes, with $v = 1/2$ representing neither pessimism nor optimism. In the event that we would want to counteract an elicited value v, we simply invert the slope by using $(1-v)$ in the place of v.

5. Application

Figure 5 shows 29 activities in a real-world electronic module development project from Moder et al. (1983). The critical path is marked by a heavy line.



Project activities and paths

Figure 5: *Reproduction of PERT network of an electronic module development project (Moder et al., 1983).*

The distribution for the total project time can be found using Monte Carlo simulation and accounting for the diagram's precedence relationships. To obtain our results, we simulated from the activity times using the beta distribution information from Figure 5 and the listed values of θ and v appearing in our results. The results are based on 10000 Monte Carlo simulations from the distributions of interest. Please observe that we report results arising from use of the beta rectangular distribution and, for completeness, use of the tilted beta distribution with $v = 1/2$ which is equivalent to the beta rectangular. Results for the two equivalent distributions are equivalent up to Monte Carlo error at the third significant digit with a few exceptions that are slightly larger such as the upper 95% confidence interval for $\theta = 1/4$ in Table 1 (62.34 versus 62.13).

With increasing θ, the distributions approach the beta distribution and they equal the beta when $\theta = 1$. Therefore we see in Table 1 that the standard deviation declines with

Table 1: *Stochastic characteristics of the total project time variable obtained by Monte Carlo simulations where Beta-R is the beta rectangular distribution and T-Beta is the tilted beta distribution.*

θ	Distribution	Mean	Stand. Dev.	Skewness	Kurtosis	95% C.I.
$\theta = 1/4$	Beta-R	51.50	5.55	0.23	2.26	(42.21, 62.34)
	T-Beta ($v = 0$)	56.56	5.42	-0.17	2.15	(46.41, 65.49)
	T-Beta ($v = 1/4$)	54.08	5.65	0.02	2.11	(44.13, 64.10)
	T-Beta ($v = 1/2$)	51.51	5.56	0.20	2.22	(42.12, 62.13)
	T-Beta ($v = 3/4$)	48.85	5.14	0.46	2.58	(40.64, 59.80)
$\theta = 1/2$	Beta-R	50.15	5.09	0.47	2.54	(42.05, 60.96)
	T-Beta ($v = 0$)	53.54	5.49	0.21	2.14	(44.38, 63.92)
	T-Beta ($v = 1/4$)	51.86	5.38	0.36	2.30	(43.17, 62.59)
	T-Beta ($v = 1/2$)	50.11	5.10	0.48	2.54	(42.18, 60.88)
	T-Beta ($v = 3/4$)	48.54	4.78	0.54	2.77	(41.03, 59.12)
$\theta = 3/4$	Beta-R	48.88	4.48	0.67	3.10	(41.96, 59.33)
	T-Beta ($v = 0$)	50.56	4.90	0.60	2.74	(43.09, 61.51)
	T-Beta ($v = 1/4$)	49.82	4.70	0.65	2.93	(42.63, 60.57)
	T-Beta ($v = 1/2$)	48.87	4.44	0.69	3.17	(41.95, 59.23
	T-Beta ($v = 3/4$)	48.09	4.21	0.60	3.07	(41.44, 57.72)

increasing θ. This is because when the tilting distribution predominates, the dispersion is increased which makes estimates wider and more conservative. For v, higher values correspond to assigning more weight to shorter, more optimistic outcomes. Accordingly the means in Table 1 are monotonically decreasing in v. It is also somewhat surprising to note that the standard deviations also are decreasing in v (except for the case when $v = 0$ and $\theta = 1/4$). Inspection of Figure 5 indicates that the judgments rendered tend to be optimistic or neutral at worst (coincidentally, this is consistent with our review in Section 2). Hence a value of $v = 1/4$ further concentrates the optimistic nature of the judgments into the shorter times, reducing the standard deviation. Larger values of v counteract this, especially when θ is low.

Graphical displays of the distributions of total project times appear in Figure 6. The most conservative results for the distribution of total project time can be seen when $v = 1/4$ and $\theta = 1/4$. This distribution is the least skewed of all those displayed, and appears approximately uniform across the middle third of its range. We observe that the distribution in the centre of the top row for the beta rectangular when $\theta = 1/2$ is equivalent to the distribution for the tilted beta with $\theta = 1/2, v = 1/2$ in the centre of the third row, and these coincide as we would expect.

Figure 7 provides another way of viewing the changes in project times as a function of distributional parameters. It plots the CDF for the simulated project times under selected values of θ and v. On the left side where $\theta = 1/4$, the tilting portion of the mixture is predominant. We see there that the optimistic assessment of $v = 3/4$ leads to

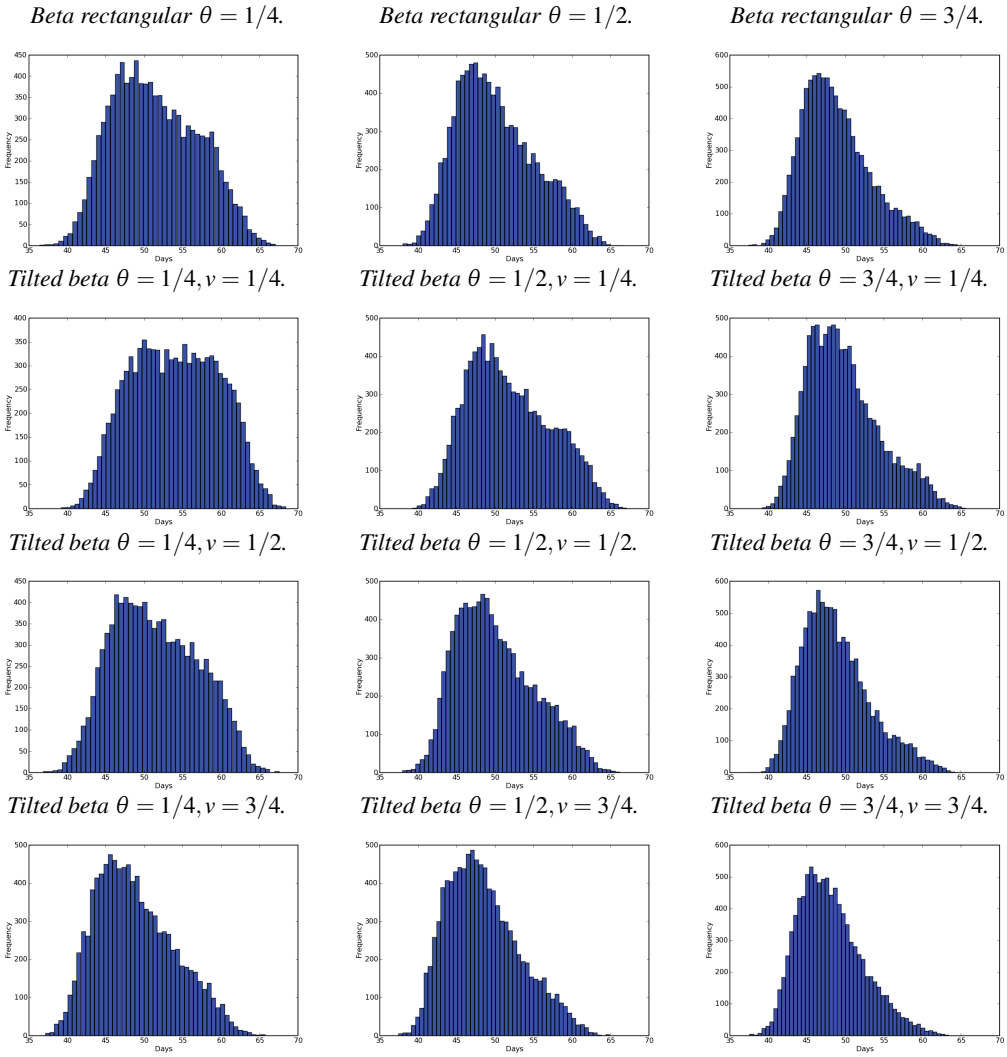


Figure 6: *Distributions of the total project time (Electronic Module Development Project).*

a relatively high cumulative probability of completion by 55 days. For less optimistic values of v, the cumulative probability of completion by 55 days (or other values we might select) falls off considerably. The right side of Figure 7 shows the CDF when the beta portion of the mixture predominates. The CDFs have some variability due to v but in general the CDFs are closer together and rise more steeply since they preserve more of the classic PERT beta influence. For completeness, we also observe that the CDF for the beta rectangular and the tilted beta with $v = 1/2$ again gives essentially the same result, as we would expect, since the solid beta rectangular line and the dashed tilted beta $v = 1/2$ line are essentially superimposed.

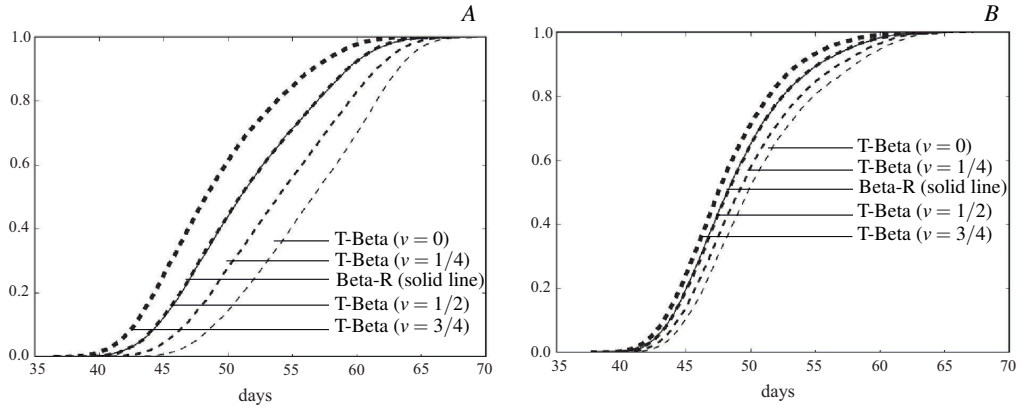


Figure 7: *Cumulative distribution function of beta rectangular distribution and tilted beta distribution for* $\theta = 1/4$ *(A) and* $\theta = 3/4$ *(B)*

Finally, Table 2 shows an example of what would constitute a rare or tail-area event under the different distributions. Here the 95% value of the CDF is obtained for v and θ taking on the values 0, 1/4, 1/2 and 3/4. With the introduction of the parameter v, the CDF of the tilted beta distribution may provide project time-to-completion estimates that can be either higher or lower than the beta rectangular distribution. The most striking comparison involves in the case when $\theta = 1$ which is the classic PERT beta case. We observe the classic PERT result would say there is a 95% chance of the project being completed by approximately 53.9 days, excluding some Monte Carlo error visible for the four values of v in the plot. Compare this results with the case of $\theta = 3/4$ where a small amount of extra-beta variability has been mixed in but the beta distribution still predominates at $\theta = 3/4$. Here even under the most optimistic case ($v = 3/4$) the 95% completion time has increased by 2 days to about 55.9 days. Hence, this example 'worst-case scenario' is two days worse than that given by the classic PERT beta. With less optimistic values of v, the time increases further. For example, with $v = 3/4$ and $\theta = 1/2$ we approach 57.7, i.e. approximately four days more than PERT.

Table 2: *The maximum time needed to complete the 95% the project for beta rectangular (Beta-R) and tilted beta (T-Beta) with* $v = 0$, $v = 1/4$, $v = 1/2$ *and* $v = 3/4$.

θ	Beta-R	T-Beta ($v = 0$)	T-Beta ($v = 1/4$)	T-Beta ($v = 1/2$)	T-Beta ($v = 3/4$)
0.25	60.98	**64.68**	63.06	60.95	**58.39**
0.50	59.58	**62.80**	61.45	59.60	**57.69**
0.75	57.75	**60.03**	59.22	**57.68**	55.98
1.00	53.90	**53.92**	**53.69**	53.78	53.83

In most cases, when the parameter v is higher than 1/2, the tilted beta distribution will be by construction more optimistic than the beta rectangular distribution and as a

consequence its estimates will indicate that a lower time will be needed to complete the project. Conversely, when v is lower than 1/2, the tilted beta distribution provides more conservative estimations.

6. Conclusion

The introduction of different activity distributions has played an important part in the PERT methodology. However, this issue divides the researchers of this field. Some authors argue against the introduction of new probability distributions into PERT (see Clark, 1962; Hajdu, 2013; Hajdu and Bokor, 2014). Conversely as shown in Section 3 other authors have applied new distributions. Regarding the current paper, this debate has parallels in statistical practice. Some authors use robust statistical methods to handle outliers while other authors adopt less formal techniques or may even naively do nothing at all.

This paper introduced the tilted beta distribution and shown how it can be used in project management. Since the classic PERT results can be reproduced, it is simple to adopt in any environment where PERT is utilized. We can easily explain to executives and decision makers that incorporating additional uncertainty can help us to arrive at new insights. Elicitation of parameters is straightforward or one could perform sensitivity analysis using several parameter values as we have done here. The tilted beta has a number of attractive computational properties such as being easy to simulate from and having closed-form moment expressions. In summary, the tilted beta distribution provides project managers with a flexible and easy to work with distribution that allows for the extensive representation of optimistic or pessimistic beliefs regarding activity times.

Past work has pointed out the need to describe a more flexible distribution which allows for varying amounts of dispersion and greater likelihoods of more extreme tail-area events (Hahn, 2008). The construction of beta rectangular distribution is characterized by greater flexibility in the variance. However, this distribution assigns the same probability density at both the high and low extreme values. With the introduction of tilted beta distribution, we have expanded the set of continuous type distributions defined on a bounded domain, with the advantage of accommodating different relative likelihoods of high versus low extreme tail-area events and, as opposed to other distributions applied in this methodology, the tilted beta has an expression of the expected value where the extreme values have different weight. As a consequence, this distribution will be more relevant for modeling a broader range of heavy tailed phenomena. In a closely related work, the elevated two-sided power distribution (García et al., 2011) also permits different relative likelihoods of high versus low extreme tail-area events; however, as described above, the current distribution is simpler to use in practice. Furthermore, note that the procedure to induce tilting can be applied to any bounded continuous distribution.

We have compared the results of the tilted beta distribution with the results of beta rectangular distribution for different values of the parameter θ and v. The results of the application show that this probabilistic model permits a risk manager to incorporate more optimistic and pessimistic scenarios than the beta rectangular distribution due to the flexibility of the tilted beta. Our literature review suggests that experts may tend to be too optimistic, and the beta distribution gives little weight to outliers in the standard PERT Type I and Type II beta conditions. The current methodology redresses these issues.

The distributions presented in this paper will be closer to the uniform since they give more weight to the tails. We think this gives some evidence that the distribution chosen is relevant by considering a recent paper by Hajdu and Bokor (2014). For larger projects, 10% can be the difference between a project being on time or late and the authors show that the uniform distribution is similar to PERT-beta + 10%. Furthermore, for small projects the authors state that it may not matter much. However, the incentive to use PERT with smaller projects is probably smaller.

There are at least two areas in which this research can be extended: first, the use of heavy-tailed distributions in the context of different activity calendars (Hajdu, 2013); second, to find more applications of these distributions fitting extreme tail-area events which are present in a great variety of fields such us finance, groundwater hydrology and atmospheric science among others.

Acknowledgements

We would like to thank referees for their detailed, encouraging and constructive reviews of our paper which clearly contributed to improving both the structure and the content of this manuscript.

References

Bevilacqua, M., F. E. Ciarapica, and G. Giacchetta (2009). Critical chain and risk analysis applied to high-risk industry maintenance: a case study. *International Journal of Project Management*, 27, 419–432.

Boulding, W., R. Morgan, and R. Staelin (1997). Pulling the plug to stop new product drain. *Journal of Marketing Research*, 34, 164–176.

Chaloner, K. M. and G. T. Duncan (1983). Assessment of a beta prior distribution: PM Elicitation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32, 174–180.

Clark, C. E. (1962). The PERT model for the distribution of an activity time. *Operations Research*, 10, 405–406.

Gallagher, C. (1987). A note on PERT assumptions. *Management Science*, 33, 1360.

García, C. B., J. García, and S. Cruz (2010). Proposal of a new distribution in PERT methodology. *Annals of Operations Research*, 181, 515–538.

García, C. B., J. García Pérez, and J. R. van Dorp (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support. *Statistical Methods and Applications*, 20, 463–482.

García Pérez, J., S. Cruz Rambaud, and C. B. García García (2005). The two-sided power distribution for the treatment of the uncertainty in PERT. *Statistical Methods & Applications*, 14, 209–222.

Garvey, P. R. (2000). *Probability Methods for Cost Uncertainty Analysis, a Systems Engineering Perspective*. New York, NY: Marcel Dekker.

Gillard, S. (2005). Managing IT projects: Communication pitfalls and bridges. *Journal of Information Science*, 31, 37–43.

Grubbs, F. E. (1962). Attempts to validate certain PERT statistics or 'Picking on PERT'. *Operations Research*, 10, 912–915.

Hahn, E. D. (2008). Mixture densities for project management activity times: a robust approach to PERT. *European Journal of Operational Management*, 188, 450–459.

Hajdu, M. (2013). Effects of the application of activity calendars on the distribution of project duration in PERT networks. *Automation in Construction*, 35, 397–404.

Hajdu, M. and O. Bokor (2014). The effects of different activity distributions on project duration in pert networks. *Procedia-Social and Behavioral Sciences*, 119, 766–775.

Herrerías, J. M., R. Herrerías, and J. R. van Dorp (2009). The generalized two-sided power distribution. *Journal of Applied Statistics*, 36, 573–587.

Herrerías, R. (1989). Utilización de modelos probabilísticos alternativos para el método PERT. Aplicación al análisis de inversiones. *Estudios de Economía Aplicada, Secretariado de Publicaciones de la Universidad de Valladolid*, 89–112.

Herrerías, R. and H. Calvete (1987). Una ley de probabilidad para el estudio de los flujos de caja de una inversión. *Libro Homenaje al Profesor Gonzalo Arnaiz Vellando*, 279–296. Madrid: ICE.

Hill, J., L. C. Thomas, and D. E. Allen (2000). Experts' estimates of task duration in software development projects. *International Journal of Project Management*, 18, 13–21.

Iacovou, C. L., R. L. Thompson, and H. J. Smith (2009). Selective status reporting in information systems projects: a dyadic-level investigation. *MIS Quarterly*, 33, 785–810.

Johnson, D. (1997). The triangular distribution as a proxy for the beta distribution in risk analysis. *The Statistician*, 46, 387–398.

Kamburowski, J. (1997). New validations of PERT times. *Omega*, 25, 323–328.

Keefer, D. L. and W. A. Verdini (1993). Better estimation of PERT activity time parameters. *Management Science*, 39, 1086–1091.

Keil, M., G. Depledge, and A. Rai (2007). Escalation: the role of problem recognition and cognitive bias. *Decision Sciences*, 38, 391–421.

Keil, M., J. Mann, and A. Rai (2000). Why software projects escalate: an empirical analysis and test of four theoretical models. *MIS Quarterly*, 24, 631–664.

Kotiah, T. C. T. and N. D. Wallace (1973). Another look at the PERT assumptions. *Management Science*, 20, 44–49.

Kotz, S. and J. R. van Dorp (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support*. Singapore: World Scientific.

Littlefield, Jr., T. K. and P. H. Randolph (1987). An answer to Sasieni's question on PERT times. *Management Science*, 33, 1357–1359.

López Martín, M. M., C. B. García, J. García, and M. A. Sánchez (2012). An alternative for robust estimation in project management. *European Journal of Operational Research*, 220, 443–451.

MacCrimmon, K. R. and C. A. Ryavec (1964). An analytical study of the PERT assumptions. *Operations Research*, 12, 16–37.

Malcolm, D. G., J. H. Roseboom, C. E. Clark, and W. Fazar (1959). Application of a technique for research and development program evaluation. *Operations Research*, 7, 646–669.

Megill, R. E. (1984). *An Introduction to Risk Analysis* (2nd ed.). Tulsa, OK: PennWell Books.

Moder, J. J., C. R. Phillips, and E. W. Davis (1983). *Project Management with CPM, PERT and Precedence Diagramming* (3rd ed.). New York: Van Nostrand Reinhold Company Inc.

Moder, J. J. and E. G. Rodgeres (1968). Judgement estimates of the moments of PERT type distributions. *Management Science*, 15, B76–B83.

Mohan, S., M. Gopalakrishnan, H. Balasubramanian, and A. Chandrashekar (2007). A lognormal approximation of activity duration in pert using two time estimates. *The Journal of the Operational Research Society*, 58, 827–831.

Nadarajah, S. and A. Gupta (2004). Products and linear combinations. In A. Gupta and S. Nadarajah (Eds.), *Handbook of Beta Distribution and Its Applications*, 55–88. New York: Marcel Dekker.

Öztaş, A. and O. Ökmen (2005). Judgmental risk analysis process development in construction projects. *Building and Environment*, 40, 1244–1254.

Pham-Gia, T. and N. Turkkan (1993). Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*, 22, 1755–1771.

Pouliquen, L. Y. (1970). *Risk Analysis in Project Appraisal*. Baltimore: John Hopkins University Press.

Powell, M. R. and J. D. Wilson (1997). *Risk Assessment for National Natural Resource Conservation Programs*. Washington: Resources for the Future.

Romero, C. (1991). *Técnicas de Programación y Control de Proyectos*. Madrid: Pirámide.

Sasieni, M. W. (1986). A note on PERT times. *Management Science*, 32, 1652–1653.

Sengupta, K., T. K. Abdel-Hamid, and L. N. van Wassenhove (2008). The experience trap. *Harvard Business Review*, 86, 94–101.

Smith, H. J., M. Keil, and G. Depledge (2001). Keeping mum as the project goes under: toward an explanatory model. *Journal of Management Information Systems*, 18, 189–227.

Snow, A. P., M. Keil, and L. Wallace (2007). The effects of optimistic and pessimistic biasing on software project status reporting. *Information & Management*, 44, 130–141.

Suárez, A. (1980). *Decisiones Óptimas de Inversión y Financiación en la Empresa*. Madrid: Pirámide.

Topp, C. W. and F. C. Leone (1955). A family of J-Shaped frequency functions. *Journal of the American Statistical Association*, 50, 209–219.

Trietsch, D., L. Mazmanyan, L. Gevorgyan, and K. R. Baker (2012). Modeling activity times by the Parkinson distribution with a lognormal core: theory and validation. *European Journal of Operational Research*, 216, 386–396.

Vãduva, I. (1971). Computer generation of random variables and vectors related to PERT problems. *Proceedings of the 4th conference on Probability Theory*, Brasov, 381–395.

Van Dorp, J. R. and S. Kotz (2006). Modeling income distributions using elevated distributions. In R. Herrerías Pleguezuelo, J. Callejón Céspedes, and J. M. Herrerías Velasco (Eds.), *Distribution Models Theory*, 1–25. Singapore; 2006: World Scientific.

Williams, T. M. (1992). Practical use of distributions in network analysis. *Journal of the Operations Research Society*, 43, 265–270.

SORT 39 (2) July-December 2015, 273-280

A note on "Double bounded Kumaraswamy-power series class of distributions"

Tibor K. Pogány[1] and Saralees Nadarajah[2]

Abstract

In a recent edition of SORT, Bidram and Nekoukhou proposed a novel class of distributions and derived its mathematical properties. Several of the mathematical properties are expressed as single infinite sums or double infinite sums. Here, we show that many of these properties can be expressed in terms of known special functions, functions for which in-built routines are widely available.

MSC: 11G55, 33C20, 33E20, 60E10.

Keywords: Double bounded Kumaraswamy-power series class of distributions, Fox Wright generalized hypergeometric function, Generalized hypergeometric function.

1. Introduction

Bidram and Nekoukhou (2013), referred to as BN from now, present a novel class of distributions referred to as *double bounded Kumaraswamy-power series distributions*. They derive various mathematical properties of the distributions, including their density functions, survival functions, hazard rate functions, quantiles, moment generating functions, moments, order statistic properties and stress strength parameter. They also discuss maximum likelihood estimation of the parameters of the distributions and provide a real data application.

Several of the expressions given in BN involve single infinite sums or double infinite sums. This is the case with the moment generating functions given in BN, Table 2; the moments given in BN, Table 2; the density of the ith order statistic given in BN, page 221; the rth moment of the ith order statistic given in BN, page 221; the stress-strength

[1] Faculty of Maritime Studies, University of Rijeka, 51000 Rijeka, Croatia
Institute of Applied Mathematics, Obuda University, 1034 Budapest, Hungary.

[2] School of Mathematics, University of Manchester, Manchester M13 9PL, UK

Received: November 2014
Accepted: June 2015

parameter given in BN, page 222; and others. This is not very convenient for practical implementation of the mathematical properties. The aim of this note is to show that many of the infinite sums and so the mathematical properties given in BN can be reduced to known special functions, functions for which in-built routines are widely available.

Let $\mathbb{N}=\{1,2,3,\ldots\}$, $\mathbb{N}_0=\mathbb{N}\cup\{0\}$, $\mathbb{Z}_0^-=\{0,-1,-2,\ldots\}$, $\mathbb{R}$ the set of real numbers, $\mathbb{R}^+$ the set of positive real numbers and $\mathbb{C}$ the set of complex numbers.

The closed form expressions in Section 2 involve several special functions. First is the gamma function defined by $\Gamma(a)=\int_0^\infty t^{a-1}e^{-t}\mathrm{d}t$ for $a\in\mathbb{R}^+$. The second is the polylogarithm function defined by $\mathrm{Li}_\nu(z)=\sum_{n\geq 1} n^{-\nu}\,z^n$ for $|z|<1$. The third is the generalized hypergeometric function ${}_pF_q[\cdot]$ defined by

$$ {}_pF_q[z]={}_pF_q\left[\begin{array}{c} a_1,\ldots,a_p \\ b_1,\ldots,b_q \end{array};z\right]=\sum_{k=0}^{\infty}\frac{(a_1)_k\cdots(a_p)_k}{(b_1)_k\cdots(b_q)_k}\frac{z^k}{k!}, \tag{1} $$

where $(\lambda)_\mu$ denotes the Pochhammer symbol defined by

$$ (\lambda)_\mu:=\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda)}=\begin{cases}1, & \big(\mu=0;\ \lambda\in\mathbb{C}\setminus\{0\}\big),\\ \lambda(\lambda+1)\cdots(\lambda+n-1), & \big(\mu=n\in\mathbb{N};\ \lambda\in\mathbb{C}\big)\end{cases} \tag{2} $$

with the convention that $(0)_0:=1$. The *Gauss hypergeometric function* ${}_2F_1(a,b:c;z)$ is the particular case of (1) for $p=2$, $q=1$. In the case $a,b\in\mathbb{Z}_0^-$ are negative integers, ${}_2F_1(a,b:c;z)$ becomes a polynomial $P_N(z)$ of degree $\deg(P_N)=N=\min(-a,-b)$.

The fourth is the Fox Wright generalized hypergeometric function ${}_p\Psi_q^*[\cdot]$ with p numerator parameters $a_1,\ldots,a_p$ and q denominator parameters $b_1,\ldots,b_q$, defined by (Kilbas et al., 2006, page 56)

$$ {}_p\Psi_q^*\left[\begin{array}{c}(a_1,\rho_1),\ldots,(a_p,\rho_p)\\(b_1,\sigma_1),\ldots,(b_q,\sigma_q)\end{array};z\right]=\sum_{n=0}^{\infty}\frac{\prod_{j=1}^{p}(a_j)_{\rho_j n}}{\prod_{j=1}^{q}(b_j)_{\sigma_j n}}\,\frac{z^n}{n!} \tag{3} $$

for $a_j\in\mathbb{C}$, $j=1,2,\ldots,p$, $b_k\in\mathbb{C}$, $k=1,2,\ldots,q$, $\rho_j\in\mathbb{R}^+$, $j=1,2,\ldots,p$ and $\sigma_k\in\mathbb{R}^+$, $k=1,2,\ldots,q$. The series in (3) converges in the whole complex z-plane when

$$ \Delta:=1+\sum_{j=1}^{q}\sigma_j-\sum_{j=1}^{p}\rho_j>0. $$

If $\Delta = 0$, then the series in (3) converges for $|z| < \nabla$, where

$$\nabla := \left(\prod_{j=1}^{p} \rho_j^{-\rho_j}\right)\left(\prod_{j=1}^{q} \sigma_j^{\sigma_j}\right).$$

The particular case of (3) for $\rho_1 = \cdots = \rho_p = 1$ and $\sigma_1 = \cdots = \sigma_q = 1$ is the generalized hypergeometric function in (1).

In-built routines for computing these special functions are widely available in packages like Maple, Matlab and Mathematica. `Gamma[z]` in Mathematica computes the gamma function, `PolyLog[`ν`,z]` in Mathematica computes the polylogarithm function, `HypergeometricPFQ[{`a_1`,. . . ,`a_p`}, {`b_1`,. . . ,`b_q`},z]` in Mathematica computes the generalized hypergeometric function, and so on. The routines allow for arbitrary precision, so computational accuracy is not an issue.

2. Closed form expressions

The closed form expressions are given by Propositions 2.1 to 2.3. Proposition 2.1 expresses $F_{i:n}(x)$, the cumulative distribution function of the ith order statistic given in BN, page 217, equation (12), in terms of the Gauss hypergeometric function. Proposition 2.2 expresses the moments given in BN, page 220, Table 2 in terms of the Fox Wright generalized hypergeometric function. These moments are to any real order, the ones given in BN were for integer orders only. Proposition 2.3 expresses R, the stress-strength parameter given in BN, page 222, in terms of the polylogarithm function.

Proposition 2.1 *For all* $1 \le i \le n$ *and for all* $x \in (0,1)$,

$$F_{i:n}(x) = \binom{n}{i} A^i \, {}_2F_1\left(-n+i, i; i+1; A\right) = P_n(A)$$

is a polynomial in A, where

$$A = 1 - \frac{C\left(\theta\left(1-x^a\right)^b\right)}{C(\theta)}.$$

Moreover, $F_{n:n}(x) = A^n$.

Proof: Follows by noting

$$F_{i:n}(x) = \frac{1}{\mathrm{B}(i, n-i+1)} \sum_{k=0}^{n-i} \binom{n-i}{k} \frac{(-1)^k}{k+i} A^{k+i}$$

$$= \frac{A^i}{\mathrm{B}(i,n-i+1)} \sum_{k=0}^{n-i} \frac{(-n+i)_k \Gamma(k+i)}{\Gamma(k+i+1)} \frac{A^k}{k!}$$

$$= \frac{A^i}{i\,\mathrm{B}(i,n-i+1)} \sum_{k=0}^{n-i} \frac{(-n+i)_k (i)_k}{(i+1)_k} \frac{A^k}{k!}$$

$$= \frac{A^i\,\Gamma(n+1)}{\Gamma(i+1)\,\Gamma(i,n-i+1)}\ {}_2F_1(-n+i,i;i+1;A)$$

$$= \frac{A^i\,n!}{i!\,(n-i)!}\ {}_2F_1(-n+i,i;i+1;A).$$

The hypergeometric function reduces to 1 when $i=n$, so $F_{n:n}(x)=A^n$. ∎

Proposition 2.2 *Let X_{KG}, X_{KP}, X_{KL} and X_{KB} be random variables following, respectively, the Kumaraswamy geometric, Kumaraswamy Poisson, Kumaraswamy logarithmic and Kumaraswamy binomial distributions defined in BN. Then, for all real $r > -a$ and $b > 0$, we have*

$$\mathbb{E}\big(X_{\mathrm{KG}}^r\big) = b(1-\theta)\mathrm{B}\left(1+\frac{r}{a},b\right)\ {}_2\Psi_1^*\left[\begin{array}{c}(1+b,b),\ (1,1)\\ \left(1+b+\dfrac{r}{a},b\right)\end{array};\theta\right], \tag{4}$$

$$\mathbb{E}\big(X_{\mathrm{KP}}^r\big) = \frac{b\,\theta}{e^{\theta}-1}\ \mathrm{B}\left(1+\frac{r}{a},b\right)\ {}_1\Psi_1^*\left[\begin{array}{c}(b,b)\\ \left(1+b+\dfrac{r}{a},b\right)\end{array};\theta\right], \tag{5}$$

$$\mathbb{E}\big(X_{\mathrm{KL}}^r\big) = -\frac{b\theta}{\log(1-\theta)}\mathrm{B}\left(1+\frac{r}{a},b\right)\ {}_2\Psi_1^*\left[\begin{array}{c}(b,b),\ (1,1)\\ \left(1+b+\dfrac{r}{a},b\right)\end{array};\theta\right] \tag{6}$$

and

$$\mathbb{E}\big(X_{\mathrm{KB}}^r\big) = -\frac{bm\theta}{(1+\theta)^m-1}\mathrm{B}\left(1+\frac{r}{a},b\right)\ {}_2\Psi_1^*\left[\begin{array}{c}(b,b),\ (1-m,1)\\ \left(1+b+\dfrac{r}{a},b\right)\end{array};-\theta\right]. \tag{7}$$

Each of these expressions is valid for all $|\theta| < 1$.

Proof: (4) follows by noting that

$$\mathbb{E}\left(X_{\mathrm{KG}}^r\right) = b(1-\theta)\sum_{n\geq 1} n\,\mathrm{B}\left(1+\frac{r}{a}, nb\right)\,\theta^{n-1}$$

$$= b(1-\theta)\Gamma\left(1+\frac{r}{a}\right)\sum_{n\geq 1}\frac{n\,\Gamma(nb)\,\theta^{n-1}}{\Gamma\left(1+\frac{r}{a}+nb\right)}$$

$$= (1-\theta)\Gamma\left(1+\frac{r}{a}\right)\sum_{n\geq 1}\frac{\Gamma(nb+1)\,\theta^{n-1}}{\Gamma\left(1+\frac{r}{a}+nb\right)}$$

$$= (1-\theta)\Gamma\left(1+\frac{r}{a}\right)\sum_{m\geq 0}\frac{\Gamma(1+b+mb)\,\theta^{m}}{\Gamma\left(1+b+\frac{r}{a}+mb\right)}$$

$$= (1-\theta)\frac{\Gamma\left(1+\frac{r}{a}\right)\Gamma(1+b)}{\Gamma\left(1+b+\frac{r}{a}\right)}\sum_{m\geq 0}\frac{(1+b)_{mb}\,(1)_m}{\left(1+b+\frac{r}{a}\right)_{mb}}\frac{\theta^m}{m!}$$

and that the infinite sum in the last step corresponds to a Fox Wright generalized hypergeometric function with $\Delta = 0$, $\nabla = 1$.

(5) follows by noting that

$$\mathbb{E}\left(X_{\mathrm{KP}}^r\right) = \frac{b}{e^{\theta}-1}\sum_{n\geq 1}\mathrm{B}\left(1+\frac{r}{a}, nb\right)\,\frac{\theta^n}{(n-1)!}$$

$$= \frac{b\,\Gamma\left(1+\frac{r}{a}\right)\,\theta}{e^{\theta}-1}\sum_{m\geq 0}\frac{\Gamma(b+mb)}{\Gamma\left(1+b+\frac{r}{a}+mb\right)}\frac{\theta^m}{m!}$$

$$= \frac{b\,\Gamma(b)\,\Gamma\left(1+\frac{r}{a}\right)\,\theta}{\left(e^{\theta}-1\right)\,\Gamma\left(1+\frac{r}{a}+b\right)}\sum_{m\geq 0}\frac{(b)_{mb}}{\left(1+b+\frac{r}{a}\right)_{mb}}\frac{\theta^m}{m!}$$

and that the infinite sum in the last step corresponds to a Fox Wright generalized hypergeometric function with $\Delta = 1$.

The proof of (6) is similar to the proof of (4).

(7) follows by noting that

$$\mathbb{E}\left(X_{\mathrm{KB}}^{r}\right) = \frac{b\theta}{(1+\theta)^{m}-1}\sum_{n\geq 1} n\binom{m}{n}\,\mathrm{B}\left(1+\frac{r}{a}, nb\right)\,\theta^{n-1}$$

$$= \frac{b\theta m}{(1+\theta)^{m}-1}\,\Gamma\left(1+\frac{r}{a}\right)\sum_{n-1\geq 0}\frac{(-1)^{n-1}(1-m)_{n-1}\Gamma\left(b+(n-1)b\right)}{\Gamma\left(1+b+\frac{r}{a}+(n-1)b\right)}\frac{\theta^{n-1}}{(n-1)!}$$

$$= \frac{b\theta m}{(1+\theta)^{m}-1}\,\frac{\Gamma(b)\Gamma\left(1+\frac{r}{a}\right)}{\Gamma\left(1+b+\frac{r}{a}\right)}\sum_{k\geq 0}\frac{(-1)^{k}(1-m)_{k}(b)_{kb}}{\left(1+b+\frac{r}{a}\right)_{kb}}\frac{\theta^{k}}{k!}$$

and that the infinite sum in the last step corresponds to a Fox Wright generalized hypergeometric function with $\Delta = 0, \nabla = 1$. ∎

Proposition 2.3 *For all* $|\theta_1| < 1$ *and* $|\theta_2| < 1$,

$$R(\theta_1,\theta_2) = \sum_{k\geq 0}\sum_{j\geq 0}\frac{(k+1)\theta_1^{k}\theta_2^{j}}{(k+j+1)^{2}(k+j+2)}$$

$$= \frac{1}{(\theta_2-\theta_1)^{2}}\Big[\mathrm{Li}_2(\theta_1) - \mathrm{Li}_2(\theta_2) + \theta_1 - \theta_2$$

$$+ (2-\theta_2/\theta_1-\theta_2)\,\log(1-\theta_1) - (1-\theta_2)\,\log(1-\theta_2)\Big]. \qquad (8)$$

Proof: The double series for $R(\theta_1,\theta_2)$ converges for all $|\theta_1| < 1$, $|\theta_2| < 1$, where it is term-wise differentiable with respect both θ_1, θ_2. To obtain a closed form expression for R, we define

$$\mathfrak{R}(\theta_1,\theta_2) = \sum_{k\geq 0}\sum_{j\geq 0}\frac{\theta_1^{k}\theta_2^{j}}{(k+j+1)^{2}(k+j+2)}.$$

By the differentiability property,

$$R = \frac{\mathrm{d}}{\mathrm{d}\theta_1}\left(\theta_1\cdot\mathfrak{R}(\theta_1,\theta_2)\right). \qquad (9)$$

By partial fractions,

$$\mathfrak{R}(\theta_1,\theta_2)=\sum_{k\geq 0}\sum_{j\geq 0}\left(\frac{1}{k+j+2}-\frac{1}{k+j+1}+\frac{1}{(k+j+1)^2}\right)\,\theta_1^k\,\theta_2^j=:S_1-S_2+S_3. \quad (10)$$

By legitimate integration-summation order exchange and the definition of gamma function,

$$S_1=\int_0^\infty e^{-2t}\sum_{k\geq 0}\left(\theta_1 e^{-t}\right)^k\sum_{j\geq 0}\left(\theta_2 e^{-t}\right)^j\,\mathrm{d}t=\int_0^\infty\frac{e^{-2t}\,\mathrm{d}t}{\left(1-\theta_1 e^{-t}\right)\left(1-\theta_2 e^{-t}\right)},$$

$$S_2=\int_0^\infty e^{-t}\sum_{k\geq 0}\left(\theta_1 e^{-t}\right)^k\sum_{j\geq 0}\left(\theta_2 e^{-t}\right)^j\,\mathrm{d}t=\int_0^\infty\frac{e^{-t}\,\mathrm{d}t}{\left(1-\theta_1 e^{-t}\right)\left(1-\theta_2 e^{-t}\right)},$$

$$S_3=\int_0^\infty te^{-t}\sum_{k\geq 0}\left(\theta_1 e^{-t}\right)^k\sum_{j\geq 0}\left(\theta_2 e^{-t}\right)^j\,\mathrm{d}t=\int_0^\infty\frac{t\,e^{-t}\,\mathrm{d}t}{\left(1-\theta_1 e^{-t}\right)\left(1-\theta_2 e^{-t}\right)}.$$

Routine but lengthy calculations show that

$$S_1=\frac{\theta_1\log(1-\theta_2)-\theta_2\log(1-\theta_1)}{\theta_1\theta_2(\theta_1-\theta_2)},$$

$$S_2=\frac{\log(1-\theta_2)-\log(1-\theta_1)}{\theta_1-\theta_2}.$$

Using the fact

$$\int_0^\infty\frac{t\,e^{-t}\,\mathrm{d}t}{1-ae^{-t}}=\frac{1}{a}\,\mathrm{Li}_2(a),$$

we can reduce

$$S_3=\frac{1}{\theta_2-\theta_1}\int_0^\infty\frac{t\,e^{-t}\,\mathrm{d}t}{1-\theta_1 e^{-t}}-\frac{1}{\theta_2-\theta_1}\int_0^\infty\frac{t\,e^{-t}\,\mathrm{d}t}{1-\theta_2 e^{-t}}=$$

$$=\frac{1}{\theta_2-\theta_1}\left[\frac{1}{\theta_1}\,\mathrm{Li}_2(\theta_1)-\frac{1}{\theta_2}\,\mathrm{Li}_2(\theta_2)\right].$$

Collecting S_j, $j = 1,2,3$, we obtain by virtue of (10) that

$$\mathfrak{R}(\theta_1,\theta_2) = \frac{\theta_2(1-\theta_1)\log(1-\theta_1) - \theta_1(1-\theta_2)\log(1-\theta_2)}{\theta_1\theta_2(\theta_2-\theta_1)}$$

$$+\frac{1}{\theta_2-\theta_1}\left[\frac{1}{\theta_1}\operatorname{Li}_2(\theta_1) - \frac{1}{\theta_2}\operatorname{Li}_2(\theta_2)\right]$$

$$= \frac{(1-\theta_1)\log(1-\theta_1)+\operatorname{Li}_2(\theta_1)}{\theta_1(\theta_2-\theta_1)} - \frac{(1-\theta_2)\log(1-\theta_2)+\operatorname{Li}_2(\theta_2)}{\theta_2(\theta_2-\theta_1)}.$$

The result follows by applying (9). ■

Acknowledgments

The authors would like to thank the Editor and the two referees for careful reading and comments which greatly improved this letter.

References

Bidram, H. and Nekoukhou, V. (2013). Double bounded Kumaraswamy-power series class of distributions. *SORT*, 37, 211–230.

Kilbas, A. A., Srivastava, H. M. and Trujillo, J. J. (2006). *Theory and Applications of Fractional Differential Equations*, North-Holland Mathematical Studies, volume 204, Elsevier (North-Holland) Science Publishers, Amsterdam, London and New York.

SORT 39 (2) July-December 2015, 281-308

Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: a simulation study

Martí Casals[1,2,3,4], Klaus Langohr[5], Josep Lluís Carrasco[1] and Lars Rönnegård[6]

Abstract

Generalized linear mixed models are flexible tools for modeling non-normal data and are useful for accommodating overdispersion in Poisson regression models with random effects. Their main difficulty resides in the parameter estimation because there is no analytic solution for the maximization of the marginal likelihood. Many methods have been proposed for this purpose and many of them are implemented in software packages. The purpose of this study is to compare the performance of three different statistical principles – marginal likelihood, extended likelihood, Bayesian analysis – via simulation studies. Real data on contact wrestling are used for illustration.

MSC: 62J12, 62P99, 62F99.

Keywords: Estimation methods, overdispersion, Poisson generalized linear mixed models, simulation study, statistical principles, sport injuries.

1. Introduction

One of the methodologies used to study disease incidence in medicine or injuries in sport research is the generalized linear model (GLM). This methodology is able to model counts and proportions besides normally distributed variables (McCullagh and Nelder, 1989). Furthermore, GLMs assume that the observations conditioned on the predictors

[1] Department of Public Health, Universitat de Barcelona, Barcelona, Spain.

[2] Epidemiology Service, Public Health Agency of Barcelona, Barcelona, Spain.

[3] Area of Biostatistics, Universitat Internacional de Catalunya, Barcelona, Spain.

[4] CIBER de Epidemiología y Salud Pública (CIBERESP), Spain.

[5] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya/ BARCELONATECH, Barcelona, Spain.

[6] Statistics Unit, Dalarna University, Falun, Sweden.

Received: December 2014
Accepted: August 2015

are independent and identically distributed. However, these assumptions may be violated in some situations, such as longitudinal studies, where there are repeated measures and, hence, correlated data. Ignoring correlation of data when fitting the model may lead to biased estimates and misinterpretation of results (Bolker et al., 2009).

Generalized linear mixed models (GLMMs) are an extension of GLMs adding random effects in the linear predictor term in a regression setting (Breslow and Clayton, 1993). The GLMM is a more flexible analysis approach for analyzing non-normal data and it is known to be useful for accommodating the overdispersion in Binomial or Poisson regression models, and modelling the dependence structure among outcome variables for longitudinal or repeated measures data (Williams, 1982; Breslow, 1984).

The main difficulty of these models is the estimation of their parameters, as it is often not viable to obtain an analytic solution that allows maximizing the marginal likelihood of the data. Due to this fact, different estimation methods based on approximation or simulation have been developed in recent years. One approximation using numerical integration is the Gauss-Hermite quadrature (GHQ) (McCulloch and Searle, 2001). However, there are alternatives to the marginal likelihood principle including Bayesian statistics and the extended likelihood principle. For example, the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) is a Bayesian implementation and the hierarchical (h-)likelihood is an implementation of the extended likelihood principle (Lee and Nelder, 1996, 2001). It is worth mentioning that the comparison between Bayesian and non-Bayesian methods is difficult to perform given that they are different principles.

Nowadays, GLMMs are implemented in most statistical software packages and several researchers have published and updated different guides and reviews of different software packages for fitting a GLMM. West et al. (2014) introduce the fitting and interpretation of several types of linear mixed models using the statistical software packages SAS, SPSS, Stata, R, and HLM. Dean and Nielsen (2007) review the theoretical background of generalized linear mixed models and the inferential techniques that have been developed for SAS, S-Plus, and contributed R packages. Bolker et al. (2009) describe the use of generalized linear mixed models for ecology and evolution and give information on available functions and packages in SAS or R. For further comparisons of statistical software for GLMMs for binary responses and frailty models, see, for instance, Zhang et al. (2011); Li et al. (2011); Hirsch and Wienke (2012); Kim et al. (2013); or Grilli et al. (2014).

The aim of this work is to compare three different statistical principles – marginal likelihood, extended likelihood, and Bayesian analysis; see, Table 1 – to estimate the parameters of a Poisson Mixed Model in R using both real and simulated data. It is structured as follows: in Section 2, we briefly review the definition of the GLMM and highlight the problem of deriving and maximizing the likelihood. In Sections 3 and 4, we give a theoretical description according to the statistical principle used. Several contributed R packages for the fit of GLMMs are presented in Section 5 and three of them are used in Section 6 for the analysis of the motivating real data set on Leonese Wrestling. These data are then used to define the settings of the simulation study presented in Section 7.

In Sections 8 and 9, the results of the simulation are presented and discussed, and recommendations are given on which statistical principle is, preferably, to be used in each of the settings under study.

2. Generalized linear mixed models

The GLMM extends the GLM by adding normally distributed random effects to the linear predictor. As Bolker et al. (2009) point out, GLMMs combine the properties of linear mixed models (LMMs) and GLMs by using link functions and exponential family distributions such as Binomial or Poisson distributions.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})'$ be a vector of m observations of the response variable of interest corresponding to subject i, $i = 1, \ldots, n$ and $\boldsymbol{u}_i$, $i = 1, \ldots, n$, be the random effects vector of the same subject. Conditional on $\boldsymbol{u}_i$, the distribution of $\boldsymbol{Y}_i$ is assumed to be from the exponential family type with density function $f(\boldsymbol{Y}_i|\boldsymbol{u}_i;\cdot)$ and with conditional mean $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{u}_i)$ and conditional variance $\mathrm{Var}(\boldsymbol{Y}_i|\boldsymbol{u}_i) = \Phi V(\boldsymbol{\mu}_i)$, where Φ is the dispersion parameter and $V(\boldsymbol{\mu}_i)$ is the variance function of the GLMM.

The definition of the GLMM is completed by introducing a monotone and differentiable function $g(\cdot)$ known as the link function (McCullagh and Nelder, 1989) and a linear predictor $\boldsymbol{\eta}$ as follows:

$$\boldsymbol{\eta}_i = g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i,\ i = 1, \ldots, n$$

where $\boldsymbol{X}_i$ (of dimension $m \times k$) and $\boldsymbol{Z}_i$ ($m \times l$) are subject i's design matrices associated with fixed and random effects, respectively. Vector $\boldsymbol{\beta}$ ($k \times 1$) is the fixed effects vector and $\boldsymbol{u}$ ($l \times 1$) is the random effects vector assumed to follow a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and unknown positive definite covariance matrix $\boldsymbol{\Sigma}$. Its density function is denoted by $f(\boldsymbol{u};\boldsymbol{\Sigma})$.

Estimation in the GLMM and theoretical description of the likelihood principle

By the local independence assumption, the conditional density of $\boldsymbol{Y}$ given $\boldsymbol{u}$ has the form

$$f(\boldsymbol{Y}|\boldsymbol{u};\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\boldsymbol{Y}_i|\boldsymbol{u}_i;\boldsymbol{\beta})$$

and the multivariate density function of $\boldsymbol{u}$ is given by

$$f(\boldsymbol{u};\boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{u}_i;\boldsymbol{\Sigma}).$$

The likelihood principle involves two kinds of objects: observed random variables (the data) and (unknown) fixed parameters. In the case of models with random effects, the estimation is based on the marginal likelihood where the random effects are integrated out (Birnbaum, 1962; Pawitan, 2001). Hence, the following likelihood function needs to be maximized in order to obtain the maximum likelihood (ML) estimates for $\boldsymbol{\beta}$ and the variance components in $\boldsymbol{\Sigma}$:

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{Y}) = f(\boldsymbol{Y}; \boldsymbol{\beta}) = \int f(\boldsymbol{Y}|\boldsymbol{u}; \boldsymbol{\beta}) f(\boldsymbol{u}; \boldsymbol{\Sigma})\, d\boldsymbol{u}. \tag{1}$$

The classical method that uses ML estimation and in which $\boldsymbol{u}$ is integrated out does not present problems with linear mixed models. The problem exists with GLMMs because of the more complicated integral (McCulloch and Searle, 2001). For this reason, one of the main interests of the research on the GLMM is to develop more efficient estimation methods for the fixed effects vector and the variance components.

Several ways to solve the integration in (1) and to obtain the marginal likelihood to estimate the parameters of a GLMM have been proposed. The Laplace method for integral approximation is considered to be a possible solution, which can be used to estimate the parameters of interests (Breslow and Clayton, 1993). Alternatives are the GHQ method or pseudo and penalized quasilikelihood methods (Aitkin, 1996). The GHQ method presents better estimation properties than the other methods because the GHQ estimates are maximum likelihood estimates. However, it is not feasible for analyses with more than two or three random effects because the speed of the GHQ decreases rapidly when increasing the number of random effects (Bolker et al., 2009).

3. Theoretical description of the extended likelihood principle

Lee and Nelder (1996, 2001) extended generalized linear models to include random effects by using their hierarchical (h-)likelihood method. This method is based on the extended likelihood principle (Bjørnstad, 1996) and is an implementation of the extended likelihood restricted by a weak canonical link for the random effects (Lee et al., 2006).

The h-likelihood is given by the log joint likelihood, that is, the extended likelihood L_E:

$$h = \log\big(L_E(y; \beta, v)\big) = \log\big(f(y; \beta|v)\big) + \log\big(f(v)\big)$$

where $\log(f(y; \beta|v)$ denotes the log of the density function with β as parameter and conditional on $v = v(\boldsymbol{u})$, where $\boldsymbol{u}$ is a vector of random effects and $\nu(\cdot)$ is an appropriate link function defining the h-likelihood. Unlike the GLMMs, the random effect is not restricted to be normal and can follow other distributions (e.g., gamma, beta, or inverse gamma).

A fundamental difference compared to classical marginal likelihood theory is that estimation and inference based on the h-likelihood includes random effects, whereas in classical likelihood theory the random effects are integrated out and a marginal likelihood is used. Hence, the use of the h-likelihood avoids the integration required for a classical marginal likelihood.

To estimate parameters (β, ν), the fixed and random effects are estimated from the score functions of the h-likelihood:

$$\frac{\partial h}{\partial \beta} = 0, \quad \frac{\partial h}{\partial \nu} = 0.$$

The variance components are estimated by maximizing the adjusted profile h-likelihood defined as

$$p_{\beta,u} = \left(h + \frac{1}{2}\log(2\pi \boldsymbol{H}^{-1})\right)\Big|_{\beta=\hat{\beta},\boldsymbol{u}=\hat{\boldsymbol{u}}},$$

where $\boldsymbol{H}$ is a Hessian matrix of the h-likelihood.

The estimates can be obtained by using iterative weighted least squares (IRWLS) as implemented in the `hglm` package (Rönnegård et al., 2010). The variance components are then estimated iteratively by applying a gamma GLM to the estimated deviances and with an intercept term included in the linear predictor and appropriate weights (Lee et al., 2006).

4. Theoretical description of the Bayesian principle

The Bayesian methods differ from the likelihood and extended likelihood principles in their philosophy as well as in the specific procedures used.

In order to implement a Bayesian principle, prior distributions are required for all parameters in the model, since under the Bayesian paradigm, all parameters are treated as random variables rather than fixed unknowns.

The Bayesian principle is based on assigning prior distributions to the parameters of the model. Thus, following the model defined in Section 2, the following prior distributions must be specified: $f(\boldsymbol{\beta}|\cdot)$, $f(\boldsymbol{u}|\cdot)$, and $f(\boldsymbol{\Sigma}|\cdot)$. These prior distributions express the beliefs on the parameters and these beliefs are modified by the data to obtain the posterior distribution of the parameters, $f(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}|\boldsymbol{Y})$, which is defined as to be proportional to the product of the prior distributions and the likelihood of the data. The posterior distribution is therefore used for inference purposes.

Here, a non-informative normal distribution is assumed as prior distribution for $\boldsymbol{\beta}$, that is, a normal distribution with a huge variance. Let $\boldsymbol{\gamma} = (\boldsymbol{u}, \boldsymbol{\beta})^{\mathsf{T}}$ denote the $G \times 1$ vector of Gaussian parameters. Concerning the random effects, we assume $\boldsymbol{u}$ to follow a multivariate normal distribution, $\boldsymbol{u}|\boldsymbol{\Gamma} \sim \mathcal{N}(0, \boldsymbol{\Gamma}^{-1})$, where the precision matrix $\boldsymbol{\Gamma} =$

$\boldsymbol{\Gamma}(\boldsymbol{\phi})$ depends on parameters $\boldsymbol{\phi}$. Let $\boldsymbol{\phi}$ also be the vector of the variance components for which the prior $\Pi(\boldsymbol{\phi})$ is assigned. However, often it is not possible to obtain an explicit expression of the posterior distribution and algorithms such as Markov chain Monte Carlo (MCMC) methods are used to generate the posterior distribution by simulation.

The Bayesian principle is attractive because it offers several advantages over the likelihood principle (e.g., it can increase the stability in small samples or in clustered binary data), but it has the difficulty of specifying prior distributions with variance components (Fong et al., 2010).

The use of MCMC methods for GLMMs is the most popular approach, but has problems in terms of convergence and computational time. These problems with Bayesian estimation have been greatly improved by Integrated Nested Laplace Approximations (Rue et al., 2009).

Integrated Nested Laplace Approximation (INLA)

INLA is a new tool for Bayesian inference based on latent Gaussian models introduced by Rue et al. (2009). The method combines Laplace approximations and numerical integration in a very efficient manner. For the GLMM described in Section 2 and using $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ as defined in the previous paragraphs, the posterior density is given by

$$\pi(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{Y}) \propto \pi(\boldsymbol{\gamma}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\prod_{i=1}^{m} p(\boldsymbol{Y}_i|\boldsymbol{\gamma},\boldsymbol{\phi}).$$

It is computed via numerical integration as

$$\pi(\boldsymbol{\gamma}|\boldsymbol{Y}) = \int \pi(\boldsymbol{\gamma}|\boldsymbol{\phi},\boldsymbol{Y})\pi(\boldsymbol{\phi}|\boldsymbol{Y})d\boldsymbol{\phi},$$

where Laplace approximation is applied to carry out the integrations required for the evaluation of $\pi(\boldsymbol{\gamma}|\boldsymbol{\phi},\boldsymbol{Y})$. For more details we refer the readers to Rue et al. (2009).

5. Contributed R packages for GLMMs

For the likelihood principle, there exist different packages in R such as `glmmML` (Broström and Holmberg, 2013), `lme4` (Bates et al., 2015), or the function `glmmPQL` in the `MASS` package (Venables and Ripley, 2002). One of the most popular and stable functions for fitting GLMMs is called `glmer` and is found within the package `lme4`. This package implements the GHQ to approximate the log-likelihood using numerical integration. By default, it uses the Laplace approximation with one quadrature point.

For the extended likelihood principle, two packages are available for fitting Hierarchical Generalized Linear Models with random effects: `hglm` (Rönnegård et al., 2010) and `HGLMMM` (Molas and Lesaffre, 2011).

Concerning packages for performing Bayesian inference on GLMM, general packages such as `glmmBUGS` (Brown and Zhou, 2010) and `R2WinBugs` (Sturtz et al., 2005), and specialized packages such as `glmmAK` (Komárek and Lesaffre, 2008), `MCMCglmm` (Hadfield, 2010), and `INLA` (Lindgren and Rue, 2015) exist. INLA substitutes MCMC simulations with accuracy and the quality of such approximations is extremely high.

For both the analysis of the wrestling data (Section 6) and the simulation (Section 7), the R packages `lme4`, `hglm`, and `INLA` were used. Note that the `lme4` package does not report standard errors for variance components. The reason is provided by the developer of the package in his book (Bates, 2010) stating that the sampling distribution of the variance is highly skewed, which makes the standard error nonsensical (Li et al., 2011). Regarding the GHQ method, we used 5 quadrature points since it was indicated that this method can give a poor approximation to the integrated likelihood when the number of quadrature points is low (Lesaffre and Spiessens, 2001). This method can be made arbitrarily accurate by increasing the number of quadrature points. We adopted a strategy of increasing the number of quadrature points until there was a negligible difference in the values of the estimators (Ormerod and Wand, 2012).

Recent changes of the packages used

The three packages studied in this paper have implemented some new features in their latest versions.

Regarding the `lme4` package, the authors of the package have been discussing new features and new versions of the package through the forum "R-sig-mixed models" (`https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`) since its version 0.99 (better known and used as lme4.0). Currently, the present version is 1.1-7, which offers some advantages with respect to the version 0.99, especially in terms of convergence and optimization. These developments can be found at the following link: `https://github.com/lme4/lme4`.

Concerning the `hglm` package, since version 2.0 it is possible to fit several random effects from different distributions (e.g., gamma or Gaussian), to fit a linear predictor for the dispersion of the random effects, to fit spatial conditional autoregressive (CAR) and spatial autoregressive (SAR) models for the random effects, and to perform a likelihood-ratio test for the dispersion parameter of the random effects (Alam et al., 2014). The method options have also been extended to include the EQL1 method which is a "HL(1,1) correction" (Lee and Lee, 2012; Noh and Lee, 2007) applied on the default EQL method. These developments can be found at the following link: `http://cran.r-project.org/web/packages/hglm/vignettes/hglm.pdf`.

Regarding the INLA package, this analytical framework includes normally distributed latent variables and thus allows for hierarchical data structure, but it is targeted towards complex applications involving temporal and spatial smoothing where MCMC estimation may be too difficult to apply (requiring specialized MCMC samplers) or prohibitively slow. On the one hand, now there is an increase of articles using this package in several application fields such as fishering (Cosandey-Godin et al., 2014) or ecology (Quiroz et al., 2014). Few applications of the INLA methodology in injury epidemiology or sport science have been published (Cervone et al., 2014), and it would be of interest to study in more detail the medical impact in terms of understanding sport injuries with this methodology. On the other hand, one of the most recent papers of Martins et al. (2014) shows a new perspective on the selection of default priors.

6. Real data example: folk wrestling data

Leonese Wrestling (LW) or Aluche is a traditional and popular sport of the province of León, in Northwestern Spain. It is registered with and recognized by three international associations: *Fédération Internationale des Luttes Associées* (FILA), *Asociacion Española de luchas tradicionales* (AELT), and International Belt Wrestling Association (IBWA), respectively. Like with all styles of wrestling, the risk of injury is always present.

The main variable of interest in epidemiological investigation of sports injuries is the incidence of injury, which is generally expressed as the proportion of injuries per fight (Ayán et al., 2010; Hägglund et al., 2010). There are few studies in the international scientific literature about the incidence of injury in combat sports and its associated factors (Klügl et al., 2010; Hewett et al., 2005). Nonetheless, in published papers, it has been found that the incidence of injury in these sports is higher than in other sports activities (Hägglund et al., 2005; Junge et al., 2009).

Concerning factors associated with the incidence of injury, it is known that this incidence is higher during wrestling matches than in training. However, there is not much information on the frequency of injuries, their incidence, and their causes to carry out prevention and control programs in this sport. This lack of knowledge has motivated this analysis of the impact and risk factors of injuries.

Data on matches and injuries of the LW summer seasons were available for 213 wrestlers during the summer seasons from 2005 through 2010. The response variable of interest was the frequency of injuries which was assumed to follow a Poisson distribution. The study design was unbalanced with different numbers of repeated measures given that not all wrestlers participated in official competitions in each of the six years from 2005 to 2010. The possible risk factors for injuries considered were: i) Winner: This variable is defined as a function of the falls during a match. It is set to 'Yes' if the wrestler had more falls in his favor than against him; otherwise, the value of Winner is set to 'No'; and ii) Weight category: a categorical variable with levels Light, Medium

(chosen as reference), Semi-heavy, and Heavy. It has been taken into account that these variables could change from one season to another.

6.1. The model under study

Let $\boldsymbol{Y}_i, i = 1,\ldots,n$, be the vector of the number of injuries per season of wrestler i. The length of $\boldsymbol{Y}_i$ depends on the number of seasons the wrestler took part in official competitions. It is assumed that the distribution of $\boldsymbol{Y}_i$ follows a Poisson distribution: $\boldsymbol{Y}_i \sim Po(\boldsymbol{\mu}_i)$. Usually, the counts are considered in relation to some differential or offset (λ) in order to obtain rates. Here, the offset is the number of the wrestler's matches per season. The Poisson generalized linear mixed model used to analyze the data links the mean of $\boldsymbol{Y}_i$ with both covariates $\boldsymbol{X}_i$ of interest —Winner and Weight category— by means of the following equation

$$\log(\boldsymbol{\mu}_i) = \log(\boldsymbol{\lambda}_i) + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{u}_i,\ i = 1,\ldots,n, \tag{2}$$

where the vector $\boldsymbol{\beta}$ contains the fixed effects parameters and $\boldsymbol{u}_i$ stands for the random effect intercept for wrestler i. Random effects are assumed to be independent and normally distributed with mean $\mathbf{0}$ and variance σ^2. Random effects for the slope parameters were not considered in order to keep the model simpler. In addition, posterior model fits including such random effects did not improve the model fit significantly at a 0.05 significance level.

The model's marginal variance (over subjects) can be expressed as

$$\mathrm{Var}(\boldsymbol{Y}_i) = \Phi \cdot \boldsymbol{\mu}_i,\ i = 1,\ldots,n,$$

where Φ is the dispersion parameter. The Poisson distribution assumes $\Phi = 1$ and if $\Phi > 1$, overdispersion is present. In that case, the data have larger variance than expected under the assumption of a Poisson distribution.

The dispersion parameter can be estimated based on the χ^2 approximation of the residual deviance or Pearson residuals. The dispersion parameter is estimated by dividing the χ^2 statistic by the residual degrees of freedom, $n - r$:

$$\hat{\Phi} = \frac{1}{n-r}\sum_{i=1}^{n}\frac{(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^2}{\hat{\boldsymbol{\mu}}_i}. \tag{3}$$

If there is high overdispersion, the negative binomial distribution is an alternative to the Poisson distribution for count data because the negative binomial distribution allows for a variance greater than mean. Apparent overdispersion is normally due to missing covariates or interactions, outliers in the response variable, non-linear effects of covariates entered as linear terms in the systematic part of the model, or the choice of

a wrong link function. These are mainly model misspecifications. Real overdispersion exists when none of the previous causes can be identified. The reason for this might be that the variation in the data is, actually, larger than the mean. Or, there may be many zero observations, clustering of observations, or correlation between observations (Hardin and Hilbe, 2007; Zuur et al., 2009).

6.2. Results

We fitted Model (2) with packages `lme4`, `hglm`, and `INLA`, and for the sake of comparison, we also analysed the data as if they were not correlated. That is, we fitted a GLM using function `glm` of the R package `stats`. All analyses were carried out with R, version 3.1.1, and the estimates obtained, together with the 95% confidence and (in the case of the `INLA` package) credible intervals, are presented in Table 2. The confidence intervals of the `lme4` package were calculated using the function `confint.merMod`, which computes a likelihood profile and finds the appropriate cutoffs based on the likelihood ratio test (by choosing `method = "profile"`).

We calculated the dispersion term based on the Pearson residuals using Equation (3) for function `glm` and packages `lme4` and `hglm`. In addition, we checked the possible overdispersion using an individual-level random variable (translating to a lognormal-Poisson model, which is qualitatively similar to a negative binomial (Elston et al., 2001)).

According to the results obtained, the estimates of the coefficients of the linear predictor are quite similar with only slight differences in the second decimal digit. The same is true for the confidence and credible intervals obtained with packages `lme4`, `hglm`, and `INLA`, whereas function `glm` provides smaller confidence intervals because it treats the data as if they were independent observations. Moreover, it can be seen that the estimate of the random effect variance and the dispersion term differ a little from each other.

Concerning both variables of interest, the positive signs of the parameter estimates corresponding to the weight category indicate a higher risk for injuries among all three weight categories as compared with the medium weight category. Nonetheless, all 95% confidence and credible intervals include 0, that is, the differences are not statistically significant at a 0.05 significance level. In the case of variable Winner, the model indicates a lower injury risk – $\hat{\beta} < 0$ and the 95% confidence and credible intervals do not include 0 – among wrestlers with more falls in their favor.

7. Simulation study

In this section, we present the simulation study designed to assess the performance of the three statistical principles using different scenarios based on the wrestling data. The aim is to compare the three methods in different settings defined by overdispersion or sample size with respect to measures of model accuracy, precision, empirical bias, and

empirical coverage of the estimators as well as computation time and possible problems of convergence. A total of 40 different simulation settings were used that can be classified into two main simulation scenarios.

7.1. Simulation scenario 1

For the first simulation scenario, we used the structure of the real data set introduced in Section 6. The values of the response variable – number of injuries – were generated as a function of the observed values of the independent variables and the number of matches of each wrestler in each of the years under study following the model expression in (2). The aim of this scenario was to closely represent the structure of this real data set.

We simulated the number of injuries using the parameter values given in Table 2 obtained with the `lme4` package ($\beta_1 = 0.24, \beta_2 = 0.1, \beta_3 = 0.4, \beta_4 = -0.46$). Concerning the model intercept (β_0) and the variance of the random effects (σ_u^2), we used combinations of both parameters that lead to three different values of overdispersion ($\Phi \in \{1.5, 3, 10\}$) that we identify as low, moderate, and high overdispersion settings; see McCulloch and Searle (2001) for technical details. Furthermore, we added the value of $\Phi = 1$, i.e., no overdispersion, to also assess how the GLMM behaves in this situation. In addition, the values of β_0 were chosen such that two different marginal means of the injury numbers were obtained ($\mu = 1$ and $\mu = 10$). As can be seen in the R code of the first simulation study in the Supplemental Material, the values of β_0 ranged from -4.8 to -1.7, those of σ_u^2 from 0 to $2.1^2 = 4.41$.

In total, with four different values of dispersion and two of the marginal mean, the number of simulation settings for the Simulation scenario 1 was $4 \cdot 2 = 8$.

7.2. Simulation scenario 2

The second simulation scenario was motivated by the goal to study the effect of different sample sizes on the parameter estimation. For this purpose, we considered two different sample sizes of $n = 30$ and $n = 100$ wrestlers and for each wrestler, a random number of seasons was generated using a discrete uniform distribution ranging from 1 through 6. In the sequel, the match numbers for each wrestler and season were generated using a Poisson distribution with parameters 60 and 100, respectively. These were the offset terms $\boldsymbol{\lambda}_i$ in Model (2). The remaining parameters were chosen similar to the first simulation scenario resulting in four dispersion parameters ($\Phi \in \{1, 1.5, 3, 10\}$) and two marginal means for the number of injuries ($\mu = 1$ and $\mu = 10$) so that the number of simulation settings was $2 \cdot 2 \cdot 4 \cdot 2 = 32$ for Simulation scenario 2.

The values of both independent variables – weight category and winner – were generated using equal probabilities for all categories and, as in Simulation scenario 1, the injury numbers were generated using the expression of the model in (2).

7.3. Evaluation criteria

For each of the $8+32=40$ simulation settings, we simulated 1000 data sets of the model under study and used the three methods to estimate the model parameters (Marginal Likelihood, Extended Likelihood, and Bayesian Analysis). In addition, we used R function `glm` treating the data as if we dealt with a GLM. Measures for the comparison of the different estimation methods were the empirical mean squared error (MSE) as a measure of model accuracy, the ratio of precision, the empirical bias, and the empirical coverage of the confidence and credible intervals, respectively. Moreover, we recorded the computation times and studied possible problems of convergence.

For each simulation setting and estimation method, the empirical bias was calculated as the mean bias over the 1000 data sets and its squared value was used together with the empirical variance to compute the empirical MSE. The rate of precision was computed as the ratio between the estimator's empirical variance and the mean of the squared standard errors. In order to calculate and compare the empirical bias and the empirical MSE in the case of the `INLA` package, the distribution of the parameters provided by `INLA` were reduced to only one value (the posterior mean).

For the likelihood and the extended likelihood principles, we used the 95% confidence interval and for the Bayesian principle, we used the 95% credible interval of parameter given by the 0.025 and 0.975 sample quantiles of the posterior parameter distributions. In the case of the `lme4` package, we used 5 quadrature points for the GHQ method and non-informative priors were assumed for the Bayesian analysis. Moreover, the random intercepts were assumed to have a normal distribution. Regarding the prior distribution for the precision, a half-normal distribution with mean 0 and precision 0.0001 was assigned to the standard deviations (Gelman et al., 2006).

The comparison was done for the two main parameters of interest: the parameter β_4, which corresponds to the variable Winner, and the variance of the random effects (σ_u^2). The former was chosen, since the analysis of the wrestling data showed a statistically significant association between the number of injuries and this variable. Whereas the value of σ_u^2 varied across the simulation settings, the value of β_4 was kept constant in all settings.

8. Results of the simulation study

For the simulation scenarios, results are presented only for the intercept (β_0), the slope (parameter β_4, corresponding to the covariate Winner), and variance of the random effect (σ_u^2). The performance of the estimation methods in terms of empirical bias, empirical MSE, precision ratio, and empirical coverage of $\hat{\beta}_4$ and $\hat{\sigma}_u^2$ is summarized in Figures 1a, 1b, 2a, and 2b. The corresponding figures for $\hat{\beta}_0$ are provided in the Supplemental Material.

In order not to mix up statistical principles, methods, and algorithms as presented in Table 1, following, we only refer to these by the names of the R packages used.

Note that the results are yielded by package INLA with the prior distribution selected, a half-normal distribution with mean 0 and precision 0.0001.

Concerning the percentage of convergence of the estimation methods, a model was considered as "not convergent" if either the estimation process did not converge or if the estimate or its standard error was not provided. For example, in some cases the parameter can be estimated but the estimation process may be unable to provide a positive definite variance-covariance matrix of the parameters (for problems with the Hessian matrix), mainly due to the instability of the model. Convergence was checked and obtained using the criteria offered in each software package. In the case of the first simulation scenario, the convergence percentages were always equal to 100% with only one exception: package hglm achieved convergence in 99.2% of all data sets in the case of $\mu = 1$ and $\Phi = 10$. The results for Simulation scenario 2 are shown in Table 3. The rate of convergence of all estimation methods was close to 100% for most of the settings. However, in the case of packages hglm and lme4, the percentage of convergence slightly decreased in some settings with $\mu = 1$ and $n = 30$ as overdispersion increased.

Regarding the empirical bias of the slope, all packages provided mostly unbiased and similar estimates. In terms of accuracy, the highest empirical MSE for the slope for all GLMM packages is given when $\Phi = 10$, $\mu = 1$, and $n = 30$. In this case, function glm is the one that presents highest values. The empirical MSE value obtained with INLA is higher than with hglm and lme4, which are both similar; see Figure 1a.

In terms of precision (upper panel of Figure 1b), we calculated the ratio of the estimator's empirical variance and the mean of the squared standard errors as a precision measure. In general, we found that almost all estimation methods presented an underestimation in the case of $\Phi = 1$ and $\Phi = 1.5$ together with $\mu = 10$. More differences between the packages were observed with sample size equal to 30. In the case of the lme4 package, the ratio was slightly larger than 1 (equivalent to 100%) especially with moderate and high overdispersion. By contrast, the values of the hglm and INLA packages were close to 100% for that sample size independently of the offset, the marginal mean, and the dispersion term. The function glm, in general, showed values far larger than 100% (and, hence, out of the range of the corresponding plots) from $\Phi = 1.5$.

The empirical coverage of the confidence intervals for the GLMM packages were close to 95% in all settings; see the lower panel of Figure 1b. The GLM appeared to have bad coverage, only acceptable for those combinations with low overdispersion. It suffered from substantial undercoverage (down to 75%) when $\Phi = 3$ and $\Phi = 10$. This result may be expected for the GLM since it does not include random effects and therefore can not assume any overdispersion. As overdispersion increased, the empirical coverage behavior became worse.

Regarding the empirical bias of the variance component (upper panel of Figure 2a), the three packages performed similarly for $\Phi = 1$ and $\Phi = 1.5$. For $\Phi = 3$, the INLA package showed the largest empirical bias with $n = 30$, whereas with $n = 100$, the differences among the packages were small for $\Phi = 3$. For $\Phi = 10$ and $n = 30$, package lme4 had the smallest empirical bias in terms of the absolute value, INLA the largest

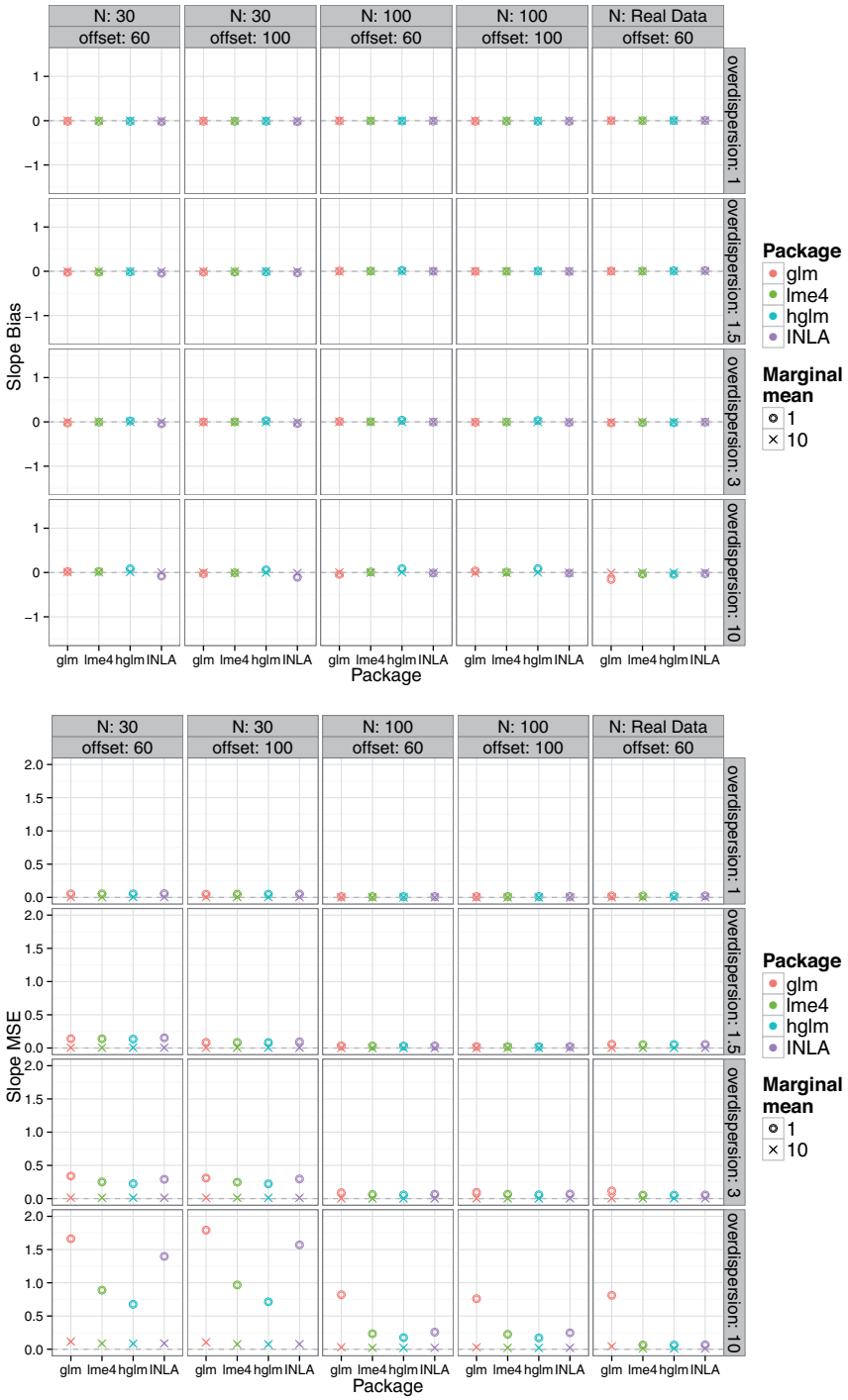


Figure 1a: *Empirical bias (upper panel) and empirical MSE of the slope estimate ($\hat{\beta}_4$) as a function of overdispersion, marginal mean, offset, and sample size.*

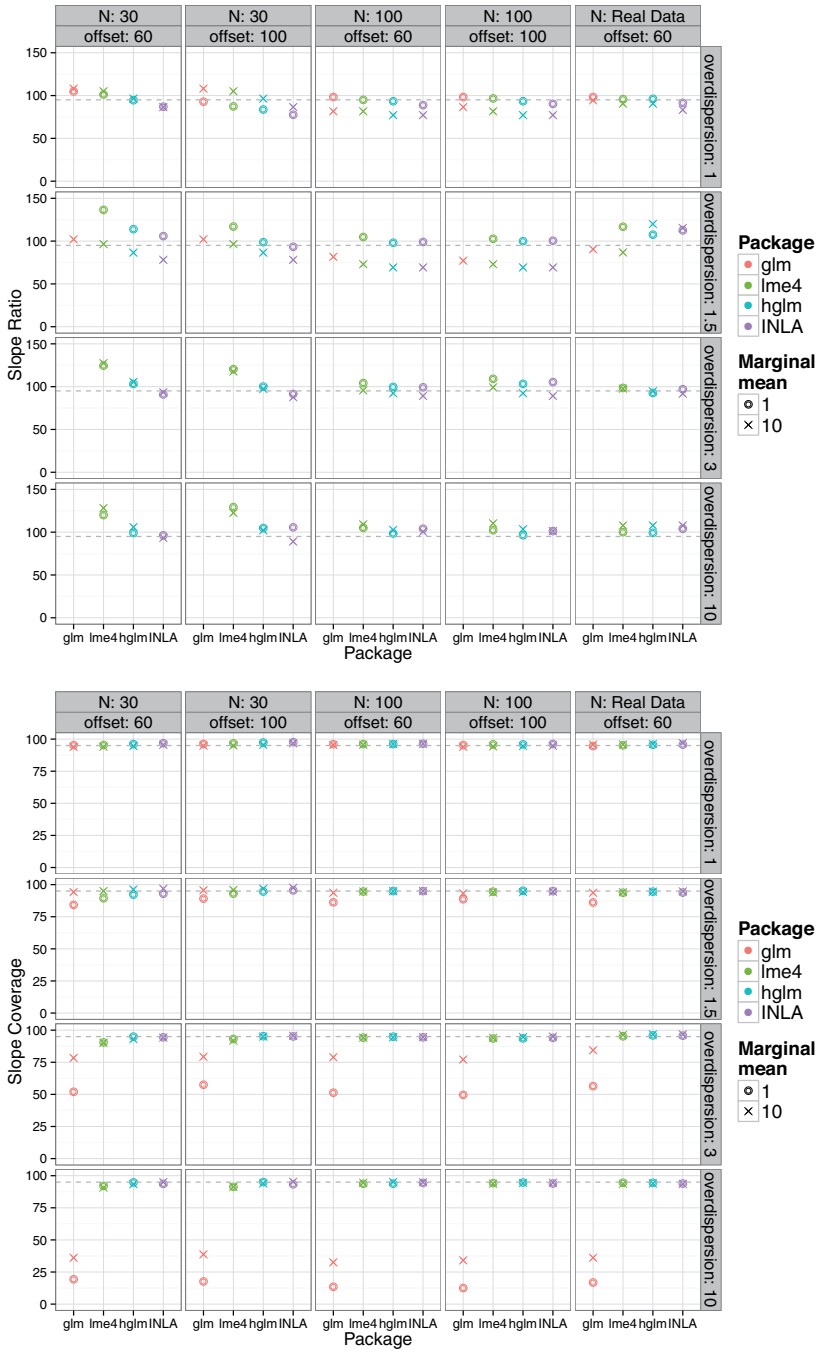


Figure 1b: *Precision (upper panel) and empirical coverage of the slope estimate ($\hat{\beta}_4$) as a function of overdispersion, marginal mean, offset, and sample size. Precision is measured as the ratio of the estimator's empirical variance divided by the average of the squared standard errors.ope estimate ($\hat{\beta}_4$) as a function of overdispersion, marginal mean, offset, and sample size.*

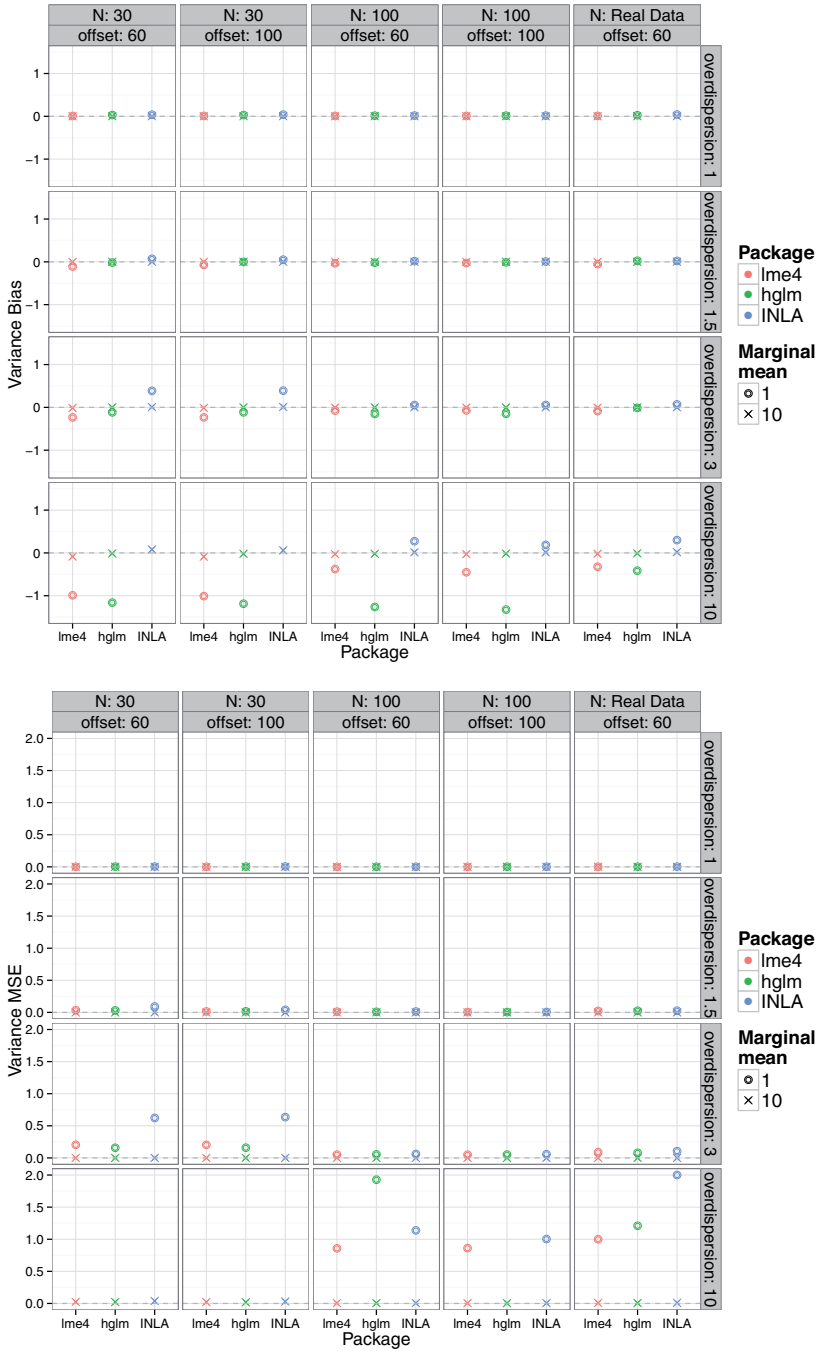


Figure 2a: *Empirical bias (upper panel) and empirical MSE of the variance component estimate ($\hat{\sigma}_u^2$) as a function of overdispersion, marginal mean, offset, and sample size.*

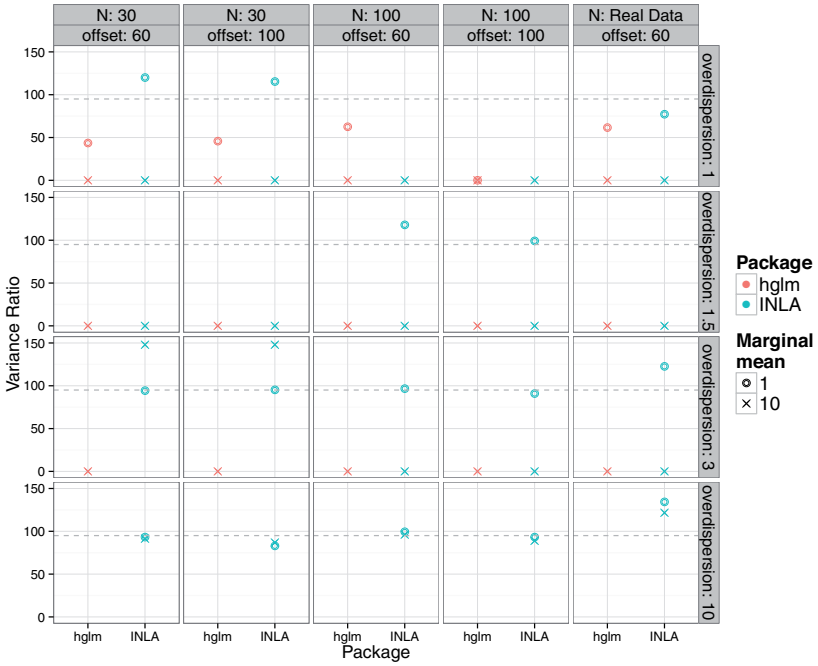


Figure 2b: *Precision of the variance component ($(\hat{\sigma}_u^2)$) as a function of overdispersion, marginal mean, offset, and sample size. Precision is measured as the ratio of the estimator's empirical variance divided by the average of the squared standard errors.*

(with values out of the range of the plot). By contrast, for $\Phi = 10$ and $n = 100$ the absolute values of the empirical bias of `lme4` and `INLA` were roughly the same. It was largest for `hglm` in that setting. Values for function `glm` do not appear in that figure since a GLM does not consider any random effects.

In terms of the accuracy of the variance component (lower panel of Figure 2a), the empirical MSEs were very similar except when $\Phi > 1.5$ and $\mu = 1$. The `INLA` package had the largest empirical MSE when $\Phi = 3$, $\mu = 1$, and $n = 30$. By contrast, in this same setting, the `lme4` and `hglm` packages had very similar values. None of the packages showed a satisfactory behavior when $\Phi = 10$, $\mu = 1$, and $n = 30$: all empirical MSEs are excessively high and, hence, out of the range of the corresponding plot. With $n = 100$, the empirical MSE of package `hglm` was still out of the plot's range, whereas that of `lme4` and `INLA` were close to 1.

Concerning the precision of the estimation of the variance component (Figure 2b), the ratio obtained with `hglm` and `INLA` was close to zero when $\Phi \neq 10$ and $\mu = 10$. For $n = 30$, in `INLA`, the values were close to 150 when $\mu = 10$, and they were close to 100 when $\Phi > 1.5$ and $\mu = 1$. For `hglm` and when $\Phi > 1$ and $\mu = 1$, the values were excessively high in most of the simulation settings. It could not be computed with `lme4` since this package does not provide the standard error of the estimation of the random effect's variance.

Contrary to overdispersion, sample size, and marginal mean, the choice of the offset did not seem to have any effect on the estimators' performance.

Finally, we also compared the computational times measured with R function `system.time`. Among the three packages that consider random effects, the average computing times of packages `lme4` and `hglm` were very similar in each setting. On average, they were four times faster than package `INLA`.

In summary, the approaches involving a random effect (`lme4`, `INLA`, and `hglm`) showed good performance on estimating the model parameters except for the estimation of the random effect's variance in the case of combinations of huge overdispersion, a small marginal mean, and a small sample size. Given that the empirical MSE and the empirical bias of the `lme4` package are close to zero for most of the simulation settings, it seems that this package, generally, outperforms the other packages, even though often only slightly.

9. Discussion

An overview of statistical principles for GLMMs is presented in this paper. The problem of selecting the best approach for estimation and inference within Poisson Mixed Models is very complex and too difficult to solve analytically. For this reason, we have carried out a simulation study that has evaluated the impact of overdispersion, marginal mean, offset, and sample size assuming Poisson Mixed Models using different statistical principles.

The fact that the bias and mean square error improve with larger sample size can be interpreted as that the estimators are consistent when overdispersion is taken into account in the model by means of random effects. By contrast, the empirical bias and MSE were larger with the GLM since this type of model does not include random effects and, hence, ignores overdispersion (Bolker et al., 2009; Milanzi et al., 2012). The results for `INLA`, in general, did hardly differ with the values obtained with both `lme4` and `hglm`; at least not with the prior distribution used.

We have found that for small sample sizes, the random effects variances are difficult to estimate, which has also been described in other studies (Li et al., 2011). In addition, the results are worse in the case of a moderate dispersion term ($\Phi = 3$) and especially with high overdispersion ($\Phi = 10$), in which none of the methods provide satisfactory results. As Zuur et al. (2009) explain, with a dispersion term up to 1.5 there are no important overdispersion problems. To solve overdispersion problems, other methods and distributions may be used, for instance, the Poisson-lognormal distribution, GEE models, or quasi-Poisson distributions (Booth et al., 2003; Bolker et al., 2009). In the case of high overdispersion with real data, it would be more reasonable to change the Poisson distribution for another distribution that does not have the restriction of the variance equalling the mean, for instance, the Negative Binomial distribution (Czado et al., 2007). Recently, Aregay et al. (2013) suggested a Hierarchical Poisson-Normal overdispersed

model (HPNOD) as an alternative using the Bayesian principle. The HPNOD performs better than a Hierarchical Poisson-Normal model for data with low, moderate, and high overdispersion.

In the simulation study, for most combinations, we observed similar performance in terms of the empirical bias and the empirical MSE whatever was the package applied. On the other hand, differences arise with respect to the precision of estimates. This fact may indicate a problem of underestimation because the methods do not capture well all the variability present in data. A bootstrap approach may be a solution to solve the standard error problem.

Regarding computational time and convergence, the `glm` function requires less time because it does not capture the presence of the random effect. The `hglm` and `lme4` packages need similar computational times, whereas the `INLA` package takes more time. Nonetheless, at least for the simulation settings under study, computation time was not excessive and results were obtained in less than 5 seconds in most of the cases. In some combinations of the simulation with small sample size ($n = 30$), huge overdispersion ($\Phi = 10$), and small marginal mean ($\mu = 1$), we found problems of convergence in the `hglm` package. To solve a convergence problem we recommend specifying other starting values.

Several studies carried out until now have compared estimation methods only for the Bayesian principle, the marginal likelihood principle or both in the GLMMs (Zhang et al., 2011; Ormerod and Wand, 2012; Li et al., 2011; Kim et al., 2013). For example, Li et al. (2011) recommend the use of the `lme4` package. The authors highlight that in case a Bayesian package is chosen, the parameter estimates might be influenced by the priors for the variances of the random effects. They also mention that when the data set is small, the random effects' variances are difficult to estimate with both frequentist and Bayesian methods. Kim et al. (2013) recommend the use of the GHQ method given that it performs well in terms of accuracy, precision, convergence rates, and computing speed. According to the authors, this is also valid with small sample sizes and for longitudinal studies with a few time points. On the other hand, there are some studies that compare the extended likelihood approach with the Bayesian principle (Pan and Thompson, 2007; Collins, 2008). However, they do not take into account the `INLA` method, which is a recently proposed approximate Bayesian approach for latent Gaussian models. According to Collins (2008), in some case studies both Bayesian and extended likelihood approach estimators of the variance component were positively biased, whereas GHQ method based estimators were not.

Our work is different from these previous studies given that we have focused our work on different estimation methods, principles, and more commonly used free software packages. Regarding the real data example, most GLMMs are used in applications of ecology, epidemiology, genetics, clinical medicine, and other applications but these models are now gaining attention in sports sciences, too (Avalos et al., 2003; Bullock and Hopkins, 2009; Sampaio et al., 2010; Casals and Martínez, 2013; Casals et al., 2014). However, there are only a few studies in epidemiology of sports injuries.

Concerning the Bayesian principle, when the sample size is small, the posterior distribution may be more influenced by the choice of prior distributions than when the sample size is moderate or large. Note that the posterior means depend on the choice of the non-informative prior of the variance component (Li et al., 2011). Bayesian algorithms such as MCMC offer different advantages over frequentist algorithms but they have problems in terms of convergence and computational time. These aspects have improved with `INLA`. However, for some combinations, the default initial value is not adequate for the data because of the use of very weak priors. For example, it is known that the inverse Gamma$(0.001, 0.001)$ prior is not always a good choice (Fong et al., 2010). Frühwirth-Schnatter and Wagner (2010b) and Frühwirth-Schnatter and Wagner (2010a) demonstrate overfitting due to Gamma-priors and suggest using a (half) Gaussian prior for the standard deviation to overcome this problem, as suggested by Gelman et al. (2006). Another interesting possibility that could be considered in the future is the use of Penalized Complexity (PC) priors that have heavier tails than the half Gaussian, but lighter tails than the half-Cauchy (Martins et al., 2014). According to the developers of the `INLA` package, the use of PC priors works pretty much identically to the half-Cauchy in practice.

There are some limitations of the present work. First, in this study we have worked with a Poisson mixed model with a random intercept, but we have not considered models with random slopes. The reasons for this decision were twofold: On one hand, the inclusion of random slopes in the real data example did not significantly improve the model fit at a 0.05 significance level. On the other hand, to study the impact of overdispersion, marginal mean, sample size, and offset, we preferred to analyze a Poisson mixed model with a random intercept due to its frequency in sports medicine research. Given the importance of more complex mixed models, future research should investigate the performances of these principles in mixed Poisson models including random slopes, cross-classified random effects and multiple membership structures that may be analyzed in future simulation studies. Second, we only examined three packages corresponding to the three estimation methods and principles. There are other R packages such as `glmmML` and `MCMCglmm` as well as the function `glmmPQL` of the `MASS` package that could be included in further simulation studies. In addition, such simulation studies could include other software packages such as SAS, STATA, or SPSS, which were not considered for the present work. We decided to focus our work on R because of its great popularity among statisticians (Muenchen, 2015) and the constant development of new packages and functions to deal with GLMMs.

In addition to the two limitations mentioned, the simulation study could have also studied other parameters of interest such as the cluster size, which, in the real data example in Section 6, is equivalent to the number of seasons of a wrestler. Indeed, several simulation studies point out that for binary responses, the performance of the estimators is influenced by the cluster size, e.g., that clusters of size two usually entail problems (Breslow and Clayton, 1993; Diaz, 2007; Kim et al., 2013; Grilli et al., 2014). Following the suggestion of one of the reviewers, we decided to assess the role of the cluster size

on empirical bias and MSE with a small additional simulation study with eight different settings defined by two values of cluster size ($m \in \{2, 5\}$), two values of overdispersion ($\Phi \in \{1.5, 10\}$), and two different marginal means ($\mu \in \{1, 10\}$). The offset ($\lambda = 60$), the sample size ($n = 100$), and the slope parameter ($\beta_4 = -0.46$) were kept constant. Regarding the results of the empirical bias and the MSE, which are shown in Tables 4 and 5 in the Supplemental Material, there were hardly any differences between cluster sizes two and five. However, the estimates obtained by the three packages had larger empirical bias and MSE when $\Phi = 10$ and $\mu = 1$. Under this scenario, the values of package `hglm` were larger than those of the `lme4` and `INLA` packages. Moreover, with respect to computing time under the settings of this additional simulation study, the `hglm` package was somewhat faster than `lme4`. The computing times of the `INLA` package were, again, much larger. Problems of convergence were not detected.

It is important to highlight that the concepts of estimation and standard error are different in the three statistical principles used. In the case of the marginal likelihood, estimation is the value that maximizes the likelihood and the standard error reflects the sample variation of the estimator. In Bayesian analysis, estimation is a summary of the posterior distribution and the standard error is a measure of the variability of this distribution. As for the classical likelihood, the extended likelihood advocates the use of standard errors computed from the Fisher information matrix of the marginal likelihood. In practice, the standard errors are computed from the matrix of second derivatives for the adjusted profile h-likelihood, which is an approximation of the marginal likelihood (Lee et al., 2006).

Parameter estimation of GLMM is nowadays possible using different statistical principles, though estimation methods as well as statistical packages are still under development (Bolker et al., 2009). The problem of selecting the most adequate approach for the estimation and inference within GLMM is very complex. In addition, the software implementations can differ considerably in flexibility, computation time and usability (Austin, 2010; Li et al., 2011). A strategy could be to carry out a simulation study that emulated the data design and to apply the different estimation methods. Although one may think that this strategy is not very practical, it would be indeed worse to use an estimation method that could provide biased and inefficient estimates.

We have shown through simulations that ignoring overdispersion in Poisson Mixed Models can have serious consequences on the parameter estimates. Available R packages can handle this problem very satisfactorily; however, care must be taken in situations with small sample size, large overdispersion, and small marginal mean. In such situations, the `lme4` package seems to have a slightly better performance than packages `hglm` and the `INLA`, which also depends on the choice of the prior (Grilli et al., 2014), especially concerning the estimation of the random effect's variance (Figure 2a). This observation coincides with the recommendation of Kim et al. (2013) to use the GHQ method under such settings and the active discussions of the package's authors (Douglas Bates, Martin Maechler, and Ben Bolker) and members of the R sig-mixed-models mailing list (`https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`). All pack-

ages under study have recently been improved in terms of convergence and optimization.

Appendix

Table 1: *Overview of statistical principles.*

Principle	**Method**	**Algorithms**
Marginal Likelihood	Maximum likelihood	Newton-Raphson (N-R), Fisher scoring, Penalized iteratively reweighted least squares (PIRLS) Adaptative Gauss Hermite Quadrature (GHQ)
Extended likelihood	h-likelihood	N-R, Iterative weighted least squares (IRWLS)
Bayesian	Posterior mean	MCMC, Integrated Nested Laplace Approximations (INLA)

Table 2: *Results from the Poisson mixed model applied to the folk wrestling data. CI stands for confidence interval and credible interval (in the case of INLA), respectively.*

	Function `glm`		Package `lme4`		Package `hglm`		Package `INLA`	
	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI	$\hat{\beta}$	95% CI
Intercept	−4.34	[−4.73, −4.0]	−4.37	[−4.82, −3.99]	−4.37	[−4.79, −3.95]	−4.41	[−4.85, −4.01]
Weight category[1]								
Light	0.25	[−0.2, 0.71]	0.24	[−0.26, 0.76]	0.25	[−0.28, 0.78]	0.25	[−0.26, 0.76]
Semiheavy	0.1	[−0.36, 0.57]	0.1	[−0.41, 0.63]	0.11	[−0.43, 0.65]	0.12	[−0.4, 0.64]
Heavy	0.39	[−0.1, 0.87]	0.4	[−0.14, 0.96]	0.4	[−0.16, 0.97]	0.41	[−0.14, 0.97]
Winner	−0.48	[−0.82, −0.15]	−0.46	[−0.82, −0.07]	−0.46	[−0.85, −0.07]	−0.44	[−0.82, −0.06]
σ_u^2		–	0.08	[0.0, 0.39]	0.08	[0.0, 0.22]	0.12	[0.01, 0.31]
Dispersion (Φ)[2]		1.45		1.29		1.35		–

[1] The reference category is Medium

[2] Obtained by means of equation (3)

Table 3: *Percentages of convergence in Simulation scenario 2 as a function of the marginal mean (μ), the average match number per season (Offset), overdispersion (OD), and the sample size.*

			glm		**lme4**		**hglm**		**INLA**	
	Offset	**OD**	$n = 30$	$n = 100$	$n = 30$	$n = 100$	$n = 30$	$n = 100$	$n = 30$	$n = 100$
$\mu = 1$	60	$\Phi = 1$	100	100	99.7	100	99.7	100	100	100
		$\Phi = 1.5$	100	100	99.4	100	99.7	100	100	100
		$\Phi = 3$	100	100	99.4	100	99.6	100	100	100
		$\Phi = 10$	100	100	98.1	99.9	98	99.8	100	100
	100	$\Phi = 1$	100	100	100	100	100	100	100	100
		$\Phi = 1.5$	100	100	99.8	99.9	99.9	100	100	100
		$\Phi = 3$	100	100	99.4	100	99.7	100	100	100
		$\Phi = 10$	100	100	97.2	100	97.9	99.9	100	100
$\mu = 10$	60	$\Phi = 1$	100	100	100	100	100	100	100	100
		$\Phi = 1.5$	100	100	100	100	100	100	100	100
		$\Phi = 3$	100	100	100	100	100	100	100	100
		$\Phi = 10$	100	100	100	100	100	100	100	100
	100	$\Phi = 1$	100	100	100	100	100	100	100	100
		$\Phi = 1.5$	100	100	100	100	100	100	100	100
		$\Phi = 3$	100	100	100	100	100	100	100	100
		$\Phi = 10$	100	100	100	100	100	100	100	100

Acknowledgements

We would like to thank Vicente Martin, Moudud Alam, Lesly Acosta, and members of the R sig-mixed-models mailing list and the INLA group discussion forum for useful comments. We are also grateful for the thorough revision of the manuscript which has helped to improve it. The work was partially supported by the grants MTM2012-38067-C02-01 of the Spanish Ministry of Economy and Competitiveness and 2014 SGR 464 from the *Departament d'Economia i Coneixement* of the *Generalitat de Catalunya* and by the *Diputación de León y la Federación Territorial de Castilla y Leon de Lucha*.

This article contains supplementary material that can be consulted at the web page: www.idescat.cat/sort/artpublished.html

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6, 251–262.

Alam, M., L. Rönnegård, and X. Shen (2014). Fitting conditional and simultaneous autoregressive spatial models in hglm. Submitted.

Aregay, M., Z. Shkedy, and G. Molenberghs (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics & Data Analysis*, 57, 233–245.

Austin, P. (2010). Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *The International Journal of Biostatistics*, 6, Article 16. doi: 10.2202/1557–4679.1195.

Avalos, M., P. Hellard, and J. C. Chatard (2003). Modeling the training-performance relationship using a mixed model in elite swimmers. *Medicine and Science in Sports and Exercise*, 35, 838–846.

Ayán, C., A. J. Molina, H. García, G. González, M. J. Álvarez, T. Fernández, and V. Martín (2010). Rules modification effect's in incidence of injuries in Lucha Leonesa (Leonesa wrestling). *Apunts. Medicina de l'Esport*, 45, 17–22.

Bates, D. (2010, February). lme4: Mixed-effects modeling with R.
URL: `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57, 269–306.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791–806.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. White (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24, 127–135.

Booth, J. G., G. Casella, H. Friedl, and J. P. Hobert (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3, 179–191.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, 33, 38–44.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.

Broström, G. and H. Holmberg (2013). *glmmML: Generalized linear models with clustering*. R package version 1.0.

Brown, P. E. and L. Zhou (2010). Mcmc for generalized linear mixed models with glmmbugs. *The R Journal*, 2, 13–17.

Bullock, N. and W. G. Hopkins (2009). Methods for tracking athletes' competitive performance in skeleton. *Journal of Sports Sciences*,27, 937–940.

Casals, M., M. Girabent-Farrés, and J. L. Carrasco (2014). Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PloS one*, 9, e112653.

Casals, M. and J. A. Martínez (2013). Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sports*, 13, 64–82.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2014). A multiresolution stochastic process model for predicting basketball possession outcomes. Submitted. arXiv:1408.0777v1.

Collins, D. (2008). *The performance of estimation methods for generalized linear mixed models*. Ph. D. thesis, School of Mathematics & Applied Statistics-Faculty of Informatics, University of Wollongong.

Cosandey-Godin, A., E. Teixeira Krainski, B. Worm, and J. Mills Flemming (2014). Applying Bayesian spatio-temporal models to fisheries bycatch in the canadian arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, 72, 1–12.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling*, 7, 125–153.

Dean, C. B. and J. D. Nielsen (2007). Generalized linear mixed models: a review and some extensions. *Lifetime Data Analysis*, 13, 497–512.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics & Data Analysis*, 51, 2871–2888.

Elston, D., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology*, 122, 563–569.

Fong, Y., H. Rue, and J. Wakefield (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11, 397–412.

Frühwirth-Schnatter, S. and H. Wagner (2010a). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics*, 9, 1–21. Oxford University Press.

Frühwirth-Schnatter, S. and H. Wagner (2010b). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154, 85–100.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1, 515–534.

Grilli, L., S. Metelli, and C. Rampichini (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 85, 2718–2726.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33, 1–22.

Hardin, J. W. and J. Hilbe (2007). *Generalized Linear Models and Extensions*. Stata Press.

Hewett, T., C. Pasque, R. Heyl, and R. Wroble (2005). Wrestling injuries. *Medicine and Sport Science*, 48, 152–178.

Hägglund, M., M. Waldén, R. Bahr, and J. Ekstrand (2005). Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *British Journal of Sports Medicine*, 39, 340–346.

Hägglund, M., M. Waldén, L. Til, and R. Pruna (2010). The importance of epidemiological research in sports medicine. *Apunts. Medicina de l'Esport*, 45, 57–59.

Hirsch, K. and A. Wienke (2012). Software for semiparametric shared gamma and log-normal frailty models: an overview. *Computer methods and programs in biomedicine*, 107, 582–597.

Junge, A., L. Engebretsen, M. L. Mountjoy, J. M. Alonso, P. A. F. H. Renström, M. J. Aubry, and J. Dvorak (2009). Sports injuries during the Summer Olympic Games 2008. *The American Journal of Sports Medicine*, 37, 2165–2172.

Kim, Y., Y.-K. Choi, and S. Emery (2013). Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *The American Statistician*, 67, 171–182.

Klügl, M., I. Shrier, K. McBain, R. Shultz, W. H. Meeuwisse, D. Garza, and G. O. Matheson (2010). The prevention of sport injury: an analysis of 12 000 published manuscripts. *Clinical Journal of Sport Medicine*, 20, 407.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis*, 52, 3441–3458.

Lee, W. and Y. Lee (2012). Modifications of REML algorithm for HGLMs. *Statistics and Computing*, 22, 959–966.

Lee, Y., J. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman & Hall/CRC.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B (Methodological)*, 58, 619–678.

Lee, Y. and J. A. Nelder (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88, 987–1006.

Lesaffre, E. and B. Spiessens (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50, 325–335.

Li, B., H. Lingsma, E. Steyerberg, and E. Lesaffre (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology*, 11, 77–87.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63, 1–25.

Martins, T. G., D. P. Simpson, A. Riebler, H. Rue, and S. H. Sørbye (2014). Penalising model component complexity: a principled, practical approach to constructing priors. Submitted. arXiv:1403.4630v3.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons.

Milanzi, E., A. Alonso, and G. Molenberghs (2012). Ignoring overdispersion in hierarchical loglinear models: possible problems and solutions. *Statistics in Medicine*, 31, 1475–1482.

Molas, M. and E. Lesaffre (2011). Hierarchical generalized linear models: the R package HGLMMM. *Journal of Statistical Software*, 39, 1–20.

Muenchen, R. A. (2015). The popularity of data analysis software.
URL: `http://r4stats.com/articles/popularity/`.

Noh, M. and Y. Lee (2007). Reml estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98, 896–915.

Ormerod, J. and M. Wand (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21, 2–17.

Pan, J. and R. Thompson (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*,51, 5765–5775.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford.

Quiroz, Z. C., M. O. Prates, and H. Rue (2014). A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics*, 71, 208–217.

Rönnegård, L., X. Shen, and M. Alam (2010). hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2, 20–28.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 71, 319–392.

Sampaio, J., E. J. Drinkwater, and N. M. Leite (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science*, 10, 141–149.

Sturtz, S., U. Ligges, and A. Gelman (2005). R2winbugs: A package for running WinBUGS from R. *Journal of Statistical Software*, 12, 1–16.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

West, B. T., K. B. Welch, and A. T. Galecki (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software* (2nd ed.). CRC Press.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, 144–148.

Zhang, H., N. Lu, C. Feng, S. W. Thurston, Y. Xia, L. Zhu, and X. M. Tu (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, 30, 2562–2572.

Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

SORT 39 (2) July-December 2015, 309-336

Multinomial logistic estimation in dual frame surveys

David Molina[1], Maria del Mar Rueda[1], Antonio Arcos[1]
and Maria Giovanna Ranalli[2]

Abstract

We consider estimation techniques from dual frame surveys in the case of estimation of proportions when the variable of interest has multinomial outcomes. We propose to describe the joint distribution of the class indicators by a multinomial logistic model. Logistic generalized regression estimators and model calibration estimators are introduced for class frequencies in a population. Theoretical asymptotic properties of the proposed estimators are shown and discussed. Monte Carlo experiments are also carried out to compare the efficiency of the proposed procedures for finite size samples and in the presence of different sets of auxiliary variables. The simulation studies indicate that the multinomial logistic formulation yields better results than the classical estimators that implicitly assume individual linear models for the variables. The proposed methods are also applied in an attitude survey.

MSC: 62D05

Keywords: Finite population, survey sampling, auxiliary information, model assisted inference, calibration.

1. Introduction

Sampling theory for finite populations usually assumes the existence of one sampling frame containing all population units. Then, a probability sample is drawn according to a sampling design and information collected is used for estimation and inference purposes. To ensure quality of the results obtained, the sampling frame must contain every single unit of population of interest (that is, it must be complete) and it must be updated as well. Otherwise, estimates could be affected by a serious bias due to the non-representativeness of the frame and, therefore, of the selected sample. Unfortunately,

[1] Department of Statistics and Operational Research, University of Granada, Spain. dmolinam@ugr.es, mrueda@ugr.es, arcos@ugr.es

[2] Department of Political Sciences, University of Perugia, Italy. giovanna@stat.unipg.it

Received: September 2014
Accepted: October 2015

this is not an easy task: populations are constantly changing, with new units entering and exiting the population frequently, so getting a good sampling frame can be difficult.

The dual frame approach tries to solve the aforementioned problems. This approach assumes that two frames are available for sampling and that, overall, they cover the entire target population. A sample is selected from each frame using a, possibly different, sampling design. Much attention has been devoted to the introduction of different ways of combining estimates coming from the different frames – see the seminal papers by Hartley (1962), Fuller and Burmeister (1972), Bankier (1986) and Kalton and Anderson (1986). However, these techniques were originally proposed to estimate means and totals of quantitative variables, and although their extension to the estimation of proportions in multinomial response variables is possible, it requires further investigation. Questionnaire items with multinomial outcomes are quite common in public opinion research, marketing research, and official surveys: estimating the proportion of voters in favour of each political party, based on a political opinion survey, is just one practical example of this procedure. Items where respondents must select one in a series of options can be modeled by a multinomial distribution. Lehtonen and Veijanen (1998) present estimators for a proportion which use logistic regression.

This paper focuses on the estimation of proportions for multinomial response variables when data come from two sampling frames. The proposed approach is motivated by a study on immigration. After describing the survey of opinions and attitudes of the Andalusian population regarding immigration, in Section 2, alternative estimators for the proportions are proposed following different approaches and their main theoretical properties are studied. A simulation study is also carried out to study their finite size sample properties. The results from the application to this dual frame attitude survey are then presented in Section 9.

2. Study background: the 2013 survey on opinions and attitudes of the Andalusian population regarding immigration

The 2013 survey on opinions and attitudes of the Andalusian population regarding immigration (OPIA) is a population-based survey conducted by the *Instituto de Estudios Sociales Avanzados* (IESA), a public scientific research institute for social sciences. The aim of the survey is to reflect the opinion of the Andalusian population with regard to various aspects of immigration and refugee policies in Spain and towards immigrants as a group. This survey is based on telephone interviews on a sample of adults drawn from both landline and mobile phone frames. Taking into account the time and budget available, 2402 interviews were performed by professional interviewers. The number of interviews to be conducted via landline and via mobile phone was determined by calculating the optimal proportion (in the sense of minimum variance) for each frame, taking into account costs and the percentage of possession of each type of device (following Hartley (1962)). As a result, final sample sizes were 1919 for landline and 483

Table 1: *Sample sizes for the OPIA survey. Landline and Mobile in the columns refer to the frame the interview comes from, while in the rows, they refer to the domain in which the units actually reside (type of user).*

Domain	Landline Sample	Mobile Sample	Total
Both	1 727	237	1 964
Mobile		246	246
Landline	192		192
Total	1 919	483	2 402

for mobile. Interviews were carried out by the Statistics and Surveys sections of IESA from April, 22 to May, 13, 2013, using Computer Assisted Telephone Interviewing (CATI) data input techniques. Sample sizes are reported in Table 1. The landline sample was also stratified by provinces in the region of Andalusia, as shown in Table 2. Cell-phone interviews were carried out with no control over the distribution by provinces owing to the difficulty of determining the location of this type of telephone. Hence, more interviews were performed in the most populated provinces than in the less populated ones.

Table 2: *Stratification in land-phone sample.*

Province	Almería	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
Population(*)	353 787	767 370	508 258	558 087	308 941	423 548	872 011	1 190 918
Sample	262	210	252	256	275	263	207	194

(*) Those estimates can be found on the INE website: http://www.ine.es/

At the time of data collection, frame sizes of landline and mobile were 4 982 920 and 5 707 655, respectively, and the total population size was 6 350 916 (source ICT-H 2012, Survey on the Equipment and Use of Information and Communication Technologies in Households, INE, National Statistical Institute, Spain). Auxiliary information about the user's sex and age is also available from the ICT-H 2012 survey. The total number of individuals in each domain (landline, mobile and both users) for every possible combination of values of the auxiliary variables is therefore known. The information about these auxiliary variables is displayed in Table 3.

One of the most important response variables in this study is related to the "attitude towards immigration". The variable is the answer to the following question: *And in relation to the number of immigrants currently living in Andalusia, do you think there are ...?: Too many, A reasonable number, Too few, No reply*. In the following sections we review approaches available in the literature to address the issue of estimating the distribution of a multiple choice type of variable in the population using a dual frame survey. We then illustrate our proposal to fully account for the nature of the response variable and the auxiliary information available.

Table 3: *Population data for variables* sex *and* age.

	Both	Landline	Mobile	Total
		Males		
18 - 29	428 750	0	188 172	616 922
30 - 44	724 435	4 259	298 416	1027 110
45 - 59	603 338	59 385	135 981	798 704
≥ 60	396 626	206 410	94 729	697 765
		Females		
18 - 29	480 151	0	115 472	595 623
30 - 44	658 984	17 673	289 106	965 763
45 - 59	601 478	39 362	141 553	782 393
≥ 60	445 897	316 172	104 567	866 636

(*) Source: Survey of Information Technologies in Households (INE)

3. Existing approaches to estimation of class frequencies in dual frame surveys

We employ the notation considered in Rao and Wu (2010). Let U denote a finite population with N units, $U = \{1,\ldots,k,\ldots,N\}$ and let A and B be two sampling-frames. Let $\mathscr{A}$ be the set of population units in frame A and $\mathscr{B}$ the set of population units in frame B. The population of interest, U, may be divided into three mutually exclusive domains, $a = \mathscr{A} \cap \mathscr{B}^c, b = \mathscr{A}^c \cap \mathscr{B}$ and $ab = \mathscr{A} \cap \mathscr{B}$. Because the population units in the overlap domain ab can be sampled in either survey or both surveys, it is convenient to create a duplicate domain $ba = \mathscr{B} \cap \mathscr{A}$, which is identical to $ab = \mathscr{A} \cap \mathscr{B}$, to denote the domain in the overlapping area coming from frame B. Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in U, A, B, a, b, ab, ba, respectively. We assume that N_A, N_B and N_{ab} are known, so the population size $N = N_A + N_B - N_{ab}$ is also known. This is also the situation in our motivating dataset.

We consider the estimation of class frequencies of a discrete response variable. Assume that we collect data from respondents who provide a single choice from a list of alternatives. We code these alternatives $1,2,\ldots,m$. Therefore, consider a discrete m-valued survey variable y. The objective is to estimate the frequency distribution of y in the population U. To estimate this frequency distribution, we define a class of indicators z_i $(i = 1,\ldots,m)$ such that, for each unit $k \in U$, $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. Our problem thus, is to estimate the proportions $P_i = N^{-1}\sum_{k\in U} z_{ki}$, for $i = 1,2,\ldots,m$. These proportions are such that

$$P_i = N^{-1}(Z_{ai} + \eta Z_{abi} + (1-\eta)Z_{bai} + Z_{bi}), \tag{1}$$

where $0 \leq \eta \leq 1$ and $Z_{ai} = \sum_{k\in a} z_{ki}$, $Z_{abi} = \sum_{k\in ab} z_{ki}$, $Z_{bai} = \sum_{k\in ba} z_{ki}$ and $Z_{bi} = \sum_{k\in b} z_{ki}$.

Two probability samples s_A and s_B are drawn independently from frame A and frame B of sizes n_A and n_B, respectively. Each design induces first-order inclusion probabilities

π_{Ak} and π_{Bk}, respectively, and sampling weights $d_{Ak} = 1/\pi_{Ak}$ and $d_{Bk} = 1/\pi_{Bk}$. The sample s_A can be post-stratified as $s_A = s_a \cup s_{ab}$, where $s_a = s_A \cap a$ and $s_{ab} = s_A \cap (ab)$. Similarly, $s_B = s_b \cup s_{ba}$, where $s_b = s_B \cap b$ and $s_{ba} = s_B \cap (ba)$. Note that s_{ab} and s_{ba} are both from the same domain ab, but s_{ab} is part of the frame A sample and s_{ba} is part of the frame B sample. Then, assuming that duplicated units (i.e. $s_A \cap s_B$) cannot be identified and that this event has a negligible chance to happen, we let $s = s_A \cup s_B$. Note that this is a reasonable assumption in the OPIA survey at hand.

The Hartley (1962) estimator of P_i, for $i = 1, 2, \ldots, m$, is given by

$$\hat{P}_{Hi}(\eta) = N^{-1}(\hat{Z}_{ai} + \eta\hat{Z}_{abi} + (1-\eta)\hat{Z}_{bai} + \hat{Z}_{bi}), \tag{2}$$

where $\hat{Z}_{ai} = \sum_{k \in s_a} d_{Ak} z_{ki}$ is the expansion estimator for the population count of category i in domain a and similarly for the other domains. If we let

$$d_k^\circ = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ \eta d_{Ak} & \text{if } k \in s_{ab} \\ (1-\eta) d_{Bk} & \text{if } k \in s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases}, \tag{3}$$

then $\hat{P}_{Hi}(\eta) = N^{-1}(\sum_{k \in s_A} d_k^\circ z_{ki} + \sum_{k \in s_B} d_k^\circ z_{ki}) = N^{-1}(\sum_{k \in s} d_k^\circ z_{ki})$. Since the population count in each domain is estimated by its expansion estimator, $\hat{P}_{Hi}(\eta)$ is an unbiased estimator of P_i for a given η.

Fuller and Burmeister (1972) proposed modifying Hartley's estimator by incorporating additional information regarding estimation of the overlap domain. The resulting estimator is:

$$\hat{P}_{FBi}(\beta_1, \beta_2) = N^{-1}(\hat{Z}_{ai} + \beta_1\hat{Z}_{abi} + (1-\beta_1)\hat{Z}_{bai} + \hat{Z}_{bi} + \beta_2(\hat{N}_{ab} - \hat{N}_{ba})) \tag{4}$$

where $\hat{N}_{ab} = \sum_{k \in s_{ab}} d_{Ak}$ and $\hat{N}_{ba} = \sum_{k \in s_{ba}} d_{Bk}$. Coefficients β_1 and β_2 are selected to minimize $V(\hat{P}_{FBi}(\beta_1, \beta_2))$. In this case, and as with Hartley's estimator, a new set of weights must be calculated for each response variable. This leads to possible inconsistencies among the estimated proportions, which is particularly relevant when dealing with multinomial outcomes. In addition, optimal values depend on covariances among Horvitz-Thompson estimators, which may be difficult to compute in practice and, finally, it is also possible to obtain values of β_1 outside the range $[0, 1]$.

Skinner and Rao (1996) propose a modification of the estimator proposed by Fuller and Burmeister (1972) for simple random sampling to handle complex designs. They introduce a pseudo maximum likelihood (PML) estimator that does not achieve optimality like the FB estimator, but it can be written as a linear combination of the observations and the same set of weights can be used for all variables of interest:

$$\hat{P}_{PMLi}(\theta) = N^{-1}\left(\frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a}\hat{Z}_{ai} + \frac{\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)}\hat{Z}_{abi}(\theta) + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b}\hat{Z}_{bi}\right) \quad (5)$$

where $\hat{Z}_{abi}(\theta) = \theta\hat{Z}_{abi} + (1-\theta)\hat{Z}_{bai}, \hat{N}_{ab}(\theta) = \theta\hat{N}_{ab} + (1-\theta)\hat{N}_{ba}$ and $\hat{N}_{ab}^{PML}(\theta)$ is the smallest root of the quadratic equation

$[\theta/N_B + (1-\theta)/N_A]x^2 - \left[1 + \theta\hat{N}_{ab}/N_B + (1-\theta)\hat{N}_{ba}/N_A\right]x + \hat{N}_{ab} = 0.$

Recently, Rao and Wu (2010) extended the Pseudo-Empirical-Likelihood approach (PEL) proposed by Wu and Rao (2006) from one-frame surveys to dual-frame surveys following a stratification approach. In particular,

$$\hat{P}_{PELi}(\theta) = (N_a/N)\hat{\bar{Z}}_{aip} + \theta(N_{ab}/N)\hat{\bar{Z}}_{abip} + (1-\theta)(N_{ba}/N)\hat{\bar{Z}}_{baip} + (N_b/N)\hat{\bar{Z}}_{bip}, \quad (6)$$

where $\theta \in (0,1)$ is a fixed constant to be specified and $\hat{\bar{Z}}_{aip} = \sum_{k\in s_a}\hat{p}_{ak}z_{ki}$, $\hat{\bar{Z}}_{bip} = \sum_{k\in s_b}\hat{p}_{bk}z_{ki}$ and $\hat{\bar{Z}}_{abip} = \sum_{k\in s_{ab}}\hat{p}_{abk}z_{ki} = \hat{\bar{Z}}_{baip}$. The p-weights maximize the pseudo empirical likelihood and verify $\sum_{k\in s_a}\hat{p}_{ak} = 1, \sum_{k\in s_{ab}}\hat{p}_{abk} = 1, \sum_{k\in s_{ba}}\hat{p}_{bak} = 1, \sum_{k\in s_b}\hat{p}_{bk} = 1$, and the additional constraint induced by the common domain mean $\hat{\bar{Z}}_{abip} = \hat{\bar{Z}}_{baip}$ (see Rao and Wu (2010) for more details). Note that (6) can be rewritten as:

$$\hat{P}_{PELi} = (N_a/N)\hat{\bar{Z}}_{aip} + (N_{ab}/N)\hat{\bar{Z}}_{abip} + (N_b/N)\hat{\bar{Z}}_{bip}, \quad (7)$$

so the estimator does not depend on explicitly on θ and its value only affects the estimator $\hat{\bar{Z}}_{abip}$ for the population mean of the overlapping domain.

Ranalli et al. (2015) used calibration procedures for estimation from dual frame sampling assuming that some kind of auxiliary information is available. For example, assuming that there are p auxiliary variables, $\boldsymbol{x}_k = (x_{1k}, ..., x_{pk})$ is the value taken by such auxiliary variables on unit k. It is assumed that the vector of population totals of the auxiliary variables, $\boldsymbol{t}_x = \sum_{k\in U}\boldsymbol{x}_k$ is also known. In this context, the dual frame calibration estimator can be defined as follows,

$$\hat{P}_{CalDFi} = N^{-1}\left(\sum_{k\in s} d_k^{\mathrm{DF}} z_{ki}\right) \quad (8)$$

where weights d_k^{DF} are chosen to be as close as possible to basic design weights and, at the same time, satisfy benchmark constraints on the auxiliary variables, i.e. they are such that

$$\min_{d_k^{\mathrm{DF}}} \sum_{k\in s} G(d_k^{\mathrm{DF}}, d_k^{\circ}), \quad \text{subject to} \quad \sum_{k\in s} d_k^{\mathrm{DF}}\boldsymbol{x}_k = \boldsymbol{t}_x,$$

with $G(\cdot,\cdot)$ a given distance measure.

When inclusion probabilities in domain *ab* are known for both frames, and not just for the frame from which the unit is selected, *single-frame* methods (Bankier (1986), Kalton and Anderson (1986)), which combine the observations into a single dataset and adjust the weights in the intersection domain for multiplicity, can also be used. To adjust for multiplicity, the weights are defined as follows for all units in frame A and in frame B,

$$\tilde{d}_k = \begin{cases} d_{Ak} & \text{if } k \in a \\ (1/d_{Ak} + 1/d_{Bk})^{-1} & \text{if } k \in ab \\ d_{Bk} & \text{if } k \in b \end{cases} .$$

In this context, BKA single frame estimator (Bankier (1986) and Kalton and Anderson (1986)) is given by

$$\hat{P}_{BKAi} = N^{-1} \left(\sum_{k \in s_A} \tilde{d}_k z_{ki} + \sum_{k \in s_B} \tilde{d}_k z_{ki} \right) = N^{-1} \left(\sum_{k \in s} \tilde{d}_k z_{ki} \right). \tag{9}$$

Single frame weights are the same for all response variables, and so estimators are internally consistent.

A calibration estimator under the *single-frame* approach can be defined as follows:

$$\hat{P}_{CalSFi} = N^{-1} \left(\sum_{k \in s} d_k^{SF} z_{ki} \right) \tag{10}$$

with weights d_k^{SF} verifying that $\min \sum_{k \in s} G(d_k^{SF}, \tilde{d}_k)$ subject to $\sum_{k \in s} d_k^{SF} \boldsymbol{x}_k = \boldsymbol{t}_x$.

The single-frame approach requires the knowledge of the design weight of a unit for both frames, not just for the one in which the unit was selected. Given this information, multiplicity can be adjusted for using sampling weights only. Therefore, unlike the dual frame methods, they do not require calculation of η. Single-frame estimators are usually more efficient than dual-frame estimators, and this can be explained by the extra-information they incorporate in the estimation process. The estimators presented in this Section can be computed using the R-package `Frames2` (Arcos et al., 2015).

4. Estimation of class frequencies using multinomial logistic regression

Auxiliary information is often available in survey sampling. This information, which may come from past censuses or from other administrative sources, can be used to obtain more accurate estimators. Then, other than the values of the variables of interest and of

the auxiliary variables for $k \in s$, assume we also know the distribution or at least some summary statistics of the auxiliary variables in the population. We consider that the population under study $\boldsymbol{y} = (y_1, ..., y_N)^\top$ is the determination of a set of super-population random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$ s.t.

$$\mu_{ki} = P(Y_k = i|\boldsymbol{x}_k) = E(Z_{ki}|\boldsymbol{x}_k) = \frac{\exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}_i)}{\sum_{r=1,...,m} \exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}_r)}, \qquad i = 1, \ldots, m,$$

that is, we use the multinomial logistic model to relate y and $\boldsymbol{x}$. Let $\boldsymbol{\beta}$ be the parameter vector $(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$. In the following sections we introduce new estimators for the population proportions P_i. To this end, as a first step, we need to consider estimation of the superpopulation parameter $\boldsymbol{\beta}$ using the sample s.

4.1. Case I: The same set of auxiliary variables is available for all population units

Suppose that for each unit in the population we have information about one vector of auxiliary variables $\boldsymbol{x}$. In this case, for each unit $k \in U$ we know the value of $\boldsymbol{x}_k$. In addition, for each unit $k \in s$, we observe the value of the main variable y_k and we denote by $(z_{k1}, z_{k2}, ..., z_{km})$ the multinomial trial observed for this unit k.

We can estimate $\boldsymbol{\beta}$ by maximizing the π-weighted log-likelihood (Godambe and Thompson (1986), Särndal et al. (1992)) given by

$$\ell_{d^\circ}(\boldsymbol{\beta}) = \sum_{i=1,...,m} \left(\sum_{k \in s_A} d_k^\circ z_{ki} \ln \mu_{ki} + \sum_{k \in s_B} d_k^\circ z_{ki} \ln \mu_{ki} \right). \tag{11}$$

This approach is usually motivated by first defining a census-level parameter $\boldsymbol{\beta}_U$, obtained by maximizing the likelihood over all units in the population, i.e. $\ell_U(\boldsymbol{\beta}) = \sum_{i=1,...,m} \sum_{k \in U} z_{ki} \ln \mu_{ki}$. Then, $\widehat{\boldsymbol{\beta}}^\circ$ obtained using the the π-weighted likelihood (11) is its design based estimate. Computing $\widehat{\boldsymbol{\beta}}^\circ$ usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work rather well. Most statistical packages include a multinomial logit procedure that can handle weights.

Given the estimate $\widehat{\boldsymbol{\beta}}^\circ$ of $\boldsymbol{\beta}$, we consider the following auxiliary variable

$$p_{ki}^\circ = \widehat{\mu}_{ki}^\circ = \frac{\exp(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}_i^\circ)}{\sum_{r=1,...,m} \exp(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}_r^\circ)}. \tag{12}$$

Please note that these p values are different from those involved in the definition of estimator (6). Since the vector $\boldsymbol{x}_k$ is known for all units of the population U, the values

p°_{ki} are available for all $k \in U$ and we propose to use such values to define a new estimator for P_i,

$$\begin{aligned}\widehat{P}^{\mathrm{DW}}_{\mathrm{ML}i} &= N^{-1}\left(\sum_{k\in U} p^{\circ}_{ki} + \sum_{k\in s_A} d^{\circ}_k (z_{ki} - p^{\circ}_{ki}) + \sum_{k\in s_B} d^{\circ}_k (z_{ki} - p^{\circ}_{ki})\right) \\ &= N^{-1}\left(\sum_{k\in U} p^{\circ}_{ki} + \sum_{k\in s} d^{\circ}_k (z_{ki} - p^{\circ}_{ki})\right).\end{aligned} \tag{13}$$

We observe that this estimator takes the same model-assisted form as the MLGREG estimator proposed in Lehtonen and Veijanen (1998), but here it is adjusted to account for the dual frame sampling setting. The subscript *ML* stands for Multinomial-Logistic and the superscript DW stands Dual frame setting and auxiliary information available from the Whole population.

Note that we cannot compute $\sum_{k\in U} p^{\circ}_{ki}$ in (13) without knowing $\boldsymbol{x}_k$ for each $k \in U$, i.e. we need the value of the auxiliary variables for each individual in the population. This assumption can be quite restrictive; nonetheless, it can be relaxed. For example, if we have two discrete or categorical variables, we only need the population counts in the two-way contingency table. In human populations, sizes of certain demographic groups are known and are used often as auxiliary information. This is also the case in the OPIA survey and this information can be retrieved from the last column in Table 3.

An important way to incorporate available auxiliary information is given by calibration estimation (Deville and Särndal (1992)), that seeks for new weights that are close (in some sense) to the basic design weights and that, at the same time, match benchmark constraints on auxiliary information. We have reviewed in the previous section extension of linear calibration to the dual frame setting. Here, using the idea of model calibration introduced by Wu and Sitter (2001a), we propose the following model calibration estimator (the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript DW stands Dual frame setting and auxiliary information available from the Whole population), given by

$$\widehat{P}^{\mathrm{DW}}_{\mathrm{MLC}i} = N^{-1}\Big(\sum_{k\in s_A} w^{\circ}_k z_{ki} + \sum_{k\in s_B} w^{\circ}_k z_{ki}\Big) = N^{-1}\Big(\sum_{k\in s} w^{\circ}_k z_{ki}\Big),$$

where w°_k minimizes $\sum_{k\in s_A} G(w^{\circ}_k, d^{\circ}_k) + \sum_{k\in s_B} G(w^{\circ}_k, d^{\circ}_k) = \sum_{k\in s} G(w^{\circ}_k, d^{\circ}_k)$ for a distance measure $G(\cdot,\cdot)$ as those considered in Deville and Särndal (1992), subject to:

$$\sum_{k\in s} w^{\circ}_k p^{\circ}_{ki} = \sum_{k\in U} p^{\circ}_{ki}, \quad \sum_{k\in s_a} w^{\circ}_k = N_a, \quad \sum_{k\in s_b} w^{\circ}_k = N_b,$$

$$\sum_{k\in s_{ab}} w_k^{\circ} = \eta N_{ab} \quad \text{and} \quad \sum_{k\in s_{ba}} w_k^{\circ} = (1-\eta) N_{ab}.$$

Suppose, now, that for each unit in the population inclusion probabilities in domain *ab* are known for both frames, and not just for the frame from which the unit is selected. In this situation, the single-frame approach can also be used to propose new multinomial logistic estimators. First, we calculate $\tilde{\boldsymbol{\beta}}$ by maximizing the π-weighted log-likelihood given by

$$\ell_{\tilde{d}}(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k\in s} \tilde{d}_k z_{ki} \ln \mu_{ki}. \tag{14}$$

We use the new auxiliary variable $\tilde{p}_{ki} = \tilde{\mu}_{ki} = \dfrac{\exp(\boldsymbol{x}_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_r)}$ to define a new estimator (the subscript ML stands for Multinomial-Logistic and the superscript SW stands Single frame setting and auxiliary information available from the Whole population):

$$\begin{aligned} \widehat{P}_{\mathrm{ML}i}^{\mathrm{SW}} &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s_A} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) + \sum_{k\in s_B} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) \right) \\ &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) \right). \end{aligned} \tag{15}$$

Note that $\tilde{d}_k$ weights are used in the formulation of the estimator (15) and also in the likelihood function (14).

Model calibration can be also used to define a single-frame estimator (the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript SW stands Single frame setting and auxiliary information available from the Whole population):

$$\widehat{P}_{\mathrm{MLC}i}^{\mathrm{SW}} = N^{-1}\Big(\sum_{k\in s_A} \tilde{w}_k z_{ki} + \sum_{k\in s_B} \tilde{w}_k z_{ki}\Big) = N^{-1}\Big(\sum_{k\in s} \tilde{w}_k z_{ki}\Big),$$

where $\tilde{w}_k$ minimizes $\sum_{k\in s_A} G(\tilde{w}_k, \tilde{d}_k) + \sum_{k\in s_B} G(\tilde{w}_k, \tilde{d}_k) = \sum_{k\in s} G(\tilde{w}_k, \tilde{d}_k)$ for a distance measure $G(\cdot,\cdot)$ satisfying the usual conditions specified in the calibration paradigm subject to:

$$\sum_{k\in s} \tilde{w}_k \tilde{p}_{ki} = \sum_{k\in U} \tilde{p}_{ki}, \quad \sum_{k\in s_a} \tilde{w}_k = N_a, \quad \sum_{k\in s_b} \tilde{w}_k = N_b \quad \text{and} \quad \sum_{k\in s_{ab}\bigcup s_{ba}} \tilde{w}_k = N_{ab}.$$

Note that when inclusion probabilities are known for both frames, it is possible to calculate single and dual frame type estimators.

4.2. Case II: Two different sets of auxiliary variables are available according the frame considered

Now we consider a different situation: the auxiliary information is available separately in each frame. In this case, for each unit $k \in \mathscr{A}$ we have an auxiliary vector $\boldsymbol{x}_{Ak}$ and for each unit $k \in \mathscr{B}$ we have another auxiliary vector $\boldsymbol{x}_{Bk}$ where the components of $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ can be different. Indeed in the OPIA survey the two sets of auxiliary variables coincide. Nonetheless, we will leave the treatment general and provide two proposals based on the dual frame approach to handle this situation as well.

In this case, we can use the available auxiliary information to fit a multinomial logistic model separately in each frame. For each $k \in \mathscr{A}$, using data from s_A we can compute

$$p_{ki}^A = \frac{\exp(\boldsymbol{x}_{Ak}^\intercal \widehat{\boldsymbol{\beta}}_i^A)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_{Ak}^\intercal \widehat{\boldsymbol{\beta}}_r^A)} \tag{16}$$

where we estimate $\boldsymbol{\beta}^A$ by maximizing $\ell_{d_A}(\boldsymbol{\beta}^A) = \sum_{i=1,\ldots,m} \sum_{k \in s_A} d_{Ak} z_{ki} \ln \mu_{ki}$. Similarly we obtain p_{ki}^B for $k \in \mathscr{B}$, and define for each $i = 1, \ldots, m$ the following regression estimator:

$$\begin{aligned}\widehat{P}_{\mathrm{ML}i}^{\mathrm{DF}} = N^{-1} \Bigg(& \sum_a p_{ki}^A + \eta \sum_{ab} p_{ki}^A + (1-\eta) \sum_{ba} p_{ki}^B + \sum_b p_{ki}^B + \\ & + \sum_{s_a} (z_{ki} - p_{ki}^A) d_{Ak} + \eta \sum_{s_{ab}} (z_{ki} - p_{ki}^A) d_{Ak} + \\ & + (1-\eta) \sum_{s_{ba}} (z_{ki} - p_{ki}^B) d_{Bk} + \sum_{s_b} (z_{ki} - p_{ki}^B) d_{Bk} \Bigg) .\end{aligned}$$

As in the previous section, the subscript ML stands for Multinomial-Logistic, while the superscript DF stands now for Dual frame setting and auxiliary information available from the Frames. To compute $\widehat{P}_{\mathrm{ML}i}^{\mathrm{DF}}$ we only need to know the total number of individuals in each domain (a, b and ab) for every possible combination of values of the auxiliary variables in the cases where discrete variables have been used as auxiliary information. In the OPIA survey this information is obtained from Table 3.

A calibration estimator in this setting can be defined under the dual frame approach as follows,

$$\widehat{P}_{\mathrm{MLC}i}^{\mathrm{DF}} = N^{-1} \Big(\sum_{k \in s_A} w_k^\star z_{ki} + \sum_{k \in s_B} w_k^\star z_{ki} \Big) = N^{-1} \Big(\sum_{k \in s} w_k^\star z_{ki} \Big), \tag{17}$$

where the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript DF stands Dual frame setting and auxiliary information available from the Frames. Weights $w_k^\star$ are such that

$$\begin{aligned}
&\min \sum_{k\in s_A} G(w_k^\star, d_{Ak}) + \sum_{k\in s_B} G(w_k^\star, d_{Bk}) \qquad \text{s.t.}\\
&\sum_{k\in s_A} w_k^\star p_{ki}^A = \sum_{k\in a} p_{ki}^A + \eta \sum_{k\in ab} p_{ki}^A,\\
&\sum_{k\in s_B} w_k^\star p_{ki}^B = (1-\eta)\sum_{k\in ba} p_{ki}^B + \sum_{k\in b} p_{ki}^B,\\
&\sum_{k\in s_a} w_k^\star = N_a, \quad \sum_{k\in s_b} w_k^\star = N_b,\\
&\sum_{k\in s_{ab}} w_k^\star = \eta N_{ab} \quad \text{and} \quad \sum_{k\in s_{ba}} w_k^\star = (1-\eta)N_{ab},
\end{aligned}$$

where p_{ki}^A are the estimated probabilities defined in (16) and p_{ki}^B are their analogous in frame B.

5. Properties of proposed estimators

To show the asymptotic properties of the proposed estimators $\hat{P}_{\text{ML}}^{\text{DW}}$, $\hat{P}_{\text{MLC}}^{\text{DW}}$, $\hat{P}_{\text{ML}}^{\text{SW}}$, $\hat{P}_{\text{MLC}}^{\text{SW}}$, $\hat{P}_{\text{ML}}^{\text{DF}}$, $\hat{P}_{\text{MLC}}^{\text{DF}}$, we adapt and place ourselves in the asymptotic framework of Isaki and Fuller (1982), in which the dual-frame finite population U and the sampling designs $p_A(\cdot)$ and $p_B(\cdot)$ are embedded into a sequence of such populations and designs indexed by N, $\{U_N, p_{A_N}(\cdot), p_{B_N}(\cdot)\}$, with $N \to \infty$. We will assume therefore, that N_{A_N} and N_{B_N} tend to infinity and that also n_{A_N} and n_{B_N} tend to infinity as $N \to \infty$. We will further assume that $N_a > 0$ and $N_b > 0$. In addition $n_{A_N}/n_N \to c_1 \in (0,1)$, where $n_N = n_{A_N} + n_{B_N}$, $N_a/N_A \to c_2 \in (0,1)$, $N_b/N_B \to c_3 \in (0,1)$ as $N \to \infty$. Subscript N may be dropped for ease of notation, although all limiting processes are understood as $N \to \infty$. Stochastic orders $O_p(\cdot)$ and $o_p(\cdot)$ are with respect to the aforementioned sequences of designs. The constant $\eta \in (0,1)$ is kept fixed over repeated sampling.

We first discuss the theoretical properties of $\hat{P}_{\text{MLC}}^{\text{DW}}$ and then move to the other estimators, because these can be dealt with using slight modifications of this more general setting. Let $\mu(\boldsymbol{x}_k, \boldsymbol{\theta}_i) = \exp(\boldsymbol{x}_k^\intercal \boldsymbol{\theta}_i)/\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^\intercal \boldsymbol{\theta}_r)$, for $i = 1, \ldots, m$. In order to prove our results, we make the following technical assumptions.

A1 *Let $\boldsymbol{\beta}_U$ be census level parameter estimate obtained by maximizing the likelihood $\ell_U(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k\in U} z_{ki} \ln \mu_{ki}$. Assume that $\boldsymbol{\beta} = \lim_{N\to\infty} \boldsymbol{\beta}_U$ exists and that $\hat{\boldsymbol{\beta}}^\circ = \boldsymbol{\beta}_U + O_p(n_N^{-1/2})$.*

A2 *For each* $\boldsymbol{x}_k$, $|\partial\mu(\boldsymbol{x}_k,\boldsymbol{\theta}_i)/\partial\boldsymbol{\theta}_i| \leq f_1(\boldsymbol{x}_k,\boldsymbol{\beta}_i)$ *for* $\boldsymbol{\theta}_i$ *in a neighborhood of* $\boldsymbol{\beta}_i$ *and* $f_1(\boldsymbol{x}_k,\boldsymbol{\beta}_i) = O(1)$, *for* $i = 1,\ldots,m$.

A3 *For each* $\boldsymbol{x}_k$, $\max_{j,j'}|\partial^2\mu(\boldsymbol{x}_k,\boldsymbol{\theta}_i)/\partial\theta_j\theta_{j'}| \leq f_2(\boldsymbol{x}_k,\boldsymbol{\beta}_i)$ *for* $\boldsymbol{\theta}_i$ *in a neighborhood of* $\boldsymbol{\beta}_i$ *and* $f_2(\boldsymbol{x}_k,\boldsymbol{\beta}_i) = O(1)$, *for* $i = 1,\ldots,m$.

A4 *The auxiliary variables* $\boldsymbol{x}$ *have bounded fourth moments.*

A5 *For any study variable* ξ *with bounded fourth moment, the sampling designs are such that for the normalized Hartley estimators of* $\bar{\xi} = N^{-1}\sum_{k\in U}\xi_k$ *a central limit theorem holds, i.e.*

$$\sqrt{n_N}(\hat{\bar{\xi}}_H - \bar{\xi}) \xrightarrow{\mathscr{L}} N(0, V(\hat{\bar{\xi}}_H)),$$

where $\hat{\bar{\xi}}_H = N^{-1}\sum_{k\in s} d_k^\circ \xi_k$ *and* $V(\hat{\bar{\xi}}_H) = V(\hat{\bar{\xi}}_a + \eta\hat{\bar{\xi}}_{ab}) + V((1-\eta)\hat{\bar{\xi}}_{ba} + \hat{\bar{\xi}}_b)$. *The latter can be consistently estimated by* $v(\hat{\bar{\xi}}_H) = v(\hat{\bar{\xi}}_a + \eta\hat{\bar{\xi}}_{ab}) + v((1-\eta)\hat{\bar{\xi}}_{ba} + \hat{\bar{\xi}}_b)$.

Assumption A1 requires consistency of parameter estimates defined by weighted estimating equations to their census level counterpart. See e.g. Binder (1983). We will first state the properties of $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$ for the Euclidean distance. In fact, in this case an analytic solution to the constrained distance minimization problem exists and is given by

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} = N^{-1}\left\{\sum_{k\in s} d_k^\circ z_{ki} + \Big(\sum_{k\in U}\tilde{p}_{ki}^\circ - \sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ\Big)^{\mathsf{T}}\hat{\boldsymbol{\alpha}}_i^\circ\right\},$$

where $\tilde{p}_{ki}^\circ = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), p_{ki}^\circ)^{\mathsf{T}}$ is a vector that contains p_{ki}° defined in (12) and a set of indicator variables – $\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b)$ – implicitly used in the benchmark constraints. In particular, $\delta_k(a)$ takes value 1 if unit $k \in U$ belongs to domain a and 0 otherwise. Then $\sum_{k\in U}\delta_k(a) = N_a$. The other indicator variables are defined similarly. In addition, $\hat{\boldsymbol{\alpha}}_i^\circ = (\sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ \tilde{p}_{ki}^{\circ T})^{-1}(\sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ z_{ki})$, i.e. it is the vector of coefficients of the generalized regression of z_{ki} on $\tilde{p}_{ki}^\circ$ similar to the case of classical model calibration for one frame only (see Wu and Sitter (2001a)). Then from calibration theory (see Deville and Särndal (1992)), it is well known that all other calibration estimators that use different distance functions are equivalent to $\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}$, under additional regularity conditions on the shape of the distance function itself.

Theorem 1 *Under assumptions A1–A5,* $\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}$ *is design* $\sqrt{n_N}$*-consistent for* P_i *in the sense that*

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} - P_i = O_p(n_N^{-1/2}),$$

and has the following asymptotic distribution

$$\frac{\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} - P_i}{\sqrt{V_\infty(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i})}} \xrightarrow{\mathscr{L}} N(0,1)$$

where $V_\infty(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}) = N^{-2}V(\hat{t}_{eiH})$ *and* $\hat{t}_{eiH} = \sum_{k\in s} d_k^\circ e_{ki}$ *is the Hartley estimator of the population total of the census-level residuals* $e_{ki} = z_{ki} - \tilde{\boldsymbol{\mu}}_{ki}^{\circ T}\boldsymbol{\alpha}_i^\circ$, *and* $\boldsymbol{\alpha}_i^\circ = (\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ\tilde{\boldsymbol{\mu}}_{ki}^{\circ T})^{-1}(\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ z_{ki})$, *where* $\tilde{\boldsymbol{\mu}}_{ki}^\circ$ *is like* $\tilde{p}_{ki}^\circ$ *but with* p_{ki}° *replaced by its population counterpart*

$$\mu_{ki}^\circ = \frac{\exp(\boldsymbol{x}_k^\intercal\boldsymbol{\beta}_{Ui})}{\sum_{r=1,\ldots,m}\exp(\boldsymbol{x}_k^\intercal\boldsymbol{\beta}_{Ur})}. \tag{18}$$

In addition, let $\hat{e}_{ki} = z_{ki} - \tilde{p}_{ki}^{\circ T}\hat{\boldsymbol{\alpha}}_i^\circ$. *Then,* $V(\hat{t}_{eiH})$ *can be consistently estimated by*

$$\begin{aligned} v(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}) &= N^{-2}v(\hat{t}_{\hat{e}iH}) \\ &= N^{-2}\Big\{v\Big(\textstyle\sum_{k\in s_a} d_{Ak}\hat{e}_{ki} + \eta\sum_{k\in s_{ab}} d_{Ak}\hat{e}_{ki}\Big) + \\ &\quad + v\Big((1-\eta)\textstyle\sum_{k\in s_{ba}} d_{Bk}\hat{e}_{ki} + \sum_{k\in s_b} d_{Bk}\hat{e}_{ki}\Big)\Big\}. \end{aligned} \tag{19}$$

Proof. Using the same approach developed in Montanari and Ranalli (2005) and similarly to Wu and Sitter (2001b), it is easy to show that by assumptions A1–A2 and A4–A5,

$$N^{-1}(\sum_{k\in s} d_k^\circ p_{ki}^\circ - \sum_{k\in U} p_{ki}^\circ) = O_p(n_N^{-1/2}),$$

using a first order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^\circ)$ at $\hat{\boldsymbol{\beta}}_i^\circ = \boldsymbol{\beta}_{Ui}$, and that $\hat{\boldsymbol{\alpha}}_i^\circ - \boldsymbol{\alpha}_i^\circ = O_p(n_N^{-1/2})$ because $\hat{\boldsymbol{\alpha}}_i^\circ$ is just a function of population means of variables with finite fourth moments, that can be consistently estimated by their Hartley counterparts. Using A1–A5 and a second order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^\circ)$ at $\hat{\boldsymbol{\beta}}_i^\circ = \boldsymbol{\beta}_{Ui}$,

$$N^{-1}(\sum_{k\in s} d_k^\circ p_{ki}^\circ - \sum_{k\in U} p_{ki}^\circ) = N^{-1}(\sum_{k\in s} d_k^\circ \mu_{ki}^\circ - \sum_{k\in U} \mu_{ki}^\circ) + O_p(n_N^{-1}).$$

Then,

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} = N^{-1}\sum_{k\in s} d_k^\circ z_{ki} + N^{-1}\Big(\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ - \sum_{k\in s} d_k^\circ\tilde{\boldsymbol{\mu}}_{ki}^\circ\Big)^\intercal\boldsymbol{\alpha}_i^\circ + O_p(n_N^{-1})$$

and the first part of the result is proven.

Now, from assumption A5, $v(\hat{t}_{eiH}) = V(\hat{t}_{eiH}) + o_p(n_N^{-1})$. Since $p_{ki}^{\circ} = \mu_{ki}^{\circ} + O_p(n_N^{-1/2})$, $\hat{e}_{ki} = e_{ki} + O_p(n_N^{-1/2})$ and $v(\hat{t}_{\hat{e}iH}) = v(\hat{t}_{eiH}) + o_p(n_N^{-3/2})$, then the argument follows. ■

Note that, given the asymptotic equivalence of calibration and generalized regression estimation, analytic variance estimator in (19) can be used to estimate the variance of $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$ also when using different distance functions.

Now, $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ can be seen as a particular case of $\widehat{P}_{\mathrm{MLC}i}^{GDF}$ in which $\tilde{p}_{ki}^{\circ}$ includes only p_{ki}°, and $\hat{\boldsymbol{\alpha}}_i^{\circ}$ is only a scalar and is set exactly equal to 1. Therefore, $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ is consistent for P_i and asymptotically normal with $V_{\infty}(\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu_{ki}^{\circ}$. Variance estimation can again be conducted by plugging sample level estimated residuals in (19) given in this case by $\hat{e}_{ki} = z_{ki} - p_{ki}^{\circ}$.

Estimator $\hat{P}_{\mathrm{MLC}}^{\mathrm{DF}}$ is in all similar to $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, the only difference is in the fact that coefficient estimates for the multinomial model are obtained separately from the two frames and, therefore, we have two separate model calibration constraints. In this case the vector of auxiliary variables used in the calibration procedure can be written as $\tilde{p}_{ki}^{A,B}$ and contains p_{ki}^{A}, p_{ki}^{B} and the other indicator variables used in the benchmark constraints: for example $\tilde{p}_{ki}^{A,B} = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab)]p_{ki}^{A}, [\delta_k(b) + \delta_k(ba)]p_{ki}^{B})^{\intercal}$.

To encompass this situation, it is enough to change assumption A1 accordingly and assume that the two sets of population parameters $\boldsymbol{\beta}^A$ and $\boldsymbol{\beta}^B$ are consistently estimated by $\hat{\boldsymbol{\beta}}^A$ and $\hat{\boldsymbol{\beta}}^B$ and that these samples fits and the finite population fits share a common finite limit. Then, it is easy to show that $\hat{P}_{\mathrm{MLC}}^{\mathrm{DF}}$ is design consistent and the variance of its asymptotic normal distribution can again be written in terms of the variance of the population total of residuals. In particular, $V_{\infty}(\widehat{P}_{\mathrm{MLC}i}^{GDF}) = N^{-2}V(\hat{t}_{eiH})$ and $\hat{t}_{eiH} = \sum_{k\in s} d_k^{\circ} e_{ki}$ is the Hartley estimator of the population total of the census-level residuals given here by $e_{ki} = z_{ki} - (\tilde{\boldsymbol{\mu}}^{A,B})_{ki}^{\intercal}\boldsymbol{\alpha}_i$, where $\tilde{\boldsymbol{\mu}}_{ki}^{A,B}$ is like $\tilde{p}_{ki}^{A,B}$ but with p_{ki}^{A} and p_{ki}^{B} replaced by their population counterparts, similarly to (18). Analytic variance estimation can be conducted by using sample level estimates of the residuals. In particular, by using $\hat{e}_{ki} = z_{ki} - (\tilde{p}_{ki}^{A,B})^{\intercal}\hat{\boldsymbol{\alpha}}_i$ in formula (19).

Now, similarly as for $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ and $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, $\hat{P}_{\mathrm{ML}}^{\mathrm{DF}}$ can be seen as a particular case of $\widehat{P}_{\mathrm{MLC}i}^{GDF}$ in which $\tilde{p}_{ki}^{\circ}$ includes only $p_{ki}^{A,B}$, with $p_{ki}^{A,B} = p_{ki}^{A}$ if $k \in s_A$ and $p_{ki}^{A,B} = p_{ki}^{B}$ if $k \in s_B$, and $\hat{\boldsymbol{\alpha}}_i^{\circ}$ is again a scalar here and its value is set exactly equal to 1. Therefore, it is consistent for P_i and asymptotically normal with $V_{\infty}(\hat{P}_{\mathrm{ML}}^{\mathrm{DF}}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu_{ki}^{A,B}$, and $\mu_{ki}^{A,B}$ is the census level fit corresponding to $p_{ki}^{A,B}$. Variance estimation can again be conducted by using sample level estimated residuals in equation (19) given by $\hat{e}_{ki} = z_{ki} - p_{ki}^{A}$ if $k \in s_A$ and $\hat{e}_{ki} = z_{ki} - p_{ki}^{B}$ if $k \in s_B$.

The calibration estimator $\hat{P}_{\mathrm{MLC}}^{\mathrm{SW}}$ is very similar to $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, the only differences are (i) in the set of basic design weights employed in the calibration procedure: for $\hat{P}_{\mathrm{MLC}}^{\mathrm{SW}}$ we use $\tilde{d}_k$, and (ii) p_{ki}° is replaced by $\tilde{p}_{ki}$ in the definition of the vector $\tilde{p}_{ki}^{\circ}$. Once these changes are incorporated across assumption A1, and assumption A5 reflects the fact that we are now dealing with Bankier-Kalton-Anderson type estimators, instead of Hartley estimators, then all the results can be proven. The variance of the asymptotic distribution

of $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ is given by $V_\infty(\widehat{P}^{\mathrm{GSW}}_{\mathrm{MLC}i}) = N^{-2}V(\hat{t}_{ei})$ and $\hat{t}_{ei} = \sum_{k\in s}\tilde{d}_k e_{ki}$ is the single-frame estimator of the population total of the census-level residuals $e_{ki} = z_{ki} - \tilde{\boldsymbol{\mu}}_{ki}^{\mathrm{T}}\boldsymbol{\alpha}_i$, and where $\tilde{\boldsymbol{\mu}}_{ki}$ is like $\tilde{p}_{ki}$ but with p_{ki} replaced by its population counterpart

$$\mu_{ki} = \frac{\exp(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta}_{Ui})}{\sum_{r=1,\ldots,m}\exp(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta}_{Ur})}.$$

In addition, let $\hat{e}_{ki} = z_{ki} - \tilde{p}_{ki}^{T}\hat{\boldsymbol{\alpha}}_i$. Then, $V(\hat{t}_{ei})$ can be consistently estimated so that $v(\widehat{P}^{\mathrm{GSW}}_{\mathrm{MLC}i}) = N^{-2}v(\hat{t}_{\hat{e}i})$.

6. Selection of the optimal weight

In the previous sections we have considered a fixed value $0 < \eta < 1$. Selection of parameter η is an important issue in dual frame estimators, because the efficiency of the estimator relies heavily on this value (see Lohr (2009) for a review). Hartley (1962) proposed choosing η to minimize the variance of the estimator in (2). Using the same idea, we can derive the optimal value of η for each proposed multinomial logistic estimator by minimizing its asymptotic variance with respect to η. However, as the optimal value for the Hartley estimator, such optimal values would depend on unknown population quantities, such as variances and covariances that, when estimated from sample data, would make the final estimator depend on the values of the variable of interest. This implies a need to recompute an optimal η for each value $i = 1,\ldots,m$ and for each variable of interest y, which will be inconvenient in practice for statistical agencies conducting surveys with several variables, other than introducing a lack in coherence among estimates that is particularly relevant when dealing with multinomial outcomes (namely, $\sum_i \widehat{P}_i$ can be $\neq 1$).

Skinner and Rao (1996) suggested choosing

$$\eta_{SR} = \frac{N_a N_B V(\hat{N}_{ba})}{N_a N_B V(\hat{N}_{ba}) + N_b N_A V(\hat{N}_{ab})},$$

or alternatively

$$\eta_{SR2} = \frac{V(\hat{N}_{ba})}{V(\hat{N}_{ba}) + V(\hat{N}_{ab})},$$

being $V(\hat{N}_{ab})$ and $V(\hat{N}_{ba})$ the variances of the estimated sizes of domain ab based on samples s_A and s_B respectively. These two proposals provide a value for η that does not depend on the sample values of y. In this way, resulting estimator uses the same η for all variables of interest, even if variances $V(\hat{N}_{ab})$ and $V(\hat{N}_{ba})$ are unknown and must be estimated from the data.

Brick et al. (2006) propose using the simple value $\eta = 1/2$ in their dual-frame study in which frame A was a landline telephone frame and frame B was a cell-phone frame. For this purpose, the value of $\eta = 1/2$ is frequently recommended (see, for example, Mecatti (2007)). Another simple choice for η is given by $\frac{N_B/n_B}{N_A/n_A+N_B/n_B}$ (see Skinner and Rao (1996) or Lohr and Rao (2000)).

7. Jackknife variance estimation

In this section we explore the possibility of using jackknife methods to estimate the variance of the proposed estimators as an alternative to the analytic variance estimators considered in Section 5. The jackknife approach is a common replication method for variance estimation that can be used in complex surveys for different types of estimators (see e.g. Wolter (2003) for an introduction to jackknife). For the sake of brevity, in this section all estimators are denoted by $\hat{P}_i, i = 1, \cdots, m$.

If we consider a non clustered and non stratified design, the jackknife estimator for the variance of $\hat{P}_i$ may be given by

$$v_J(\hat{P}_i) = V_J^A + V_J^B = \frac{n_A - 1}{n_A} \sum_{g \in s_A} (\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{P}_i^B(j) - \overline{P}_i^B)^2 \qquad (20)$$

where $\hat{P}_i^A(g)$ is the value taken by estimator $\hat{P}_i$ after dropping unit g from s_A and $\overline{P}_i^A$ is the average of $\hat{P}_i^A(g)$ values. Each value $\hat{P}_i^A(g)$ is computed by fitting a new model that does not consider the $g-th$ sample unit. $\hat{P}_i^B(j)$ and $\overline{P}_i^B$ are defined similarly.

In the case of a stratified design in both frames, let frame A be divided into H strata and let stratum h has N_{Ah} observation units of which n_{Ah} are sampled. Similarly, frame B has L strata, stratum l has N_{Bl} observation units of which n_{Bl} are sampled. Then, a jackknife variance estimator of $\hat{P}_i$ is given by

$$v_J^{st}(\hat{P}_i) = V_J^{stA} + V_J^{stB} =$$

$$= \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}} \sum_{g \in s_{Ah}} (\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2, \qquad (21)$$

where $\hat{P}_i^A(hg)$ is the value taken by estimator $\hat{P}_i$ after dropping unit g of stratum h from sample s_{Ah}, $\overline{P}_i^{Ah}$ is the average of these n_{Ah} values; $\hat{P}_i^B(lj)$ and $\overline{P}_i^{Bl}$ are defined similarly.

In case of a non stratified design in one frame and a stratified design in the other one, previous methods can be combined to obtain the corresponding jackknife estimator of the variance.

Alternatively, a finite-population correction can be considered, as described in Ranalli et al. (2015), resulting in the following jackknife variance estimators:

$$v_{Jc}(\hat{P}_i) = \frac{n_A - 1}{n_A}(1-\overline{\pi}_A)\sum_{g\in s_A}(\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B}(1-\overline{\pi}_B)\sum_{j\in s_B}(\hat{P}_i^B(j) - \overline{P}_i^B)^2 \quad (22)$$

for non stratified designs in frames, where $\overline{\pi}_A = \frac{1}{n_A}\sum_{k\in s_A}\pi_{Ak}$ and similarly for $\overline{\pi}_B$, and

$$\begin{aligned} v_{Jc}^{st}(\hat{P}_i) = &\sum_{h=1}^{H}\frac{n_{Ah}-1}{n_{Ah}}(1-\overline{\pi}_{Ah})\sum_{g\in s_{Ah}}(\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2 \\ &+\sum_{l=1}^{L}\frac{n_{Bl}-1}{n_{Bl}}(1-\overline{\pi}_{Bl})\sum_{j\in s_{Bl}}(\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2 \end{aligned} \quad (23)$$

for a stratified design in each frame, where $\overline{\pi}_{Ah} = \frac{1}{n_{Ah}}\sum_{k\in s_{Ah}}\pi_{Ak}$ and similarly for $\overline{\pi}_{Bl}$.

A non clustered sampling design is assumed subsequently. No new principles are involved in the application of jackknife methodology to clustered samples. We simply work with the ultimate cluster rather than elementary units (see e.g. Wolter (2003)).

8. Monte Carlo simulation experiments

For our simulation study we use the `hsbdemo` data set (`http://www.ats.ucla.edu/stat/data/hsbdemo.dta`). The data set contains variables on 200 students. The outcome variable is `prog`, program type, a three-level categorical variable whose categories are `academic`, `general`, `vocation`. The predictor variables are social economic status, `ses`, a three-level categorical variable and a mathematical score, `math`, a continuous variable. We estimate a multinomial logistic regression model. We create a new data set with 50 copies of the predictor variables `ses` and `math` and with the predicted values for the variable `prog` (the category with highest probability). The simulated populations, namely POP1, have, therefore, dimension $N = 10\,000$.

Units are randomly assigned to the two frames, A and B, according to three different scenarios depending on the overlap domain size N_{ab}. We first generate N normal random numbers, $\varepsilon_k, k = 1,\ldots,N$ and data is sorted by such random numbers. Then, the first N_a records of the ordered dataset are considered as the values of the domain a, the N_b subsequent records as the values belonging to domain b and the last N_{ab} records as the values of the domain ab. The first scenario has a *small* overlap domain size N_{ab}=1 000 and the resulting sizes of the two frames are N_A=6 000 and N_B=5 000. The second and the third scenario have respectively *medium* and *large* overlap domain size. The resulting frame sizes in the second scenario are given by N_A=6 000 and N_B=7 000 and the overlap domain size is N_{ab}=3 000, while for the third scenario we have N_A=8 000, N_B=7 000 and N_{ab}=5 000. In POP1, we compute all estimators using as auxiliary information `ses` and `math`.

On the other hand, POP2 is built first by assigning units to the frames and second by fitting a multinomial logistic regression model separately in each frame. In frame A, `ses`

and `math` have been considered as auxiliary variables and in frame B the auxiliary variables are `ses` and `write` (a score in writing). To be able to fit a separated model in each frame we consider that the units composing the overlap domain can be equally divided into two groups, each one coming from a frame. So half of the overlap domain units are used to fit a multinomial logistic regression model in frame A and the remaining ones are considered when fitting the multinomial logistic model in frame B. POP2 is built with the predicted values from the two multinomial logistic model. In this population, we compute $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ estimators using as x-variable `ses` (Case I), and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$ estimators using as x_A-variables `ses` and `math` and as x_B-variables `ses` and `write` (Case II).

Samples of schools from frame A are selected by means of Midzuno sampling, with inclusion probabilities proportional to the size of the school the student belongs to. All students in the selected schools are included in the sample. The variable `cid` is an indicator of school. Samples from frame B are selected by means of simple random sampling. For each scenario, we draw a combination of sample sizes for frame A and frame B, as follows: $n_A = 180$ and $n_B = 232$.

We have two populations, three sizes of the overlap domain and different sets of auxiliary variables.

We compute the BKA estimator in (9), for the purpose of comparison. The Pseudo Empirical Likelihood estimator (PEL) proposed in Rao and Wu (2010) and the dual frame and the single frame calibration estimator ($\hat{P}_{\mathrm{CalDF}}$ and $\hat{P}_{\mathrm{CalSF}}$) proposed in Ranalli et al. (2015) are also computed using the auxiliary information as previously mentioned (in POP1 `ses` and `math` for both estimators and in POP2 as x_A-variable `ses` and `math` and as x_B-variable `ses` and `write` for $\hat{P}_{\mathrm{CalDF}}$ estimator and as x-variable `ses` for $\hat{P}_{\mathrm{CalSF}}$ estimator). When needed (and for comparative purposes) the value of η has been estimated using $\eta = v(\hat{N}_{ba})/(v(\hat{N}_{ab}) + v(\hat{N}_{ba}))$ (see for example Rao and Wu (2010)) for all compared estimators, where $v(\hat{N}_{ab})$ is an estimate of the variance of the Horvitz-Thompson estimator $\hat{N}_{ab}$ for the size of overlap domain, and similarly for $v(\hat{N}_{ba})$.

For each estimator, we compute the percent relative bias $\mathrm{RB}\% = 100 * E_{\mathrm{MC}}(\hat{Y} - Y)/Y$, the percent relative mean squared error $\mathrm{RMSE}\% = 100 * E_{\mathrm{MC}}[(\hat{Y} - Y)^2]/Y^2$, based on 1000 simulation runs, for each category of the main variable `prog`.

The percent relative biases are negligible in all cases (the results on RB are not included for brevity), so efficiency comparisons can be based on variances. Table 4 displays the relative efficiency of proposed estimators with respect to BKA estimator. From this table we can see that, consistently with theoretical findings, the performance in terms of efficiency of the estimators is essentially driven by the model employed. When the auxiliary varibles are used in a calibration process using a linear model ($\hat{P}_{CalSF}$, $\hat{P}_{CalDF}$) or through a pseudo-empirical likelihood method (PEL), the efficiency increases with respect to the BKA estimator, which does not use auxiliary information or any model. As expected, a most effective situation arises when the auxiliary variables are also used through a multinomial model ($\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$).

Table 4: *Relative efficiency (respect to the BKA estimator) of compared estimators. POP1 and POP2.*

	POP1			POP2		
	acad.	gen.	voc.	acad.	gen.	voc.
			Medium			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	149.94	142.21	132.30	152.77	145.10	129.26
$\hat{P}_{\text{PEL}}$	217.89	135.87	177.26	175.94	146.75	148.75
$\hat{P}_{\text{CalDF}}$	213.91	134.83	175.14	175.03	146.84	147.59
$\hat{P}^{\text{DW}}_{\text{ML}}$	347.02	181.43	252.42	204.46	194.97	148.32
$\hat{P}^{\text{DW}}_{\text{MLC}}$	356.87	181.05	258.60	209.29	192.64	153.29
$\hat{P}^{\text{SW}}_{\text{ML}}$	348.12	181.25	252.44	205.63	194.71	148.82
$\hat{P}^{\text{SW}}_{\text{MLC}}$	358.10	180.97	258.85	210.22	192.32	153.70
$\hat{P}^{\text{DF}}_{\text{ML}}$	350.18	187.65	257.22	207.83	251.93	147.44
$\hat{P}^{\text{DF}}_{\text{MLC}}$	358.93	186.31	263.52	214.76	250.13	153.44
			Small			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	155.30	137.56	140.60	152.77	142.46	137.70
$\hat{P}_{\text{PEL}}$	232.55	147.36	198.25	179.24	149.26	158.30
$\hat{P}_{\text{CalDF}}$	210.50	134.54	179.08	182.73	150.09	160.65
$\hat{P}^{\text{DW}}_{\text{ML}}$	331.43	163.16	247.64	165.45	146.32	157.70
$\hat{P}^{\text{DW}}_{\text{MLC}}$	353.76	163.06	265.66	176.59	146.83	166.11
$\hat{P}^{\text{SW}}_{\text{ML}}$	331.75	163.33	248.08	166.09	146.83	157.60
$\hat{P}^{\text{SW}}_{\text{MLC}}$	353.77	163.17	265.85	176.78	146.99	165.93
$\hat{P}^{\text{DF}}_{\text{ML}}$	343.94	164.70	257.75	170.24	150.15	154.31
$\hat{P}^{\text{DF}}_{\text{MLC}}$	365.15	163.94	275.28	184.50	150.24	164.51
			Large			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	147.60	130.53	138.13	152.25	121.61	125.29
$\hat{P}_{\text{PEL}}$	193.48	124.99	173.21	163.71	142.12	149.74
$\hat{P}_{\text{CalDF}}$	192.10	125.72	170.56	165.55	153.62	161.09
$\hat{P}^{\text{DW}}_{\text{ML}}$	354.00	161.79	256.45	303.59	118.57	269.38
$\hat{P}^{\text{DW}}_{\text{MLC}}$	371.74	161.23	266.64	307.98	123.76	282.16
$\hat{P}^{\text{SW}}_{\text{ML}}$	356.73	161.87	257.40	302.59	119.33	269.14
$\hat{P}^{\text{SW}}_{\text{MLC}}$	375.21	161.38	267.54	306.81	124.75	281.93
$\hat{P}^{\text{DF}}_{\text{ML}}$	362.07	168.39	265.88	344.86	130.46	370.90
$\hat{P}^{\text{DF}}_{\text{MLC}}$	376.11	167.22	274.78	348.03	137.80	379.38

Table 5: *Length reduction (in percent, %) of proposed estimator with respect to linear calibration estimators using the same amount of auxiliary information ($\hat{P}^{DW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{SW}_{ML}$ and $\hat{P}^{SW}_{MLC}$ have been compared with $\hat{P}_{CalSF}$ and $\hat{P}^{DF}_{ML}$ and $\hat{P}^{DF}_{MLC}$ have been compared with $\hat{P}_{CalDF}$). Coverage (in percent, %) of jackknife confidence intervals. POP1.*

	Length reduction			Cov		
	acad.	gen.	voc.	acad.	gen.	voc.
			Medium			
$\hat{P}^{DW}_{ML}$	10.31	25.44	30.91	94.5	93.9	94.9
$\hat{P}^{DW}_{MLC}$	9.90	28.28	32.78	95.2	93.9	94.5
$\hat{P}^{SW}_{ML}$	10.59	25.73	31.18	94.8	94.1	95.0
$\hat{P}^{SW}_{MLC}$	9.95	28.34	32.82	95.0	93.8	94.5
$\hat{P}^{DF}_{ML}$	8.83	33.04	16.41	95.8	96.0	95.5
$\hat{P}^{DF}_{MLC}$	8.11	35.23	18.24	95.9	95.3	95.1
			Small			
$\hat{P}^{DW}_{ML}$	9.14	23.76	28.25	95.0	93.2	95.2
$\hat{P}^{DW}_{MLC}$	8.78	26.86	30.41	94.1	93.4	93.6
$\hat{P}^{SW}_{ML}$	9.43	24.04	28.52	94.5	93.5	94.0
$\hat{P}^{SW}_{MLC}$	8.81	26.89	30.43	94.8	92.5	94.2
$\hat{P}^{DF}_{ML}$	6.98	24.64	13.09	96.3	95.0	95.9
$\hat{P}^{DF}_{MLC}$	6.30	27.15	15.32	96.6	94.6	95.1
			Large			
$\hat{P}^{DW}_{ML}$	10.11	25.45	30.71	94.2	93.5	93.9
$\hat{P}^{DW}_{MLC}$	9.34	28.24	32.38	94.1	93.4	93.6
$\hat{P}^{SW}_{ML}$	10.64	25.94	31.14	94.5	93.5	94.0
$\hat{P}^{SW}_{MLC}$	9.71	28.51	32.62	94.8	92.5	94.2
$\hat{P}^{DF}_{ML}$	10.18	35.37	17.96	96.3	95.0	95.9
$\hat{P}^{DF}_{MLC}$	9.29	37.39	19.45	96.6	94.6	95.1

In general, the best results in efficiency are achieved by the $\hat{P}^{DF}_{MLC}$ estimator and the efficiency increases as the size of the overlap domain increases, particularly for POP2. As a consequence of the ignorability of the frames the units belong to when modelling the relation between the response and the auxiliary variables, there is not a relevant difference in efficiency between estimators using a multinomial model in the whole population and estimators using a multinomial model in each frame.

We now turn to the evaluation of the precision of the proposed estimators by means of confidence intervals. We obtain the 95% confidence intervals based on a normal distribution and the jackknife variance estimator proposed in Section 7 with finite-population correction. Table 5 shows the average length reduction of 95% confidence intervals and

Table 6: *Relative efficiency (respect to the BKA estimator) of compared estimator for* $\hat{\eta}_{SR2} = v(\hat{N}_{ba})/(v(\hat{N}_{ab}) + v(\hat{N}_{ba}))$, $\hat{\eta}_{SR} = N_a N_B v(\hat{N}_{ba})/(N_b N_A v(\hat{N}_{ab}) + N_a N_B v(\hat{N}_{ba}))$ *and* $\eta_{1/2} = \frac{1}{2}$. *Overlap domain size* Medium.

		POP1			POP2		
		acad.	gen.	voc.	acad.	gen.	voc.
$\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$	$\hat{\eta}_{SR2}$	347.02	181.43	252.42	204.46	194.97	148.32
	$\hat{\eta}_{SR}$	348.45	181.32	252.88	205.14	194.69	148.71
	$\eta_{1/2}$	347.27	181.30	252.57	204.69	194.91	148.32
$\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$	$\hat{\eta}_{SR2}$	356.87	181.05	258.60	209.29	192.64	153.29
	$\hat{\eta}_{SR}$	358.65	181.01	259.21	209.78	192.36	153.62
	$\eta_{1/2}$	357.11	180.91	258.76	209.48	192.54	153.26
$\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$	$\hat{\eta}_{SR2}$	350.18	187.65	257.22	207.83	251.93	147.44
	$\hat{\eta}_{SR}$	351.57	187.70	257.90	207.85	249.31	147.45
	$\eta_{1/2}$	350.34	187.45	257.33	208.03	251.91	147.50
$\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$	$\hat{\eta}_{SR2}$	358.93	186.31	263.52	214.76	250.13	153.44
	$\hat{\eta}_{SR}$	360.76	186.46	264.35	214.57	247.50	153.26
	$\eta_{1/2}$	215.02	250.07	153.52	182.44	148.19	163.36

the empirical coverage probability over 1000 simulation runs in each category of the main variable. The confidence interval lengths of proposed estimators have been compared with the confidence interval lengths of their linear calibration counterparts using the same amount of auxiliary information. That is, $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ have been compared with $\hat{P}_{\mathrm{CalSF}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$ have been compared with $\hat{P}_{\mathrm{CalDF}}$.

From Table 5 we conclude that all the proposed estimators considerably reduce the length of the confidence intervals obtained, with respect to the linear calibration estimators. The empirical coverage is very close to the nominal level. It is observed that the estimates based on the joint estimation of the parameter $\boldsymbol{\beta}$ ($\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$) have a somewhat lower coverage than the others.

Looking at the effect of the choice of η (in relative bias and relative mean squared error), we have repeated the simulation study (for all populations and scenarios) using alternative values for η. In particular, other than that used previously, i.e.

$$\eta_{SR2} = \frac{v(\hat{N}_{ba})}{v(\hat{N}_{ba}) + v(\hat{N}_{ab})},$$

we have considered a fixed value $\eta = \frac{1}{2}$ and one estimated following Skinner and Rao (1996)

$$\eta_{SR} = \frac{N_a N_B v(\hat{N}_{ba})}{N_a N_B v(\hat{N}_{ba}) + N_b N_A v(\hat{N}_{ab})}.$$

See Section 6. for details and guidelines on choosing a value for η. Table 6 shows (only when the overlap domain size is *Medium*, for space reason) that there is a little effect of these three different estimates for η on the behaviour of the considered estimators. We can conclude that the available auxiliary information and the way in which it is included in the estimation procedure play a much more relevant role than the choice of a value for η.

9. Application to the survey on opinions and attitudes of the Andalusian population regarding immigration (OPIA) 2013

To examine the performance of the proposed estimation methods in practice, we have applied them to the dataset from the OPIA survey. The main variable in this study is related to the "attitude towards immigration". The variable is the answer to the following question: *And in relation to the number of immigrants currently living in Andalusia, do you think there are ...?: Too many, A reasonable number, Too few, No reply*.

We have considered the same set of auxiliary variables (`sex` and `age`) in the two frames. To incorporate information about sex into estimation process two indicator variables (one for males and another one for females) were created. Similarly, four age classes were established and each respondent was assigned to one of them. Corresponding indicator variables were used, then, for the analysis. Necessary population information about these variables for calculating proposed estimators is displayed in Table 3. Note that both auxiliary variables `sex` and `age` are available from the two frames. In this case, the population counts in the two-way contingence table are known in each domain.

Table 7 shows point and jackknife confidence estimation for proposed estimators. Length reduction in jackknife confidence interval for each estimator regarding same interval for BKA estimator is also displayed. In keeping with results obtained from simulation experiments, reduction is quite significative for all estimators whatever the category of the main variable. The calibration approach achieves most important reductions in length, with single frame calibration presenting the best results. On the other hand, using $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ estimators the length reduction is less noticeable.

Table 8 shows point estimation for proposed estimators by sex and age. Analyzing results by gender, it is noticeable that there are more males than females thinking that there are too many immigrants in Andalusia and that females are more reticent to answer the question than males.

On the other hand, it is worth noting that perception that there are too many immigrants in Andalusia increases together with age. So, while most of the people in the 18-29 age group think that the number of immigrants in Andalusia is reasonable, most part of people aged 45 years or over think that there are too many. The age group where the non-response is higher is the one including people aged 60 years or over.

Table 7: *Point and 95% confidence level estimation of proportions using several methods for Jackknife variance estimation. Length reduction (in percent, %) respect to the BKA estimator. Main variable: "Amount of immigration".*

In relation to the number of immigrants currently living in Andalusia, do you think there are ...?					
Estimator	PROP	LB	UB	LEN	Length reduction
			Too many		
$\hat{P}^{DW}_{ML}$	42.75	39.76	45.74	5.98	14.33
$\hat{P}^{DW}_{MLC}$	41.23	38.78	43.68	4.90	29.80
$\hat{P}^{SW}_{ML}$	42.89	39.94	45.84	5.90	15.47
$\hat{P}^{SW}_{MLC}$	41.41	39.03	43.79	4.76	31.81
$\hat{P}^{DF}_{ML}$	42.61	39.64	45.58	5.94	14.90
$\hat{P}^{DF}_{MLC}$	41.16	38.67	43.65	4.98	28.65
			A reasonable number		
$\hat{P}^{DW}_{ML}$	45.24	42.27	48.20	5.93	12.28
$\hat{P}^{DW}_{MLC}$	46.57	44.11	49.03	4.92	27.22
$\hat{P}^{SW}_{ML}$	45.09	42.17	48.01	5.84	13.61
$\hat{P}^{SW}_{MLC}$	46.40	44.02	48.78	4.76	29.59
$\hat{P}^{DF}_{ML}$	45.45	42.49	48.41	5.92	12.43
$\hat{P}^{DF}_{MLC}$	46.68	44.17	49.18	5.01	25.89
			Too few		
$\hat{P}^{DW}_{ML}$	6.06	4.55	7.58	3.03	15.36
$\hat{P}^{DW}_{MLC}$	5.77	4.58	6.97	2.39	33.24
$\hat{P}^{SW}_{ML}$	6.05	4.56	7.54	2.98	16.76
$\hat{P}^{SW}_{MLC}$	5.76	4.61	6.91	2.30	35.75
$\hat{P}^{DF}_{ML}$	6.13	4.62	7.64	3.02	15.64
$\hat{P}^{DF}_{MLC}$	5.63	4.46	6.80	2.34	34.64
			No reply		
$\hat{P}^{DW}_{ML}$	5.95	4.65	7.25	2.60	12.75
$\hat{P}^{DW}_{MLC}$	6.43	5.27	7.58	2.31	22.48
$\hat{P}^{SW}_{ML}$	5.96	4.67	7.25	2.58	13.42
$\hat{P}^{SW}_{MLC}$	6.43	5.30	7.56	2.26	24.16
$\hat{P}^{DF}_{ML}$	5.80	4.51	7.10	2.59	13.09
$\hat{P}^{DF}_{MLC}$	6.54	5.33	7.74	2.41	19.13

Table 8: *Point estimation of proportions by sex and age. Main variable: "Amount of immigration".*

	In relation to the number of immigrants currently living in Andalusia, do you think there are ...?						
Estimator	ALL	MALES	FEMALES	18-29	30-44	45-59	≥ 60
			Too many				
$\hat{P}^{DW}_{ML}$	42.75	46.46	39.15	32.46	44.29	46.03	45.14
$\hat{P}^{DW}_{MLC}$	41.23	43.64	38.97	30.97	42.07	43.31	46.58
$\hat{P}^{SW}_{ML}$	42.89	46.74	39.11	32.76	43.89	46.44	45.85
$\hat{P}^{SW}_{MLC}$	41.41	43.79	39.19	31.55	41.61	43.87	45.77
$\hat{P}^{DF}_{ML}$	42.61	44.45	39.16	31.99	41.69	43.56	48.13
$\hat{P}^{DF}_{MLC}$	41.16	43.55	38.96	30.01	42.14	43.28	48.56
			A reasonable number				
$\hat{P}^{DW}_{ML}$	45.24	42.31	48.10	59.82	40.71	40.72	44.47
$\hat{P}^{DW}_{MLC}$	46.57	44.39	48.74	61.97	44.44	42.72	43.25
$\hat{P}^{SW}_{ML}$	45.09	42.04	48.11	59.62	40.90	40.68	43.70
$\hat{P}^{SW}_{MLC}$	46.40	44.14	48.63	61.49	44.67	42.64	43.61
$\hat{P}^{DF}_{ML}$	45.45	44.02	48.35	60.42	43.98	42.81	42.11
$\hat{P}^{DF}_{MLC}$	46.68	44.59	48.78	63.21	44.46	42.56	41.65
			Too few				
$\hat{P}^{DW}_{ML}$	6.06	6.75	5.35	3.77	9.84	6.18	2.82
$\hat{P}^{DW}_{MLC}$	5.77	6.68	4.92	3.29	7.58	6.73	2.80
$\hat{P}^{SW}_{ML}$	6.05	6.64	5.47	3.79	9.89	6.12	2.83
$\hat{P}^{SW}_{MLC}$	5.76	6.67	4.92	3.39	7.62	6.66	2.95
$\hat{P}^{DF}_{ML}$	6.13	6.58	5.11	3.50	8.17	6.37	2.39
$\hat{P}^{DF}_{MLC}$	5.63	6.46	4.81	2.92	7.46	6.77	2.35
			No reply				
$\hat{P}^{DW}_{ML}$	5.95	4.47	7.39	3.95	5.16	7.06	7.56
$\hat{P}^{DW}_{MLC}$	6.43	5.28	7.37	3.76	5.91	7.24	7.37
$\hat{P}^{SW}_{ML}$	5.96	4.58	7.31	3.83	5.32	6.76	7.62
$\hat{P}^{SW}_{MLC}$	6.43	5.41	7.26	3.57	6.10	6.84	7.67
$\hat{P}^{DF}_{ML}$	5.80	4.95	7.38	4.09	6.15	7.25	7.36
$\hat{P}^{DF}_{MLC}$	6.54	5.39	7.45	3.86	5.93	7.39	7.44

10. Conclusions

Data collected from surveys are often organized into discrete categories. Analyzing such categorical data from a complex survey often requires specialized techniques. To improve the accuracy of estimation procedures, a survey statistician often makes use of the auxiliary data available from administrative registers and other sources.

Generalized regression is a popular design-based method used in the production of descriptive statistics from survey data. Although the generalized regression estimator is design-consistent regardless of the form of the assisting model, a linear model is not the best choice for multinomial response variables. For such variables we introduce a class of multinomial logistic generalized regression estimators when data are obtained from samples from different frames.

We introduce a new approach to the model-assisted estimation of population class of frequencies in dual frame surveys. We propose a class of logistic estimators based on multinomial logistic models describing the joint distribution of the category indicators in the total population or in each frame separately. We also consider different ways of combining estimates coming from the two frames.

The type of sample design used in practice drives the user to choose between Dual-Frame or Single-Frame approaches. The Single-Frame approach requires additional information in the overlapping domain that is not always easy to take in practical applications.

As for calibration, it seems clear that the better for efficiency is to incorporate it, regardless of whether or not a logistics model is used. As for the model, apart from the advantage provided by the fact that the estimates of proportions for each category add to one, our simulation study suggests that it is preferable to use it. As for the type of model, in most practical applications it will be almost entirely forced, depending on the auxiliary information available and, more specifically, on the availability of auxiliary variable totals for domains, for frames or for the entire population.

To compute the proposed estimators, we have assumed to know the values of auxiliary variables for each individual in the population, which can be quite a restrictive assumption. Indeed, to compute the proposed estimators we need to know the count of each value of the auxiliary variable vector in the population. This is a very frequent situation that arises, for example, when categorical variables (as the gender or the professional status of the individual) or quantitative categorized variables (as the age of the individual, grouped in classes) are used as auxiliary information in a survey. In this context, we do not have a complete list of individuals but still the proposed estimators can be computed since the population information needed can be found in databases of national statistical organisms. In fact, in this case, we only need to know the population count in the multi-way contingency table. This is also the situation in the application to data from the survey on opinions and attitudes of the Andalusian population regarding immigration explored in Section 9.

Here we have considered two frames. The extension to more than two frames is under study as well. One important issue when dealing with more than two frames is that of using a proper notation (see Lohr and Rao (2006) and Singh and Mecatti (2011)). A first simple way around is the one, also considered in Rao and Wu (2010), in which weights from the multiplicity estimator of Mecatti (2007) are used as starting weights and calibration is applied straightforwardly. More complicated is the issue of accounting for different levels of frame information, although we believe that Singh and Mecatti (2011) may provide a good starting point.

Acknowledgements

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2012-35650 and FPU grant program, Spain), by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain), and under the support of the project PRIN-SURWEY (grant 2012F42NS8, Italy). The authors thank the Editor and the reviewers for their helpful suggestions.

References

Arcos, A., D. Molina, M. Rueda, and M. G. Ranalli (2015). Frames2: A package for estimation in dual frame surveys. *The R Journal*, 7, 52–72.

Bankier, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074–1079.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, 279–292.

Brick, J. M., S. Dipko, S. Presser, C. Tucker, and Y. Yuan (2006). Nonresponse bias in a dual frame survey of cell and landline numbers. *Public Opinion Quarterly*, 70, 780–793.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Fuller, W. A. and L. F. Burmeister (1972). Estimators for samples selected from two overlapping frames. *Proceedings of social science section of The American Statistical Asociation*.

Godambe, V. P. and M. E. Thompson (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127–138.

Hartley, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

Isaki, C. T. and W. A. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89–96.

Kalton, G. and D. W. Anderson (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General)*, 149, 65–82.

Lehtonen, R. and A. Veijanen (1998). On multinomial logistic generalized regression estimators. Technical Report 22, Department of Statistics, University of Jyväskylä.

Lohr, S. and J. Rao (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019–1030.

Lohr, S. L. (2009). Multiple-frame surveys. *Handbook of Statistics*, 29, 71–88.

Lohr, S. L. and J. N. K. Rao (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271–280.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology*, 33, 151–157.

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.

Ranalli, M., A. Arcos, M. Rueda, and A. Teodoro (2015). Calibration estimation in dual-frame surveys. *Statistical Methods and Applications First online: 01 September 2015*, 1–29.

Rao, J. N. K. and C. Wu (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494–1503.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Singh, A. C. and F. Mecatti (2011, 12). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of official statistics*, 27, 1–19.

Skinner, C. J. and J. N. K. Rao (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349–356.

Wolter, K. (2003). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Wu, C. and J. N. K. Rao (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359–375.

Wu, C. and R. R. Sitter (2001a). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185–193.

Wu, C. and R. R. Sitter (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289–307.