Proceedings of the

33rd International Workshop on Statistical Modelling

Volume II

July 16-20, 2018

Bristol, UK

Contents

Box–Cox response transformations for random effect models

Amani Almohaimeed[1][2], Jochen Einbeck[1]

[1] Durham University, UK
[2] Qassim University, Saudi Arabia

E-mail for correspondence: `jochen.einbeck@durham.ac.uk`

Abstract: For the linear model with random effects of unspecified distribution, we develop methodology for simultaneous response transformation and estimation of regression parameters. This is achieved by extending the "Nonparametric Maximum Likelihood" towards a "Nonparametric Profile Maximum Likelihood" technique. The methods allow to deal with overdispersion as well as two–level data scenarios.

Keywords: Box–Cox transformation, variance component model, EM algorithm

1 Introduction

For data with a two–level structure, such as longitudinal data, correlation of responses within upper–level units can be induced by adding a random effect z_i to the linear predictor $x_{ij}^T\beta$, with the upper-level indexed by $i = 1, \dots, r$, and the lower-level indexed by $j = 1, \dots, n_i$, $\sum n_i = n$. Conditional on the random effect, the responses y_{ij} are independently distributed with mean function

$$E(y_{ij}|z_i) = x_{ij}^T\beta + z_i, \tag{1}$$

which is also known as a variance component model.

The Box–Cox transformation has been widely used in applied data analysis. The objective of the transformation is to select an appropriate parameter λ which is then used to transform the responses such that they follow a normal distribution more closely than the untransformed data. Under the scenario of model (1), the transformation by Box and Cox (1964) can be written as

$$y_{ij}^{(\lambda)} = \begin{cases} \frac{y_{ij}^{\lambda}-1}{\lambda} & \lambda \neq 0, \\ \log y_{ij} & \lambda = 0 \end{cases}$$

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

for $y_{ij} > 0$, $i = 1, ..., r$ and $j = 1, ..., n_i$. It is assumed that there is a value of λ for which

$$y_{ij}^{(\lambda)} | z_i \sim N(x_{ij}^T \beta + z_i, \sigma^2)$$

where z_i is a random effect with an unspecified mixing distribution $g(z_i)$. Taking account of the Jacobian of the transformation from y_{ij} to $y_{ij}^{(\lambda)}$, the conditional density function of y_{ij} given z_i is

$$f(y_{ij}|z_i) = \frac{y_{ij}^{\lambda-1}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_{ij}^{(\lambda)} - x_i^T \beta - z_i)^2\right].$$

The goal is to estimate λ and β under the presence of the random effect.

2 Estimation

Under the NPML estimation approach, the distribution of the random effect will be approximated by a discrete distribution at mass points $z_1, \dots, z_K$, which can be considered as intercepts for the different unknown subgroups on the upper level. Hence, the likelihood in relation to the original observations can be approximated as (Aitkin et al, 2009)

$$L(\lambda, \beta, \sigma^2, g) = \prod_{i=1}^{r} \int \left[\prod_{j=1}^{n_i} f(y_{ij}|z_i)\right] g(z_i) dz_i \approx \prod_{i=1}^{r} \sum_{k=1}^{K} \pi_k m_{ik}, \qquad (2)$$

where $m_{ik} = \prod_{j=1}^{n_i} f(y_{ij}|z_k)$. Defining indicators $G_{ik} = 1$ if case i stems from cluster k and 0 otherwise, the complete log–likelihood takes the shape

$$\ell^* = \log L^* = \sum_{i=1}^{r} \sum_{k=1}^{K} [G_{ik} \log \pi_k + G_{ik} \log m_{ik}]$$

where $L^* = \prod_{i=1}^{r} \prod_{k=1}^{K} (\pi_k m_{ik})^{G_{ik}}$. Of course, ℓ^* depends on λ. For fixed λ, one proceeds via a standard EM algorithm, where in the E-step expectations of G_{ik} are obtained via $w_{ik} = \frac{\pi_k m_{ik}}{\sum_\ell \pi_\ell m_{i\ell}}$, and in the M-step the expected complete likelihood is maximized, yielding

$$\begin{aligned}
\hat{\beta}^{(\lambda)} &= \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} x_{ij}^T\right)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} \left(y_{ij}^{(\lambda)} - \sum_{k=1}^{K} w_{ik} \hat{z}_k\right), \\
\hat{\sigma}^{2(\lambda)} &= \frac{\sum_{i=1}^{r} \sum_{k=1}^{K} w_{ik} \left[\sum_{j=1}^{n_i} (y_{ij}^{(\lambda)} - x_{ij}^T \hat{\beta} - \hat{z}_k)^2\right]}{\sum_{i=1}^{r} n_i}, \\
\hat{z}_k^{(\lambda)} &= \frac{\sum_{i=1}^{r} w_{ik} \left[\sum_{j=1}^{n_i} (y_{ij}^{(\lambda)} - x_{ij}^T \hat{\beta})\right]}{\sum_{i=1}^{r} n_i w_{ik}}, \quad \hat{\pi}_k^{(\lambda)} = \frac{\sum_{i=1}^{r} w_{ik}}{r}.
\end{aligned}$$

Replacing the results into m_{ik} and then into the right–hand term of equation (2) we get the non–parametric profile likelihood function $L_P(\lambda)$, or its logarithmic version $\ell_P(\lambda) = \log(L_P(\lambda))$. The non–parametric profile maximum likelihood (NPPML) estimator is therefore given by

$$\hat{\lambda} = arg \max_{\lambda} \ell_P(\lambda),$$

which can be found through a grid search over λ.

3 Example: Oxboys data

In order to demonstrate this methodology, we consider a data set available as part of the **R** package **nlme** (Pinheiro et al, 2017), which consists of measurements of `age` and `height` for 26 boys in Oxford, yielding a total of 234 observations. The response variable `height` is defined as the height of the boy in (cm), associated with the covariate `age` that is the standardized age (dimensionless). We fitted a variance component model

$$E(y_{ij}|z_i) = \texttt{age}_j + z_i$$

where z_i is boy–specific random effect and $\texttt{age}_j$ is the j-th standardized age measurement, $j = 1, \ldots, 9$, which is equal for all boys for fixed j.

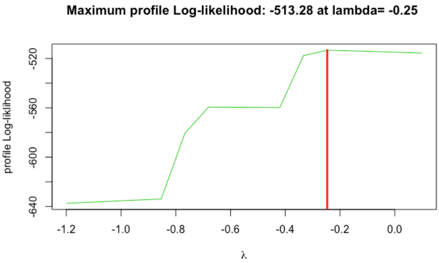


FIGURE 1. For the Oxboys data, a grid search over λ, with $K = 6$.

From Figure 1, it can be seen that the best estimate of λ that maximizes $\ell_P(\lambda)$ is $\hat{\lambda} = -0.25$, suggesting that some transformation need to be carried out to make the response distribution more normal. The results before and after applying the response transformation are summarized in Table 1. As can be seen from this table, comparing the Akaike Information Criterion (AIC) values of the untransformed model fit ($\lambda = 1$) and our method using $K = 5, 6$ and 7, respectively, showed a better performance of the NPPML approach. In other words, using the response after applying the transformation leads to a better fitting model than the original data.

TABLE 1. Comparison of results from original & transformed data, using $K = 5, 6$ and 7

	$K=5$		$K=6$		$K=7$	
	$\hat{\lambda} = -0.51$	$\lambda = 1$	$\hat{\lambda} = -0.25$	$\lambda = 1$	$\hat{\lambda} = -0.25$	$\lambda = 1$
$-2\log L$	1119.3	1132.8	1026.2	1048.3	1024.2	1132.8
AIC	1141.3	1154.9	1052.2	1074.3	1054.2	1162.9

4 Simulation Study

We are interested in examining the method's ability to estimate the true parameter values. Therefore, we first simulate data by applying the Box–Cox transformation 'backwards' to a dataset that follows a normal distribution using a set of λ values. Specifically, for each of four given values λ_ℓ, $\ell = 1, 2, 3, 4$, we generate 1000 datasets with 100 observations as follows,

$$\begin{aligned}
\zeta_{ij\ell} &= \tilde{y}(\eta_{ij}, \lambda_\ell),\ \ i = 1, ..., 20,\ j = 1, ..., 5 \qquad (3)\\
\tilde{y}(\eta_{ij}, \lambda_\ell) &= \begin{cases} \left(1 + \lambda_\ell \eta_{ij}\right)^{\lambda_\ell} & (\lambda_\ell \neq 0), \\ e^{\eta_{ij}} & (\lambda_\ell = 0) \end{cases}\\
\eta_{ij} &= 3\,x_{ij} + z_i + \varepsilon_{ij}\\
x_{ij} &\sim U(-4, 4)\\
\varepsilon_{ij} &\sim N(0, 0.5)\\
\lambda_1 = 0,\ \ \lambda_2 &= 0.5,\ \ \lambda_3 = 1,\ \ \lambda_4 = 2.
\end{aligned}$$

Note that $\tilde{y}(\cdot)$ denotes the 'backward' Box–Cox–transformation, and that the generated data possess a variance component structure due to the random effect terms z_i, which are generated by a discrete distribution with mass-points $z_k = (15, 20, 30, 35)$ and masses $\pi_k = 1/k, k = 1, ..., 4$.

In the estimation step, we estimate λ and β simultaneously, yielding for each (true) value of λ a total of 1000 estimates of $\hat{\lambda}$ and $\hat{\beta}$. Figure 2 shows the boxplots for the regression and transformation parameter estimates, respectively. The reference lines in the figures indicate the actual values of the parameters. It is clear that the median of the estimated β and λ is approximately equal to the true value in each plot. There are some outliers in each of the plots; in fact the outliers in the transformation estimates cause the outliers in the regression estimates as they shift the scale of the linear predictor. The means and medians of the estimated β and λ parameters are also provided in Table 2; we see that the medians for the transformation parameters sit exactly at their true values, and those of the regression parameters approximately so.

We also investigate the standard errors of the regression parameter estimates. An empirical but robust measure of spread of the estimated β can be obtained by computing the IQR of (the non–logarithmic version of) each of the four columns in Figure 2 (top). Via normal reference, the IQR can be mapped back to the scale of the standard deviations by division through 1.349. We call the resulting robust estimate of standard deviation RESD($\hat{\beta}$). Table 2 displays RESD($\hat{\beta}$) values along with means and medians of EM–based standard errors, $SE(\hat{\beta})$, which were obtained by extraction from the model fitted in the last M–step. It is conceptually clear that such EM–based standard errors cannot be 'correct' as they ignore the variation caused by the EM algorithm itself, but we see from Table 2 that they are still satisfyingly close to their empirical counterparts.

TABLE 2. Summary of simulation results.

True values	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
β	3	3	3	3
Mean($\hat{\lambda}$)	0	0.5026	1.003	2.0049
Median($\hat{\lambda}$)	0	0.5	1	2
Mean($\hat{\beta}$)	2.9996	3.0901	3.0770	3.1090
Median($\hat{\beta}$)	3.0003	3.0001	3.0003	3.0006
RESD($\hat{\beta}$)	0.0246	0.0251	0.0255	0.0335
Mean($\hat{SE}(\hat{\beta})$)	0.0256	0.0267	0.0264	0.0268
Median($\hat{SE}(\hat{\beta})$)	0.0214	0.0214	0.0214	0.0214

5 Implementation

The methodology is implemented in **R** package **boxcoxmix** (Almohaimeed and Einbeck, 2017) which is available on CRAN. This package features further variants and capabilities which have not been introduced here, such as a version for simple 'overdispersion' models (where $n_i \equiv 1$), and several routines to select the starting points for the EM algorithm.

Acknowledgments: The first author is grateful to Qassim University for financial support.

References

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). *Statistical Modelling in R.* Oxford: University Press.

Almohaimeed, A. and Einbeck, J. (2017). boxcoxmix: Response transformations for random effect and variance component models. URL `https://CRAN.R-project.org/package=boxcoxmix`

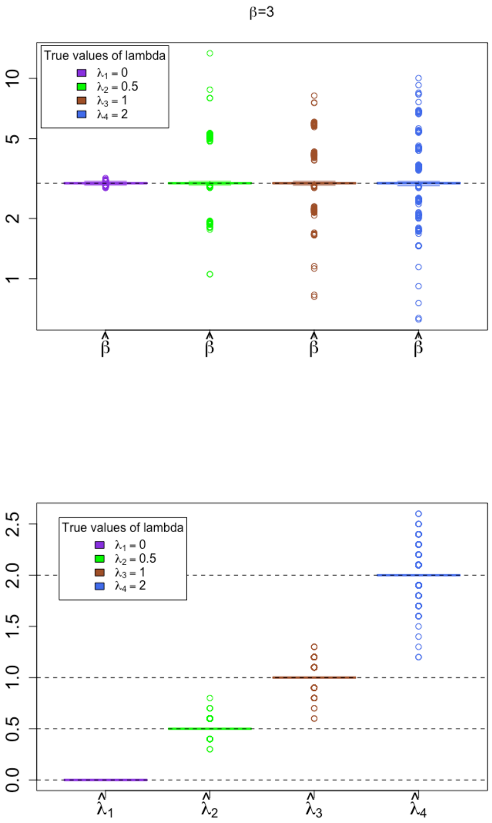


FIGURE 2. Simulation results: Estimates $\hat{\beta}$ (top) and $\hat{\lambda}$ (bottom), in each plot for true $\lambda_\ell = 0, 0.5, 1, 2$ (from left to right). The vertical axis in the upper plot is given on log–scale. Horizontal lines indicate the true values.

Box, G.E. and Cox, D. (1964). *Journal of the Royal Statistical Society. Series B*, **26**, 211 – 252.

Pinheiro, D. et al (2017). nlme: Linear and nonlinear mixed effect models. `https://CRAN.R-project.org/package=nlme`

Fitting a spatial-temporal rainfall model using Approximate Bayesian Computation

Nanda Aryal[1], Owen D. Jones[2]

[1] University of Melbourne, Australia
[2] Cardiff University, UK

E-mail for correspondence: `joneso18@cardiff.ac.uk`

Abstract: We fit a stochastic spatial-temporal model to high-resolution rainfall radar data for a single rainfall event. Approximate Bayesian Computation (ABC) is used to fit a model of Cox, Isham and Northrop, previously fitted using the Generalised Method of Moments (GMM). We then show that ABC readily adapts to more general, and thus more realistic, variants of the model. The Simulated Method of Moments (SMM) is used to initialise the ABC fit.

Keywords: Spatial-temporal; spatiotemporal; rainfall; Approximate Bayesian Computation.

1 Introduction

The Cox-Isham-Northrop (C-I-N) rainfall model is a spatial-temporal stochastic model for a rainfall event, constructed using a cluster point process. The cluster process is constructed by taking a primary process, called the storm arrival process, and then attaching to each storm center a finite secondary point process, called a cell process. To each cell center we then attach a rain cell, with an associated area, duration and intensity. The storm and cell centers all share a common velocity. The total rainfall intensity at point (x, y) and time t is then the sum of the intensity at (x, y) of all cells active at time t. (Cox & Isham 1988, Northrop 1998.)

The storm arrival process is taken to be a Poisson process in $\mathbb{R}^2 \times [0, \infty)$ with homogeneous rate λ. Let $\mathbf{v} = (v_x, v_y)$ be the velocity of the rainfall event, so if a storm center arrives at $(\mathbf{u}, s)$ then at time $s + t$ it will be at $(\mathbf{u} + t\mathbf{v}, s + t)$. Storm durations are random with an $\exp(\gamma)$ distribution. While a storm is active it produces cells at a rate β in time, starting with a cell at the moment the storm center begins. If the storm arrives at $(\mathbf{u}, s)$

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

and produces a cell at time $s+t$, the cell will be centered at $\mathbf{u}+t\mathbf{v}+\mathbf{w}$, where $\mathbf{w}$ comes from a Gaussian distribution with mean $\mathbf{0}$ and covariance Σ. The cell centre then also moves with velocity $\mathbf{v}$. We parameterise Σ using its size σ^2, eccentricity e and orientation Θ. σ^2 is assumed to have an inverse-gamma distribution, where $1/\sigma^2$ has mean ξ_μ and coefficient of variation ξ_{CV}.

Individual cells have random durations, distributed as $\exp(\eta)$, and random sizes. Rain cells are elliptical, with the same eccentricity e and orientation Θ as the storms. The size is given by the major axis, which is distributed as a $\Gamma(\alpha_1, \alpha_2)$. Note that we can re-express α_2 using α_1, e, and μ_A (the mean area of a rain cell). The intensity of a rain cell is constant over the shape and duration of the cell, with an exponential distribution mean μ_X. The displacements, durations, sizes, and intensities of a cell are all independent, and independent of other cells.

The C-I-N model is stationary and is used to model the 'interior' of a rainfall event. We will suppose that we have observations of the rainfall in some finite space-time window $A \times [0,T]$, where T is chosen so that the leading and trailing edges of the rainfall event are not observed. For this study we used radar data collected at Laverton, Melbourne, on 24th September 2016, calibrated by the Australian Bureau of Meteorology using rain-gauge data. The data gives rainfall intensity averaged over 1 km square pixels every 6 minutes, over an area of 180×180 pixels for a period of 3 hours. A contour plot of the spatial rainfall intensity at a single time-point is given in Figure 1(a).

Because it has an intractible likelihood function, the C-I-N model has been fitted using the Generalized Method of Moments (GMM) (Wheater et al. 2006). The puprose of this paper is firstly to show that Approximate Bayesian Computation (ABC) can be used to fit a Bayesian version the C-I-N model, and secondly to use ABC to fit a generalisation of the C-I-N model that is too much for GMM to cope with. GMM fitting matches theoretical and observed moments of the process, and thus is restricted to moments for which you have an analytic expression. ABC fitting compares the observed process to simulations, and thus places no restrictions on the statistics used to compare them. The penalty we pay for this increased flexibility is an increase in computational time.

1.1 Approximate Bayesian Computation

ABC was introduced by Pritchard et al. (1999), and was later extended to incorporate Markov Chain Monte Carlo (MCMC) by Marjoram et al. (2003), or alternatively Sequential Monte Carlo (SMC) (Sisson et al. 2007 and 2009, Beaumont et al. 2009). We will use the ABC-MCMC methodology.

We suppose that we have an observation D from some model $f(\cdot|\boldsymbol{\theta})$, depending on parameters $\boldsymbol{\theta}$, and that we are able to simulate from f. Let π

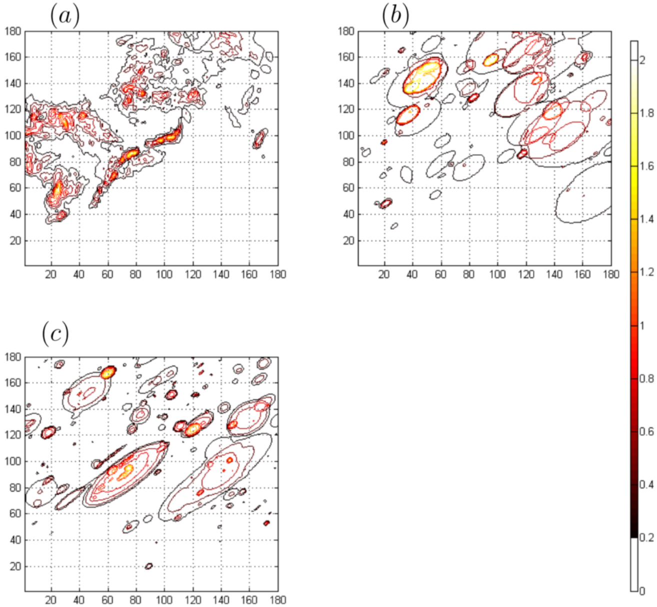


FIGURE 1. (a) Calibrated rainfall radar data, courtesy of the Australian Bureau of Meteorology. (b) Simulation from the C-I-N model fitted using ABC. (c) Simulation from the modified C-I-N model.

be the prior distribution for $\boldsymbol{\theta}$ and $S = S(D)$ a vector of summary statistics for D, then ABC generates samples from $f(\boldsymbol{\theta}|\rho(S(D^*), S(D)) < \epsilon)$, where $D^* \sim f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \pi$, and ρ is some distance function. If S is a sufficient statistic, then as $\epsilon \to 0$ this will converge to the posterior $f(\boldsymbol{\theta}|D)$. ABC-MCMC adds a proposal chain with density q and a rejection step, to generate a sample $\{\boldsymbol{\theta}_i\}$. The algorithm is as follows:

```
FOR $i = 1$ to $N$
    1 Given current state $\boldsymbol{\theta}_i$ propose a new state $\boldsymbol{\theta}^*$ using $q(\cdot|\boldsymbol{\theta}_i)$
    2 Put $\alpha = \min\{1, (\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*))/(\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i))\}$
    3 Go to 4 with probability $\alpha$, otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ and return to 1
    4 Simulate data $D^* \sim f(\cdot|\boldsymbol{\theta}^*)$
    5 If $\rho(S(D^*), S(D)) \le \epsilon$ then set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$
END FOR
```

Note that the MCMC rejection at step 3 comes before the ABC comparison in step 5. This is to avoid unnecessarily running the simulation in step 4.

2 Fitting the C-I-N model using ABC

Following Wheater et al. (2006), the velocity $\mathbf{v}$, eccentricity e and orientation Θ were all estimated ad hoc using temporal and spatial autocovariance estimates, and then fixed.
The remaining parameters were transformed to reduce dependence and skewness. For the ABC step we used $\log(\lambda/\gamma)$, $\log(\lambda\gamma)$, $\log(\beta/\eta)$, $\log(\beta\eta)$, $\log(\mu_X/\mu_A)$, $\log(\mu_X\mu_A)$, $\log(\alpha_1)$, $\log(\xi_\mu)$, and $\log(\xi_{CV})$. Vague normal priors are used for all the transformed parameters, and for the proposal chain we used a random walk with $N(0, 0.2^2 I)$ steps.
The choice of summary statistics S and distance metric ρ plays a large part in the performance of ABC. Ideally S should be sufficient, but certainly it should reflect those aspects of the real process considered most important. However choosing S too large reduces the efficiency of ABC, though this can be mitigated to some extent using post hoc analysis of the significance of each component (Beaumont et al. 2002).
We used 23 summary statistics:

- The overall mean and standard deviation of rainfall, taken over all pixels and all times.
- The spatial-temporal auto-correlation, with lags of (x, y, t), where x and y are measured in pixels and t is in units of 6-minutes. We take $t = 0$, $x \in \{-1, 0, 1\}$, $y \in \{-1, 0, 1\}$, and $t = 1$, $x \in \{-1, 0, 1\} + v_x$, $y \in \{-1, 0, 1\} + v_y$. Here v_x and v_y are the velocity components, in units of pixels per 6-minutes. Note that the lag $(0, 0, 0)$ auto-correlation is not used because it is just variance.
- The probability of an arbitrary pixel and time being dry.
- The ratio of dry/wet area and mean and standard deviation of wet area, averaged over time.

For the distance function ρ we used a weighted sum of squares $\rho(S(D^*), S(D)) = \sum_i w_i(S^*(i) - S(i))^2$, where $S^*(i)$ and $S(i)$ are the i-th components of $S(D^*)$ and $S(D)$ respectively. We found empirically, as have other authors, that a good choice is to take w_i inversely proportional to the variance of $S^*(i)$ under the posterior.
Plots of our fitted posteriors are given in Figure 2, and a simulation from the fitted model is given in Figure 1(b), for a single time-point. For the simulation the posterior means were used as point estimates for the parameter values.

2.1 Starting ABC using SMM

The Simulated Method of Moments (SMM) is a variant of the Generalised Method of Moments (GMM) that uses Monte-Carlo estimates of moments, rather than analytic expressions (McFadden 1989). Thus, like ABC, using

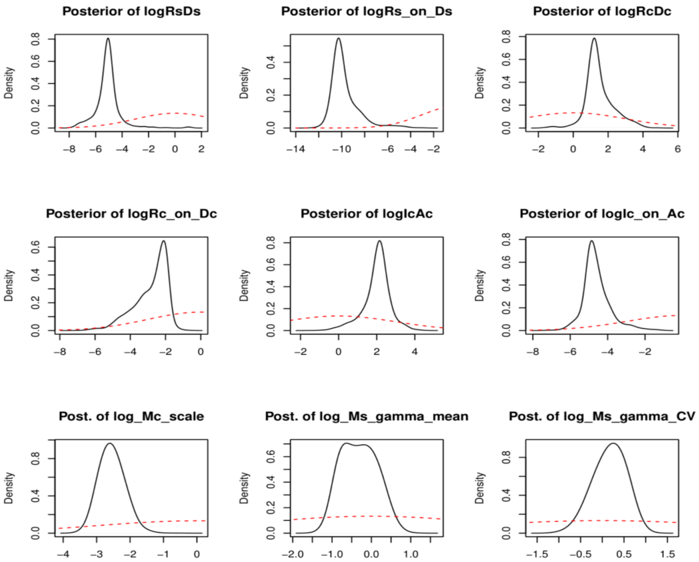


FIGURE 2. Some nice looking posteriors. Priors are given by the red dashed lines. See the text for details.

SMM we have much more freedom in the choice of moments used to fit the model to the data.

When applying ABC, we found it advantageous to 'jump-start' the algorithm by choosing the initial parameter selection $\boldsymbol{\theta}_0$ using an SMM fit, using the same summary statistics S that we chose for the ABC fitting. There are two main benefits to this step. The first is that if $\boldsymbol{\theta}_0$ has very small posterior probability, then ABC-MCMC requires a prohibitively large burn-in period. The second is that it gives us a distribution for $S(D^*)$ that we can use to estimate the weights w_i of the distance function ρ.

Previous authors have suggested using a separate ABC step to estimate $\boldsymbol{\theta}$; we found that using SMM instead requires much less computation time.

3 Extending the C-I-N model

There are many ways in which the C-I-N model can be extended. If you do so, however, GMM is no longer suitable for estimation, as it becomes too difficult to obtain analytic expressions for the moments. Fortunately this does not apply to ABC, which can be applied much as before. For the example below we did not even have to modify the set of summary statistics S.

We leave a comprehensive generalisation of the C-I-N model to future work, and just consider the following modifications:

- Randomised cell eccentricity.
- Rainfall intensity that increases continuously from the edge to the centre of each cell, rather than acting as a step function.
- Heavy tailed distributions for cell intensity and area.
- Correlated cell intensity and area.

Using the posterior distribution of $S(D^*)$ we can show that the modified model gives a better fit. A simulation from the fitted model is given in Figure 1(c), for a single time-point; qualitatively it also looks to be doing a better job

References

Beaumont, M.A., J.-M. Cornuet, J.-M. Marin, and C.P. Robert (2009). Adaptive approximate Bayesian computation. *Biometrika 96*(4):983–990.

Beaumont, M.A., W. Zhang, and D.J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*(4):2025–2035.

Cox, D.R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proc. Roy. Soc. A*, 415(1849):317–328

Northrop, P. (1998). A clustered spatial-temporal model of rainfall. *Proc. Roy. Soc. A*, 454(1975):1875–1888.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States 100*(26):15324–15328.

Pritchard, J.K., M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*:1791–1798.

Sisson, S., Y. Fan, and M.M. Tanaka (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States 104*:1760–1765.

Sisson, S., Y. Fan, and M.M. Tanaka (2009). Correction for "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Sciences of the United States 106*:16889.

Wheater, H.S., V.S. Isham, R.E. Chandler, C.J. Onof and E.J. Stewart (2006). Improved methods for national spatial-temporal rainfall and evaporation modelling for BSM. Joint Defra/EA Flood and Coastal Erosion Risk Management R&D Programme, Technical Report F2105/TR.

Study of the convergence of Morris extension method for evaluating the combined influences of model inputs

Majdi Awad [1], Tristan Senga Kiesse[2], Zainab Assaghir [3], Anne Ventura [4]

[1] University of Nantes, GeM, Institute of Research in civil engineering and Mechanics-CNRS UMR 6183-Chair civil engineering eco-construction, France
[2] UMR SAS, INRA, AGROCAMPUS OUEST, 35000 Rennes, France
[3] Lebanese University, Faculty of sciences, Beirut, Lebanon
[4] Bretagne Loire University (UBL), French Institute of Sciences and Technical Transports Networking (IFSTTAR / MAST /GPEM), France

E-mail for correspondence: Majdi.awad@etu.univ-nantes.fr

Abstract: This work concerns the Morris' extension method to evaluate the influence of combined variations of a pair of model inputs at a low computational cost. There is a lack of studies on this method concerning the crucial choice of the adequate number of trajectories to obtain stable results to distinguish groups of influential and non-influential pairs of parameters, and to rank the pairs of parameters according to their relative importance. The Morris' extension method is studied according to the previous issues via an application on a complex model from civil engineering. In addition, the median of mixed elementary effects (MEE) is calculted as a robust statistic to calculate sensitivity indices, in comparison with the classical mean of MEE. Results showed that the sensitivity indices based on the median of MEE median were more appropriate than those based on the mean of MEE, to obtain stable results of the influence of pairs of inputs. Moreover, results on the relative importance of pairs of inputs according to their combined influences are similar to those obtained when using total interaction indices of Sobol.

Keywords: Sensitivity analysis; Combined action; Input pairs; Screening; Ranking.

1 Introduction

This work is interested in Morris' extension method (Campolongo and Braddock(1999)) to select and classify the influential pairs of model inputs at low computational cost. The study of the influence of the combined variation of parameters is important in various fields, especially in civil engineering. For instance, the Morris' extension method is implemented in a study of building energy models, to evaluate the relative influence of pairs of model inputs that reflect the coupling between phenomena such as occupation, microclimate and envelope of building (Menberg et al. (2016), Sanchez et al. (2014)). However, there is a lack of studies on the crucial choice of the adequate number of trajectories to obtain stable results to (i) distinguish groups of influential and non-influential pairs of model inputs and (ii) rank the pairs of parameters according to their relative important influence. This work is a contribution to the assessment of the robustness of results of the Morris' extension method through an illustration on a complex civil engineering model. A comparison is made with total interaction indices of Sobol ((Fruth et al. (2014)), which are useful to determine and quantify the total contribution of the influence of combined variations of parameters (including interactions of order > 2). The stability of screening and ranking results provided by the Morris' extension method is studied by simulating indices based on the median of mixed elementary effects (MEE), compared with classical indices based on the mean of MEE.

2 Sensitivity analysis methods

Morris' extension method. Consider a given value of the input vector x of the input parameter space Q_n and the model $y = f(x)$. The influence of combined variations of the inputs X_i and X_j on the output y is studied by calculating the following partial derivative:

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = dd_{ij}(x) = EE_{ij} - \frac{1}{\Delta_i} E_i - \frac{1}{\Delta_j} E_j,$$

with $E_i = \partial f(x)/\partial x_i$ and $EE_{ij}(x) = [f(x + e_i\Delta_i + e_j\Delta_j) - f(x)]/\Delta_i\Delta_j$, $1 \leq i \leq j \leq n$, where $\Delta = (\Delta_1, \ldots, \Delta_n)$ is a predetermined vector such that $x + e_i\Delta_i + e_j\Delta_j \in Q_n$ and $f(x + e_i\Delta_i + e_j\Delta_j) = f(x_1, \ldots, x_{i-1}, x_i + \Delta_i, x_i + 1, \ldots, x_{j-1}, x_j + \Delta_j, x_{j+1}, \ldots, x_n)$. The MEE dd_{ij} are calculated using the sampling strategy based on the *Handcuffed prisoners* which aims to extract a sample of r elements $dd_{ij}^1, \ldots, dd_{ij}^r$, for each input pair, in order to estimate descriptive statistics of MEE (Campolongo and Braddock(1999)). Then, the following indices are calculated: the median γ_{ij}^* of absolute values $|dd_{ij}(l)|$ and their standard deviation σ_γ in relation to γ_{ij}^*. A comparison is made with classical indices such as the mean $\mu_{ij}^* = (1/r)\sum_{l=1}^r |dd_{ij}^l|$ and standard deviation $\sigma_\mu = \sqrt{(1/r)\sum_{l=1}^r (dd_{ij}^l - \mu_{ij}^*)^2}$, where r is the number of trajectories. The indices γ^*_{ij} and μ^*_{ij} provide information on relative importance of combined influence of a pair of inputs, while σ_μ and σ_γ indicate non-bilinear effect and/or interaction of order ≥ 3.

Total interaction indices of Sobol. The Total Interaction Index (TII) are based on the decomposition of variance of the model output (Fruth et al. (2014)). They are estimated using the classical Monte Carlo sampling method The TII of a pair of inputs (X_i, X_j) is defined as: $V^{super}_{ij} = \sum_{I \supseteq \{i,j\}} V_I$, with $I \subseteq \{1, \ldots, n\}$.

3 Application

This application aims to robustly identify the joint influence of environmental and technological factors on service life of concrete structures using a meta-model of carbonation (detailed in Ta et al. (2016)).

Corrosion model. The studied model calculates the carbonation depth x_{CO_2} in concrete structures as follows:

$$x_{CO_2} = A\sqrt{t}, \tag{1}$$

with the exposure time t and the carbonation coefficient defined as a function of environmental parameters (relative external humidity RH, ambient temperature T, CO_2-concentration in the air) and technological parameters (cement content C, water to cement ratio W/C, sand to gravel ratio S/G, maximum aggregate size S_{max}, cement type CEM, cement strength class f_{cem}, concrete cover depth d, initial curing period t_c). The service life $t_{ser} = d^2/A^2$ of concrete structures is reached when the carbonation depth x_{CO_2} is equal to the concrete cover depth d in Eq.1.

Results. The Morris' extension method and the total interaction indices of Sobol were applied by incrementally increasing the number of trajectories from 100 to 4600 and size of Monte-Carlo samples from 100 to 9100, respectively (Figure 1).

The influence of pair $(T, W/C)$ was ranked first showing the combined action of environmental and technological factors on the service life of concrete strcutures. The mean μ^*_{ij} enabled to distinguish groups of influential and non-influential input pairs (Figure 1(b)), while the median γ^*_{ij} enabled to prioritize the first most influential input pairs (Figure 1 (c)). The two sensitivity indices globally revealed the same most influential input pairs but not the same ranking of their influence. The number of trajectories r_{req} required to obtain stable results of the sensitivity indices was higher when using the mean μ^*_{ij} ($r_{req} \geq 4600$, Figure 1 (b)) than the median γ^*_{ij} ($r_{req} \approx$ 2600 trajectories, Figure 1 (c)), reflecting the statistical properties of the median as a more robust indicator. Finally, (i) the indice μ^*_{ij} provided also information on the total relative importance of input pairs, through similar ranking of influential input pairs than the Sobol total interaction indices (Figure 1 (d)), and (ii) non-bilinear effects and / or combined actions of

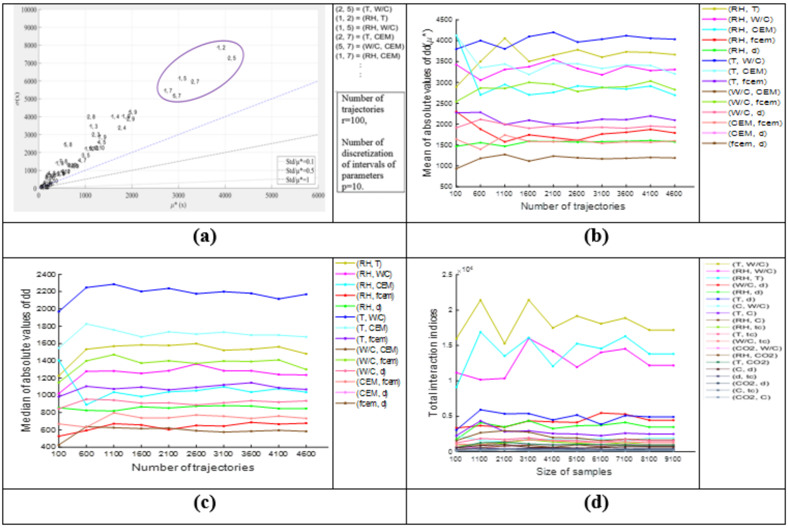


FIGURE 1. Indices μ^*_{ij} $(a)(b)$ and γ^*_{ij} (c) of Morris' extension method and total interaction indices of Sobol (d), for input pairs of corrosion model (Ta et al.(2016))

order > 2 were indicated by the calculation of ratio σ_{ij}/μ^*_{ij} (> 0.5) for input pairs (Figure 1(a)).

References

Campolongo, F. and Braddock, R. (1999). *The use of graph theory in the sensitivity analysis of the model output: a second order screening method*. Australia: Reliability Engineering & System Safety.

Fruth, J. and Roustant, O. and Kuhnt, S. (2014). *Total interaction index: A variance-based sensitivity index for second-order interaction screening*. Germany: Journal of Statistical Planning and Inference.

Menberg, K. and Heo, Y. and Choudhary, R. (2016). *Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information*. Cambridge: Energy and Building.

Sanchez, D.G. and Lacarrire, B. and Musy, M. and Bourges, B (2014). *Application of sensitivity analysis in building energy simulations: Combining first-and second-order elementary effects methods*. France: Energy and Buildings.

Ta, V-L. and Bonnet, S. and Senga Kiesse, T. and Ventura, A. (2016). *A new meta-model to calculate carbonation front depth within concrete structures*. France: Construction and Building Materials.

A Three-Component Lee-Carter approach to decompose and forecast human mortality

Ugofilippo Basellini[1,2], Carlo G. Camarda[1]

[1] Institut national d'études démographiques (INED), Paris, France
[2] Department of Public Health, University of Southern Denmark, Odense, Denmark

E-mail for correspondence: `ugofilippo.basellini@ined.fr`

Abstract: The Lee-Carter model is an elegant and powerful methodology to model and forecast mortality based on a log-bilinear form for the hazard function. We propose a novel extension of the model that overcomes its drawback of a fixed age-specific rate of mortality improvements. This new approach improves goodness-of-fit and offers an innovative perspective in terms of forecasting by decomposing mortality into childhood, early-adulthood and senescent components. The three components are estimated via an iterative series of penalized composite link models. We illustrate the approach on Swiss male mortality data.

Keywords: Mortality forecasting; Mortality decomposition; Composite Link Model; Additive components; Penalized likelihood.

1 Introduction

Mortality forecasting has received growing attention in recent decades as worldwide ageing populations pose increasing challenges to the sustainability of social security and public health policies. In 1992, Lee and Carter proposed a seminal methodology to model and forecast mortality hazard rates, which nowadays is presumably the best known mortality forecasting procedure. Nonetheless, the model has some limitations, the main one being the assumption of a fixed age-specific rate of mortality improvement. In the following, we propose a novel extension of the Lee-Carter model that overcomes this issue by decomposing the hazard of mortality into three components that operates principally upon childhood, middle and old ages, respectively. We illustrate the proposed approach on Swiss male mortality data from the Human Mortality Database (2018).

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 The Three-Component Lee-Carter model

We suppose that we have mortality data, deaths, and exposures to the risk of death, arranged in two matrices, $\boldsymbol{Y} = (y_{ij})$ and $\boldsymbol{E} = (e_{ij})$, each $m \times n$. We assume that the number of deaths y_{ij} at age i in year j follows a Poisson distribution with mean $\mu_{ij} = e_{ij}\,\theta_{ij}$, i.e. product of exposures and actual force of mortality or hazard rate (Brillinger, 1986). In matrix notation, the Lee-Carter model assumes that the log of the force of mortality, $\boldsymbol{\Theta} = (\theta_{ij})$, is given by:

$$\ln \boldsymbol{\Theta} = \boldsymbol{\alpha}\,\mathbf{1}_n^{\mathrm{T}} + \boldsymbol{\beta}\,\boldsymbol{\kappa}^{\mathrm{T}} \tag{1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are m-dimensional vectors describing the average shape of the age profile and the rate of mortality improvement at age i, respectively, and $\boldsymbol{\kappa}$ is a n-dimensional vector capturing the general level of mortality at each year j.

The assumption of a single shape of mortality improvement over age described by $\boldsymbol{\beta}$ can be often strong in mortality data. To overcome this issue, we decompose mortality into three components that operates principally upon childhood, middle and old ages. Each component will be described by a distinct Lee-Carter model.

We arrange the matrix of deaths by column order into a vector $\boldsymbol{y}$. Likewise we arrange the matrix of exposures $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$. Expected values $\boldsymbol{\mu}$ can be written as $\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}^{\mathrm{T}} = (\boldsymbol{\gamma}_C, \boldsymbol{\gamma}_A, \boldsymbol{\gamma}_S)^{\mathrm{T}}$, representing childhood, early-adulthood and senescent mortality, respectively. The matrix $\boldsymbol{C}$ additively combines the $\boldsymbol{\gamma}_i$ and also incorporates the exposures:

$$\boldsymbol{C} = \mathbf{1}_{1,3} \otimes \mathtt{diag}(\boldsymbol{e})\,. \tag{2}$$

Each $\boldsymbol{\gamma}_i \in \mathbb{R}_+^{mn}$ is defined as a component-specific Lee-Carter:

$$\ln \boldsymbol{\gamma}_i = \mathtt{vec}(\boldsymbol{\alpha}_i\,\mathbf{1}_n^{\mathrm{T}} + \boldsymbol{\beta}_i\,\boldsymbol{\kappa}_i^{\mathrm{T}}) \quad \text{for } i = C, A, S\,. \tag{3}$$

Hence we call this a Three-Component Lee-Carter (3C-LC) model. Moreover, to avoid irregular patterns in fitted and projected life tables, we enforce smoothness in the shape of $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\kappa}_i$.

3 A Composite Link Model approach

The 3C-LC model can be viewed as a Composite Link Model (CLM, Thompson and Baker, 1981) and as a special case of the Sum of Smooth Exponentials model (SSE, Camarda et al., 2016), in which each component follows the parametric Lee-Carter assumption. Unlike the SSE, each component of the 3C-LC model is not linear with respect to all unknown parameters. Nevertheless, we can linearize the system of equations with respect to each series of parameters and iteratively solve the following penalized IWLS algorithm (Eilers, 2007):

$$(\breve{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}} + \boldsymbol{P}_{\boldsymbol{a}})\tilde{\boldsymbol{a}} = \breve{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\,, \tag{4}$$

where $\breve{\boldsymbol{X}} = \boldsymbol{C}_{\boldsymbol{a}}\,\frac{\tilde{\gamma}_{\boldsymbol{a}}}{\tilde{\boldsymbol{\mu}}}\,\boldsymbol{X}_{\boldsymbol{a}}$, $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \tilde{\boldsymbol{\eta}}_{\boldsymbol{a}}$.
We start by fixing all triplets of parameters $\boldsymbol{\kappa}_i$ and $\boldsymbol{\beta}_i$ and estimate $\boldsymbol{a}^{\mathrm{T}} = (\boldsymbol{\alpha}_C, \boldsymbol{\alpha}_A, \boldsymbol{\alpha}_S)^{\mathrm{T}}$. Composite and design matrices for solving (4) are thus given by:

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{\alpha}} &= [\boldsymbol{d}_C(\boldsymbol{1}_n \otimes \boldsymbol{I}_m) : \boldsymbol{d}_A(\boldsymbol{1}_n \otimes \boldsymbol{I}_m) : \boldsymbol{d}_S(\boldsymbol{1}_n \otimes \boldsymbol{I}_m)] \\ \boldsymbol{X}_{\boldsymbol{\alpha}} &= \boldsymbol{I}_{3m} \end{aligned} \tag{5}$$

where $\boldsymbol{d}_i = \mathtt{diag}[\boldsymbol{e}\,\mathtt{vec}(\exp(\boldsymbol{\beta}_i\,\boldsymbol{\kappa}_i^T))]$ for $i = C, A, S$.
Analogously, we then fix the other triplets of parameters and we solve (4) by changing composite and design matrices:

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{\beta}} &= \boldsymbol{C}_{\boldsymbol{\kappa}} = [\boldsymbol{u}_C : \boldsymbol{u}_A : \boldsymbol{u}_S] \\ \boldsymbol{X}_{\boldsymbol{\beta}} &= \mathtt{diag}[\boldsymbol{\kappa}_C \otimes \boldsymbol{I}_m,\, \boldsymbol{\kappa}_A \otimes \boldsymbol{I}_m,\, \boldsymbol{\kappa}_S \otimes \boldsymbol{I}_m] \\ \boldsymbol{X}_{\boldsymbol{\kappa}} &= \mathtt{diag}[\boldsymbol{I}_n \otimes \boldsymbol{\beta}_C,\, \boldsymbol{I}_n \otimes \boldsymbol{\beta}_A,\, \boldsymbol{I}_n \otimes \boldsymbol{\beta}_S] \end{aligned} \tag{6}$$

where $\boldsymbol{u}_i = \mathtt{diag}[\boldsymbol{e}\,\mathtt{vec}(\exp(\boldsymbol{\alpha}_i \boldsymbol{1}_n^T))]$ for $i = C, A, S$.
Smoothness of the parameters is achieved by the penalty matrix $\boldsymbol{P}_{\boldsymbol{a}}$ which is specific for each triplets of parameters, though it is always a block diagonal matrix:

$$\boldsymbol{P} = \mathtt{diag}(\boldsymbol{P}_{\boldsymbol{a}_C}, \boldsymbol{P}_{\boldsymbol{a}_A}, \boldsymbol{P}_{\boldsymbol{a}_S})$$

where $\boldsymbol{P}_{\boldsymbol{a}_i} = \lambda_{\boldsymbol{a}_i}\boldsymbol{D}_i^T\boldsymbol{D}_i$ and $\boldsymbol{D}_i$ is the matrix that computes d^{th}-order differences for the coefficients $\boldsymbol{a}_i$ for $i = C, A, S$. In general, we choose second order differences, except for α_A, whose log-concave shape suggests using $d = 3$.
The smoothing parameter $\lambda_{\boldsymbol{a}_i}$ controls the roughness of the vector $\boldsymbol{a}_i$. To determine the optimal values of the $\lambda_{\boldsymbol{a}_i}$, we minimize the Bayesian information criterion (BIC). Minimization is achieved by performing a multidimensional grid search over different combinations of $\lambda_{\boldsymbol{a}_i}$. To reduce the computational burden, we assume that the smoothing parameter is the same within each triplet of parameters, i.e. $\lambda_{\boldsymbol{a}_C} = \lambda_{\boldsymbol{a}_A} = \lambda_{\boldsymbol{a}_S}$ for $\boldsymbol{a} = \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}$. However, we do not impose restrictions on $\lambda_{\boldsymbol{a}_i}$ across different triplets, i.e. in general $\lambda_{\boldsymbol{\alpha}_i} \neq \lambda_{\boldsymbol{\beta}_i} \neq \lambda_{\boldsymbol{\kappa}_i}$.
In demography age 0 is commonly treated differently. We incorporate this feature allowing discontinuity of the first age of infant mortality in $\boldsymbol{\alpha}_C$ and $\boldsymbol{\beta}_C$, i.e. the corresponding first coefficient is not penalized. Finally the choice of the starting values is not as crucial as the non-linearity of the 3C-LC model would suggest: the parametric structure within each component ensures convergence of the proposed iterative penalized CLM algorithm.

4 Swiss males application

Figure 1 shows the estimated parameters of the original Lee-Carter (LC) and of the proposed Three-Component Lee-Carter (3C-LC) model on Swiss male mortality during 1970-2014. Average shapes of each component of the 3C-LC model (left panel) distinctly decompose the mortality age-pattern. The middle panel shows component-specific rate of mortality improvement by age. Unlike the original LC, here we allow for different shapes of mortality improvement, whose smoothness avoids jaggedness of the fitted and forecast age profiles. The right panel presents time-trends for each of these component age-patterns, which we forecast by standard time-series procedures.

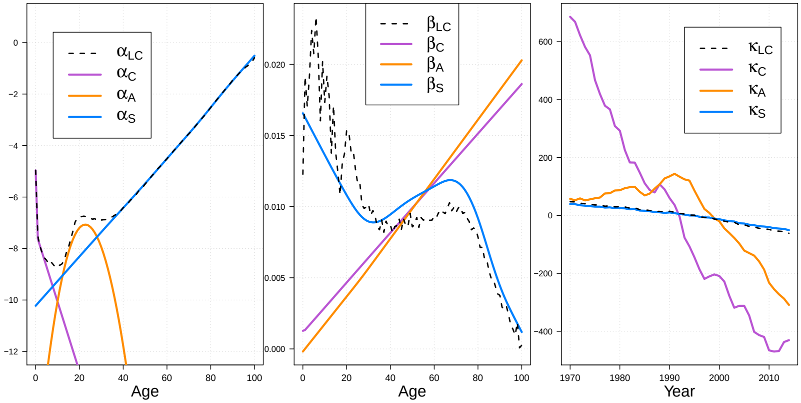


FIGURE 1. Estimated parameters of the original Lee-Carter (LC, dashed black lines) and of the proposed Three-Component Lee-Carter model for Swiss males aged 0-100 during 1970-2014.

Figure 2 shows actual, estimated and forecast mortality rates. For comparison purposes, we include estimates from a smooth extension of the Lee-Carter model (Delwarde et al., 2007).

To formally assess and compare the goodness-of-fit of the LC, LC smooth and 3C-LC models, we compute and compare their Deviance, Effective Dimension and BIC measures. Table 1 below reports the corresponding measures.

Our proposed model outperforms the other two in observed years. In particular, neither the LC nor the LC smooth are able to capture the increase in mortality at young adult ages during 1970-1990 due to the HIV epidemic (see patterns at age 25 on right panel in Figure 2). Conversely, the increasing trend of κ_A in the 3C-LC model allows to capture this relevant mortality development for Swiss males, which translates into a lower Deviance measure.

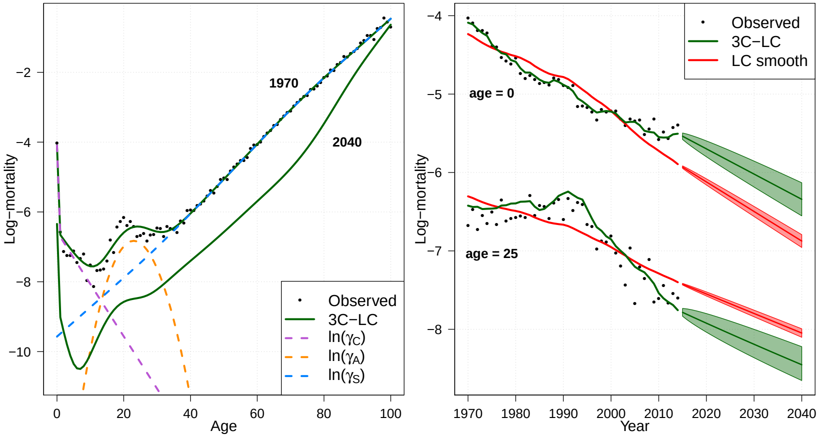


FIGURE 2. Observed, estimated and forecast mortality rates (with 80% prediction intervals in the right panel) from the Three-Component Lee-Carter and the Lee-Carter smooth models at selected years and ages for Swiss males.

TABLE 1. Poisson Deviance, Effective Dimensions (ED) and BIC measures for the LC, LC smooth and 3C-LC models. Lower values of the Deviance, ED and BIC (in bold) correspond to better fit, more parsimony and better model, respectively.

	Deviance	ED	BIC
Original LC	7193	247	9273
Smooth LC	7159	**130**	8257
Proposed 3C-LC	**6627**	141	**7813**

With respect to forecast rates, the 3C-LC provides smooth and reasonable future age patterns, which can be clearly decomposed over age and time into interpretable components. An additional advantage of the 3C-LC model is that it produces wider prediction intervals than the LC model, which has been criticized for producing too narrow intervals.

Finally, Figure 3 shows the observed and forecast life expectancy at birth (e_0) and the Keyfitz's entropy of the life table ($\mathcal{H}$) for the LC smooth and 3C-LC models. While e_0 is the mean age at death of the population, $\mathcal{H}$ is a relative measure of variability of the associated distribution and it captures the degree of lifespan inequality within a population. Although point forecasts of the two models are very close, the wider prediction intervals of the 3C-LC methodology clearly emerge from Figure 3.

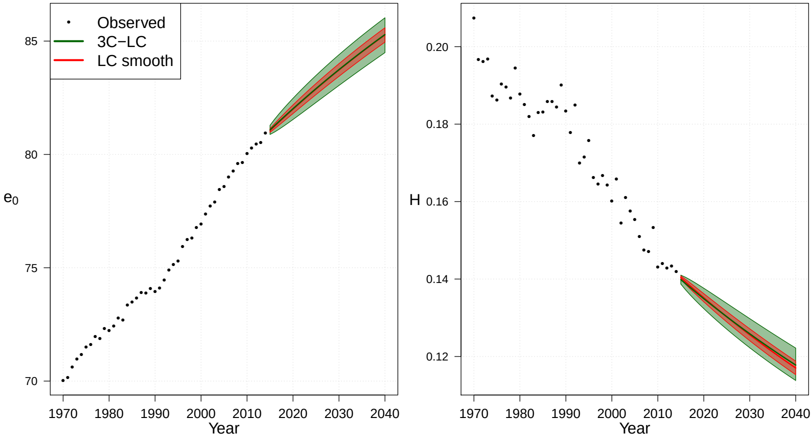


FIGURE 3. Observed and forecast with 80% prediction intervals life expectancy at birth (e_0, left panel) and Keyfitz's entropy of the life table ($\mathcal{H}$, right panel) from the Three-Component Lee-Carter and the Lee-Carter smooth models for Swiss males, 1970-2040.

References

Brillinger, D.R. (1986). A biometrics invited paper with discussion: the natural variability of vital rates and associated statistics. *Biometrics*, 693 – 734.

Camarda, C.G., Eilers, P.H., and Gampe, J. (2016). Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling*, **16(4)**, 279 – 296.

Delwarde, A., Denuit, M., and Eilers, P.H. (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Statistical Modelling*, **7**, 29 – 48.

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

Human Mortality Database (2018). University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org`. Data downloaded on February 2018.

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

Thompson, R. and Baker, R.J. (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics*, **30**, 125 – 131.

KOALA: Estimating coalition probabilities in multi-party electoral systems

Alexander Bauer[1], Andreas Bender[1], André Klima[1], Helmut Küchenhoff[1]

[1] Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany

E-mail for correspondence: Alexander.Bauer@stat.uni-muenchen.de

Abstract: Common election poll reporting is often misleading as sample uncertainty is either not covered at all or only insufficiently. For a more comprehensive coverage, we propose shifting the focus towards reporting survey-based probabilities for specific election outcomes. We present such an approach for multi-party electoral systems, focusing on probabilities of coalition majorities. A Monte Carlo based Bayesian Multinomial-Dirichlet model is used for estimation. The method utilizes published opinion polls and is accompanied by a pooling approach to summarize multiple current surveys, accounting for dependencies between polling agencies. Sample uncertainty-based probabilities are estimated, assuming the election was held today. An implementation in `R` is freely available.

Keywords: Election analysis; Opinion polls; Election reporting; Multinomial-Dirichlet; Pooling.

1 Introduction and data

Election polls try to represent the public opinion based on a finite sample. Current reporting on such surveys is most often limited to the observed shares, while sample uncertainty is usually ignored. Often e.g., a coalition – i.e. a union of multiple parties, formed to reach a governing majority – is stated to "lose" its majority just because the joint poll share drops under 50% (cf. "Umfrage zur Bundestagswahl", 2017). In our opinion, the focus in survey reporting in multi-party electoral systems should be shifted towards *how probable* an events is. We present our KOALA (Coalition Analysis) approach to estimate such probabilities to bring more value to opinion poll-based reporting. Prior to the German federal elections 2013 and 2017,

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

results based on (an earlier iteration of) our approach already entered general media reporting (cf. "Serie: Wahlistik", 2013, or Gelitz, 2017).
We use data from established polling agencies, quantifying the electoral behavior *if an election was held today.* Our approach is to be differentiated from prediction-aimed methods (cf. Graefe, 2017 or Norpoth & Gschwend, 2010) as potential shifts until election day are not taken into consideration. A Bayesian Multinomial-Dirichlet model with Monte Carlo simulations is used for estimation. Also, a pooling approach is presented to summarize multiple current opinion polls to reduce sample uncertainty.
All methods were implemented in `R` and are available in the open-source package `coalitions` (Bender & Bauer, 2018). An interactive `shiny`-based (Chang et al., 2017) website `koala.stat.uni-muenchen.de` visualizes the results and is used for communication to the general public. The process of fetching new polls, updating the website and sending out Twitter messages based on the newest results is automated.

2 Calculation of probabilities

In the last opinion poll conducted before the German federal election 2013 (Forsa, 2013), special interest was on whether CDU/CSU-FDP (also "Union-FDP") would obtain enough votes to form the governing coalition:

TABLE 1. Observed voter shares in the Forsa opinion poll published September 20th, 2013 with $n = 1995$ respondents

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
40%	26%	10%	5%	9%	2%	4%	4%

The German election system mandates a 5% votes share for parties to enter the parliament. Votes for parties below this threshold are redistributed (proportionally) to parties above it. Here, Union-FDP with its 45% raw voter share would get exactly 50% of parliament seats after redistribution. Thus, ingoring uncertainty one would conclude that a majority is slightly missed. However, it is clear that this only holds with a certain probabilitiy and particularly depends on whether FDP and/or AfD pass the 5% hurdle.
To estimate coalition probabilities, we choose a Multinomial-Dirichlet model with uninformative prior for the true party shares θ_j (Gelman et al., 2013):

$$(\theta_1, \ldots, \theta_k)^{\mathrm{T}} \sim \mathit{Dirichlet}(\alpha_1, \ldots, \alpha_k), \quad \text{with} \quad \alpha_1 = \ldots = \alpha_k = \frac{1}{2}$$

Given one (pooled) survey, the posterior also is a Dirichlet distribution with $\alpha_j = x_j + \frac{1}{2}$ for each party j and its observed vote counts x_j.
Using Monte Carlo simulations of election outcomes, one can obtain specific event probabilities by taking their relative frequency of occurence. As vote

shares are usually rounded before publication, we adjust the available data by adding random noise to x_j before calculating the Bayesian posterior. To visualize the development of such probabilities together with the underlying uncertainty for a specific coalition we recommend using ridgeline plots (Wilke, 2017) for the simulated seat distributions (Fig. 1). Looking at the probabilities based on the last opinion poll before the German election 2013, the posterior distribution is bimodal, based on the distinction whether FDP and/or AfD pass the 5% hurdle. The resulting probability for a Union-FDP majority is 27.2%, based on $10,000$ simulations.

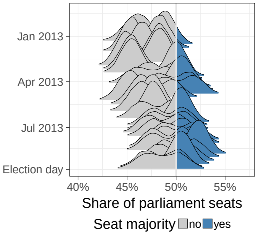


FIGURE 1. Development of simulated parliament seat share densities for the coalition Union-FDP before the German federal election in September 2013 based on Forsa opinion polls. The parts of the densities encoding for seat majorities are colored blue.

3 Pooling approach

Pooling is used to summarize multiple polls to reduce sample uncertainty. To reliably estimate the current public opinion, we use polls published within the past 14 days, only using the most recent survey per polling agency. As vote counts X_{ij} of party j in poll i are multinomially distributed, so are the summed number of votes $\sum_i X_{ij}$ when pooling multiple independent polls. Further analyses, however, show that polls from different polling agencies are correlated. Therefore, we adjust the distribution by using an *effective sample size* (Hanley et al. ,2003). Party-specific correlations were estimated based on 20 surveys of polling agencies Emnid and Forsa, using

$$Cov(X_{Aj}, X_{Bj}) = \frac{1}{2} \cdot (Var(X_{Aj}) + Var(X_{Bj}) - Var(X_{Aj} - X_{Bj})),$$

with $Var(X_{Aj})$, $Var(X_{Bj})$ the theoretical variances of binomial distributions and $Var(X_{Aj} - X_{Bj})$ estimated from the party share differences. For simplicity, we set the correlation to a fixed value of 0.5. The effective sample size n_{eff} is then defined as the ratio between the estimated variance for the pooled sample and the theoretical variance for a sample of size one:

$$n_{\text{eff}} = \frac{Var(\text{pooled})}{Var(\text{sample of size one})}.$$

For convenience, this calculation is based on the party with most votes, as the specific party choice only marginally affects the results.

4 Conclusion

We presented an approach to estimate probabilities for specific election outcomes based on publicly available opinion polls. Pooling allows for the inclusion of information from multiple surveys. Visualizing the results on a publicly available website for chosen elections, our long-term goal is to make proper uncertainty assessment in general opinion poll-based reporting the rule, rather than an exception.

References

Bender, A. and Bauer, A. (2018). coalitions: Coalition probabilities in multi-party democracies. *Journal of Open Source Software*, **3(23)**, 606, `https://doi.org/10.21105/joss.00606`.

Chang, W. et al. (2017). *shiny: Web Application Framework for R.* R package version 1.0.5. URL `https://CRAN.R-project.org/package=shiny`

Forsa (2013, September 20). Last retrieved 15/02/18, `http://archive.is/f9vse`.

Gelitz, C. (2017, September 20). *Können die aktuellen Umfragen noch falschliegen?*. Last retrieved 15/02/18, `http://archive.is/JydHd`.

Gelman, A. et al. (2013). *Bayesian Data Analysis, 3rd edition.* Boca Raton, FL: CRC press.

Graefe, A. (2017). The PollyVote's long-term forecast for the 2017 German federal election. *PS: Political Science & Politics*, **50.3**, 693 – 696.

Hanley, J. A. et al. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology*, **157(4)**, 364 – 375.

Norpoth, H. and Gschwend, T. (2010). The chancellor model: Forecasting German elections. *International Journal of Forecasting*, **26(1)**, 42 – 53.

Serie: Wahlistik (2013, September 17). Last retrieved 15/02/18, `http://archive.is/1SU1I`.

Umfrage zur Bundestagswahl: Schwarz-Gelb verliert die Mehrheit (2017, August 9). Last retrieved 15/02/18, `http://archive.is/SuXVt`.

Wilke C.O. (2017). *ggridges: Ridgeline Plots in 'ggplot2'.* R package version 0.4.1. URL `https://CRAN.R-project.org/package=ggridges`

Significance Tests for Gaussian Graphical Models Based on Shrunken Densities

Victor Bernal [12], Victor Guryev[3], Rainer Bischoff[2], Peter Horvatovich[2], Marco Grzegorczyk [1]

[1] Johann Bernoulli Institute, University of Groningen, Groningen, NL.

[2] Department of Pharmacy, Analytical Biochemistry, University of Groningen, Groningen, NL.

[3] Universitair Medisch Centrum Groningen (UMCG), ERIBA, University of Groningen, Groningen, NL.

E-mail for correspondence: v.a.bernal.arzola@rug.nl

Abstract: Gaussian Graphical Models (GGMs) are important probabilistic graphical models in Statistics. Inferring a GGM's structure from data implies computing the inverse of the covariance matrix (i.e. the precision matrix). When the number of variables p is larger than the sample size n, the (sample) covariance estimator is not invertible and therefore another estimator is required. Covariance estimators based on shrinkage are more stable (and invertible), however, classical hypothesis testing for the "shrunk" coefficients is an open challenge. In this paper we present an exact null-density that naturally includes the shrinkage, and allows an accurate parametric significance test that is accurate and computationally efficient.

Keywords: Gaussian Graphical Models; Shrinkage; Genetic Networks, "small n, large " problem.

1 Introduction

Gaussian Graphical Models (GGMs) are important network models in Statistics. A GGM is represented as a network where each variable is a node, and an edge is present between a pair of nodes if their respective partial correlation is (statistically) significant. Partial correlations measure linear dependences between a pair of variables adjusted for all other nodes. Inferring the matrix of pair-wise partial correlations (i.e. the GGM's structure) demands the estimation of the covariance matrix $\hat{\mathbf{C}}$ and its inverse,

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

therefore the importance that the covariance estimator is invertible, and well-conditioned (i.e. numerical errors are not magnified). The sample covariance estimator $\hat{\mathbf{C}}_{sm}$ with p variables and n samples is not invertible if $n \ll p$, thus another estimator must be employed. This is a common scenario in systems biology (e.g large number of genes with few measurements), and is usually refered to as the "small n, large p" problem, symbolically $n \ll p$.

Covariance estimators based on shrinkage produce a more stable estimator by using a (convex) linear combination of $\hat{\mathbf{C}}_{sm}$ with a target estimator $\mathbf{T}$ (e.g. a diagonal matrix). The result is a well-conditioned estimator, and its inverse can be used to compute the "shrunk" partial correlations. A significance test have been developed by Schäfer, J. and Strimmer, K. (2005) but it does not take the shinkage intensity into account. This is an open and important challenge as the reconstruction is a multiple testing problem (testing $\frac{p(p-1)}{2}$ edges), thus an slight bias would translate into an error repeated systematically during the inference. In this work we aim to obtain an exact density that includes the shrinkage effects. Our empirical results in Section 3 demonstrate that this leads to a substantial improvement over the earlier approach.

2 Shrinkage based Gaussian Graphical Models

Partial correlations are a measure of linear dependence between two variables adjusting the effects coming from all other variables. GGMs are undirected graphical models represented by a matrix of partial correlations. The matrix entry ρ_{ij} in a GGM represents the partial correlation between the variables i and j and can be computed from the inverse $\mathbf{C}^{-1}$ of the covariance matrix $\mathbf{C}$,

$$\rho_{ij} = -\frac{\mathbf{C}_{ij}^{-1}}{\sqrt[2]{\mathbf{C}_{ii}^{-1}}\sqrt[2]{\mathbf{C}_{jj}^{-1}}} \tag{1}$$

where $\mathbf{C}$ needs to be estimated from the data. However, when $n \leq p$ the sample covariance estimator $\hat{\mathbf{C}}_{sm}$ is ill-conditioned and cannot be used. Instead, the shrinkage based estimator $\hat{\mathbf{C}}^{[\lambda]}$ is a linear combination of $\hat{\mathbf{C}}_{sm}$ with a target matrix $\mathbf{T}$ in the form $\hat{\mathbf{C}}^{[\lambda]} = \lambda\mathbf{T} + (1-\lambda)\hat{\mathbf{C}}_{sm}$, where $\lambda \in [0, 1]$. The resulting estimator is well-conditioned, and is implemented in the widely used R package *GeneNet* (see. Schäfer, J. and Strimmer, K. (2005)) where λ is chosen following an optimization criteria. Moreover, significance is tested with the density of the standard partial correlation $f(\rho, k)$.

In the same way the correlation matrix $\mathbf{R}$ (i.e. the standarized $\hat{\mathbf{C}}_{sm}$) can be combined with (or shrunk towards) the identity matrix $\mathbf{I}$. In this case the diagonal elements of $\mathbf{R}$ (i.e. the variances) remains equal to 1, and the off-diagonal r_{ij} (i.e. the pair-wise correlations coefficients) are scaled by a factor of $(1-\lambda)$. Therefore, the probability density function (pdf) of the "shrunk" correlation $r^{[\lambda]}$ can be found via the transformation $r^{[\lambda]} = (1-\lambda)r$,

$$f(r^{[\lambda]},k) = \frac{[(1-\lambda)^2 - (r^{[\lambda]})^2]^{(\frac{k-3}{2})}}{Beta(\frac{1}{2},\frac{k-1}{2})(1-\lambda)(1-\lambda)^{\frac{k-3}{2}}} \tag{2}$$

where k denotes the degrees of freedom. We now use a classical result from Fisher (1924) to obtain the probability density of the "shrunk" partial correlation $f(r^{[\lambda]},k)$. Here it was proved that ρ and r have the same density differing only in the value of k. The main idea is to study the problem in *subject space* where each random variable is represented with a vector, and probabilistic relationships can be interpreted geometrically (see Wickens, T. D. (2014)). For the purpose of illustration, lets consider three random variables X, Y, and Z with expectation equal to zero (i.e. $E[X] = E[Y] = E[Z] = 0$). Given n data points for each variable, the corresponding random vectors $\vec{x}$, $\vec{y}$, and $\vec{z}$ are in an space of dimension n. The correlation between X and Y can be writen as

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{j=1}^{n} y_j^2}} = cos(\angle\vec{x},\vec{y}) \tag{3}$$

where the last equality comes from the product $\vec{x}\cdot\vec{y} = ||\vec{x}||||\vec{y}||cos(\angle\vec{x},\vec{y})$ under the Euclidean norm. The rationale behind the proof is that r is related to the angle between the vectors (see Eq 3), and that this angle is invariant under rotations of the coordinate axes. Therefore, the rotation can be peformed in such a way that one of the axis coincides with $\vec{z}$, and conditioning on the random variable Z is equivalent to decreasing k by one. This procedure can be continued by rotating again, and conditioning over a new variable. The same idea can be used for $r^{[\lambda]}$ (as it is a scaled correlation), and$f(\rho^{[\lambda]},k)$ is the probability density of $\rho^{[\lambda]}$.

To test the null hypothesis H_0 : (the "shrunk" partial correlation is zero) with $f(\rho^{[\lambda]},k)$ we propose the following approach: Suppose the data D consists of p variables and sample size n.

1. For D find the optimal shrinkage λ_{opt}, and estimate $\rho_{ij}^{[\lambda_{opt}]}$ (Schäfer, J. and Strimmer, K. (2005)).
2. Estimate k under H_0:
   (a) Simulate data of length n from H_0 (i.e. the precision matrix is the p x p identity), and using λ_{opt} (from step 1) infer the null-hypothetic coefficients $\rho_{0_{ij}}^{[\lambda_{opt}]}$.
   (b) Find $\hat{k}$ by maximizing the likelihood of the $\rho_{0_{ij}}^{[\lambda_{opt}]}$ with Eq 2.

3. Test the coefficients $\rho_{ij}^{[\lambda_{opt}]}$ from step 1 with $f(\rho_{ij}^{[\lambda_{opt}]}, \hat{k})$.

We will refer to this approach as "Shrunk MLE" in the following section.

3 Results

In this section we provide empirical evidence that the proposed "Shrunk MLE" approach is significantly superior to *GeneNet* 1.2.13. We cross-compare the methods on synthetic, and real gene expression data by testing the null hypothesis H_0 : (the "shrunk" partial correlation is zero). The Positive Predictive Values (PPVs) are compared on (i) syntethic data were the true structure is known, and (ii) on real data were we use MC (a computationally expensive approach) to generate a reliable goldstandard network. To simulate GGMs with a fixed percentage of edges δ we used *GeneNet* (for the algorithm see Schäfer, J. and Strimmer, K. (2005)). Figure 1 shows the $PPV = \frac{TP}{(TP+FP)}$ for different samples sizes n.

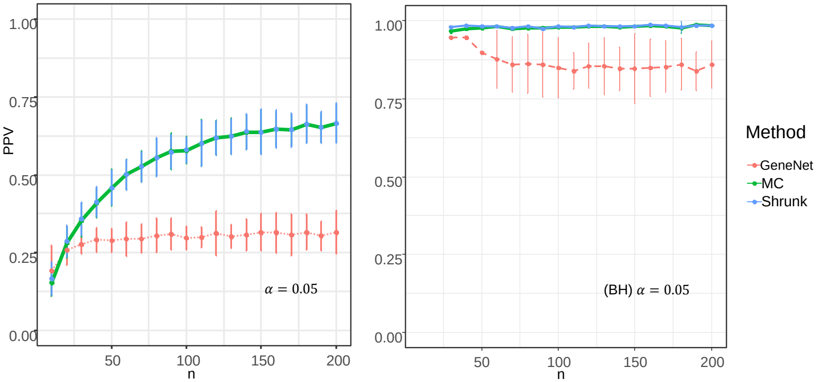


FIGURE 1. **Positive predictive value.** *On the left*: GGM simulation for $p = 100$, and n ranging from 10 to 200 in steps of size 10. The simulated GGM structure has 148 correlations (i.e. $\delta = 0.03$). The Positive predictive values (PPV) are computed using p-values at $\alpha = 0.05$. Dots (and bars) represent the average PPV (± 2 standard errors) over 25 repeated simulations, and MC was performed 15 times. Three methods are displayed:*GeneNet* (in dashed red), MC (green), and Shrunk MLE (thick blue). Note that the green and blue curves are superposed.*On the left*: The PPV are computed using Benjamini Hochberg adjusted p-values at $\alpha = 0.05$.

The results show a close agreement between the PPV obtained by MC, and with "Shrunk MLE". On the other hand, *GeneNet* has a lower PPV suggesting that it learns too many False Positives (FPs) Figure 2. We also analyzed Escherichia Coli microarray data from Schmidt-Heck, W. et al. (2004), consisting of stress temporal response of 102 genes in 9 time points after IPTG (induction of the recombinant protein SOD). Figure 2 shows a Venn diagram for the edges found by each method, here we can observe that ”Shrunk MLE” learns nearly the same edges as MC.

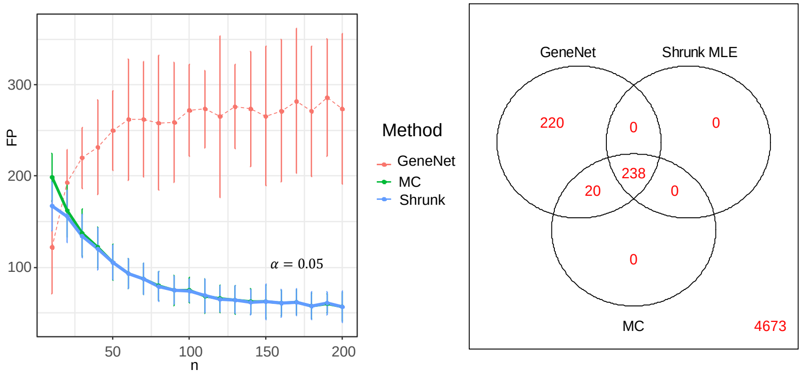


FIGURE 2. **False positives and Empirical results.** *On the left*: GGM simulation for $p = 100$, and n ranging from 10 to 200 in steps of size 10. The simulated GGM structure has 148 correlations (i.e. $\delta = 0.03$). The False Positives (FPs) are computed using p-values at $\alpha = 0.05$. Dots (and bars) represent the average FPs (± 2 standard errors) over 25 repeated simulations, and MC was performed 15 times. Three methods are displayed: *GeneNet* (in dashed red), MC (green), and Shrunk MLE (thick blue). Note that the green and blue curves are superposed. *On the right*: Venn diagram for the inferred edges in *E. coli*. Taking MC as a gold standard *GeneNet*'s sensitivity is 258/258=1, with a low PPV of 258/478 $\approx$ 0.54. Shrunk MLE has a slightly decreased sensitivity of 238/258 $\approx$ 0.92, but yields a perfect PPV of 1.

A GO enrichment analysis (http://geneontology.org/) with False Discovery Rate (FDR< 0.05) shows that the connected genes belong significantly to stress response (FDR= $2.02E^{-}02$), in contrast with *GeneNet* (FDR=$7.73E^{-}02$). This suggests a dillution of the GO's significance due higher rate of FPs. The strongest connections were lacA–lacZ, lacY– lacZ, and lacA–lacY related to the lac operon (known to be triggered by IPTG).

4 Conclusions

Gaussian Graphical Models (GGMs) are an important tool for network learning. Reconstructing the network demands the estimation of the covariance matrix, which is ill-conditioned if the sample size is smaller than the number of variables. Covariance estimators based in shrinkage make the covariance matrix invertible, however, for an accurate (parametric) significanc tests the shrinkage value needs to be included, otherwise the inference will have a systematic error (e.g. biased p-values). In this paper a new shrunk density was introduced, and to our knowledge is the only test that includes the regularization effects. In the "small n, large p" scenario the new density allows an accurate inference for any shrinkage value.

References

Fisher, R. (1924). The Distribution of the Partial Correlation Coefficient. *Metron*, **3**, 329 – 332.

Schäfer, J. and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Funtional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1 – 30.

Schmidt-Heck, W. et al. (2004). Reverse Engineering of the Stress Response during Expression of a Recombinant Protein. *Proceedings of the EUNITE symposium*, Aachen, 10 – 12.

Wickens, T. D. (2014). *The Geometry of Multivariate Statistics.* Psychology Press.

Bayesian Additive Quantile Regression Treed

Mauro Bernardi[1][2], Paola Stolfi[2]

[1] Department of Statistical Sciences, University of Padova, Padova, Italy
[2] Institute for applied mathematics "Mauro Picone" - CNR, Roma, Italy

E-mail for correspondence: mauro.bernardi@unipd.it

Abstract: Decision trees and their population counterparts are becoming promising alternatives to classical linear regression techniques because of their superior ability to adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their popularity, those methods have been developed for classification and mean regression, while often the conditional mean would not be enough to provide a complete picture of data that strongly deviate from the Gaussian assumption. The approach proposed in this paper instead considers the conditional quantile at level $\tau \in (0, 1)$ of the response variable as a sum of regression trees and is particularly valuable when skewness, fat–tails, outliers, truncated and censored data, and heteroskedasticity, can shadow the nature of the dependence between the variable of interest and the covariates.

Keywords: Regression trees, Bayesian methods, quantile regression.

1 Introduction

In empirical studies, researchers are often interested in analysing the behaviour of a response variable given the information on a set of covariates. The typical answer is to specify a linear regression model where unknown parameters are estimated by OLS, thereby leading to the approximation of the mean function. Although the mean describes the average response path as a function of the covariates, it provides little o no information about the behaviour of the conditional distribution on the tails. As far as the entire distribution is concerned, quantile regression methods adequately characterise the behaviour of the response variable at different confidence levels providing a complete picture of the relationship with the covariates. Moreover, the quantile analysis is particularly suitable when the conditional distribution strongly deviates from the Gaussian assumption because it displays heterogeneity, asymmetry or fat–tails, see, e.g., Koenker (2005). Linear quantile regression models have been extensively applied in different areas, such as, finance, engineering, econometrics and environmetrics, as a direct approach to quantify the level of risk of a given event, social sciences and quantitative marketing to find appropriate and effective solutions to specific segments of customers, and many other related fields see, Koenker et al (2017). However, despite their relevance and widespread application in empirical studies, linear quantile regression models provide only a rough "first order" approximation of the relationship between the τ–level quantile of the response variable and the covariates. Indeed, as first recognised by Koenker (2005), quantiles are linear functions only within a Gaussian world, thereby stimulating many recent attempts to overcome this limitation. Chen et al (2009), for example, consider the copula–based approach to formalise nonlinear and parametric conditional quantile relationships. Although quite flexible in fitting marginal data, the copula approach forgets to consider nonlinear interactions among the covariates. Classification and regression trees (CART, Breiman et al (1984)) and their population counterparts (Breiman (2001)) extensively use recursive partitioning algorithms to perform nonparametric regression and variable selection. The attractive feature of decision trees methods rely in their ability to partition the covariates space into disjoint hyperrectangles, thereby improving the local fit. Therefore, CART adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their extensive use in a wide variety of fields, those methods have been mainly developed for classification and mean regression. In this paper, we adopt the Bayesian point of view and we extend the additive regression trees (BART) approach of Chipman et al (2010) to deal with conditional quantiles estimation. Quantile estimation have been previously extended within the related context of random forest by Meinshausen (2006). However, unlike random forests, the Bayesian approach to decision trees learning, being likelihood–based, provides a complete inferential tool for model assessment and selection.

2 Quantile regression tree

The linear quantile regression framework for i.i.d. data models the conditional τ–level quantile of the response variable Y, with $\tau \in (0,1)$, as a linear function of the vector of dimension $(q \times 1)$ of exogenous covariates $\mathbf{X}$, i.e., $\mathcal{Q}_\tau\left(Y \mid \mathbf{X} = \mathbf{x}\right) = \mathbf{x}'\boldsymbol{\beta}$, thereby avoiding any explicit assumptions about the conditional distribution of $Y \mid \mathbf{X} = \mathbf{x}$. This is equivalent to assume an additive stochastic error term ϵ for the conditional regression function $\mu\left(\mathbf{x}\right) = \mathbf{x}'\boldsymbol{\beta}$ to be i.i.d. with zero τ–th quantile, i.e, $\mathcal{Q}_\tau\left(\epsilon \mid \mathbf{x}\right) = 0$, and constant variance. Following Yu and Moyeed (2001), the previous condition is implicitly satisfied by assuming that the conditional distribution of the response variable Y follows an Asymmetric Laplace (AL) distribution located at the true regression function $\mu(\mathbf{x})$, with constant scale $\sigma > 0$ and shape parameter τ, i.e., $\epsilon \sim \mathsf{AL}(\tau, \mu(\mathbf{x}), \sigma)$. The resulting quantile regression model assumes the AL distribution as a misspecified working likelihood that correctly identify the conditional quantile function. Similarly to the Bayesian Additive Regression Tree approach of Chipman et al (2010) for the conditional mean, the quantile regression tree extends the linear quantile model by assuming a sum–of–trees ensemble for the regression function $\mu(\mathbf{x})$. Specifically, the Bayesian Additive Quantile Regression Tree (BAQRT) model can be expressed as

$$\begin{aligned} Y &= \mu(\mathbf{x}) + \epsilon && (1) \\ &\approx \mathcal{T}_1^{\mathcal{M}}(\mathbf{x}) + \mathcal{T}_2^{\mathcal{M}}(\mathbf{x}) + \cdots + \mathcal{T}_m^{\mathcal{M}}(\mathbf{x}) + \epsilon, && (2) \end{aligned}$$

where $\epsilon \sim \mathsf{AL}(\tau, 0, \sigma)$. The assumption about the error term in equation (2) implies that $\mu(\mathbf{x}) = \mathcal{Q}_\tau(Y \mid \mathbf{X} = \mathbf{x})$. Furthermore, in equation (2) we assume that the quantile of the response variable is an additive function of $m \geq q$ regression trees, each composed by a tree structure, denoted by $\mathcal{T}$, and the parameters of the terminal nodes (also called leaves), denoted by $\mathcal{M}$. Therefore, the j–th tree for $j = 1, 2, \ldots, m$, denoted by $\mathcal{T}_j^{\mathcal{M}}$, represents a specific combination of tree structure $\mathcal{T}_j$ and tree parameters $\mathcal{M}$, i.e., the regression parameters associated to its terminal nodes. The tree structure $\mathcal{T}_j$ contains information on how any observation y_i, in a set of n i.i.d. observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, recurses down the tree specifying a splitting rule for each non–terminal (internal) node. The splitting rule has the form $x_k \leq c$ and consists of the splitting variable x_k and the splitting value $c \in \mathbb{R}$. The observation y_i is assigned to the left child if the splitting rule is satisfied and to the right child, otherwise, until a terminal node is reached and the value of the leaf of that terminal node is assigned as its predicted value. Therefore, the quantile prediction corresponding to y_i assigned by the sum of regression tree specified in equation (2) is the sum of the m leaf values. Hereafter, we denote by $\mathcal{M}_j = \{\mu_{j,1}, \mu_{j,2}, \ldots, \mu_{j,b_l}\}$ the set of parameters associated to the b_j terminal nodes of the j–th tree, where $\mu_{j,l}$, for $l = 1, 2, \ldots, b_l$ denotes the conditional quantile predicted by the model. The additive quantile regression tree specified in equation (2) provides a natural framework for likelihood–based inference on the set of quantile regression parameters, i.e., the location parameters associated to the terminal nodes of each tree belonging to the ensemble. However, additional prior information should be imposed in order to infer the structure of the each tree.

3 Application to Boston housing data

We analyse the Boston Housing data first considered by Harrison and Rubinfeld (1978) to study the influence of pollution on house prices. In par-

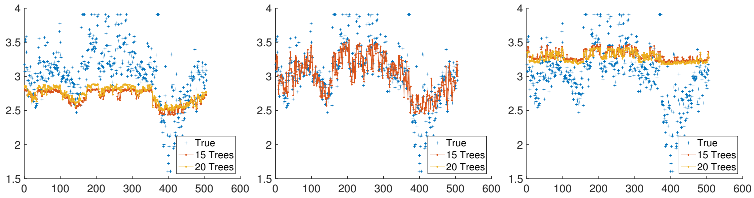


FIGURE 1. First panel: estimated quantile regression for $\tau = 0.25$; second panel: estimated quantile regression for $\tau = 0.5$; third panel: estimated quantile regression for $\tau = 0.75$.

ticular, we consider the dataset corrected by Li et al (2010). A complete description of the data and the covariates can be found in Li et al (2010). Figure 1 plots the conditional quantiles for tree different quantile levels.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. *Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.*

Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *Econometrics Journal*, 12:S50–S67.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81 – 102.

Koenker, B. (2005). Quantile Regression. *Cambridge University Press, Cambridge.*

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). Handbook of Quantile Regression. *CRC Press.*

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Anal.*, 5(3):533–556.

Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999.

Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.

Functional-Coefficient Autoregressive and Linear Regression Mixed Model for Nonlinear Time Series

Zhiqiang Cao[1], Hui Li[2], Man-Yu Wong[1]

[1] The Hong Kong University of Science and Technology, Hong Kong
[2] Beijing Normal University, China

E-mail for correspondence: zcaoae@connect.ust.hk

Abstract: For nonlinear time series, we study a kind of functional-coefficient autoregressive and linear regression mixed model. This model can handle autocorrelation and heteroscedasticity of a time series and we use the B-spline approach to estimate parameters in the model. A real data about Lake Shasta inflow is used for illustration of the model. From the performance of 6-step forward predictions, our study model performs better than functional-coefficient autoregressive regression model as well as regression and autoregressive mixed model.

Keywords: Functional-coefficient autoregressive regression; Linear regression; B-spline; Nonlinear time series.

1 The Proposed Model and Estimated Approach

1.1 The Proposed Model

The regression and autoregressive mixed (RAM) model (Box et al., 2008) is one of widely used models to study the relationship between a time series sample and some explanatory covariates, which has the following form,

$$Y_t = \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \beta_0 + \beta_1 X_{1t} + \cdots + \beta_q X_{qt} + \varepsilon_t, \qquad (1)$$

where Y_t is a time series, p is the order of autoregressive part, X_{jt} are observed covariates for $j = 1, \cdots, q$ and q is the number of interested covariates. It is often assumed that the error item ε_t is white noise, i.e., $\varepsilon_t \sim WN(0, \sigma^2)$, and ε_t is independent with Y_s when $s < t$.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

However, the coefficients of model (1) are assumed constant. And this assumption limits its application among complicated time series data, especially when data have heteroscedasticity, breaking point, trend and so on. Broaden the limitation of constant autoregressive coefficients, we study the following functional-coefficient autoregressive and linear regression mixed (FALRM) model.

For a time series Y_t and covariates, $X_{1t}, \cdots, X_{qt}$, our study mixed model has the following form,

$$Y_t = \alpha_1(Z_t)Y_{t-1} + \cdots + \alpha_p(Z_t)Y_{t-p} + \beta_0 + \beta_1 X_{1t} + \cdots + \beta_q X_{qt} + \varepsilon_t, \quad (2)$$

where Z_t is a variable depending on time t and $\varepsilon_t \sim WN(0, \sigma^2)$. The forms of autoregressive coefficient, $\alpha_j(Z_t)$ for $j = 1, 2, \cdots, p$ are unknown but their functions are smoothing enough. $\beta_j, j = 0, \cdots, q$, are constant unknown parameters.

Because the autoregressive coefficients in model (2) are functions of variable Z_t, OLS method can not be applied for obtaining their estimators. Therefore, some nonparametric estimation methods should be used instead. In this paper, we use the B-spline approach (De Boor, 2001) .

1.2 The B-spline Approach

Suppose the degree of the B-spline is k and number of knots is ν. Based on the $\nu + k + 1$ basis functions $B_1(Z_t), \ldots, B_{\nu+k+1}(Z_t)$, $\alpha(Z_t)$ can be approximated by a linear combination of the basis functions,

$$\alpha(Z_t) \approx \gamma_1 B_1(Z_t) + \cdots + \gamma_{\nu+k+1} B_{\nu+k+1}(Z_t) = \mathbf{B}^{\mathrm{T}}(Z_t)\Gamma,$$

where $\Gamma = (\gamma_1, \ldots, \gamma_{\nu+k+1})^{\mathrm{T}}$ and $\mathbf{B}(Z_t) = \{B_1(Z_t), \ldots, B_{\nu+k+1}(Z_t)\}^{\mathrm{T}}$. Note that all $\alpha_1(Z_t), \cdots, \alpha_p(Z_t)$ can be decomposed similarly, that is, $\alpha_1(Z_t) \approx \mathbf{B}^{\mathrm{T}}(Z_t)\Gamma_1, \cdots, \alpha_p(Z_t) \approx \mathbf{B}^{\mathrm{T}}(Z_t)\Gamma_p$ with $\Gamma_1 = (\gamma_{1,1}, \ldots, \gamma_{1,\nu+k+1})^{\mathrm{T}}$ and $\Gamma_p = (\gamma_{p,1}, \ldots, \gamma_{p,\nu+k+1})^{\mathrm{T}}$. After substituting them into model (2), then model (2) can be approximated by a linear regression model, that is,

$$\begin{aligned} Y_t \approx\ & [\mathbf{B}^{\mathrm{T}}(Z_t)Y_{t-1}]\Gamma_1 + \cdots + [\mathbf{B}^{\mathrm{T}}(Z_t)Y_{t-p}]\Gamma_p \\ & + \beta_0 + X_{1t}\beta_1 + \cdots + X_{qt}\beta_q + \varepsilon_t. \end{aligned} \quad (3)$$

Once constant parameter vectors $\Gamma_1, \cdots, \Gamma_p$ are estimated, estimations of corresponding functional-coefficients, i.e., $\hat{\alpha}_1(Z_t), \cdots, \hat{\alpha}_p(Z_t)$ can be obtained easily.

We know that the performance of B-spline depends on the degree, number and locations of knots, as discussions in Kim (2007). In the estimation procedure, we can use the B-spline with equally spaced knots, and select the degree k and the number of knots ν based on average mean squared error. This criterion called *AMS* was studied in Cai et al. (2000), which is essentailly a modified multifold cross-validation method.

2 A Real Study

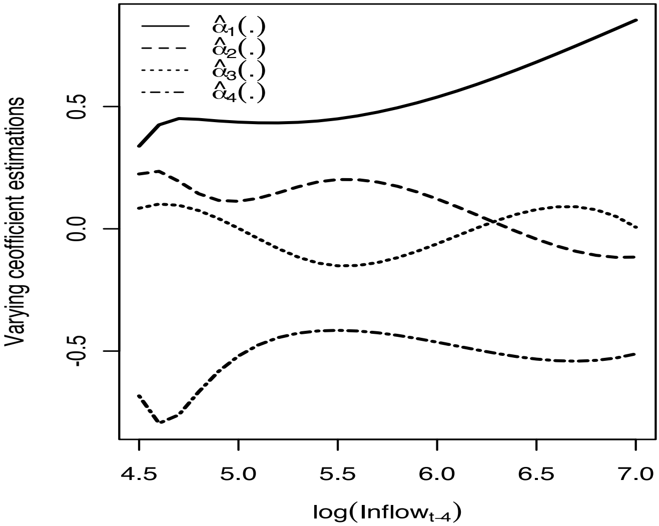


FIGURE 1. Varying coefficient estimations.

As an illustration, we apply the model (2) to Lake Shasta inflow data, which has been studied in Shumway and Stoffer (2017). The data are 454 months of measured values for the climatic variables: air temperature (Temp), dew point (DewPt), cloud cover (CldCvr)), wind speed (WndSpd), precipitation (Precip), and inflow (Inflow) at Lake Shasta, California. Our interested problem is to predict the inflow to Lake Shasta based on the climatic factors.

When using model (2) to analyze data, we put $\log(\mathrm{Inflow_t})$ as Y_t, and denote Z_t as Y_{t-d}, where d can be any integer value between 1 and p. Let $\mathrm{Temp_t}$, $\mathrm{DewPt_t}$, $\mathrm{CldCvr_t}$, $\mathrm{WndSpd_t}$ and $\mathrm{Precip_t}$ be $X_{qt}, q = 1, 2, 3, 4, 5$, respectively. Inspired by the idea of *AMS*, we can modify this criterion to not only select k and ν of B-spline, but also determine p, d and covariates in model (2). According to the modified *AMS* criterion, $k = 3$, $\nu = 2$, $p = 4$, $d = 4$ and selected covariates are $\mathrm{CldCvr_t}$, $\mathrm{WndSpd_t}$ and $\mathrm{Precip_t}$. Thus, denote $X_{1t} = \mathrm{CldCvr_t}$, $X_{2t} = \mathrm{WndSpd_t}$ and $X_{3t} = \mathrm{Precip_t}$, note that $Y_t = \log(\mathrm{Inflow_t})$, the final target model is

$$\begin{aligned} Y_t &= \alpha_1(Y_{t-4})Y_{t-1} + \alpha_2(Y_{t-4})Y_{t-2} + \alpha_3(Y_{t-4})Y_{t-3} \\ &\quad +\alpha_4(Y_{t-4})Y_{t-4} + \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t. \end{aligned} \tag{4}$$

To evaluate the performance of the model (4), we use the first 448 data points to estimate paramters, leaving out last 6 points for prediction. The estimated functions $\alpha_j(.)(1 \leq j \leq 4)$ are plotted in Figure 1, estimates of $\beta_j(j = 0, 1, 2, 3)$ are summarized in Table 1 and $\hat{\varepsilon}_t \sim WN(0, 0.045)$. Table 2 reports the absolute relative errors $(|\frac{\widehat{\log(\mathrm{Inflow_t})}-\log(\mathrm{Inflow_t})}{\log(\mathrm{Inflow_t})}|)$ of 6-step forward predictions from model (4) (FALRM).
For comparison, the function-coefficient autoregressive regression (FAR(p)) model (Cai et al., 2000) is used to fit $\log(\mathrm{Inflow_t})$ (the optimal model is achieved with $h = 0.17$, $p = 3$ and $d = 1$ based on AMS criterion), Table 2 also reports the corresponding predictions of the FAR(p) model. Besides, we use the regression and autoregressive mixed (RAM) model to study $\log(\mathrm{Inflow_t})$, and apply AR(p) model to handle the autocorrelation of residuals, its relative errors of 6-step forward predictions are reported in Table 2. It can be seen that model (4) performs better than other two models.

3 Conclusion

In this paper, we studied the functional-coefficient autoregressive and linear regression mixed model, which can be regarded as the extension of functional-coefficient autoregressive regression model. Our study model can handle autocorrelation and heteroscedasticity of time series. Based on the B-spline appraoch, the varying-coefficient and constant parameters can be estimated easily and fast. Through analyzing Lake Shasta inflow data, our study model performs better than functional-coefficient autoregressive regression model as well as regression and autoregressive mixed model.

TABLE 1. Estimation results of constant coefficients in model (4).

Coefficient	Estimate	Std. Error	T value	P value
β_0	4.262	0.132	32.375	< 0.001
β_1	0.491	0.094	5.223	<0.001
β_2	0.203	0.073	2.788	0.006
β_3	2.023×10^{-3}	9.069×10^{-5}	22.301	<0.001

References

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008). *Time Series Analysis: Forecasting and Control (4th ed).* New Yor: Wiley.

TABLE 2. The relative predictive error of $\log(\mathrm{Inflow}_t)$.

Forward step	True value	FALRM	FAR	RAM
1	4.849	0.065	0.064	0.057
2	4.575	0.038	0.116	0.162
3	4.495	0.052	0.143	0.085
4	4.301	0.026	0.224	0.100
5	4.436	0.055	0.212	0.029
6	4.535	0.090	0.198	0.109
average		0.054	0.159	0.090

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficients regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**, 941 – 956.

De Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

Kim, M. O. (2007). Quantile regression with varying coefficients. *Annals of Statistics*, **35**, 92 – 108.

Shumway, R.H. and Stoffer, D.S. (2017). *Time Series Analysis and Its Applications: With R Examples (4th ed)*. New York: Springer.

Multivariate Statistical Study on Chocolate

Mason Chen

[1] Stanford Online High School, Palo Alto, CA, USA

E-mail for correspondence: mason05@ohs.stanford.edu

Abstract: Many people like eating chocolate, but may have some concerns on health risk, especially to people with Cardiovascular or Neurovascular diseases. Chocolate, made from cocoa beans, contains flavonoids which contain antioxidants. Flavonoids are the most abundant polyphenols in human diet. Polyphenols have antioxidant properties which can prevent aging and is also beneficial to heart disease and diabetes patients. People with heart diseases should eat less of saturated fat, trans fat, sodium, and cholesterol. They should eat more dietary fiber. Cocoa flavanols promote healthy blood flow circulation from head to toe. The heart, brain, and muscle depend on a healthy circulatory system. Multivariate correlation study has found that (1) strong negative correlation between Cocoa and Sugar, and (2) strong positive correlation between Diet Fiber and Iron. Most dark chocolate contains more cocoa, and less sugar. Dietary fiber and iron are high in correlation because of the high cocoa percent. The above two correlations can be further explained by conducting the Hierarchical Clustering Analysis on separating the Dark Chocolate, Milk Chocolate and White Chocolate. The Cocoa and Calcium are the deciding factors to separate these three Chocolates.

Keywords: STEM, Flavonoids, Chocolate, Statistics, Antioxidant

1 Introduction and Literature Research

The objective of this paper are to find out if eating chocolate is unhealthy, especially what diseases can be prevented by eating chocolate? Will the nutrition composition patterns indicate which chocolate type is healthier? Chocolate is a powerful source of antioxidant. Chocolate, made from cocoa beans, contains flavonoids which contain rich antioxidants. Antioxidants prevent human aging and is also beneficial to heart disease and diabetes patients particularly. Flavonoids are the most abundant polyphenols in human diet, representing 2/3 of those digested. Polyphenols are compounds found abundantly in natural food sources that have antioxidant properties.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

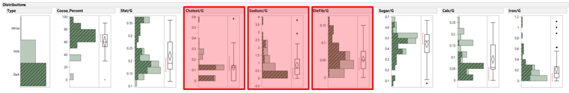


FIGURE 1. Dark Chocolate Distribution

Flavonoids have the general structure of a 15- carbon skeleton: (1) consists of two phenyl rings (A and B) and heterocyclic ring (C), and (2) this carbon structure is abbreviated C6-C3-C6. Chocolate flavonoids are flavanols.

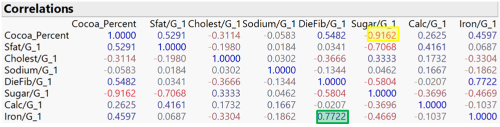
2 Experimental Setup and Graphical Analytical Analysis

Target was chosen as the main chocolate retailer since it had plenty chocolate products. 60+ different types of chocolates were collected, and each had 20 variables. In order to eliminate the bias of the central tendency and spread, the raw data was transformed to become Z- Standardized Data. After Z-transformation, all variables have new sampled distributions of mean at 0 and standard deviation of 1. The objective of this transformation is to eliminate any larger variation bias in building the statistical modeling of deriving the chocolate health index. JMP interactive graphical analysis (Figure 1) was conducted to uncover the comprehensive chocolate nutrition distributions.

After looking at the interactive graphical mode of nutrition distribution (only Dark Chocolate was selected), some interesting correlations were found. Dark chocolates mostly in common have low cholesterol, low sodium, and high dietary fiber. This helps prove that the hypothesis (dark chocolate is healthier than milk chocolate) may be correct. Milk chocolate, on the other hand, does not show any significant correlation patterns among the variables analyzed. Most sampled distributions are near random (white noise). This observation may indicate there is no health requirement on formulating the chocolate nutrition ingredients for the milk chocolate. The distribution contrast between Dark Chocolate and Milk Chocolate has provided the first-hand information on how to derive the Chocolate Health Index.

3 Multivariate Statistical Analysis

The objective of this paper is to study how the Chocolate manufacturers chose healthier nutrition facts for particular healthier chocolate types. A JMP multivariate correlation study shown in Figure 2 was further done to see if any chocolate type has strong correlation(s) between healthier nutrition and/or negative correlations between unhealthy nutrition. Correlations



Correlations

	Cocoa_Percent	Sfat/G_1	Cholest/G_1	Sodium/G_1	DieFib/G_1	Sugar/G_1	Calc/G_1	Iron/G_1
Cocoa_Percent	1.0000	0.5291	-0.3114	-0.0583	0.5482	-0.9162	0.2625	0.4597
Sfat/G_1	0.5291	1.0000	-0.1980	0.0184	0.0341	-0.7068	0.4161	0.0687
Cholest/G_1	-0.3114	-0.1980	1.0000	0.0302	-0.3666	0.3333	0.1732	-0.3304
Sodium/G_1	-0.0583	0.0184	0.0302	1.0000	-0.1344	0.0462	0.1667	-0.1862
DieFib/G_1	0.5482	0.0341	-0.3666	-0.1344	1.0000	-0.5804	-0.0207	0.7722
Sugar/G_1	-0.9162	-0.7068	0.3333	0.0462	-0.5804	1.0000	-0.3696	-0.4669
Calc/G_1	0.2625	0.4161	0.1732	0.1667	-0.0207	-0.3696	1.0000	-0.1037
Iron/G_1	0.4597	0.0687	-0.3304	-0.1862	0.7722	-0.4669	-0.1037	1.0000

FIGURE 2. Multivariate Correlations Table

between <-0.75 and >0.75 threshold were set to identify any nutrition correlation pattern.

Sugar and cocoa pair has a strong negative correlation of -0.9162. This shows that the higher the chocolate percent is, the lower the sugar percent. Since dark chocolate has a high chocolate percent, it also has low sugar. This indirectly indicates that dark chocolate is healthier with higher cocoa percent and lower sugar percent. The other identified strong positive correlation is between dietary fiber and iron. One science research has shown that most dark chocolate products with 70%-85% are rich in Fiber and Iron Recommended Daily Allowance (RDA). Dietary fiber and iron are high in positive correlation because of the dark chocolates high cocoa percent. Both graphical analyses have further provided why dark chocolate is healthier due to certain skewed nutrition preference.

Hierarchical Clustering Analysis was used to analyze and uncover evidence of correlation patterns. In data mining and statistics, hierarchical clustering analysis (HCA) is a cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types. Agglomerative: this is a ”bottom up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: this is a ”top down” approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Generally, the computing time of the Agglomerative approach is faster than the Divisive approach. Optimal efficient agglomerative methods have been developed to significantly improve the computing algorithm for large data sets. The main objective of this analysis was to search for the degree of similarity among nutrition variables, and to search for patterns (and trends) of similarity. The Agglomerative approach can identify a clustering pattern faster and more accurately from bottom-up approach by progressively merging clusters based on a defined distance metric. The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations. After grouping the first pair, JMP software calculated the center of the new formed group and found the next strongest affinity pair until the pairs were broken down as shown in the Dendrogram. Dendrogram has identified three clusters as color-coded shown in Figure 3. Number of clusters are chosen optimally by JMP algorithm based on the curve of the scree plot. In Figure 4, the three clusters identified: from, both 1st and 2nd clusters are from dark

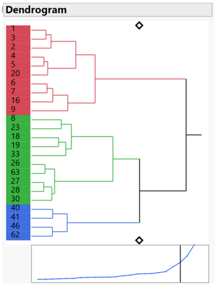


FIGURE 3. Hierarchical Dendrogram Analysis

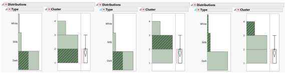


FIGURE 4. Cluster Analysis and Chocolate Type

chocolate products, and the 3rd cluster is from white chocolate and milk chocolate. The HCA is un-supervised bottom-up grouping algorithm. There is no preliminary condition/assumption to group both white chocolate and milk chocolate in the third cluster. The other interesting point is why dark chocolate products are split into two distinguished clusters.

4 Conclusions and Further Research

JMP software tools such as cluster analysis, correlation analysis, and distribution analysis were conducted to analyze Chocolate nutrition. Cocoa science, such as cocoa production, flavonoids, antioxidants, flavanol benefits, and the different types of chocolate, was learned throughout this paper. Commercial chocolate products can be categorized into three clusters based on Chocolate Nutrition amount. Dark chocolate with higher Cocoa nutrition is healthier than other chocolate products. Healthier chocolate products can prevent heart disease due to rich anti-oxidant flavonoids. This multivariate statistical analysis may provide Chocolate Producers more insight information on how to make better Chocolate Products which may help patient with Heart Disease. The same multivariate statistical approach can be further applied to the other Cardiovascular or Neurovascular diseases to extend the scope of this paper.

Regression modelling of zero-inflated multinomial counts

Alpha Oumar Diallo[1], Aliou Diop[2], Jean-François Dupuy[3]

[1] LERSTAD, CEA-MITIC, Gaston Berger University, Saint Louis, Senegal and Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, France
[2] LERSTAD, CEA-MITIC, Gaston Berger University, Saint Louis, Senegal
[3] Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, France

E-mail for correspondence: `Jean-Francois.Dupuy@insa-rennes.fr`

Abstract: Zero-inflated regression models for count data are often used in health economics to analyse the demand for medical care. Much of the recent literature on the topic has focused on univariate health-care utilization measures, such as the number of doctor visits. However, health service utilization is usually measured by a number of different counts (*e.g.*, numbers of visits to different health-care providers). In this case, zero-inflation may jointly affect several of the utilization measures. We propose a zero-inflated regression model for multinomial counts with joint zero-inflation and apply it to a set of health-care utilization data.

Keywords: Excess zeros; Health-care utilization; Multinomial logit.

1 Motivation and data description

A sample of count data is *zero-inflated* when the proportion of observed zeros is much larger than expected under standard count models. This issue arises, in particular, in the analysis of health-care utilization, as measured by the number of doctor visits.

For example, Deb and Trivedi (1997) investigate the demand for medical care by elderlies in the USA. Their analysis is based on data from the National Medical Expenditure Survey (NMES, 1987-1988). Several measures of health-care utilization are reported, such as the numbers of visits to a doctor in an office setting, visits to a non-doctor health professional (such as a nurse, optician…) in an office setting, visits to a doctor in an outpatient setting, visits to an emergency service…

A feature of these data is the high proportion of zero counts observed for some of the health-care utilization measures, i.e., there is a high proportion

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

of non-users of the corresponding health-care service over the study period. Deb and Trivedi (1997) analyse separately each measure of health-care utilization. However, several studies suggest that these measures are not independent. Therefore, we suggest to analyse them jointly, by fitting a multinomial logistic regression model adapted to zero-inflation.

We illustrate the proposed model by considering three measures of health-care utilization, namely the numbers: i) Z_1 of consultations with a non-doctor in an office setting (denoted by *ofnd*), ii) Z_2 of consultations with a non-doctor in an outpatient setting (*opnd*) and iii) Z_3 of consultations with a doctor in an office setting (*ofd*).

If m_i denotes the total number of consultations for the i-th individual and $\mathbf{X}_i$ is a vector of covariates, we let $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ and assume that Z_i has a multinomial distribution $\text{mult}(m_i, \mathbf{p}_i)$, where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and $p_{1i} = \mathbb{P}(Z_{1i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofnd* (similar interpretations hold for $p_{2i} = \mathbb{P}(Z_{2i} = 1|\mathbf{X}_i)$ and $p_{3i} = \mathbb{P}(Z_{3i} = 1|\mathbf{X}_i))$.

Frequencies of zero in *ofnd*, *opnd* and *ofd* are 62.7%, 81.3% and 1.5% respectively (calculated over the 3224 surveyed individuals). Frequencies of zeros occuring simultaneously in the pairs (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1%. That is, 51.7% of the subjects did not use any services associated with counts Z_1 and Z_2. This high frequency and the very low frequency of zero counts for *ofd* suggest that there exist some permanent non-users of *ofnd* and *opnd*. In other words, there is an excess of observations $(0, 0, m_i)$ in the data set.

2 Zero-inflated multinomial regression model

To accommodate these observations, we propose the zero-inflated multinomial (ZIM) regression model, defined as:

$$\forall i = 1, \dots, n, \quad Z_i \sim \begin{cases} (0, 0, m_i) & \text{with probability } \pi_i, \\ \text{mult}(m_i, \mathbf{p}_i) & \text{with probability } 1 - \pi_i, \end{cases} \tag{1}$$

where π_i represents the probability that the i-th individual is a permanent non-user of health-care services of the type *ofnd* and *opnd*. We model the probabilities p_{1i}, p_{2i} and p_{3i} via multinomial logistic regression:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}},\ p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}},\ p_{3i} = 1 - p_{1i} - p_{2i}.$$

The probability π_i of $(0, 0, m_i)$-inflation may depend on covariates $\mathbf{W}_i$ ($\mathbf{W}_i$ may overlap with $\mathbf{X}_i$ or be distinct) and can be modeled through a logistic regression, i.e. $\text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i$.

Based on a set of independent observations $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \dots, n$, the parameter $(\gamma, \beta_1, \beta_2)$ can be estimated by maximum likelihood (ML). The MLE is consistent and asymptotically normal, see Diallo et al. (2018), who also report results of a comprehensive simulation study.

3 An application to NMES data

Several covariates are available in the NMES data set, including: gender (1 for female, 0 for male), age (in years/10), marital status (1 if married, 0 if not), number of years of education, income (in ten-thousands of dollars), number of chronic diseases, self-perceived health level (poor, average, excellent) and an indicator of coverage by the health insurance "medicaid" (1 if covered, 0 otherwise). Self-perceived health is re-coded as "health1" (1 if health is perceived as poor, 0 otherwise) and "health2" (1 if excellent, 0 otherwise).

All parameters in model (1) are estimated by ML and a backward elimination procedure based on AIC is carried out to select relevant covariates. Results are reported in Table 1. Some interpretations are as follows (see Diallo et al. (2018) for a detailed analysis).

Age, gender, educational level and medicaid status are identified as the most influencing factors for being a permanent non-user of *ofnd* and *opnd*, with medicaid recipients being more likely to be permanent non-users. Moreover, medicaid status does not affect *ofnd* utilization, which may be explained by the fact that part of the decision of (not) using *ofnd* by medicaid recipients was captured in the model for π_i. The probability of using *opnd* is lower for medicaid recipients than for non-recipients. All this confirms previous findings in the literature that medicaid recipients tend to favor doctor visits in an office setting over non-doctor visits.

Educational level is an important determinant of the decision of being a permanent non-user of *ofnd* and *opnd*. But once an individual has chosen to use eventually these health-care services (with a probability that increases with level of education), our results suggest that schooling does not tend to favor a specific kind of health-care service.

Income does not affect utilization of medical care. This is consistent with previous findings and is explained in the literature by the fact that income may affect quality of care rather than visits number.

All these results confirm previous findings in the literature and additionally, unable us to rank the various forms of medical care by order of utilization.

References

Deb, P., and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, **12**, 313 – 336.

Diallo, A., Diop, A., and Dupuy, J.F. (2018). Analysis of multinomial counts with joint zero-inflation, with an application to health economics. *Journal of Statistical Planning and Inference*, **194**, 85 – 105.

TABLE 1. ZIM regression model fitted to NMES data.

	covariate	estimate	s.e.	Wald test	Pr(>t)	
$\beta_{1,1}$	intercept	-1.248440	0.327153	-3.816	0.000136	***
$\beta_{1,2}$	health1	-0.396752	0.071734	-5.531	3.19e-08	***
$\beta_{1,3}$	health2	0.307117	0.078294	3.923	8.76e-05	***
$\beta_{1,4}$	numchron	-0.128615	0.016425	-7.830	4.87e-15	***
$\beta_{1,5}$	age	0.021925	0.039900	0.550	0.582655	
$\beta_{1,6}$	gender	0.184974	0.046684	3.962	7.43e-05	***
$\beta_{1,7}$	fstatus	0.204095	0.046834	4.358	1.31e-05	***
$\beta_{1,8}$	school	0.007483	0.006577	1.138	0.255254	
$\beta_{1,9}$	income	-0.009338	0.006506	-1.435	0.151202	
$\beta_{1,10}$	med	-0.034313	0.090432	-0.379	0.704366	
$\beta_{2,1}$	intercept	2.164690	0.465538	4.650	3.32e-06	***
$\beta_{2,2}$	health1	0.130806	0.097185	1.346	0.178319	
$\beta_{2,3}$	health2	-0.405494	0.166065	-2.442	0.014615	*
$\beta_{2,4}$	numchron	-0.037862	0.024914	-1.520	0.128586	
$\beta_{2,5}$	age	-0.553335	0.058814	-9.408	$<$ 2e-16	***
$\beta_{2,6}$	gender	-0.028528	0.073176	-0.390	0.696649	
$\beta_{2,7}$	fstatus	-0.239131	0.072781	-3.286	0.001018	**
$\beta_{2,8}$	school	-0.017676	0.010526	-1.679	0.093101	.
$\beta_{2,9}$	income	0.011618	0.009440	1.231	0.218424	
$\beta_{2,10}$	med	-0.397013	0.148050	-2.682	0.007327	**
γ_1	intercept	-0.605712	0.548682	-1.104	0.269619	
γ_2	health1	0.249759	0.137186	1.821	0.068670	.
γ_3	numchron	-0.053888	0.035305	-1.526	0.126922	
γ_4	age	0.165217	0.069555	2.375	0.017532	*
γ_5	gender	-0.269703	0.091465	-2.949	0.003191	**
γ_6	school	-0.073233	0.012854	-5.697	1.22e-08	***
γ_7	med	0.543147	0.160838	3.377	0.000733	***

A new diffusion model for competition among three actors

Claudia Furlan, Cinzia Mortarino, Mohammad Salim Zahangir

[1] Department of Statistical Sciences, University of Padova, Italy

E-mail for correspondence: furlan@stat.unipd.it

Abstract: The aim of this paper is to propose a model to describe the mutual interactions among the lifecycles of three competing products acting simultaneously within a common market. The model is applied to real data in the energy context and its performance is compared to the results obtained with models already available in the literature for two competitors. For the examined datasets, the new model shows a relevant improvement in terms of forecasting performance.

Keywords: Competition; Nonlinear diffusion models; Forecasting accuracy.

1 Introduction

Diffusion of innovations has a long tradition within the literature (Peres et al., 2010), but the largest number of the contributions still approach the problem with separate analyses for specific products or for a global category. Only in the last years (Guseo and Mortarino, 2014 and references therein cited), some applicable models were made available to jointly describe the diffusion of two competitors simultaneously spreading into the same category niche. The relevance of building a *joint* model is due to the need for simultaneously estimating the peculiarities of each product and their mutual interaction that may generate competition or cooperation. For more than two competitors, however, there aren't published models to be feasible for applications. In other words, the extension from two to more than two actors is only theoretically included into the current literature, but high parameter dimension and complexity structure of the interactions among competitors prevent this extension to be a real tool. For this reason, in applications, practitioners, to obtain a bivariate structure, are forced to aggregate data pertaining to the more similar products or to describe the market only through the two leading actors. This, of course, leads to hiding the specific peculiarities of some of the actors thus wasting rich information. The aim of this paper is to give a contribution to the topic of modelling diffusion of innovations to describe a market where three actors compete for the same customers, illustrating how rich the description of their mutual interactions could be to accurately represent the market's features. Analyzing real data in the energy context, we will show how a three-competitor model (3CM) can be fitted. To the same dataset we will also fit a reduced model for two competitors (2CM), where, as it is often done in practice, the data for two competitors are aggregated. The aim of the comparison is to show that using the reacher dataset through the 3CM allows us obtaining better results in terms of forecasting accuracy and of reduction of prediction confidence band width.

2 Model

Diffusion models are usually defined through a differential representation, which may admit or not a closed-form solution. The main advantage relies on parsimonious descriptions of real adoption processes based on interpretable parameters. Let $z_i(t)$ be the cumulative sales of the i-th competitor, $i = 1, 2, 3$, and $z(t) = \sum_i z_i(t)$ be the category cumulative sales of all the competitors in the market. Let $z_i'(t) = dz_i(t)/dt$ be the instantaneous sales of the i-th product. Since the products represent a homogeneous category competing for the same customers, we assume a common market potential, m and correspondingly, a common residual market, $m - z(t)$. We focus here on situations where two products exist in the market from the beginning, while the third product enters the market at time $t = c_2$, with $c_2 > 0$ (t=0 represents the time origin for the first two competitors). The 3CM here proposed, as an extension of the 2CM by Guseo and Mortarino (2014), can be expressed with the following system of differential equations, where $R(t) = 1 - z(t)/m$ represents the relative category residual market:

$$\begin{aligned}
z_1'(t) =& m\Bigg\{\left[p_{1\alpha} + (q_{1\alpha} + \delta_\alpha)\frac{z_1(t)}{m} + q_{1\alpha}\frac{z_2(t)}{m}\right](1 - I_{t>c_2}) + \\
&+ \left[p_{1\beta} + (q_{1\beta} + \delta_\beta)\frac{z_1(t)}{m} + q_{1\beta}\frac{z_2(t)}{m} + q_{1\beta}\frac{z_3(t)}{m}\right] I_{t>c_2}\Bigg\} R(t)x_1(t) \\
z_2'(t) =& m\Bigg\{\left[p_{2\alpha} + (q_{2\alpha} - \delta_\alpha)\frac{z_1(t)}{m} + q_{2\alpha}\frac{z_2(t)}{m}\right](1 - I_{t>c_2}) + \\
&+ \left[p_{2\beta} + q_{2\beta}\frac{z_1(t)}{m} + (q_{2\beta} + \delta_\beta)\frac{z_2(t)}{m} + q_{2\beta}\frac{z_3(t)}{m}\right] I_{t>c_2}\Bigg\} R(t)x_2(t) \\
z_3'(t) =& m\Bigg\{\left[p_3 + (q_3 - \delta_\beta)\frac{z_1(t)}{m} + (q_3 - \delta_\beta)\frac{z_2(t)}{m} + q_3\frac{z_3(t)}{m}\right] I_{t>c_2}\Bigg\} R(t)x_3(t) \\
m =& m_\alpha(1 - I_{t>c_2}) + m_\beta I_{t>c_2} \\
z(t) =& z_1(t) + z_2(t) + z_3(t)I_{t>c_2}.
\end{aligned} \tag{1}$$

System (1) describes a competition among three products in two phases. During the first phase, until $t \leq c_2$, it is assumed that the first two products are characterized separately by three parameters each (denoted with subscript α). The parameters of the first product are $(p_{1\alpha}, q_{1\alpha} + \delta_\alpha, q_{1\alpha})$, and the parameters of the second product are $(p_{2\alpha}, q_{2\alpha}, q_{2\alpha} - \delta_\alpha)$. At time $t = c_2$, when the competition extends from two to three products, we allow the first two products to be characterized by new parameters (denoted with subscript β): $(p_{1\beta}, q_{1\beta} + \delta_\beta, q_{1\beta})$ for the first competitor, and $(p_{2\beta}, q_{2\beta}, q_{2\beta} - \delta_\beta)$ for the second. This is an important feature, since it is very common that a new competitor's launch affects the diffusion dynamics of previously existing products. The third competitor is characterized by parameters $(p_3, q_3 - \delta_\beta, q_3)$. Parameters δ_j, $j \in \{\alpha, \beta\}$, serve the purpose to differentiate between within–brand word-of-mouth (the effect on the future adopters of a product due to its own past adoptions) and cross–brand word-of-mouth (the effect on the future adopters of a product due to the past adoptions of its competitors). The relevant issue in this research topic is to build a large set of models to describe the different characteristics of the diffusion process. Confirmation or rejection of the assumptions underlying each model is then attained by fitting available observed data and comparing the models' performances. In particular, restricted models where δ_α and/or δ_β equal zero may be applied whenever data support this constraint. The common market potential, m is equal to m_α, in the first phase and is allowed to change to m_β, in the second phase.

The model may also describe specific exogenous changes in the diffusion speed of each competitor through the intervention functions $x_i(t)$, $i = 1, 2, 3$, (Bass et al., 1994). These functions are flexible structures (Guseo et al., 2005) whose parameters are estimated simultaneously with the diffusion parameters. As an example, an exponential shock could be modelled through $x(t) = 1 + ce^{b(t-a)} I_{[t \geq a]}$, where a denotes the starting time of the shock, b indicates how rapidly the shock decays towards 0, and c denotes the intensity of the shock (either positive or negative). For details, about inference, see Seber and Wild (2003) and, in particular for tests to compare nested models and for confidence bands construction, see Guseo and Mortarino (2015).

3 Application and results

For the application, we considered the yearly energy consumption (provided by British Petroleum, in Mtoe) for Switzerland and Sweden of Coal, Gas and Oil (CGO), the Renewables, and Nuclear. Data cover the period 1965-2015, except for Nuclear energy, which enters the market in 1969 in Switzerland and in 1972 in Sweden. Here, due to space limitations, only results about Switzerland will be displayed.

We first applied the bivariate model (2CM) by Guseo and Mortarino (2014). To do this, data for Nuclear were added to CGO as a unique competitor

TABLE 1. Switzerland. Estimation results for the 3CM with the constraint $\delta_\alpha = 0$. Parameters with subscript 1 refer to CGO, subscript 2 denotes the Renewables and subscript 3 is used for Nuclear.

Par.	Estimate	Standard error	Par.	Estimate	Standard error
m_α	$5.6640*10^{2}$	$2.5184*10^{-8}$	$q_{2\alpha}$	$2.6544*10^{-2}$	$1.4468*10^{-2}$
$p_{1\alpha}$	$1.6178*10^{-2}$	$7.4599*10^{-4}$	$p_{2\beta}$	$2.3531*10^{-3}$	$1.3310*10^{-4}$
$q_{1\alpha}$	$3.8451*10^{-2}$	$1.4468*10^{-2}$	$q_{2\beta}$	$1.4661*10^{-2}$	$3.1345*10^{-3}$
m_β	$2.6052*10^{3}$	$1.3844*10^{2}$	c_2	$1.2913*10^{-1}$	$6.3375*10^{-2}$
$p_{1\beta}$	$4.8964*10^{-3}$	$2.4657*10^{-4}$	b_2	$8.7119*10^{-2}$	$1.9123*10^{-1}$
$q_{1\beta}$	$2.5030*10^{-2}$	$5.8230*10^{-3}$	a_2	47.5553	$7.1297*10^{-4}$
δ_β	$-2.5424*10^{-2}$	$1.0066*10^{-2}$	p_3	$-2.6718*10^{-4}$	$1.5258*10^{-4}$
c_1	$2.6752*10^{-1}$	$1.5234*10^{-1}$	q_3	$-1.1736*10^{-2}$	$8.2234*10^{-3}$
b_1	-1.8912	1.8674	c_3	$4.1973*10^{-1}$	$1.2489*10^{-1}$
a_1	8.6553	$7.7069*10^{-2}$	b_3	$-1.0201*10^{-1}$	$5.2892*10^{-2}$
$p_{2\alpha}$	$1.0018*10^{-2}$	$7.4599*10^{-4}$	a_3	20.0000	$5.3474*10^{-3}$
$R^2 = 0.987098$					

with respect to the Renewables. Then, we applied the 3CM to CGO, Nuclear, and the Renewables. Due to the fact that period α (when only CGO and Renewables compete) is very short, we chose to fit a simpler model with $\delta_\alpha = 0$. Table 1 thus shows the results. The model chosen includes a positive shock for CGO in $1965+\hat{a}_1 \simeq 1974$, a positive one for Renewables starting in $1965+\hat{a}_2 \simeq 2013$, and a positive one for Nuclear in $1965+\hat{a}_3 \simeq 1985$. Notice that the choice among alternative nested models cannot be performed by looking at marginal confidence intervals for single parameters, since, due to the curvature of the nonlinear parameter space, each confidence interval represents only a specific section of the space and could be very misleading. The choice of the best model, conversely, is performed by evaluating tests to compare nested models based on the global fitting. The significance of exogenous shocks is tested with the same approach.

Figure 1 shows fitted values, predictions and confidence bands for the 3CM. We appreciate from the plot that in this country nuclear consumptions start slowly from 1969 and reach their long-term level around 1984, thus their evolution is very different from the profile observed for other fossil sources. This suggests of course that data aggregation to reduce to 2 competitors involves an important information reduction both for CGO and Nuclear. Our interest is to examine in depth the consequences on Renewables description. Since the two models under comparison (2CM and 3CM) use different data, it is not appropriate to make a direct comparison of global goodness-of-fit measures. We thereby decided to evaluate the improvement of the 3CM with respect to the 2CM, focusing on the Renewables.

Forecasts are made up to 5 years ahead, which is a reasonable forecasting horizon in the energy market. Figure 2 highlights the final part of the Renewables time series and the fitted values with 5-step-ahead forecasts

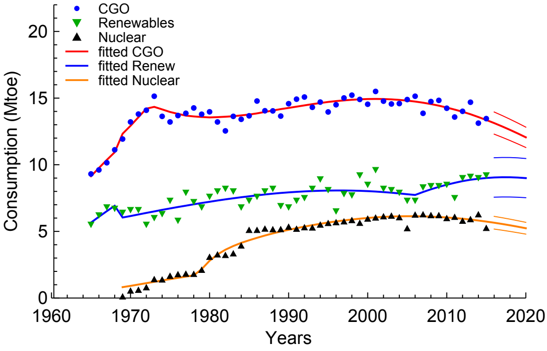


FIGURE 1. Switzerland. Data, fitted values, and prediction confidence bands for the 3CM.

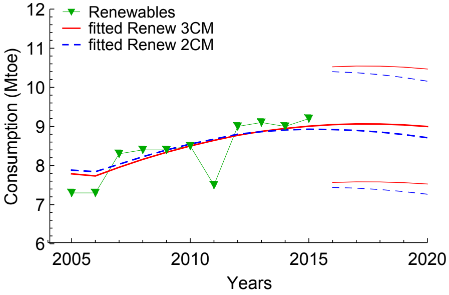


FIGURE 2. Switzerland. Predictions and confidence bands for the Renewables with 3CM and 2CM.

obtained with 3CM and 2CM. Table 2 shows confidence band width for the two models. The 3CM gives a reduction in terms of confidence band width for forecasts from 1-step-ahead to 5-step-ahead.

Forecasting accuracy analysis results are proposed in Table 3. A wide set of measures has been evaluated here: RMSE, MAPE, sMAPE, MASE (Hyndman and Koehler, 2006), UMBRAE (Chen et al., 2017), and %Better. Results are uncontroversial, since all the evaluated measures have smaller values for the 3CM if compared with the 2CM at each of the 5 steps. Analogous results have been obtained for Sweden.

In summary, this application shows the feasibility of the 3CM and highlights that a better description of the Renewables' competitors obtained by separating CGO and Nuclear data, results in a improved forecasting performance for the Renewables.

TABLE 2. Switzerland. Comparison between 2CM and 3CM: confidence band width for forecasts from 1-step-ahead to 5-step-ahead.

	step 1	step 2	step 3	step 4	step 5
2CM	3.0497	3.0799	3.1160	3.1585	3.2081
3CM	2.9349	2.9478	2.9618	2.9770	2.9936

TABLE 3. Switzerland. Comparison between 2CM and 3CM: forecasting accuracy measures for Renewables predictions.

	2CM					3CM				
	step 1	step 2	step 3	step 4	step 5	step 1	step 2	step 3	step 4	step 5
RMSE	1.604	1.976	2.260	2.763	2.368	1.287	1.653	1.891	2.225	1.959
MAPE	0.059	0.075	0.100	0.149	0.150	0.043	0.064	0.085	0.121	0.124
sMAPE	0.061	0.078	0.106	0.162	0.162	0.044	0.066	0.088	0.129	0.132
MASE	0.831	1.066	1.446	2.213	2.220	0.590	0.905	1.210	1.796	1.836
UMBRAE	2.095	2.042	1.146	1.713	2.044	0.994	1.759	0.779	1.404	1.689
% Better	29%	33%	40%	0%	0%	71%	33%	60%	25%	0%

References

Bass, F.M., Krishnan, T.V., and Jain, D.C. (1994). Why the Bass model fits without decision variables. *Marketing Science*, **13**, 203 – 223.

Chen, C., Twycross, J., and Garibaldi, J.M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, **12**, 1 – 23.

Guseo, R. and Dalla Valle, A. (2005). Oil and gas depletion: diffusion models and forecasting under strategic intervention. *Statistical Methods and Applications*, **14**, 375 – 387.

Guseo, R. and Mortarino, C. (2014). Within-brand and cross-brand word-of-mouth for sequential multi-innovation diffusions. *IMA Journal of Management Mathematics*, **25**, 287 – 311.

Guseo, R. and Mortarino, C. (2015). Modeling competition between two pharmaceutical drugs using innovation diffusion models. *Annals of Applied Statistics*, **9**, 2073 – 2089.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**, 679 – 688.

Peres, R., Muller, E., and Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, **27**, 91 – 106.

Seber, G.A.F. and Wild, C.J. (2003). *Nonlinear Regression.* New York: Wiley.

Mean survival time by ordered fractions of population with censored data

Celia García-Pareja[1], Matteo Bottai[1]

[1] Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

E-mail for correspondence: celia.garcia.pareja@ki.se

Abstract: We propose a novel approach for estimating mean survival time in the presence of censored data, in which we divide the population under study into ordered fractions defined by a set of proportions, and compute the mean survival time for each fraction separately. Our approach provides a detailed picture of the distribution of the time variable while preserving the appealing interpretation of the mean. Our measure proves to be of great use in applications, particularly those where we are able to detect differences in mean survival across groups for certain fractions of the population that would have been overlooked using other available methods.

Keywords: Quantiles; Ordered Data; Mean Survival; Censoring; Kaplan-Meier

1 Brief introduction

Survival data analysis methods are the cornerstone of a wide range of statistical applications. While mean survival time is of utmost relevance, e.g., in health economics (Paltiel et al., 2009) or oncology studies (Zhao et al., 2001), its estimation might be hindered in the presence of censoring, where the time variable is only observed until a certain quantile. In practice, censoring is almost always present, calling for specialised estimation techniques. Several approaches have been considered to overcome this problem, the most used amongst them, the restricted mean, computes mean survival time up to a specific cut-off time point (Irwin, 1949). Estimation of the restricted mean, however, might be heavily affected by the presence of censored observations, which will result in a loss of estimation accuracy. Moreover, clinical relevance and interpretation of restricted mean estimates remains unclear.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

We present a novel mean survival measure based on observed quantiles that divides the population in ordered fractions in which the mean survival can be estimated separately. Interpretation of the estimates is straightforward, as they refer to mean survival times for the specified fractions of the population. Similarly to the restricted mean, we estimate mean survival up to a specific cut-off point, that we set to the last observed p-th fraction of population to experience the event of interest. Our approach exploits that the distribution of observed and censored events imposes differences in the estimation accuracy of specific quantiles, i.e., those that are close to observed events can be more precisely estimated than those located after the occurrence of censored events. Therefore, estimates for certain fractions can be really precise, which allows quantifying significant mean survival differences across groups, even in scenarios where state-of-the-art methods are unable to detect them.

2 Mean survival by ordered fractions

Let T be a non-negative random variable with $\mathrm{E}[T] < \infty$ and let $S(\cdot)$ and $Q(\cdot)$ denote its survival and quantile functions, respectively. An expression for the expectation of T in terms of $Q(\cdot)$ is

$$\mu = \mathrm{E}[T] = \int_0^\infty S(t)\mathrm{d}t = \int_0^1 Q(p)\mathrm{d}p. \tag{1}$$

Given a grid of proportions $\{\lambda_0, \lambda_1, \ldots, \lambda_K\}$ with $\lambda_{k-1} < \lambda_k$ for all $k \in \{1, \ldots, K\}$, we can divide μ into separate components as follows

$$\mu = \sum_{k=1}^{K} \mu_k, \text{ where } \mu_k = \int_{\lambda_{k-1}}^{\lambda_k} Q(p)\mathrm{d}p, \lambda_0 = 0 \text{ and } \lambda_K = 1. \tag{2}$$

If we now weight each μ_k by its corresponding inverse proportion, we obtain

$$\overline{\mu}_k = \frac{\mu_k}{\lambda_k - \lambda_{k-1}},$$

where $\overline{\mu}_k$ is the mean survival time for a specific fraction of population delimited by $(\lambda_{k-1}, \lambda_k)$. For example, if we consider $(\lambda_0, \lambda_1) = (0, 0.5)$, $\overline{\mu}_1$ quantifies mean survival time for the first half of the population to experience the event of interest.

In the presence of a censoring variable C, when $Y = \min(T, C)$ is observed instead, the decomposition shown in (2) is of utmost convenience because λ_K can be set to the largest proportion of observed events, that is, the one corresponding to the last observed quantile. Note that when $\lambda_K < 1$, the mean survival time for the λ_K-th fraction of the population observed to experience the event, does not correspond to the restricted mean computed up to the last observed quantile $y^\star = Q(\lambda_K)$. Indeed, while

$$\overline{\mu}_K = \frac{1}{\lambda_K}\int_0^{\lambda_K} Q(p)\mathrm{d}p$$

can be easily interpreted in terms of the population under study, the corresponding

$$\mu^\star = \int_0^{y^\star} S(y)\mathrm{d}y,$$

does not prove as informative.

3 Estimation and simulation results

In the presence of censoring, estimation of μ_k is possible via the Kaplan-Meier estimator of the underlying survival function, $\widehat{S}(\cdot)$. Given $\widehat{S}(\cdot)$ and the grid of proportions $\{\gamma_0, \gamma_1, \dots, \gamma_K\} = \{1-\lambda_0, 1-\lambda_1, \dots, 1-\lambda_K\}$, an estimator for μ_k follows easily from equations (1) and (2), with

$$\begin{aligned}\widehat{\mu_k} &= \sum_{j=1}^{J_k} y_j[\min\{\widehat{S}(y_{j-1}), \gamma_{k-1}\} - \max\{\widehat{S}(y_j), \gamma_k\}] \\ &= \sum_{j_k=1}^{J_k} \widehat{Q}(p_j)(p_j - p_{j-1}),\end{aligned}$$

where y_j denote observed event times such that $\widehat{S}(y_j) \in [\gamma_k, \gamma_{k-1}]$ for all $j \in \{1, \dots, J_k\}$, and $\widehat{S}(y_0) \geq \gamma_{k-1}$ and $\widehat{S}(y_{J_k}) \leq \gamma_k$. In this case, we obtain a step-wise constant estimator of the quantile function $\widehat{Q}(\cdot)$, in which observed times y_j play the role of estimated quantiles $\widehat{Q}(p_j)$ of order $p_j = \widehat{S}(y_j)$.

We tested the performance of $\widehat{\mu_k}$ in different scenarios, all yielding analogous conclusions. In Table 1 we present results for a simulation study of $5,000$ data sets with 200 samples each, generated from a time variable following a log-logistic distribution with scale $\alpha = 1$ and shape $\beta = 2$. The censoring variables were sampled independently from a uniform distribution in $(0, 7/3)$, yielding an average censoring rate of 50%. Estimated average upper and lower bounds for $\widehat{\mu_k}$ where computed integrating over equal precision confidence bands for the Kaplan-Meier estimator (Nair, 1984). We observed that our estimates' precision decreased with increasing k (that is, the bands widened with increasing k), which was expected, as the proportion of censored observations also increased with k and fewer events were observed. In this sense, we might say that some μ_k can be *more precisely* estimated than others, which proves a highly useful tool in application settings.

TABLE 1. Results for $5,000$ simulations of 200 samples from a log-logistic model with scale $\alpha = 1$ and shape $\beta = 2$ with censoring variable uniform in (0,7/3), corresponding to an average censoring rate of 50%. True (μ_k) and average estimated ($\widehat{\mu_k}$) values, with average lower ($\widehat{\mu_k^L}$) and upper ($\widehat{\mu_k^U}$) bounds for 5 fractions of population, % of simulations (nsim$_k$) in which $\widehat{\mu_k}$ could be computed and average number of observed events (d$_k$) are reported. The average bound marked with * had finite values in 75% of the simulations.

k	λ_k	μ_k	$\widehat{\mu_k}$	$\widehat{\mu_k^L} - \widehat{\mu_k^U}$	nsim$_k$	d$_k$
1	0.20	0.064	0.064	$0.044 - 0.086$	100%	34
2	0.40	0.131	0.132	$0.101 - 0.175$	100%	29
3	0.60	0.201	0.202	$0.156 - 0.264^\star$	100%	23
4	0.80	0.311	0.304	$0.226 - \infty$	70.7%	14
5	0.95	0.420	0.307	$0.239 - \infty$	5.80%	4

4 Application example: Survival after bone marrow transplant in lymphoma patients

We analysed data on 35 patients with lymphoma that received either an allogenic or an autologous bone marrow transplant, that is, they received marrow either from a a compatible donor or their own after chemotherapy treatment and cleansing, respectively (Avalos et al., 1993). The aim of the study was to find differences between lymphoma-free survival after having received either type of transplant. After 2.5 years of follow-up, 26 patients had died or relapsed and the censoring rate was 25.7%. The estimated survival curves for both treatments are shown in Figure 1.

While restricted mean survival estimates did not detect any significant difference in mean survival between the allogenic and autologus transplant groups (restricted mean difference of 146.5 days, with 95% confidence interval $(-29.71, 322.7)$)), our approach showed that among earlier failures that difference was actually significant. In particular, considering the weakest 10% of the patients, that is, the first 10% to die or relapse after receiving the transplant, mean survival difference was estimated at 32.15 days (95% CI $(13.98 - 50.31)$) favouring those who received the autologus transplant. In Table 2 we show the results of mean survival time differences after receiving a bone marrow transplant by deciles of population up to the 80th percentile (last fraction commonly observed in both groups). Our estimates could detect an improvement on lymphoma-free survival for the autologus transplant group amongst at least the weakest 20% of patients, providing a useful guide for effective decision-making in further studies.

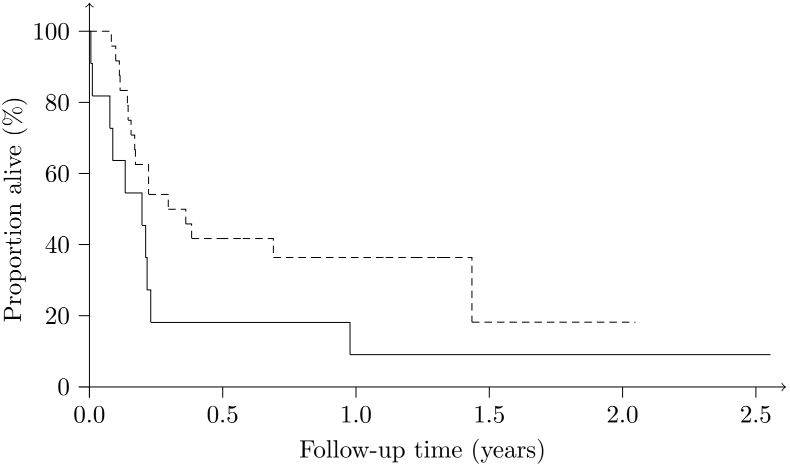


FIGURE 1. Estimated Kaplan-Meier survival curves after bone marrow transplant for lymphoma patients that received allogenic (solid line) or autologus (dashed line) transplant.

TABLE 2. Estimates for mean survival differences between allogenic ($\widehat{\mu^0{}_k}$) and autologus ($\widehat{\mu^1{}_k}$) bone marrow transplants with bootstrapped 95% confidence intervals by ordered deciles of population.

k	λ_k	$\widehat{\mu^1{}_k} - \widehat{\mu^0{}_k}$	95% CI
1	0.1	32.15	13.98 – 50.31
2	0.2	36.72	2.843 – 70.60
3	0.3	26.23	−19.94 – 72.43
4	0.4	28.98	−40.51 – 98.48
5	0.5	32.80	−124.5 – 190.1
6	0.6	80.60	−1283 – 1444
7	0.7	349.4	−446.4 – 1145
8	0.8	441.4	−130.3 – 1013

5 Final remarks

Our approach for quantifying mean survival time takes advantage of the information contained in the data and deals with the censoring hurdle. Our proposed measures are easily interpretable, providing a useful alternative to the restricted mean, which poses interpretation difficulties. By dividing the study population in ordered fractions, we provide a detailed picture of the underlying probability distribution and are able to detect mean survival differences across groups that are often undetected by other state-of-the-art methods. Results from a simulation study show good performance of our proposed estimation strategy, and support the idea that mean survival can be more accurately estimated in some fractions of the population. While estimation of the mean survival time presents several difficulties in the presence of censoring, that same quantity can be precisely estimated for certain fractions of population. In the analysis of survival data from a bone marrow transplant study, our method detected differences in mean survival between given transplants for certain fractions of population, while those differences were overlooked when using restricted mean estimates instead.

Acknowledgments: This work was partially supported by the KID funding doctoral grant from Karolinska Institutet

References

Avalos, B.R., Klein, J.L., Kapoor, N., Tutschka, P.J., et al. (1993). Preparation for Marrow Transplantation in Hodgkin's and non-Hodgkin's Lymphoma Using Bu/CY. *Bone Marrow Transplantation*, **12**, 133 – 138.

Irwin, J. O. (1949). The Standard Error of an Estimate of Expectation of Life, with Special Reference to Expectation of Tumourless Life in Experiments with Mice. *Journal of Hygiene*, **47**, 188 – 189.

Nair, V.N. (1984). Confidence Bands for Survival Functions with Censored Data: A Comparative Study *Technometrics*, **26**, 265 – 275.

Paltiel A.D., Freedberg K.A., Scott C.A., Schackman B.R., Losina E. et al. (2009). HIV Preexposure Prophylaxis in the United States: Impact on Lifetime Infection Risk, Clinical Outcomes, and Cost-Effectiveness. *Clinical Infectious Diseases*, **48**, 806 – 815.

Zhao Y., Zeng D., Socinski M.A., Kosorok M.R. (2011). Reinforcement Learning Strategies for Clinical Trials in non-Small Cell Lung Cancer. *Biometrics*, **67**, 1422 – 1433.

Modelling atmospheric dispersion: Uncertainty management of height and strength of the release

Ali S. Gargoum [1]

[1] UAE University, United Arab Emirates.

E-mail for correspondence: alig@uaeu.ac.ae

Abstract: One of the important pieces of information that are needed to inform very early decision-making immediately after a nuclear or chemical accident is how experts (plant designers and safety engineers) believe source emission will develop over time. To address this issue it is essential first to code as much expert opinion as possible about the types and profiles of release, and secondly, to modify these opinions - which are often very uncertain- in the light of any observations which do become available. In this article we present an uncertainty management procedure for the height release at source which is a key parameter in modeling the subsequent dispersal of contamination (e.g. the higher the release goes, the faster it spreads). When setting the initial parameters of the model, it is difficult to estimate the height of the release and this will obviously affects the consequences. This procedure reduces the risk of setting an erroneous height value by running mixed model. That is, we include several models in our analysis, each with a different release height. The Bayesian methodology assigns probabilities to each model representing its relative likelihood and updates these probabilities in the light of monitoring data. This has the effect that the data gives most weight to the most likely model, and thus models which consistently perform badly can be discarded. An illustration, based on running the sequential learning with an atmospheric dispersion model, is given on a real site under real atmospheric conditions but with simulated observational data.

Keywords: Dispersion models; Puff models; Bayesian forecasting.

1 Introduction

Atmospheric deterministic dispersion models are widely used for forecasting toxic contamination and obtaining results in real time with varying degrees of accuracy. The large degree of uncertainty associated with their

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

predictions is one of the most significant problems. These uncertainties may lead to a destabilization of the decision process when environmental survey results disagree with the model results.

This article is based on a Bayesian statistical model described in Smith and French (1993). The statistical model is carried out within a Bayesian paradigm Box and Taio (1973), French (1986) and West and Harrison (1997).

2 Atmospheric dispersion models: Puff Models

The basic principle for a computational puff model for predictions of atmospheric dispersion is the simulation of the continuous emission from the source by a proper distribution of discrete sequence of small puffs of different sizes Leelossy (2014). These are released at regular time intervals and then diffuse and disperse independently.

A Bayesian model based on generalization of the puff model has been adopted both to combine the puff model with expert judgments and monitoring data, and to provide an evaluation of the uncertainty associated with the forecasts.

3 The statistical model

Following Smith and Frensh (1993), the puffs are indexed such as puff i is $Q(i)$, i.e. $Q(i)$ is uncertain quantity which represents the total number of contaminated particles under the i^{th} puff. We define $Q_t = (Q(1), \ldots, Q(t))^T$ which approximates the release profile of the source term.

The spatial concentration of contamination from the i^th puff at time t and location s is given by $F_t(i, s)Q(i), the$, the stochastic multiplier. This multiplier determines how that emission is distributed over the space and time. It is a proportional of the total contaminated particles under the i^th puff at site s and time t. Typically, $F_t(i, s)Q(i).$ is a complicated deterministic function of parameters, themselves calculated from uncertain meteorological inputs. One of the simplest of such dispersal models is a Gaussian puff Pasler-Sauer (1985), which sets

$$F_t(., s) = \frac{1}{(2\pi)^{3/2}\sigma_t(1)\sigma_t(2)\sigma_t(3)} \exp\{-\frac{1}{2}[\Sigma_{j=1}^{2} \frac{(s_j - u_t(j))^2}{\sigma_t^2(j)} + \frac{(s_3 - h)^2}{\sigma_t^2(3)}]\}$$

where $(u(1), u(2))$ is a wind velocity vector possibly depending on t, and h is the height of the emission. The radial growth of puffs during dispersion as a result of internal turbulence is described by the parameters $(\sigma_t(1)\sigma_t(2))$ and $\sigma_t(3)$ which denote puff sizes in horizontal and vertical directions respectively.

Initially the stochastic multipliers $F_t(i, s)$ are assumed to be known, and we only consider uncertainty on masses. However, in the rest of the paper we will address uncertainty on certain parameters of $F_t(i, s)$ such as the release height of the emission. Instant concentrations at monitoring sites are linear functions of Q_t. Let $Y(t, s)$ denote an observation taken under some overlapping puffs at time t at location s. Here $Y(t, s)$ represents the total number of contaminated particles, i.e. the concentration of contamination which is simply the sum of concentrations of all puffs where the i^{th} puff contributes a proportion $F_t(i, s)$ of its total mass $Q(i)$. Thus $Y(t, s)$ will be a noisy function of the true contamination $\theta(t, s)$ originating from $Q(1), \ldots, Q(t)$. This needs to be stochastically modeled as

$$\theta(t, s) = \Sigma_{i=1}^{t} F_t(i, s) Q(i) + \epsilon(t, s)$$

For simplicity we assume that $\epsilon(t, s)$ are Gaussian with mean zero and variance $V(t, s)$, and $\epsilon(t, s_1)$, $\epsilon(t, s_2)$ are independent of sites s_1, s_2. Simply $(Y(t, s)|\theta(t, s))$ is defined to have a Gaussian distribution with mean $\theta(t, s)$ and a fixed variance $V(t, s)$ where $\theta(t, s)$ is assumed to be known and represents the observation and modelling error.
Now conditioning on everything else other than masses, the model provides elegant algorithms to: update distributions of the source term in time; predict contamination over space and time; and hence obtain predictive distributions of data and also to admit data assimilation. However, in practice many of the variables conditioned on will be unknown. For example, we may be uncertain about parameter like the release height and we know that this parameter is very important in the stochastic multipliers $F_t(i, s)$. This problem will be discussed in the next section.

4 Uncertainty about the release height

As we stated in the previous section, the true source emissions $Q(1), Q(2), \ldots,$ where $Q(i)$ denotes the mass of contamination under the i^{th} emitted puff, can be modelled as a dynamic linear model(DLM) West and Harrison (1997), with state $\theta_t = Q_t^T$. Explicitly,

$$Q_t | Q_{t-1} \sim N(GQ_{t-1}, W).$$

Where G (observational matrix) and W (evolution matrix) are fixed square matrices.
The DLM can be combined with a puff model to estimate the source term profile and predict the contamination spread as long as we believe the model.
The height of release at source is a key parameter in the subsequent dispersal of contamination. (e.g. the higher the release goes, the faster it spreads) . When setting the initial parameters of the model, it is difficult to estimate the release height and this will obviously effect the consequences Gargoum(2001). Here we suggest one solution to this problem, which reduces

the risk of setting an erroneous height value, by running mixed models. That is we include several models in our analysis, each with a different release height. The Bayesian methodology assigns probabilities to each model representing its relative likelihood and updates these probabilities in the light of monitoring data. This has the effect that the data give most weight to the most likely model, and thus models which consistently perform badly can be discarded.

4.1 The Bayesian updating algorithm

Suppose that we have m dispersal models $M^{(h_i)}, (i = 1, \ldots, m)$, where the dispersal algorithms were the same but whose parameters were different (e.g. the initial height parameter h of source emission). Suppose that one of the models (as yet uncertain to us) is assumed to be true. Let

$$p(M^{(h_i)}(\text{is true})) = \mathrm{p}_{\mathrm{t-1}}^{(\mathrm{h_i})} = \mathrm{p}(\mathrm{M}_{\mathrm{t-1}}^{(\mathrm{h_i})}|\mathrm{D}_{\mathrm{t-1}}) = \mathrm{\pi_i}.$$

Where $\Sigma_{i=1}^{m} = 1, \pi_i > 0, 1 \leq i \leq m$ and D_{t-1} represents data available up to time $t-1$. Here h_i and π_i are chosen to give approximations in the prior of the release height.Then the probability of an event A (e.g. A might be an observation of contamination at site s lies in the interval $[a,\, b]$ is given by

$$p(A) = \Sigma_{i=1}^{m} \pi_i p_i(A).$$

Where $p(A)$ is the probability attributed to the event A by the model $M^{(h_i)}, i = 1, \ldots, m$. Note that if $\theta(t, s)$ is the density of contamination at site$s$ and time t, then

$$E[\theta(t, s)] = \Sigma_{i=1}^{m} \pi_i E_i[\theta(t, s)]$$

where $E_i[\theta(t, s)$ is the expected contamination under model $M^{(h_i)}, (i = 1, \ldots, m)$, at site s and time t.

The Bayesian algorithm allows the updating of $\pi = (\pi_1, \ldots, \pi_m)$ in a simple manner, following the principles of parallel processing of multi-process models, Class I, as introduced in Harrison and Stevens (1976) and West and Harrison (1997). Suppose an event $B = \{Y = y\}$ has a density value under model $M^{(h_i)}, (i = 1, \ldots, m)$ of $p_i(y)$. Then, for an arbitrary event$A$,

$$p(A \cap B) = \Sigma_{i=1}^{m} \pi_i p_i(A \cap B)$$

where p is the probability of the combined model and p_i is the probability coming from $M^{(h_i)}, (i = 1, \ldots, m)$. So

$$p(A|Y = y)p(y) = \Sigma_{i=1}^{m} \pi_i p_i(A|Y = y)p_i(y)$$

where

$$p(y) = p(B) = \Sigma \pi_i p_i(y)$$

This implies that, given $Y = y$, our updated probability $p^*(A) = p(A|Y = y)$ A is given by

$$p^*(A) = \Sigma\pi_i^* p_i^*(A)$$

where
$pi_i^* = \frac{p_i(y)\pi_i}{\Sigma_{j=1}^m p_j(y)\pi_j}$
This procedure can be implemented where certain heights are taken as representing the a priori plausible range of height values and their initial probabilities are assigned, then the posterior probabilities for these heights are calculated each with its corresponding expected dispersal. This implementation is briefly discussed in the following subsection.

4.2 Illustration

Bayesian updating of dispersal contamination, based on running the sequential learning with RIMPUFF atmospheric dispersion model, is used on a real site (Lundtofte Nord1) under real atmospheric conditions but with simulated observational data.
Assuming that $\mathcal{A} = \{h_1 = 200m, h_2 = 400m, h_3 = 600m\}$ are taken as representing the a priori plausible range of height values their initial probabilities are assigned as $\pi_i = \frac{1}{3}, \quad (i = 1, 2, 3)$. The posterior probabilities for the three heights were calculated. The models were rather different: Model 3 with height $600m$ had a higher posterior probability = 0.51 at the time of interest compared to the other two models, 0.16 for model 1 and 0.32 for model 2. Note that if one of the models had a very high posterior probability at the time of interest, then it could be adopted alone for inference. Otherwise, the full unconditional mixture would be used for this purpose.

5 Conclusion

Expert judgments about key parameters, such as source height and wind direction, in dispersal models are extremely informative and can be accommodated into Bayesian uncertainty management in puff models. Even where the parameters of the profile are extremely uncertain a priori, the forecasting systems quickly fit to their empirically values. To improve the model effectiveness and manage uncertainty about the release height we proposed to include several models in our analysis to reflect potential errors in this parameter. Hence, the model as described in this paper estimates and provides distributions for source term and release height at the source. Simply, we place a discrete distribution over a set of values of the height, using expert judgments to assign prior probabilities over the set of values. The dispersal model is run for each vector of parameter values for the set of observations that had been taken. Bayes rule then allowed us to update the probabilities and as new information arrive we could update again and so on.

References

Box, G.E.P. and Taio, G.C. (1973). *Bayesian inference in statistical analysis.* Massachusetts: Addison-Wesley.

French, S. (1986). *Decision theory: An introduction to the mathematics of rationality.* Chichester: Ellis Horwood.

Gargoum, A. S. (2001), Use of Bayesian dynamic models for updating estimates of contaminated material *Environmetrics* **12**, issue 8,775 – 783

Harrison, P.J. and Stevens, C.F. (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Series B* , **38**, 205 – 247.

Leelossy, A. , Molnar, F., Lzsak, F., Havasi, A., Lagzi, I., and Meszaros, R. (2014). Dispersion modeling of air pollutants in the atmosphere: a review. *Cent. Eur. J. Geosci*, **6(3)**, 257 – 278.

Pasler-Sauer, J. (1985). Atmospheric dispersion in accident consequence assessments. Present modelling, future needs and comparative calculations. *Proceedings of the workshop on methods for assessing the off-site consequences of nuclear accidents.* Luxembourg, CEC, EUR-Report.

Smith, J.Q. and French, S. (1993). Bayesian updating of atmospheric dispersion models for use after an accidental release of radioactivity. *The Statistician* **42**, 501 – 511.

West, M. and Harrison, P.J. (1997). *Bayesian forecasting and dynamic linear models.* Springer-Verlag.

Comparison of MCMC approaches with an application to volcano earthquake processes

Anastasia Ignatieva[1], Andrew F. Bell[2], Bruce J. Worton[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK
[2] School of GeoSciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: Bruce.Worton@ed.ac.uk

Abstract: In this paper we consider statistical modelling of volcanic earthquake data. In particular, we investigate the use of Bayesian analysis with Markov Chain Monte Carlo (MCMC) to estimate the parameters of point process models, and make inferences on the models, applied to data collected from the Tungurahua volcano in Ecuador.

Keywords: Bayesian modelling; Eruption forecasting; Point processes.

1 Introduction

This paper aims to use statistical modelling to describe the occurrence of volcanic earthquakes. The main approach taken is that of using Bayesian analysis with Markov Chain Monte Carlo (MCMC) to fit point process models to the available data, collected from the Tungurahua volcano in Ecuador.

2 Dataset and modelling

This dataset was recorded in July 2013, and consists of a series of event times which were picked from a stretch of seismic data to identify the individual earthquakes. The dataset was examined in a study by Bell et al. (2018).

The events started at 6:00 on 13 July, and the eruption occurred at 11:46 on 14 July. The event rate grew increasingly up until eruption. Plots of the data show that the event rate grows at an increasing rate up to the

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

eruption, with the inter-spike interval (ISI) duration changing from over 10 minutes to below 30 seconds. The ISIs are “quasi-periodic”, being more regular than would be seen if the events followed a Poisson process, and thus not independent (Bell et al., 2018).

Applying a material failure approach to describe the physical processes leading a volcanic system to an eruption, the accelerating rate of earthquakes is described by a power law relationship (Bell et al., 2018):

$$\lambda(t) = k(t_f - t)^{-p},$$

where k is a constant (related to the amplitude of the signal), t_f is the time of eruption, and $p = \frac{1}{a-1}$ is a parameter describing the non-linearity of acceleration. At time t_f , the rate becomes instantaneously infinite, representing the eruption (Bell et al., 2018). In the model, $\lambda(t)$ is the intensity used in the inhomogeneous gamma (IG; parameter α) point process.

Details of the MCMC implementation in PyMC3 were investigated, including the sampling method used and the initialisation process. Attributes of the MCMC chain, such as convergence, were examined. Posterior checks were performed using simulated data, to sense check whether the model appears appropriate. The fit of the model was assessed further using statistical methods.

The MCMC approaches considered included:

- No-U-Turn sampler (Hoffman and Gelman, 2014);
- Metropolis;
- Slice sampling (Neal, 2003).

Alternative models were also investigated, and their fits compared to that of the given inhomogeneous gamma model: inhomogeneous Poisson (IP), inhomogeneous inverse Gaussian (IIG) and inhomogeneous Weibull (IW) models.

3 Results

Figure 1 gives comparison of the MCMC trace plots for the methods for the IG model. Figures 2 and 3 show posterior plots using MCMC sampled values for the IG model. A Kolmogorov-Smirnov goodness of fit approach (Barbieri et al., 2001; Ogata, 1988) gives an effective method of comparison of various possible models.

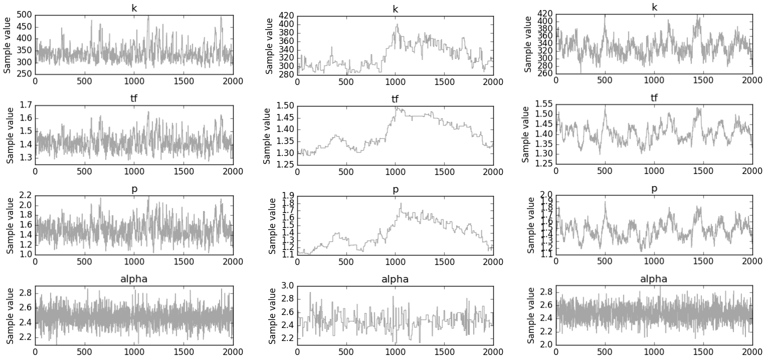


FIGURE 1. Trace plots for 2,000 iterations: IG model. No-U-Turn sampler (left), Metropolis (middle), Slice (right).

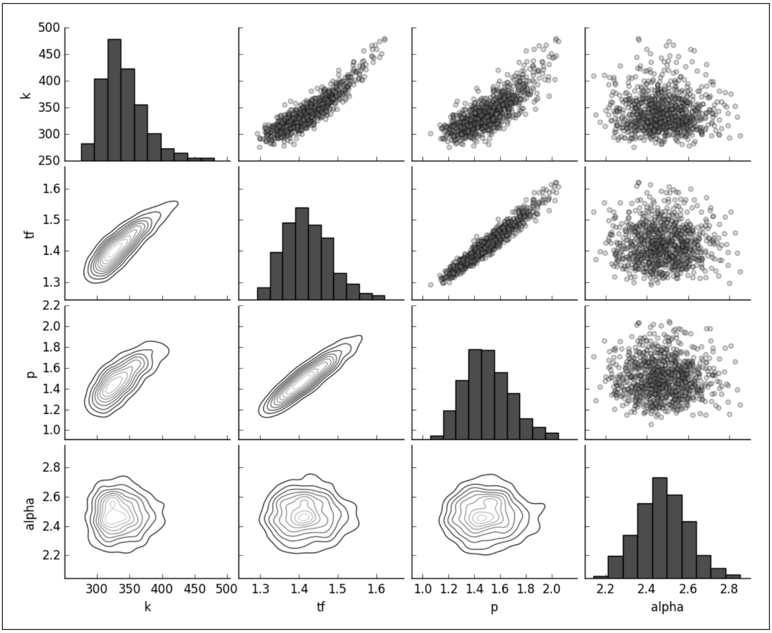


FIGURE 2. Posterior plots using MCMC sampled values: IG model.

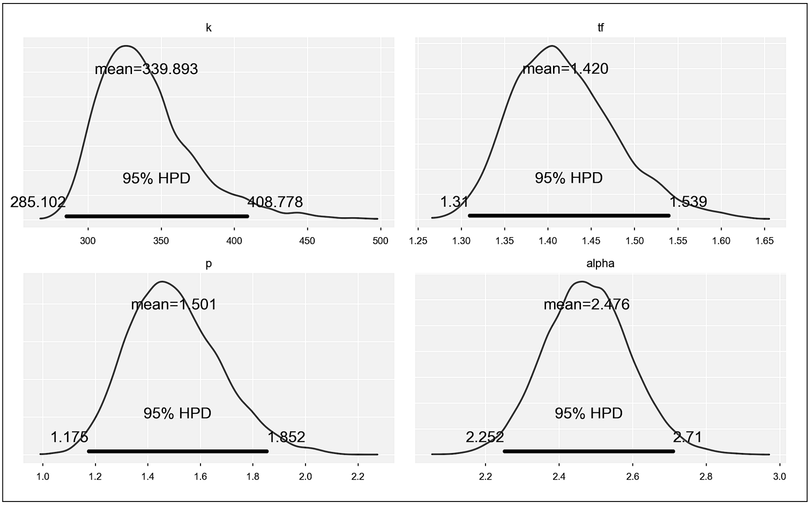


FIGURE 3. Posteriors with HPDs: IG model.

4 Conclusions

An IG model was found to produce satisfactory results. It was demonstrated that the MCMC chain appears to converge to the correct stationary distribution, providing reasonable posterior estimates. From review of simulated data, and Q-Q and K-S plots, it was found that the IG model fits the July 2013 data very well. A small number of outliers (around 5% of the data) was noted, and found to correspond to spikes with long preceding ISIs. Some lack of fit was also found in the middle quantiles of the K-S plot, however this only slightly breached the 95% error bounds.

Acknowledgments: We are particularly grateful to The Instituto Geosico of the Escuela Politecnica Nacional (IGEPN) of Ecuador for all their hard monitoring work and providing the data.

References

Barbieri, R., Quirk, M.C., Frank, L.M., Wilson, M.A., and Brown, E.N. (2001). Construction and analysis of non-Poisson stimulus-response models of neural spiking activity. *Journal of Neuroscience Methods*, **105**, 25 – 37.

Bell, A.F., Naylor, M., Hernandez, S., Main, I.G., Gaunt, H.E., Mothes, P., and Ruiz, M. (2018). Volcanic eruption forecasts from accelerating rates of drumbeat long-period earthquakes. *Geophysical Research Letters*, **45**, 1339 – 1348.

Hoffman, M.D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593 – 1623.

Neal, R.M. (2003). Slice sampling. *Annals of Statistics*, **31**, 705 – 741.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, **83**, 9 – 27.

A risk calculator to inform prostate cancer diagnosis in Ireland

Amirhossein Jalali[1], Robert W Foley[1], Robert M Maweni[1], Keefe Murphy[2 3], Dara J Lundon[1], Thomas Lynch[4], Richard Power[5], Frank O'Brien[6 7], Kieran J O'Malley[8], David J Galvin[8 9], Garrett C Durkan[10 11], T Brendan Murphy[2 3], R William Watson[1]

[1] UCD School of Medicine, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland.
[2] UCD School of Mathematics and Statistics, University College Dublin, Dublin, Ireland.
[3] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland.
[4] Department of Urology, St. James University Hospital, Dublin, Ireland.
[5] Department of Urology, Beaumont Hospital, Dublin, Ireland.
[6] Department of Urology, University Hospital Waterford, Waterford, Ireland.
[7] Department of Urology, Cork University Hospital, Cork, Ireland.
[8] Department of Urology, Mater Misericordiae University Hospital, Dublin, Ireland.
[9] Department of Urology, St Vincents University Hospital, Dublin, Ireland.
[10] Department of Urology, University Hospital Galway, Galway, Ireland.
[11] Department of Urology, University Hospital Limerick, Limerick, Ireland.

E-mail for correspondence: amir.jalali@ucd.ie

Abstract: Prostate cancer (PCa) represents a significant healthcare problem due to the dilemmas associated with its detection and treatment especially with the projected increase in its incidence in Ireland and internationally. The critical clinical question driven by the urologist and patients is the need for a biopsy. An Irish prostate cancer risk calculator created from a national collection of patients can allow for individualised risk stratification and can be used to improve clinical decision making in Irish men under investigation for PCa. Relative statistical approaches using the current clinical parameters considered building a risk calculator based on the Irish dataset which aims to inform clinicians and patients as to the need for a biopsy to diagnose PCa. The use of this risk calculator will impact on the patients' outcome and quality of life but also alleviate the pressures on our already overburdened healthcare sector by reducing the need for biopsies.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

Keywords: Risk calculator; Logistic regression model; Prostate cancer; R shiny.

1 Introduction

Patients and clinicians are faced with the dilemmas associated with the detection and treatment of Prostate cancer (PCa). One such dilemma is in the early stages of diagnosis when men are referred by their GP for suspicion of prostate cancer due to their an elevated Prostate-specific antigen (PSA) value or suspicious Digital rectal examination (DRE), but it is not clear if they need a biopsy. This is because PSA is not specific for PCa which has led to the overdiagnosis and treatment of disease exposing men to unnecessary biopsies, worry about their diagnosis and treatment impacts on their quality of life. Accurate risk stratification of patients before biopsy would help to reduce overdiagnosis and lead to better clinical decision making. The European Randomised Study of Screening for Prostate Cancer (ERSPC) and the Prostate Cancer Prevention Trial (PCPT) are two well-known international risk calculators available for diagnosis of PCa. They have been tested in an Irish population and proved to be beneficial; however, we hypothesis that the predictive accuracy will be significantly improved when built with an Irish population. This study aims to build an Irish risk calculator and compare to the PCPT risk calculator.

2 Methods

A national dataset including the routinely used clinical information of 4801 patients from the eight Irish tertiary referral rapid access clinical centres were collected. A risk calculator for the diagnosis of PCa (and high-grade PCa) was created using a logistic regression model including linear and non-linear effects of components such as age, digital rectal examination, family history of PCa, prior negative biopsy and PSA level. The calibration curve is used to assess whether the models are well calibrated, and models are validated using Cross-validation.

The discriminate ability of the model was compared with the current biomarker indicator (PSA) and PCPT risk calculator using various graphical and numerical performance outcome summaries. The Receiver operating characteristic (ROC) curves, decision curve analysis are standard graphical tools, and sensitivity, specificity, Positive predictive value (PPV), Negative predictive value (NPV), Youden index and Area Under the ROC Curve (AUC) are available numerical summaries. The visual and numerical comparisons for predicting PCa are represented in Figure 1 using the ROC curve and decision curve, and Table 1 using AUC, where they show the superiority of the proposed risk calculator to the current biomarker indicator (PSA) and PCPT risk calculator.

TABLE 1. Area Under the ROC Curve, 95% confidence interval and p-value of Delong test to compare three approaches.

Models	AUC	95% CI	p-value
PSA	0.59	0.58-0.61	<0.001
PCPT	0.63	0.62-0.65	<0.001
IPRC	0.67	0.66-0.69	-

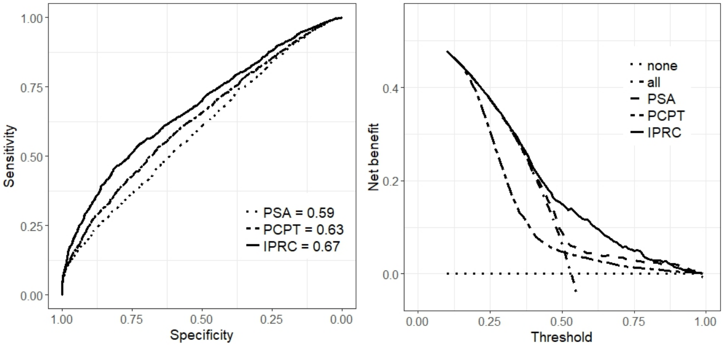


FIGURE 1. ROC curve (at the left) and decision curve (at the right) for predicting prostate cancer.

The diagnostic ability of the IPRC is compared with the PSA blood test in Figure 2. It shows a 13% reduction in unnecessary biopsy and 6% in total biopsies.

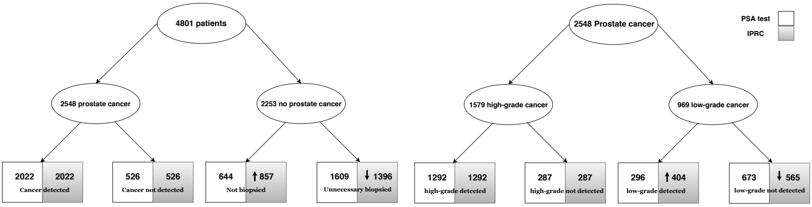
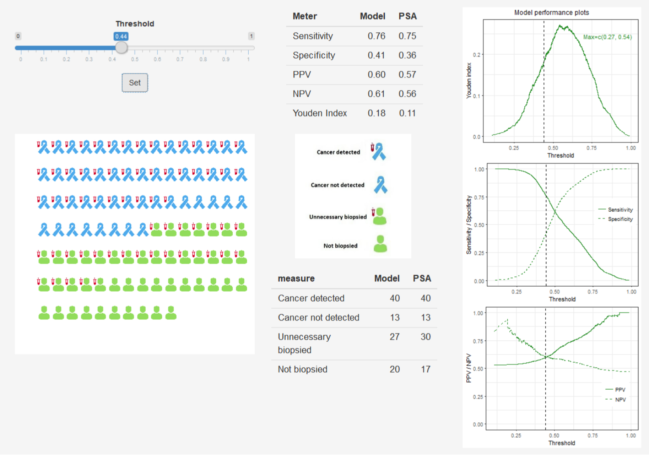


FIGURE 2. The diagnostic ability of PSA compared to IPRC.

In the year 2010, 14291 men went for a prostate biopsy, where only 3287 (23%) were diagnosed with cancer (Burns 2010). Our test could have reduced the number of biopsies by 858. This reduction in annual biopsies will also result in a significant economic savings of about €371,000 excluding the cost of dealing with biopsy complications.

The proposed model calculates the risk of having prostate cancer as a probability; however, an optimal threshold needed to be chosen to make the best clinical decision. The selection of this threshold could be challenging as it depends on a trade-off between a more sensitive test or a more specific test. For this reason, an interactive Shiny application is created to be presented to clinicians and decision makers which combine the graphical and numerical summarises to convey the result of this risk calculator in the most translated way. A screenshot of this application is given in Figure 3.



Meter	Model	PSA
Sensitivity	0.76	0.75
Specificity	0.41	0.36
PPV	0.60	0.57
NPV	0.61	0.56
Youden Index	0.18	0.11

measure	Model	PSA
Cancer detected	40	40
Cancer not detected	13	13
Unnecessary biopsied	27	30
Not biopsied	20	17

FIGURE 3. A Visualisation tool of the threshold selection for the prostate cancer model.

3 Discussion

An Irish PCa risk calculator created from a national collection of patients in Ireland can allow for individualised risk stratification and can be used to improve clinical decision making in Irish men under investigation for PCa. The development of the Irish Prostate Cancer Risk Calculator has shown a significant reduction of unnecessary biopsies, without affecting prostate cancer detection or significant disease, outperforming the current approach. This will reduce the number of men requiring a biopsy and their exposure to its side effects, as well as lightening the pressure on our already overburdened healthcare system and have economic savings.
We are currently investigating the integration of current and novel biomarkers to increase the sensitivity and specificity of the risk calculator, in order to further reduce the number of men needing a biopsy in Ireland.

References

Ankerst, Donna P., et al. (2014). Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer. *Urology*, **83.6**, 1362–1368.

Burns, Richal M., et al. (2017). The burden of healthcare costs associated with prostate cancer in Ireland. *Global & Regional Health Technology Assessment.* grhta-5000249.

DeLong, Elizabeth R., et al. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Schrder, Fritz H., et al. (1995). European randomized study of screening for prostate cancer. Progress report of Antwerp and Rotterdam pilot studies. *Cancer*, **76.1**, 129-134.

Vickers, Andrew J., and Elena B. Elkin. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, **26.6**, 565-574.

Optimized Road Closure Control in a Triathlon Event

Zonghuan(Jason) Li[1], Mason Chen[2]

[1] San Mateo High School, San Mateo, CA
[2] Stanford Online High School

E-mail for correspondence: `jasonlzh2001@gmail.com`

Abstract: The Triathlon is an Olympic event that affects the entire city. This STEM paper is to manage the road closure time of a triathlon event in any City who may have a Triathlon event. In order to quantify the Road Closure Impact to meet 5.5 hours of closure time, authors have developed three different models: (1) Ideal Scenario, (2) Practical Scenario, and (3) Potential Scenario. In the first Ideal Scenario, authors applied Minitab descriptive statistics and box plots to optimize the event schedule and sequence among three Triathlon sports (Swim, Run, and Bike). In the second Practical Scenario, authors have managed the risks such as: pre-event preparation time, award ceremony time, dependency on Gender, Age, Category factors, 2000+ players, Safety/Health, etc. In the last Potential Scenario, we have further optimized the Zone Close and Open period. We can minimize the Zone wasted usage (waiting for players) based on the Worst Case Scenario. The biggest selling point of this paper is not just on the Closure Cost reduction, but also on the Safety/Health protection as well as promoting customer service and local business. This event may be completed on the event day, but our excellent customer service can attract more visitors to join the next city event.

Keywords: Triathlon, Descriptive Statistics, Minitab, Model, Non-Parametric

1 Introduction

Triathlon is a multisport event consisting of sequential swim, cycle, and run disciplines performed over a variety of distances1. The Purpose of this paper is to build an accurate model to manage a triathlon event in a city. There is one concern / restriction; the road closure time for the triathlon event has to be less than or equal to 5.5 hours in order to minimize the impact to the local people and business. We used data from 2017 HiMCM

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

competition. In order to make that happen, we need to build a model that could tell us critical information like when we should hold the event, how are we going to arrange the sequence among the three sports etc. We will apply Minitab descriptive statistics, and three models to optimize the Road Closure time and area.

2 Building three Models

Three different models were constructed for three different scenarios; one being the ideal scenario, assuming that no unexpected events such as injuries, weather concerns occur. Therefore, the first ideal model may not be practical and needs certain improvements.
The second model is the practical scenario, in which we consider the possible risks such as injuries, effect on local traffic and life overall, weather concerns, age dependency etc. We will take those into account and consider possible changes to the arrangement of the event in a practical sense.
The third model is the potential model, where we will figure out smarter ways to further optimize road closure time by sharing the streets between two events simultaneously to ensure we can finish the event and possibly add introduction and award ceremony and still being able to hold the road closure time to under 5.5 hours.
In these models, Minitab software and descriptive statistics will be applied.

3 Results

There are so many factors and constraints hidden in the project. There are so many alternative solutions which can resolve some problems if not all. There is no perfect scenario and all decisions are down to risk assessment and risk management. Based on the hypotheses, we run the raw data and build three models accordingly. The objective is to find a model which can meet 5.5 hours road closure time and also take care of safety and health concerns during the events.

3.1 Model 1: Ideal Scenario

This is an ideal model, assuming that everything runs perfectly and smoothly. First, we will apply one rule; we won?t invite any player who cannot finish the triathlon within 4:16:46 based on upper outlier criteria in order to control the total event time within 5.5 hours as shown in Figure 1. We used upper outlier criteria statistically in order to set a reasonable bar fairly.
Upper Outlier Criteria = Q3 -1.5* IQR IQR= Q3 - Q1 Q3= 75th Percentile Q1= 25th Percentile
The reason why we chose a box-plot is because it shows us the distribution with many outliers. After applied Rule No.1, we re-plot the Box-plot as

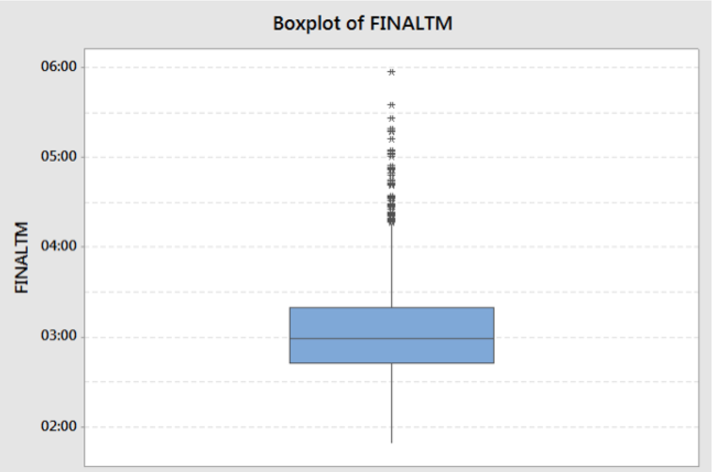


FIGURE 1. Box-plot for total time to finish the triathlon

shown in Figure 2. We have excluded all the players with record beyond the upper outlier limit at 04:16:46. Total 59 players (2.8 percent) got impacted. The new distribution looks a normal distribution.

The next step is to determine how will we arrange the sport sequence. We need to make sure there is no significant overlap between the Bike and Run for safety reason if any collision risk happened between Bike Player and Run Player. We will first look at the distribution of three sports. As shown in Figure 3 Box-plot, we have observed the individual distribution of three sports. The Bike event has both the highest mean and the widest range.

This boxplot shows the time distribution of the three sports, as swim takes the shortest time and bike takes the longest. We should start Swim Event since the event is happened in the Ocean. We should take shorter Run event first over the longer and wider Bike event. Also, Bike event will take larger closure area. Therefore, the Bike event should be after the Run event.

We will also apply the second rule to disqualify any player who cannot complete both Swim and Run event within 01:48:32 based on the upper outlier criteria. Similar to first rule, the second criteria is necessary in order to ensure the entire road closure time is within 5.5 hours.

3.2 Model 2: Practical Scenario

This is the practical model that takes in more consideration on whether we should reserve any event preparation time in the beginning and hold an award ceremony in the end; how to handle players impacted by the rules listed in previous model 1? And also should we separate male and female participants to shorten the total road closure time?

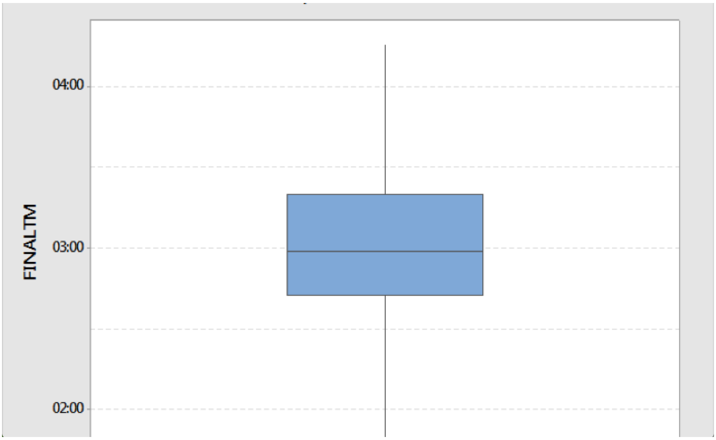


FIGURE 2. Box-plot for total time to finish triathlon after excluding outliers.

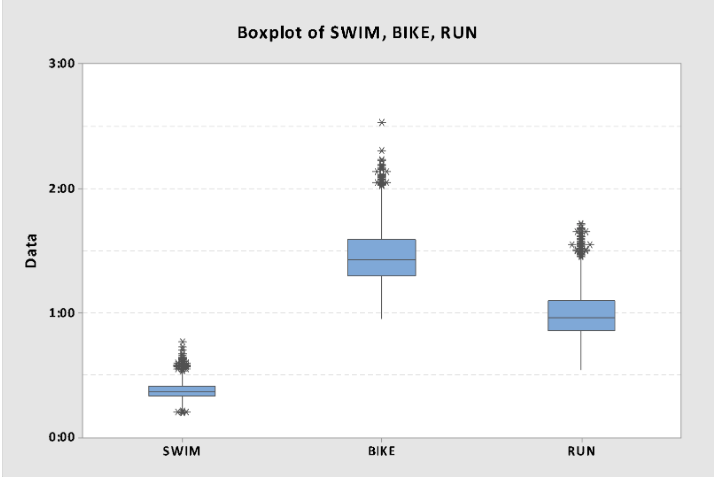
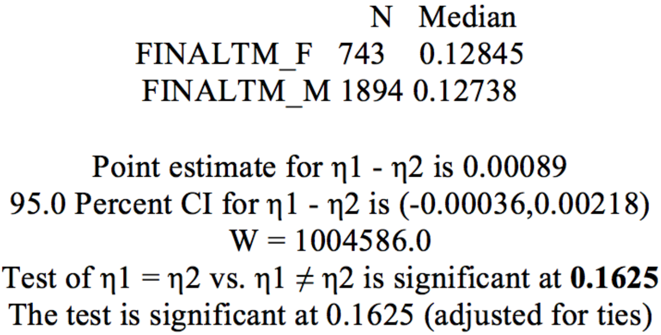


FIGURE 3. Boxplot on distribution of swim, bike and run.



```
                    N   Median
FINALTM_F   743    0.12845
FINALTM_M 1894 0.12738

Point estimate for η1 - η2 is 0.00089
95.0 Percent CI for η1 - η2 is (-0.00036,0.00218)
W = 1004586.0
Test of η1 = η2 vs. η1 ≠ η2 is significant at 0.1625
The test is significant at 0.1625 (adjusted for ties)
```

FIGURE 4. Results from Minitab software.

We constructed the Non-Parametric Mann-Whitney Test for Median as shown in figure 4 since our raw data distribution is not normal and we have too many outliers which will distort the mean location. The reason why it is not normal is because pacing strategies during triathlon are highly influenced by distance and discipline.
This shows that there is no significance (p value significance is above 0.05) on separating genders in the Triathlon event.

3.3 Model 3: Potential Scenario

Model 3 is the potential model that considers: How do we further optimize road closure time? Should we place these senior players in the back line when starting the Swim event? Can we claim any age dependency?
In order to answer these questions, we constructed a scatterplot and regression model by Minitab software as shown in Figure 6.
We chose scatterplot because it may show us any correlation between the total time and age. As shown in the graph, the R-Square is only at 0 percent (random pattern), which means Age is not a factor that affects the participants? finish time.

4 Conclusion

We will start Triathlon event with swim, then run, and bike based on the cycle time distribution and safety consideration. In order to meet 5.5 hours of road closure time requirement, we need to apply 2 disqualify rules to shorten the event duration. We will also hold the award ceremony in the end, and we will not separate male and female players or by age factor, as our model shows that there is no significance in doing so. We will take safety risk as top priority.

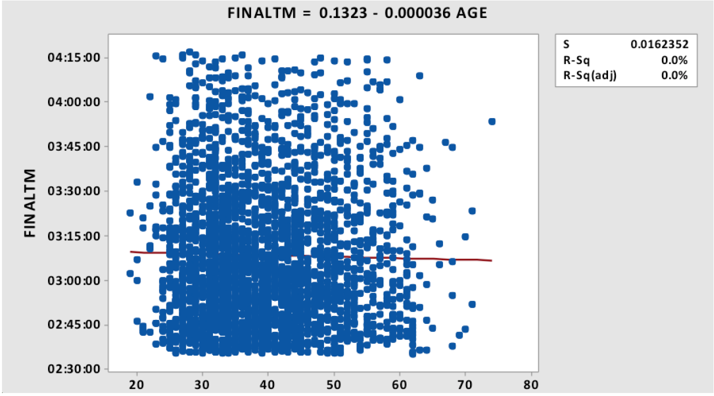


FIGURE 5. Scatterplot on age dependency.

Acknowledgments: Special Thanks to Dr. Charles Chen for Statistics Guidance. We would also like to thank our moms Fenglin Cao and Chia Lin for supporting our STEM Projects.

References

Wu, Sam SX, Jeremiah J. Peiffer, Jeanick Brisswalter, Kazunori Nosaka, and Chris R (16 Sept. 2014. Web. 23 July 2017). *Factors Influencing Pacing in Triathlon.*. Open Access Journal of Sports Medicine. Dove Medical Press.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Henderson, C.R. (1973). Sire evaluation and genetic trends. In: *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. L. Lush*, Champaign, Illinois, 10 – 41.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619 – 678.

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

Statistical JAVA Gaming Simulation

Timothy Liu[1], Joseph Jang[1], Mason Chen[2]

[1] Morrill Learning Center, California

E-mail for correspondence: timothys.new.email@gmail.com

Abstract: Our team has designed a special game on which we apply Statistics, Probability, and Java to simulate each game move and predict the winning scenario. We applied binomial probability distribution to build a predictive model that could simulate the gaming sequence between two players. The sample size was determined based on two hypotheses: (1) playing sequence and (2) winning patterns. In this project, we identified four winning patterns and used Java to code these patterns and determine the gaming sequence and consequence based on conditional probability. The Java results were then compared to the Predictive Model to conduct objective root cause analysis for further improvement and optimization. Human behavior was also considered to study the beginner level to the more advanced level. Based on the 2- Proportions Tests, team has achieved $> 95\%$ confidence that the optimum model can accurately predict the gaming sequence and winning probability which are verified and validated by Java simulation. Team has been through a systematic Six Sigma DMAIC process, and typical Team Building Cycle (Forming, Storming, Norming, and Performing). This is a good STEM Project for teaching kids on learning and applying Statistics, Java Programming, Problem Solving, and Team Building Dynamics.

Keywords: JAVA; Statistics; Probability; Predictive Modeling.

1 Section 1

The purpose of this project is to design and implement a program that can be used in the future as basis for the development of an AI for use in medical research. The use of simulation has been theorized by scientists Toupo and Strogatz to predict evolution in nature, and by Fu and Hauert to predict changes in social behavior. Because there is an increasing amount of medical data available, we decided to design a JAVA program that could later use this data for medical purposes. By setting different conditions on the game we designed, we were able to uncover different player patterns.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

1.1 Section 1.1

The 3-chips game we designed has 4 rules:
1. There are three groups of chips with different number and color in each group as the initial game condition (e.g. 10 Red chips, 8 Yellow chips, 6 Green chips). The initial condition can be randomly assigned as long as there is NO identical number of chips in any two groups such as (X, X, Y).
2. There are two types of players (Player Type A and Player Type B) who will play each other. One player will go first, and then two players will take turns until completed the game. Player Type A is not aware of any game rules. Player Type B is aware of all four game rules.
3. During each round, the player will decide one group (could be Red, Yellow, or Green) and remove at least one chip up to all of the remaining chips from that particular group.
4. The player picked the overall last chip will be the loser of the game.
We have brainstormed two hypothesis
1. Can we use basic probability to predict the winning probability among players who know the above rules or who don't know any of the rules?
2. Will the playing sequence (who goes first) impact the winning probability?
In order to test these to hypotheses, we set up 4 cases: Case I(Player A v Player A), Case II(Player B v Player B), Case III(Player A v Player B), and Case IV(Player B v Player A). Player A doesn't know the winning patterns, but Player B does.

1.2 Section 1.2

In order to save time in collecting our data, we wrote a JAVA program that would play the game for us. After playing the game several times on our own, we discovered some winning patterns that would be the basis for Player B's programming. After running each case 209 times, we recorded and analyzed our results.

2 Section 2

Case I: two Type A players played each other. Among 209 samples: first Player A won 107 times and the second Player A won 102 times. We will conduct 2-sided 1-Proportion Test (not 2-Proportions Test) since all the data was from one Sample. We conducted a Minitab 1-Proportion Test in Table 3, and the Null Hypothesis H0 : Player A Winning Probability= 0.5 (50%). Team used Normal Approximation method to conduct 1-Proportion Z test.
The P-value is $0.729 > 0.05$, which failed to reject the Null Hypothesis. This result has indicated the playing sequence has not made significant

impact on the winning probability between two Type A players.

Case II: two Type B players played each other. Among 209 samples: first Player B won 109 times and the second Player won 100 times. We will conduct 2-sided 1-Proportion Test (not 2-Proportions Test) since all the data was from one Sample.
We conducted a Minitab 1-Proportion Test, and the Null Hypothesis H0 : Player B Winning Probability= 0.5 (50 percent). Team used Normal Approximation method to conduct 1-Proportion Z test.
The P-value is $0.534 > 0.05$, which failed to reject the Null Hypothesis. This result has indicated the playing sequence has not made significant impact on the winning probability between two Type B players.

Case III:Player A (Go First) played with Player B (Go Second). Among 209 samples: Player A only won 7 times and Play B won 202 times. Based on our Case III prediction, we would predict Player A should win 23.3%. Team has conducted 1-Proportion Test and the Null Hypothesis H0 : Player A Winning Probability= 0.233 (23.3%) in Table 4. P-Value is 0.000 and we should reject Null Hypothesis which has indicated our Case III prediction model is not validated through our Java simulation.

Case IV: Player A (Go Second) played with Player B (Go First). Among 209 samples: Player A only won 6 times and Play B won 203 times. Based on our Case IV prediction, we would predict Player A should win 15.7%. Team has conducted 1-Proportion Test and the Null Hypothesis H0 : Player A Winning Probability= 0.157 (15.7%) in Table 6. P-Value is 0.000 and we should reject Null Hypothesis which has indicated our Case III prediction is not validated through our Java simulation. We will address this issue later.

Java results have supported our CASE I and CASE II non-bias result on which player would GO First. There is no significant bias observed regarding the playing sequence would impact the winning probability. However, in Case III and Case IV, our prediction model is not very reliable to predict the Java results. The biggest reason of failing the prediction is that we assumed each game will be completed within four rounds. If both players are very conservative, this assumption will be very questionable. In order to further improve the prediction capability, we will expand current four-round modeling to five-round or six-round to improve our prediction capability.

Team has successfully built a predictive model to simulate the winning probability on four Cases. There is no significant evidence showing the playing sequence would impact the winning result. This result is making sense since we are assuming all events are independent. This independency should be more accurate when we have more chips in the pool. Player B (knowing four rules) has a much higher winning probability (Less than 95% chance) over Player A (playing blindly). Our predictive model can accurately predict the winning probability if we can take 5 or 6 rounds. Team has conducted the sample size calculation in order to draw a statistical conclusion to verify the two hypotheses. Developing a Java programming has significantly reduced our effort to collect data to validate our predictive model.

Acknowledgments: Special Thanks to Dr.Charles Chan and Dr. Ying Huang for their guidance and mentorship through out this project

References

Toupo, D., Strogatz, S. (2015). *Nonlinear dynamics of the rock-paper-scissors game with mutations, Physical Review E.*

Maciejeweski, W., Fu, F., and Hauert, C. (2014). *Evolutionary Game Dynamics in Populations with Heterogenous Structures, PLOS Computational Biology*

Gokhale, D.V.; Kullback, Solomon (1978). *The Information in Contingency Tables. Marcel Dekker. ISBN 0- 824-76698-9*

Montgomery D., Runger G. (2003). *Applied Statistics and Probability for Engineers. ISBN 9780471745891. 2-Proportions Test*, 379 – 383.

Generalized Linear Mixed Effect Models to assess efficacy of a Group-based Parent Skills Training Intervention

Little F[1], Kassanjee R[1], Nhapi R[1], Ward C[1] Wessels I[1], Lachman J [2], Gardner F [2], Hutchings J [3], Cluver L [1], 2

[1] University of Cape Town, South Africa
[2] University of Oxford, UK
[3] Bangor University,Wales, UK

E-mail for correspondence: Francesca.Little@uct.ac.za

Abstract: We present a longitudinal analysis of the efficacy of a parenting intervention program using Generalized Linear Mixed Effect Models.

Keywords: GLMM; Gaussian; Negative Binomial; Poisson; Parenting Skills Intervention

1 Introduction

The Sinovuyo Caring Families Programme (SCFP) aimed to measure intervention effects of a group-based parent skills training intervention for primary caregivers of children aged 2 to 9, immediately post-test and at 12-month follow-up on four sets of primary endpoints and several secondary endpoints that describe different aspects of parent and child behaviour.
Two hundred and ninety six child-carer dyads were randomised into one of the two study arms. In the intervention arm, primary caregivers were invited to participate in the SCFP, a 12-session group-based parent skills training. In the control arm, primary caregivers were only invited to access standard of care services. The SCFP was conducted in two independent waves corresponding to different time intervals and different residential areas. Within each wave, each participant that was allocated to the intervention arm was also assigned to a group in which the intervention was administered. The intervention was assessed through several composite scores, derived by summing either Likert scale assessments of the intensity

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

of a behaviour, or binary indicators indicate the presence of a trait. It was assumed that these pseudo-continuous outcomes will followed conventional parametric distributions.

2 Model Structure

There were two types of models fitted in analysing the SCFP data: (1) the binary-intervention models and (2) the dose-response models. The former focused on comparing the control arm to the intervention arm. The latter focused on comparing whether the frequency of participants attendance to the group sessions (in the intervention arm) made signicant differences. The underlying distributions of the composite scores were modelled assuming either Gaussian (for sums of many individual items with symmetric empirical distributions), Negative Binomial (for over-dispersed count outcomes) or Poisson(for count outcomes) distributions. A log link was used for all models.

2.1 Binary intervention models

For child-caregiver dyad j in group i assessed at time k the model for response Y_{ijk} is

$$\begin{aligned}\log(E(Y_{ijk})) = {} & (\beta_0 + b_{0,ij}) + \beta_{t1}\delta_{1,ij} + \beta_{t2}\delta_{2,ij} + (\beta_{at1} + b_{at1,i})\delta_{arm,ij}\delta_{1,ij} \\ & +(\beta_{at2} + b_{at2,i})\delta_{arm,ij}\delta_{2,ij} + \beta_w\delta_{wave,ij} + \beta_s\delta_{sex,ij} + \beta_g\delta_{age,ij}\end{aligned}$$

where $i = 0, 1, 2, \ldots, 11$ denoting the group, with all participants in the control arm in group 0 and participants in the intervention arm falling in one of groups 1 to 11; $j = 1, 2, \ldots, n_i$, the number of participants in group i; $k = 0, 1, 2$, corresponding to times 0, 1 and 2, respectively; $\delta_{k,ij} = 1$ if time$=k$, for $k = 1, 2, = 0$ otherwise; $\delta_{arm,ij} = 0, 1$ for control and intervention arms,respectively; $\delta_{wave,ij} = 0, 1$ for waves 1 and 2, respectively; $\delta_{sex,ij} = 0, 1$ for females and males, respectively; $\delta_{age,i,j} = 0, 1$ for child age interval between $2 - 5$ and $6 - 9$ years, respectively; the β parameters are treated as fixed effects; $b_{0,ij} \sim\ N(0; \sigma_0)$ is a dyad-specific random effect for the intercept; $b_{at1,i} \sim N(0; \sigma_{t1})$ and $b_{at2,i} \sim N(0; \sigma_{t2})$ are group-specific random effects. Random effects were initially correlated.
The size of the intervention effects at times 1 and 2 are estimated by $\exp(\beta_{at1})$ and $\exp(\beta_{at2})$, respectively, which measure the proportional difference in the change from baseline in the $E(Y_{ijk})$ at times 1 and 2 for the intervention group compared to the control group.

2.2 Dose-response models

Due to the variable attendance of group session among participants in the intervention arm, a second model was fitted to assess whether the frequency of attendance impacted on the efficacy of the intervention program. These dose-response models included a stratification by the presence of inter-partner violence at baseline in addition to the stratifications by wave, sex and age group.

$$\begin{aligned}\log(E(Y_{ijk})) = (\beta_0 + b_{0,ij}) + \beta_{t1}\delta_{1,ij} + \beta_{t2}\delta_{2,ij} + (\beta_{At1} + b_{At1,i})\delta_{Att,ij}\delta_{1,ij} \\ +(\beta_{At2} + b_{At2,i})\delta_{Att,ij}\delta_{2,ij} + \beta_w\delta_{wave,ij} + \beta_p\delta_{ipv,ij} + \beta_s\delta_{sex,ij} + \beta_g\delta_{age,ij} \\ +\beta_{wt1}\delta_{wave,ij}\delta_{1,ij} + \beta_{wt2}\delta_{wave,ij}\delta_{2,ij} + \beta_{pt1}\delta_{ipv,ij}\delta_{1,ij} + \beta_{pt2}\delta_{ipv,ij}\delta_{2,ij} \\ +\beta_{pAt1}\delta_{ipv,ij}\delta_{Att,ij}\delta_{1,ij} + \beta_{pAt2}\delta_{ipv,ij}\delta_{Att,ij}\delta_{2,ij} \\ +\beta_{wAt1}\delta_{wave,ij}\delta_{Att,ij}\delta_{1,ij} + \beta_{wAt2}\delta_{wave,ij}\delta_{Att,ij}\delta_{2,ij}\end{aligned}$$

where $\delta_{Att,ij} = 0, 1, \ldots, 12$ for the number of group sessions attended with controls having a value of 0 by design; $\delta_{ipv,ij} = 0, 1$ for absence or presence,respectively, of ipv experienced by caregiver; $b_{At1,i} \sim N(0; \sigma_{At1})$ and $b_{At2,i} \sim N(0; \sigma_{At2})$ are group-specific random effects that could be correlated; all other terms as defined above for binary model.
The impact on the proportional change from baseline to times 1 and 2 of a unit increase in the number of group sessions attended is measured by $\exp(\beta_{At1})$ and $\exp(\beta_{At2})$, respectively for dyads in wave 1 whose caregivers did not experience ipv. The modifications due to wave and presence of ipv are measured by $\exp(\beta_{wAt1})$ and $\exp(\beta_{wAt2})$ and $\exp(\beta_{pAt1})$ and $\exp(\beta_{pAt2})$, respectively.

3 Results

The table below summarizes the interaction terms from the binary model for three selected outcomes.

TABLE 1. Estimated parameters for binary intervention models

Effect	Exp(beta)	95% CI	p-value
ECBI intensity			
Arm*time1vs0	0.9636	(0.9152;1.0147)	0.1592
Arm*time2vs0	0.9645	(0.9092;1.0232)	0.2296
Physical Discipline			
Arm*time1vs0	0.6772	(0.5430;0.8446)	0.0005
Arm*time2vs0	0.9530	(0.7379;1.2308)	0.7121
Positive Child Behaviour			
Arm*time1vs0	1.0644	(0.9717;1.1660)	0.1796
Arm*time2vs0	1.1254	(1.0410;1.2166)	0.0030

Adjusted for wave, sex and age.

The graphs in figure 1 plot the proportional change, as measured by the exponent of the sum of the relevant combination of estimated β coefficients, in the three outcomes for increasing levels of the intervention. Point-wise confidence intervals are indicated with vertical lines.

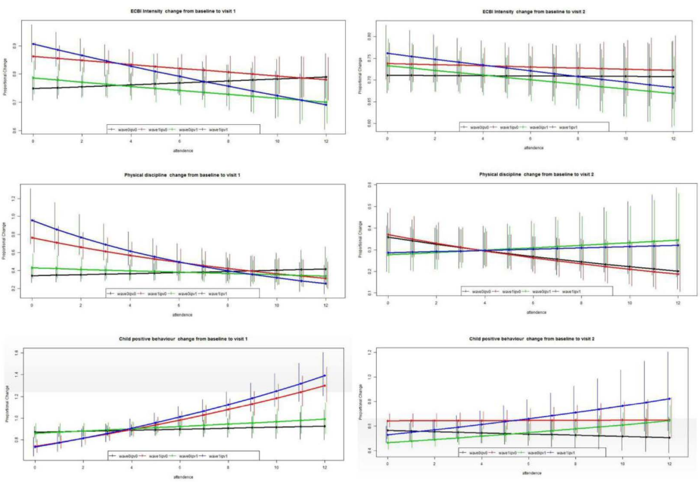


FIGURE 1. Effect of an increasing level of intervention

4 Discussion

Table 1 shows that a marginally larger, though not significant, decrease in ECBI intensity was observed for the intervention group compared to the control arm. For Physical Disicipline, the intervention arm showed a significantly larger decrease in scores compared to the control group immediately after the intervention at time 1, but 12 months later the beneficial effect of the intervention seemed to have largely disappeared. The intervention arm resulted in a 6% higher incidence of Positive Child Behaviour immediately after the intervention and a 12% higher 12 months later compared to the control arm, with only the latter increase being statistically significant.
The graphed profiles in figure 1 show a larger reduction in ECBI frequency and for Physical Discipline, and larger increases in Observed Positive Child Behaviour, at time 1 immediately after the intervention period relative to baseline for increasing number of visits attended. Non-parallel lines correspond to larger effect sizes for the three-way interaction terms in the dose-response models, indicating the modification of the effects due to the different waves or due to the presence of ipv. The profiles for the relative changes from baseline to the 12 months follow-up visit are near to horizontal and the point-wise confidence intervals overlap.

Hurdle hyper-Poisson regression model

Ana María Martínez-Rodríguez[1], Antonio Conde-Sánchez[1], Antonio José Sáez-Catillo[1], María José Olmo-Jiménez[1] and José Rodríguez-Avi[1]

[1] Department of Statistics and Operational Research, Universidad de Jaén, Spain

E-mail for correspondence: ammartin@ujaen.es

Abstract: A hurdle model is a two-part model where a binomial model is used to model the process for zero counts and a truncated one for positive counts. In this paper a hyper-Poisson truncated model is proposed for modelling the positive counts in the number of caries in 9-years-olds and it is compared with a hurdle negative binomial regression model.

Keywords: Hurdle model; Hyper-Poisson regression; Count data.

1 Introduction

A hurdle model (Mullahy, 1986; Cameron and Trivedi, 2013) is a modified count model in which there are two processes, one generating the zero counts and one generating the positive counts. The two models are not constrained to be the same. The concept underlying the hurdle model is that a binomial probability model governs the binary outcome of whether a count variable has a zero or a positive value. If the value is positive, the "hurdle is crossed" and the conditional distribution of the positive values is governed by a zero-truncated count model. Traditional models for the positive counts are the Poisson and NB truncated distributions. In these classical models, regressors are introduced to explain the mean of the non-truncated distribution which in fact cannot be observed. In addition, the interpretation of the coefficients from the zero-truncated model no longer corresponds directly to changes in the unconditional rate. For all these reasons it would be interesting to introduce the explanatory variables in the truncated mean so that the regression coefficients could explain the performance of the true observed mean (Martínez-Rodríguez et al.).

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

In this work a hyper-Poisson regression model for zero-truncated count data based on the hyper-Poisson distribution (Sáez-Castillo et al., 2013a) is used for modelling the positive counts. The truncated hyper-Poisson regression model introduces the regressors in the equation of the mean of the truncated distribution and additionally, regressors can also be introduced in the equation of the dispersion parameter in order to model under- and over- dispersion.

2 Hurdle hyper-Poisson regression model

A hurdle model has two parts: one zero part which models the zeroes and one truncated count part that models the positive counts. Formally,

$$P\left(Y=y\right)=\left\{\begin{array}{ll} f_1(0) & y=0 \\ \left(1-f_1(0)\right)f_2(y) & y>0 \end{array}\right.$$

where $f_1(0) = P\left(Y=0\right)$ and $f_2(y)$ is the p.m.f. of the corresponding zero-truncated distribution. In this work we consider the truncated hyper-Poisson distribution. Specifically,

$$f_2(y)=\frac{1}{{}_1F_1\left(1;\gamma;\lambda\right)-1}\frac{\lambda^y}{\left(\gamma\right)_y}, \quad y=1,\,2,...,$$

where $\gamma,\lambda>0$, $\left(a\right)_r=a\left(a+1\right)...\left(a+r-1\right)$ is the Pochhammer symbol and

$${}_1F_1\left(a;b;c\right)=\sum_{k=0}^{\infty}\frac{\left(a\right)_k}{\left(b\right)_k}\frac{c^k}{k!}$$

is the confluent function. γ is a dispersion parameter which determines that the hyper-Poisson distribution is over-dispersed if it is greater than one, under-dispersed if it is lower than one and matches with the Poisson if it is equal to one; λ is interpreted as a location parameter.

An expression of the mean of the truncated hyper-Poisson (Sáez-Castillo et al., 2013b) is

$$\mu=\lambda+f_2(1)-\left(\gamma-1\right)\left(1-f_2(1)\right), \tag{1}$$

where

$$f_2(1)=\frac{1}{{}_1F_1\left(1;\gamma;\lambda\right)-1}\frac{\lambda}{\gamma}.$$

Let be y_i the value of the response variable of the $i-th$ individual of the sample and $\mathbf{x}_i^T=(1,x_{i1},x_{i2},\ldots,x_{ik})$ the observed covariates in this $i-th$ individual. Let us consider,

$$\mu_i = e^{\mathbf{x}_i^T\beta}$$

In this work the parameter λ will be determined by (1) from the values of μ_i and γ. The estimation of the regression coefficients β is carried out maximizing the log-likelihood function that is given by

$$\log L(\gamma, \lambda) = -\sum_{i=1}^{n} \log \Gamma(\gamma + y_i) + \log(\lambda) n\bar{y} \\ + n\left(\log \Gamma(\gamma) - \log\left({}_1F_1(1; \gamma; \lambda) - 1\right)\right).$$

3 Application

We shall illustrate the hurdle hype-Poisson (HhP) regression model and compare it with a hurdle negative binomial (HNB) regression model. Specifically, we use data from Hofstetter et al. (2016). The dataset contains observations of 396 nine-year-old children in the Netherlands. The dependent variable is the number of caries in the primary teeth. The explanatory variables (their levels in brackets) are: education level of the mother (high education, low education), gender (male, female), ethnicity (natives, immigrants), frequency of brushing teeth (less than twice a day, at least twice a day), frequency of having breakfast (not daily, daily), frequency of food and drinks per day in addition to the three main meals (maximum 7 times daily, more than 7 times daily) and the score on Corah's Dental Anxiety Questionnaire (lower than 13, higher than or equal to 13).

Table 1 shows the coefficient estimates and standard errors of HNB and HhP regression models together with the obtained AIC.

The comparison of AIC included in Table 1 shows that HhP regression model fits slightly better than the HNB regression model, but what is more interesting is that the parameter estimates in the HhP model are more accurate as their s.e. are lower. Also, it has to be emphasized that as in the HhP model the explanatory variables are included in the truncated mean the regression coefficients could explain the performance of the true observed mean and not the one of the non-truncated distribution.

References

Cameron, A. C. and Trivedi, P. K. (2013). Regression Analysis of Count Data. Cambridge University Press, New York.

Hofstetter, H., Dusseldorp, E., Zeileis, A. and Schuller, A.A. (2016). Modeling Caries Experience: Advantages of the Use of the Hurdle Model. *Caries Research*, **50**, 517 – 526.

Martínez-Rodríguez, A.M., Conde-Sánchez, A. and Olmo-Jiménez, M.J. A new approach to truncated regression for count data: application to a hurdle model. *Manuscript submitted for publication.*

TABLE 1. Coefficient estimates and standard errors of HNB and HhP fitted model.

	Count model coefficients			
	Truncated NB		Truncated hP	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	1.293	0.131 ***	1.424	0.102 ***
educationlow	0.304	0.130 *	0.231	0.107 *
gendermale	0.050	0.125	0.075	0.103
ethnicityimmigrant	0.336	0.158 *	0.275	0.122 *
brushing(< 2)	0.378	0.143 **	0.292	0.115 **
breakfast(< 7)	0.138	0.185	0.148	0.151
fooddrink(> 7)	-0.083	0.197	-0.092	0.161
corah(≥ 13)	0.385	0.221 .	0.302	0.176 .
Log(θ)/Log(γ)	0.512	0.187 **	4.401	0.447 ***
AIC	1710.477		1704.009	
	Zero hurdle model coefficients			
	Estimate	Std. Error		
(Intercept)	-0.327	0.198 .		
educationlow	0.404	0.214 .		
gendermale	-0.044	0.213		
ethnicityimmigrant	0.488	0.291 .		
brushing(< 2)	0.262	0.268		
breakfast(< 7)	1.257	0.476 **		
fooddrink(> 7)	0.984	0.457 *		
corah(≥ 13)	16.147	865.497		
. p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001				

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341 – 365.

Sáez-Castillo, A.J. and Conde-Sánchez, A. (2013a). A hyper-Poisson regression model for overdispersed and underdispersed count data *Computational Statistics and Data Analysis*, **61**, 148 – 157.

Sáez-Castillo, A.J., Conde-Sánchez, A., Martínez-Rodríguez, A.M., Olmo-Jiménez, M.J. and Rodríguez-Avi, J. (2013b). A hyper-Poisson regression model for zero-truncated count data. *Proceedings of the 28th International Workshop on Statistical Modelling*, 761 – 764.

Item Response Theory modelling assessment of oral health in a Uruguayan population study.

Fernando Massa[1], Gabriel Camaño[2],Ramón Álvarez-Vaz[1]

[1] Unidad de Biometría,Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, Uruguay
[2] Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, Uruguay

E-mail for correspondence: `fmassa@iesta.edu.uy`

Abstract: In epidemiological studies it is common practice to work with binary variables that reflect the presence of certain diseases, which in turn may be associated with another set of variables, that in general are assumed as risk factors of the former. In the field of epidemiological studies referred to oral health, it is common to inquire about the relationship between the presence of some pathologies and certain characteristics of the study participants through generalized linear models (GLM). However, this type of analysis is usually carried out for each variable of interest separately and at no time is a measure obtained that summarizes the status of each participant. In this study we propose the use of item response theory (IRT) models (specifically the Rasch model) since they allow the joint analysis of a set of variables obtaining an individual assessment as a by-product, which in this case is interpreted as "sickness proneness". On the other hand , the analysis presented here extends the Rasch model including a linear predictor that allows to investigate about the possible effect of several factors on the propoensity of the individuals to suffer the different pathologies. Our results found evidence of an effect of gender, physical activity and age on general proneness to oral diseases.

Keywords: Rasch model; Epidemiological studies; Non Communicable diseases; Oral health.

1 Introduction

The epidemiological study of the most common oral pathologies, decay (D), loss of attachment (LoA), periodontal pockets (PP) and functional

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

dentition (FD), can be carried out through different indicators, the most accepted of them is through binary variables representing the presence/absence of each pathology. In the field of health population surveys, it is a common practice for the epidemiological analysis investigate the factors that propitiate the occurrence of such pathologies using GLMs. In this way it is possible to determine what are the conditions for a certain disease, however, these simple models are not able to carry out this analysis simultaneously in several pathologies. For this reason, we propose to use IRT models in the epidemiological field because:

- they are capable of jointly analyze a set of outcomes and
- provide an assessment of each individual.

However, the most frequently used IRT models provide indicators that describe the behavior of each variable without considering the possible effect of other set of explanatory variables. To overcome this difficulty we propose to model the outcomes through a Rasch model where the behavior of subject parameters is determined by a normal distribution whose mean is modeled by a linear predictor (see Figure 1).

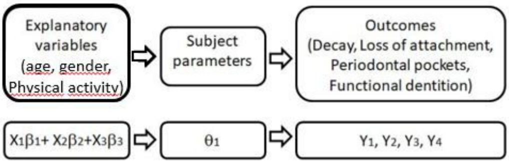

Explanatory variables (age, gender, Physical activity)
Subject parameters
Outcomes (Decay, Loss of attachment, Periodontal pockets, Functional dentition)
X1β1+ X2β2+X3β3
θ1
Y1, Y2, Y3, Y4

FIGURE 1. Flow diagram of the proposed model.

The data used comes from a study of people demanding attention in Dentistry Faculty, of the Universidad de la República, Montevideo, Uruguay, during the period 2015-2016. There were 602 participants where the presence/absence of four oral diseases was studied taking into account gender, age and physical activity of each participant.

2 Statistical Analysis

As mentioned in the previous section, given the binary nature of the variables used for each oral disease, the Rasch model appears as a natural starting point for the analysis. We extended the model considering a set of predictors in the mean of the random effect used to model the behavior of each individual as is shown in equation 1.

$$\begin{array}{rclr} \mathbb{P}(Y_{ij}=1|\theta_i,\delta_j) & = & \frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}} & j=1,2,3,4 \\ \theta_i & \sim & N(X_i^T\beta,1) & i=1,\ldots,n=602 \end{array} \tag{1}$$

Where Y_{ij} represents the occurrence of the disease j on participant i, θ_i is the subject parameter (which in this context can be interpreted as the sickness proneness of each participant), δ_j is the difficulty parameter of each variable (that here is related to the prevalence of each pathology), $X_i^T\beta$ is a linear predictor that accounts for the effect of gender, age and physical activity. Equation 2 expresses the likelihood function.

$$\mathcal{L}(Y|X,\beta,\delta)=\prod_{i=1}^{n}\int_{\mathbb{R}}\prod_{j=1}^{j=3}\mathbb{P}(\theta,\delta_j)^{Y_{ij}}(1-\mathbb{P}(\theta,\delta_j))^{(1-Y_{ij})}\frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta-X_i\beta)^2}{2}}d\theta \tag{2}$$

The optimization of the likelihood function is carried out numerically and, through it's Hessian matrix, an approximation of the variance of each estimator is obtained. All calculations are carried out through the software R.

3 Results

In Table 1 can be observed taking into account the negative values of the difficulty parameters, the four pathologies have relatively high prevalence in the studied population. Figure 2 presents the expected value of "sickness proneness" as a non-linear function of age with different intercept for gender and physical activity status. It can be seen that men present higher proneness to oral diseases than women and, regarding insufficient physical activity, it decreases the mean value of "sickness proneness".
Regarding the effect of age, the restricted cubic spline shows a significant increase effect (with negative concavity). Finally, a significant effect of physical activity, as well as gender, was found in the propensity to oral diseases.

4 Conclusion

Through the proposed model, it was possible to determine the prevalence level of the studied pathologies as well as the effect of some covariates of interest. It was observed that physical activity status had a significant effect on "sickness proneness".There was also detected a significant difference between men and women, while a non-linear (and increasing) effect of age was observed on the tendency to suffer from oral diseases.

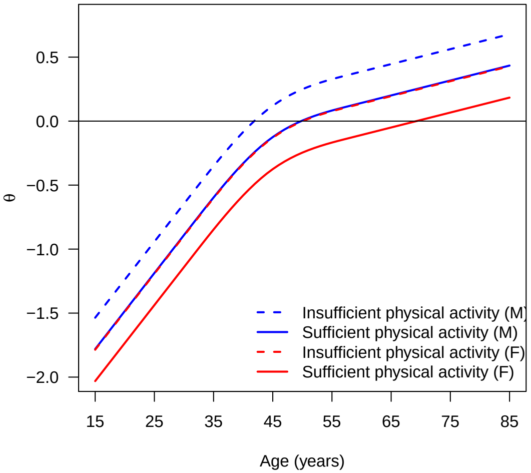


FIGURE 2. Sickness proneness according to age, gender and physical activity.

TABLE 1. Estimates of the Rasch model with covariates

Prevalence parameters				
Pathologies	Estimate	S.E.	Z core	p-value
PP	-0.989	0.093	-10.65	<0.001
D	-1.802	0.106	-17.06	<0.001
FD	-0.021	0.075	-0.286	0.775
LoA	-1.256	0.097	-13.00	<0.001
Linear predictor estimates				
gender (F)	-0.250	0.113	-2.211	0.027
spline age C0	1.225	0.129	9.506	<0.001
spline age C1	-0.646	0.121	-3.354	<0.001
physical activity (ins)	0.245	0.116	2.108	0.035

References

Baker, F. and Kim, S. (2017). *The Basics of Item Response Theory Using R*. Springer Publishing Company.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Durrelman, S. and Simon, R. . *Flexible regression models with cubic splines.* Statistics in Medicine,1989 May;8(5):551-61.

An R package for Inference and Prediction in an Illness-Death Model

Luís Meira-Machado[1], Marta Sestelo[2]

[1] Department of Mathematics and Applications, University of Minho, Campus de Azurem, 4800-058 Guimarães, Portugal.
[2] SiDOR Research Group and CINBIO, University of Vigo, Spain.

E-mail for correspondence: `lmachado@math.uminho.pt`

Abstract: Multi-state models are a useful way of describing a process in which an individual moves through a number of finite states in continuous time. The illness-death model plays a central role in the theory and practice of these models, describing the dynamics of healthy subjects who may move to an intermediate 'diseased' state before entering into a terminal absorbing state. In these models one important goal is the modeling of transition rates which is usually done by studying the relationship between covariates and disease evolution. However, biomedical researchers are also interested in reporting other interpretable results in a simple and summarized manner. These include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, prevalence and the sojourn time distributions. An **R** package was built providing answers to all these topics.

Keywords: Illness-death model; Kaplan-Meier; Landmark approach; Nonparametric estimation; Survival analysis.

1 Introduction

Multi-state models are very useful for describing complex event history data. These models may be considered a generalization of survival analysis where survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. For instances, in most biomedical applications, besides the 'healthy' initial state and the absorbing 'dead' state, one may observe intermediate (transient) states based on health conditions, disease stages, clinical symptoms, etc. The illness-death model is probably the most popular one in the medical literature. The irreversible version of this model

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

(Figure 1), describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. Many time-to-event data sets from biomedical studies with multiple events can be reduced to this generic structure. Recent reviews on this topic may be found in the papers by Putter et al. (2007), Meira-Machado et al. (2009), Meira-Machado et al. (2011) and Meira-Machado and Sestelo (2018).

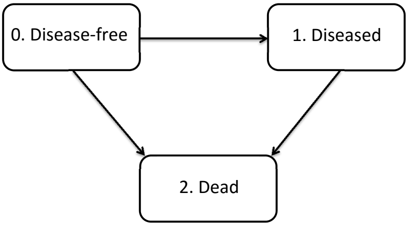


FIGURE 1. Illness-death model.

One important goal in multi-state modelling is to relate the individual characteristics with the intensity rates through a covariate vector but biomedical researchers are also interested in reporting interpretable results in a simple and summarized manner. These include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, prevalence and the sojourn time distributions. The development of **survidm R** package has been motivated by several recent contributions that account for these problems; in particular the newly developed methods based on landmarking. The current version of the package provides seven different approaches to estimate the transition probabilities, three methods for the sojourn distributions and one approach for the cumulative incidence functions. In addition, these probabilities can also be estimated conditionally on covariate measures. The package also allows the user to perform multi-state regression where the estimation of the covariate effects is achieved using Cox regression in which different effects of the covariates are assumed for different transitions.

2 survidm in practice

This software enables both numerical and graphical outputs to be displayed for several methods. This software is intended to be used with the **R** statistical program. Our package is composed of 13 functions that allow users to obtain estimates for all proposed methods. Details on the usage of the functions (described in Table 1) can be obtained with the corresponding help pages.

It should be noted that to implement the methods described in the methodology section one needs the following variables of data: `time1`, `event1`,

Function	Description
`survIDM`	Create a `survIDM` object.
`coxidm`	Fits proportional hazards regression models for each transition.
`tprob`	Nonparametric estimation of the transition probabilities.
`CIF`	Nonparametric estimation of the cumulative incidence functions.
`sojourn`	Nonparametric estimation of the sojourn distributions.
`plot.survIDM`	Plot for an object of class `survIDM`.
`print.survIDM`	Print for an object of class `survIDM`.
`summary.survIDM`	Summary for an object of class `survIDM`.
`KM`	Computes the Kaplan-Meier product-limit of survival.
`PKM`	Computes the presmoothed Kaplan-Meier product-limit of survival.
`Beran`	Computes the conditional survival probability of the response, given the covariate under random censoring.
`KMW`	Returns a vector with the Kaplan-Meier weights.
`PKMW`	Returns a vector with the presmoothed Kaplan-Meier weights.
`LLW`	Returns a vector with the local linear weights.
`NWW`	Returns a vector with the Nadaraya-Watson weights.

TABLE 1. Summary of functions in the **survidm** package.

`Stime` and `event`. A single covariate can also be included (it is only necessary for `IPCW` methods). The variable `time1` represents the observed time to the first event of interest, and `event1` the corresponding status/censoring indicator (if the survival time is a censored observation, the value is 0 and otherwise the value is 1). The variable `Stime` represents the total survival time. If `event1` = 0, then the total survival time is equal to the observed time to the first event. The variable `event` is the final status of the individual (takes the value 1 if the final event of interest is observed and 0 otherwise).

For illustration purposes we will use data of 929 patients affected by colon cancer that underwent a curative surgery for colorectal cancer. In this study, 468 developed recurrence and among these 414 died. 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up. Besides the two event times (time to recurrence and time to death) and the corresponding indicator statuses a vector of covariates including `age`, `sex` and number of lymph nodes (`nodes`) are also available.

One important goal in multi-state modeling is to study the relationships between the different predictors and the outcome. To relate the individual characteristics to the intensity rates several models have been used in literature. A common simplifying strategy is to decouple the whole process into various survival models, by fitting separate intensities to all permitted transitions using semi-parametric Cox proportional hazard regression models, while making appropriate adjustments to the risk set. This can be obtained using the following input commands:

```
library(survidm)
data(colonIDM)
fit.cmm <- coxidm(survIDM(time1, event1, Stime, event) ~ age
       + sex + nodes, data = colonIDM)
summary(fit.cmm)
```

Results obtained from the above input commands (not shown) reveal that multi-state regression models provide detailed information of the disease process, revealing how the different covariates may affect the various permitted transitions. For instances, it revealed `age` as an important predictor on the mortality transitions (with and without recurrence) but not on the recurrence incidence, whereas `sex` only revealed a significant effect on the mortality transition after recurrence.

The patients course over time may also be studied through other quantities such as the transition probabilities. To obtain these estimates (for a model with no covariates), the following input command must be typed:

```
res <- tprob(survIDM(time1, event1, Stime, event) ~ 1, s=365,
       method = "LM", conf=TRUE, data = colonIDM)
summary(res, time=365*1:6)
plot(res)
```

Figure 2 reports estimated transition probabilities ($P_{ij}(s,t)$) for a fixed value of $s = 365$ (days), along time. Results were obtained using the Landmark method (`method = "LM"`) proposed by de Uña-Álvarez and Meira-Machado (2015). It is worth mention that function `tprob` implements eight distinct methods including the possibility of estimating these quantities conditional on covariates.

Estimates and plots for the cumulative incidence (of recurrence) (Geskus 2011) and for the sojourn time distribution quantities can also be obtained. The following input commands provide the corresponding numerical and graphical output for the two quantities:

```
res.cif <- CIF(survIDM(time1, event1, Stime, event) ~ 1,
           data = colonIDM, conf = TRUE)
summary(res.cif, time = 365*1:7)
plot(res.cif, ylim=c(0, 0.6))

res.soj <- sojourn(survIDM(time1, event1, Stime, event) ~ 1,
           data = colonIDM, conf = TRUE, conf.level = 0.95)
summary(res.soj, time = 365*1:6)
plot(res.soj)
```

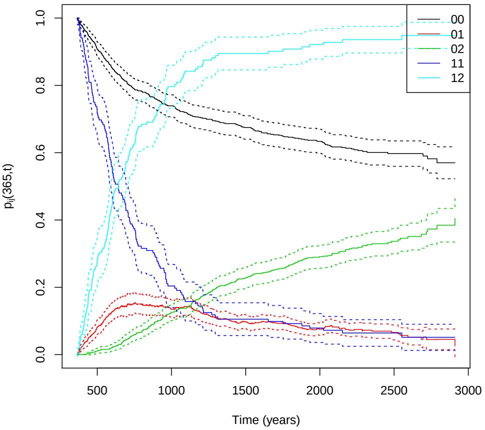


FIGURE 2. Estimates of the transition probabilities using the landmark method. Colon cancer data.

References

Moreira, A., de Uña-Álvarez, J. and Meira-Machado, L. (2013). Presmoothing the Aalen-Johansen estimator in the illness-death model. *Electronic Journal of Statistics*, **7**, 1491 – 1516.

Meira-Machado, L., de Uña-Álvarez, J. and Somnath, D. (2015). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Computational Statistics*, **30(2)**, 377 – 397.

de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study. *Biometrics*, **71**, 364 – 375.

Putter, H. and Spitoni, C. (2016). Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator. *Statistical Methods in Medical Research*, 1 – 12.

Geskus, R.B. (2011). Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. *Biometrics*, **67**, 39 – 49.

Meira-Machado, L. (2011). Inference for non-Markov multi-state models: an overview. *REVSTAT - Statistical Journal*, **9 (1)**, 83 – 98.

Meira-Machado, L. (2016). Smoothed landmark estimators of the transition probabilities. *SORT-Statistics and Operations Research Transactions*, **40**, 375 – 398.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18**, 195 – 222.

Meira-Machado, L., Sestelo, M. (2018). Estimation in the progressive illness-death model: a non-exhaustive review. *submitted.*

Statistical modelling with missing data in eHealth databases

Denny Meyer[1],Madawa W. Jayawar[1],Sandun Silva[1]

[1] Swinburne University of Technology, Australia

E-mail for correspondence: dmeyer@swin.edu.au

Keywords: engagement; random forest; mixed models

1 Introduction

Evaluating engagement with any intervention program is important for understanding its efficacy (Scherer et al., 2017). In such programs missing assessment data is often closely linked to level of engagement, resulting in the potential for informative missingness. But not always. As explained by Scherer et al. (2017), participants who stop providing assessment data do not always drop out of a program, so it is wrong to assume complete disengagement. Models that can accommodate informative missingness are required for unbiased inference when analysing the effectiveness of eHealth interventions. Various methods have been developed to address this problem in the evaluation of such interventions. In particular evaluation models are commonly fitted, controlling for the estimated probability of a missing response, which is usually obtained from a logistic or logit regression analysis using baseline data. However, this method only avoids bias when the model for non-response addresses the mechanism causing non-response (Bell et al., 2013). But what about when sample sizes are very large and attrition rates are very high as is the case for many eHealth interventions?

2 Objective

The objective in this paper is to investigate more powerful approaches for estimating the probability of missingness with big eHealth data and then, using these probabilities, to evaluate the effect of missingness on a primary

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

outcome measure, without and then with the inclusion of more direct measures of engagement, namely participant engagement with specific program modules and features.

3 Methods

3.1 Data

The data utilized in this study was kindly provided by a global Software as a Service (SaaS) vendor, Virgin Pulse, for their Global Challenge (VPGC) program run from May to September 2016. This is a workplace health and exercise program which consists of a 100-day virtual journey. Employees are placed in teams of seven, provided with an activity tracker and given access to an application through a web browser or mobile device. Teams compete with one another to accumulate steps, measured by the activity trackers and entered online or sync'ed on a daily basis. The program is gamified to encourage healthy habits through education, goal setting and positive reinforcement using special program features such as personal statistics, competitions and virtual trophies. In addition to the Physical Activity module the program includes modules addressing Sleep, Nutrition and a module called Balance which addresses psychological wellbeing.

3.2 Assessment data

Three sources of assessment data were used in this analysis. Self-reported health, psychological wellbeing and performance data were collected from participants at the start and end of the 100 Day Journey. Finally, experience data were collected from the participants approximately two weeks after the 100 Day Journey ended. These data recorded the levels of engagement with the various program modules and features.

3.3 Participants

Starting with an initial enrolment of 178350, the response rate was 84.6 percent for the first assessment, 48.4 percent for the second assessment and 10.5 percent for the final satisfaction assessment. For the purposes of this analysis we will focus on the people who completed the third assessment, endeavouring to estimate the probability of this event on the basis of the information collected for the first and second assessments.

3.4 Measures

The measures of interest for the first two assessments consisted of a well validated five item measure for psychological wellbeing (WHO5) as well as perceptions on 0-6 ordinal scales for Felt Happy, Awareness Physical Activity, Sleep Quality, Stress at Work, Awareness Nutrition, Overall Health and Productivity at Work. The WHO5 will be regarded as the primary outcome measure for this study.

3.5 Estimation of probabilities for completion of the particpant experience survey

The participants were randomly split into training, validation and test data sets in a 40:30:30 ratio. In order to allow for the lack of balance in the data a one dollar profit was assumed for each correctly identified missing survey and a ten dollar profit for each correctly identified completed survey. The classification accuracy of the various tools were compared using the ROC and Gini Indices. The random forest was of particular interest because this method was able to accommodate cases with varying degrees of missingness without losing any of the data. The forest was designed to have 100 trees, each constructed using a random sample consisting of 60 percent of the data, with Chi-Squared tests used to choose the optimum splitting variable for each split. These tests were used in order to avoid overfitting which is a commonly reported problem with random forests.

3.6 Models for psychological wellbeing

In order to accommodate the varying degrees of missingness in the collected data, mixed model analyses were conducted, controling for age and gender. The first model was used to assess to what extent missingness related to improvements in psychological wellbeing. A second model was then fitted, testing directly for engagement effects on psychological wellbeing, utilising participant engagement reports with each of the program modules and gamification features. However, the importance of missingness was evaluated at the same time, in order to determine whether any additional information could be provided by this estimated survey completion probability.

4 Results

With test data the random forest produced an area under the ROC curve of 0.776 compared to values of 0.731 for Gradient Boosting, 0.596 for a neural network with three hidden nodes and 0.586 for a binary logistic regression. The estimated survey completion probabilities obtained from the random forest ranged from 0.55 to 1.0 with a similar distribution for participants who did and did not complete the survey. However, 54.5 percent of the participants who completed the participant experience survey received an estimated probability of above 95 percent with only 21.6 percent of those who did not complete this survey receiving an estimate probability of above 95 percent as shown in Figure 1.

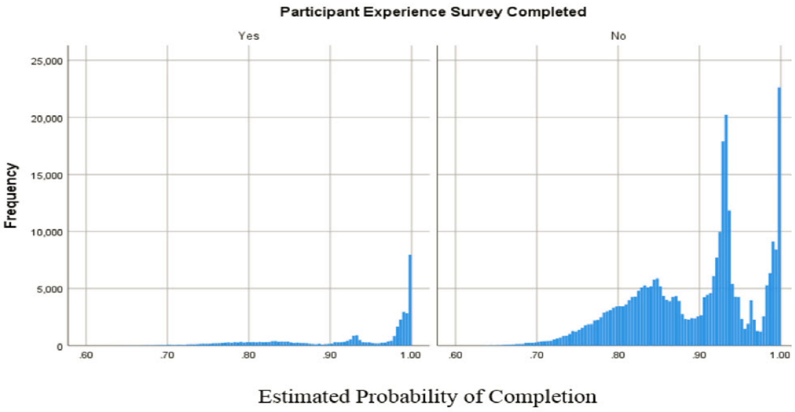


FIGURE 1. Distribution Estimated Probability of Completion for Final Survey

The most important variables for predicting the probability of survey completion were height and weight (due to low survey completion rates for people who failed to supply this information), and the WHO5 collected in the second assessment. The mixed model analysis showed that improvements in psychological wellbeing were significantly associated with the estimated probability of survey completion. However, when module and feature engagement were added to the model the effect of the probability of survey completion weakened, becoming non-significant (z=.369, p=.712). As shown in Figure 2, two of the modules, the Nutrition and the Balance modules had a very significant relationship with improvements in psychological wellbeing. As shown in Figure 3 the significance of the Mini-Challenges, Competitions, Community Chatroom and personal statistics features were confirmed, with only the personal statistics association shown to be negative.

5 Conclusions

The random forest performed better than more typically used methods for predicting the likelihood of a completed participant experience survey because, unlike these methods, it could utilise records with missing data. This is an important distinction when working with eHealth data because missing data is so common, suggesting that random forests have a special role to play with data such as this. The distribution of the estimated probabilities showed support for the Scherer et al. (2017) claim that there are participants in eHealth programs for whom failure to complete assessment is unrelated to disengagement.

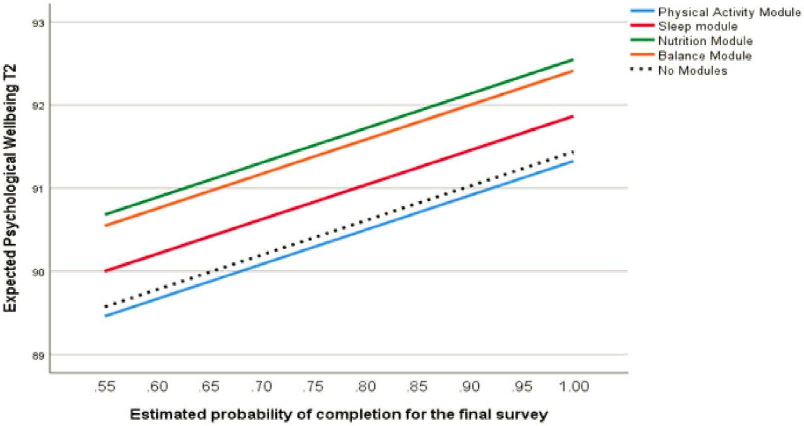


FIGURE 2. Effect of Modules on the Relationship Between Psychological Wellbeing and Estimated Probability for the Final Survey

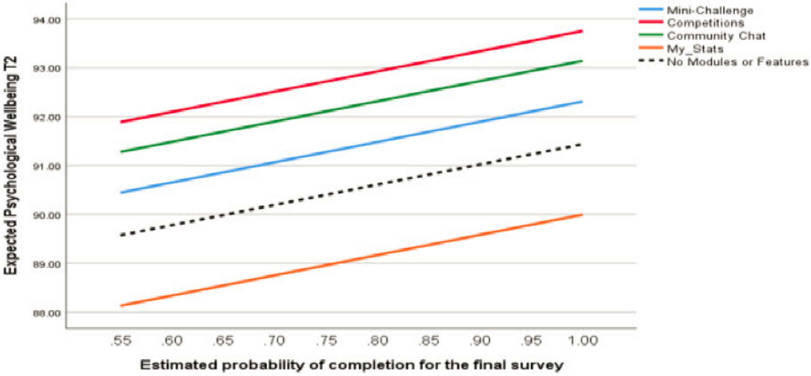


FIGURE 3. Effect of Modules on the Relationship Between Psychological Wellbeing and Estimated Probability for the Final Survey

Mixed model analysis showed that psychological wellbeing improvements were lower in the case of participants with low pedicted survey completion rates, however, this effect became non-significant when direct engagement with program modules and features were taken into consideration. This suggests that, alhough missingness was related to the level of engagement with this online program, it was not a source of bias in models for the outcome measure when direct measures of engagement were also included. Clearly direct measures for the level of engagement are critical for the evaluation of online interventions, with these measures ideally pinpointing modifiable aspects of these interventions.

References

Bell, M.L., Kenward, M.G., Fairclough, D.L. and Horton, N.J. (2013). Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ*, **346**, e8668.

Scherer, E.A., Ben-Zeev, D., Li, Z. and Kane, J.M. (2017). Analyzing mHealth Engagement: Joint models for intensively collected user engagement data. *J Med Internet Res*, **13(4)**, e1.

Predicting the stage of the prostate cancer to personalized the treatment in an Irish cohort

Shirin Moghaddam[1], Keefe Murphy[23], Lisa Murphy[1], Laura O'Gorman[1], Thomas Lynch[4], Richard Power[5], Kieran J O'Malley[6], T Brendan Murphy[23], R William Watson[1]

[1] UCD School of Medicine, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland.

[2] UCD School of Mathematics and Statistics, University College Dublin, Dublin, Ireland.

[3] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland.

[4] Department of Urology, St. James University Hospital, Dublin, Ireland.

[5] Department of Urology, Beaumont Hospital, Dublin, Ireland.

[6] Department of Urology, Mater Misericordiae University Hospital, Dublin, Ireland.

E-mail for correspondence: shirin.moghaddam@ucd.ie

Abstract: Prostate cancer is the most common malignancy among men in developed countries. The single biomarker test for prostate-specific antigen (PSA) has decreased the number of deaths from prostate cancer. However, it is controversial due to low specificity and inability to identify aggressive forms of cancer which has lead to overdiagnosis and treatment. The main dilemma faced by the patient and clinician once prostate cancer has been detected is how best to treat it. Here we are using optimised logistic regression and multiple new biomarker-based diagnostics to enable accurate staging and grading of prostate cancer and guidance for appropriate choices of treatment selection.

Keywords: Prediction models; Logistic regression; Prostate cancer.

1 Introduction

Current practice in prostate cancer detection and staging leads to inaccurate assessments often resulting in many unnecessary treatments that impact negatively on patients quality of life.

The single biomarker test for prostate-specific antigen (PSA), recommended by the American Cancer Society for early screening has decreased the number of deaths from prostate cancer but is controversial due to low specificity and inability to identify aggressive forms of cancer.

1/3 of men who undergo surgery are found to have a non-organ confined disease (extension of a tumour beyond the capsule of the gland) and will have to undergo additional radiation and hormone ablation therapy. These patients will also have to live with the quality of life issues associated with their primary treatment as well the side effects of other treatment approaches. Better and more accurate diagnosis of the stage and grade of disease will impact very significantly on the patients outcome and quality of life.

The main goal of this study is to test whether the detection of a panel of serum protein biomarkers can be used to accurately detect and establish the stage and grade of prostate cancers.

2 Statistical modelling

A panel of nine serum protein biomarkers were assessed by the MSD platform based from previous discovery studies (Oon et al.). In a cohort of 150 Irish patients undergoing a radical prostatectomy for localized prostate cancer as part of the Prostate Cancer Research Consortium bioresource. Discrimination between pathological stages and grades were investigated using logistic regression models.

Model evaluation was carried out by examining calibration, discrimination and decision curve analysis. The calibration of the models was measured using calibration curves. The discriminate ability of the models were compared by the area under the curve (AUC) values.

3 Predicting the stage in prostate cancer

Predictive tools based on standard clinicopathologic variables have been developed for prostate cancer including look-up tables and nomograms. One such highly regarded tool is the Partin tables which was developed using multivariate logistic regression. The partin tables developed to preoperatively predict pathological stage of prostate cancer and thus inform the likelihood of progressive disease to help identify men who will benefit from surgery.

While the Partin tables are well validated and used by clinicians, studies in Ireland were not able to validate the findings with the discriminate ability of less than 70% for the organ-confined and non-organ confined disease (Boyce et al).

The Partin table uses clinical variables based on the digital rectal exam (DRE), Gleason score (GS) of prostate needle biopsy and PSA.

3.1 Results

By using the stepwise logistic regression model, the biomarkers with the significant effects are PSA, GS, CD14, IGFBP3, GcGlobulin, ZAG, IGF1. Figure 1 shows the discriminate ability measured using ROC curves and AUC values for the prediction based on the Partin table and also new significant biomarkers. Furthermore, Figure 2 shows the decision curve for these two predictive models. Both graphs show that the predictive model proposed where new biomarkers included has the highest AUC and net benefit, which means it is a superior approach in predicting the stage of the prostate cancer.

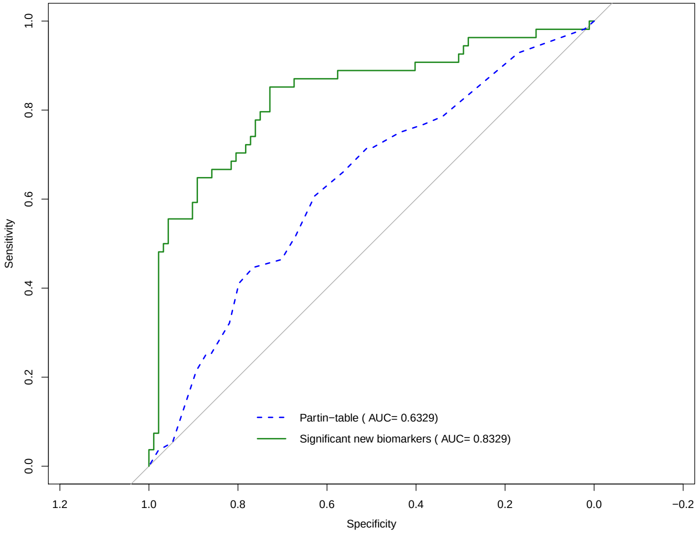


FIGURE 1. ROC curves based on the Partin table and new biomarkers for predicting stage of the prostate cancer.

We tried to validate our model using Australian dataset including 177 patients. Figure 3 shows the ROC curve for the Australian dataset comparing our final model based on significant biomarkers to the Partin table. Although the model including significant biomarkers has the highest AUC in the Irish dataset, it can not be validated using the Australian data.

4 Predicting the grade in prostate cancer

Our secondary goal is to predict the pathological grade using biopsy Gleason grade and serum biomarkers to determine if the cancer is indolent or

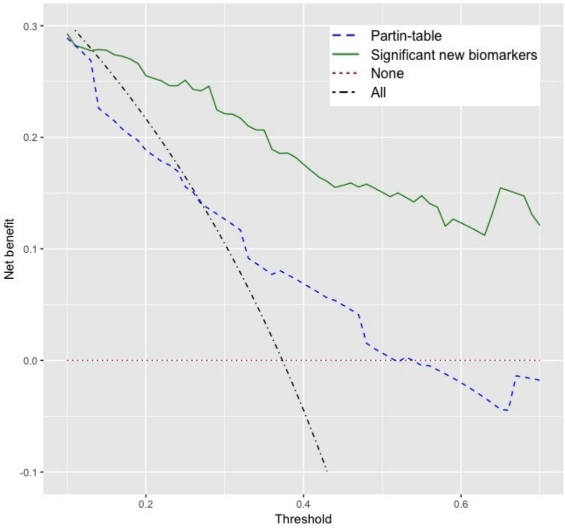


FIGURE 2. Decision curves based on the Partin table and logistic regression model based on new biomarkers for predicting the stage of the prostate cancer. Here ”None” means assuming no one has Non-organ confined prostate cancer and ”all” means assuming everyone has Non-organ confined prostate cancer.

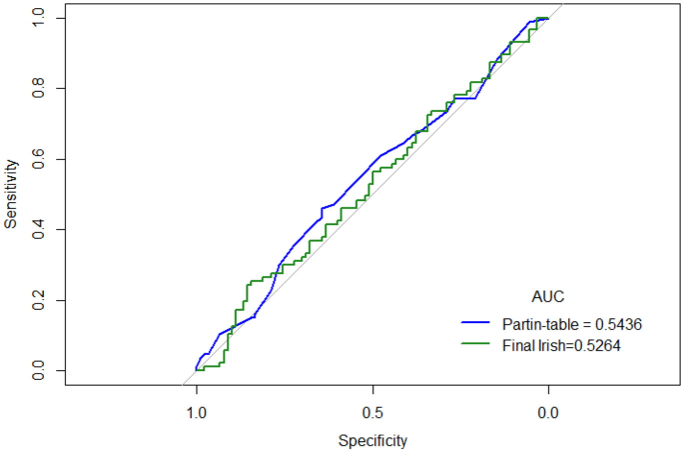


FIGURE 3. ROC curve based on the Partin table and logistic regression model based on new biomarkers for predicting the stage in the Australian dataset.

aggressive. Only a partial representation of the entire prostate can be sampled in a biopsy, and there is a chance of missing small but significant areas leading to a sampling error. As patients who have Gleason grade below 6 may benefit from active surveillance and monitoring of their disease determining the correct grade has important implications to treatment decisions.

4.1 Results

We built a model based on the Irish cohort and applied it to the Australian dataset. By using the stepwise logistic regression model, the clinical variables and new biomarkers with the significant effects are PSA, GS, VEGFD, IGFBP3, IGF1. The predictive model proposed here using three biomarkers combined with the clinical variables has the highest AUC which means it is a superior approach in predicting the grade of the prostate cancer. Based on Figure 5 this model was also validated in the Australian dataset.

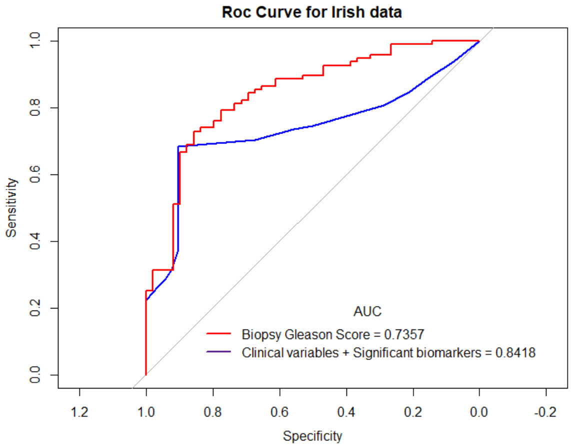


FIGURE 4. ROC curve based on the biopsy gleason score and logistic regression model based on new biomarkers for predicting the grade in the Irish dataset.

5 Conclusion

Multiple biomarkers when combined with clinical variables resulted in a model that predicted stage and grading of prostate cancer and guidance for appropriate choice of therapy. A logistic regression model is fitted to the Irish dataset and then applied to the Australian dataset. Only the model

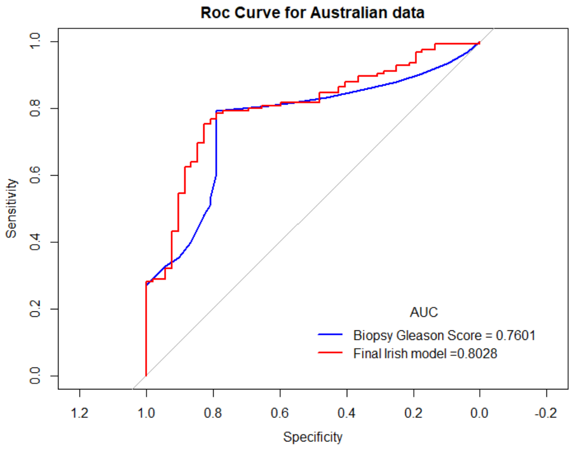


FIGURE 5. ROC curve for the biopsy gleason score and logistic regression model based on Irish dataset for predicting the grade in the Australian dataset.

for predicting grade was validated in the Australian model and current research is ongoing to further validate it in an independent Irish and Austrian cohort.

References

Boyce S et.al (2013). Evaluation of prediction models for the staging of prostate cancer. *BMC medical informatics and decision making*, **13**, 126.

DeLong, Elizabeth R., et al. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837 – 845.

Oon SF, et al. (2012). The identification and internal validation of a preoperative serum biomarker panel to determine extracapsular extension in patients with prostate cancer. *The Prostate* , **72**, 1523 – 1531.

Tosoian, Jeffrey J, et al. (2017). Prediction of pathological stage based on clinical stage, serum prostate-specific antigen, and biopsy Gleason score: Partin Tables in the contemporary era *BJU international* , **119**, 676 – 683.

Vickers, Andrew J., and Elena B. Elkin. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* , **26.6**, 565 – 574.

Statistical modelling of brain connectivity in prodromal Alzheimer's disease

Reza Mohammadi[1], Martin Dyrba [2]

[1] University of Amsterdam, Netherland
[2] German Center for Neurodegenerative Diseases (DZNE), Germany

E-mail for correspondence: a.mohammadi@uva.nl

Abstract: Several neuroimaging markers have been established for the early diagnosis of Alzheimer's disease, among them amyloid-β deposition, glucose metabolism, and grey matter volume. Up to now, these imaging modalities were mostly analyzed separately from each other, and little is known about the regional interrelation and dependency of these markers. Gaussian graphical models (GGMs) provide a probabilistic framework that estimates the conditional dependency between individual random variables. We applied GGMs for studying the interregional associations and dependency between multimodal imaging markers in prodromal Alzheimers disease. GGMs were estimated using a Bayesian framework and for each individual diagnosis, graph-theoretical statistics were calculated to determine structural changes associated with disease progression. Several clusters were obtained for highly inter-correlated regions, e.g. adjacent regions in the same lobes, but included only regions *within* the same imaging modality. Hardly any associations were found *between* different modalities, indicating almost no conditional dependency of brain regions across modalities when considering the covariance explained all other regions. Network measures clustering coefficient and path length were significantly altered across diagnostic groups.

Keywords: Brain connectivity; Bayesian inference; Graphical models.

1 Introduction

Alzheimer's disease is characterized by a range of pathological brain alterations that can be assessed in vivo using various neuroimaging methods, including MRI and PET. With the advent of multimodal imaging being applied in large samples on a regular basis, the need for adequate analysis methods has been arisen.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

We propose the application of Gaussian graphical models (GGM), which are able to reliably estimate the *partial* correlation between various multicollinear predictors. This makes them an interesting candidate for studying statistical associations between various brain regions. The partial correlation derived from GGMs is conceptually similar to the partial correlation obtained from a series of linear regression models providing the statistical association of the dependent and independent variables controlling for the confounding variables. In GGMs, this concept is being extended such that the association between each pair of variables is estimated controlled for all other variables included in the model. Technically, GGMs are naively realized by matrix inversion of the covariance matrix or, in more robust and efficient approaches, apply efficient sampling schemes Mohammadi and Wit (2015, 2017).

In this paper we assessed the statistical associations within and between three main imaging markers of Alzheimer's disease using GGMs based on a whole-cortex parcellation of the brain estimating the regional interdependency of amyloid-β deposition (florbetapir/AV45-PET), glucose metabolism (FDG-PET), and gray matter volume (T_1-weighted MRI). Based our previous results with only six representative brain regions Dyrba et al. (2017), we hypothesized that regional amyloid deposition has low contribution to neurodegeneration, whereas hypometabolism was expected to be stronger related to neurodegeneration. Further, we expected a high local association within each region following a local evolution of the disease and, additionally, few hub-nodes to influence pathology in other regions as well. For graph-theoretical measures, we expected a linear trajectory of decreasing clustering coefficient and increasing path length with stronger disease severity.

2 Material and Methods

Graphical models provide an effective way for describing statistical patterns in multivariate data and for estimating the conditional dependency between the various brain regions and imaging modalities based on GGMs Mohammadi (2015). For data following a multivariate normal distribution, undirected GGMs are commonly used. In these graphical models, the graph structure is directly characterized by the precision matrix, i.e. the inverse of the covariance matrix: non-zero entries in the precision matrix show the edges in the conditional dependency graph. Notably, simple inversion of the covariance matrix usually does not work in real world data sets, as already slight noise or selection bias in the empirical data causes the precision matrix to contain almost no zero entries. To overcome this problem, regularization techniques or efficient sampling algorithms have been proposed that additionally employ a sparsity assumption to reduce the effect of noise and to only detect the most probable conditional dependencies. For our analyses, we employed a computationally efficient Bayesian framework implemented in the R package BDgraph, more specifically a continuous-time birth-death Markov process, for estimating the most probable graph structure and edge weights that correspond partial correlations Mohammadi and Wit (2015, 2017). For this study, BDgraph was substantially extended by multithreaded parallel processing and marginal pseudo-likelihood approximation to speed up computations.

3 Results

The conditional dependency matrix obtained using the GGM approach by Mohammadi and Wit (2015), using the BDgraph package (Mohmmadi and Wit, 2017) . For amyloid-β deposition and glucose metabolism, brain regions directly adjacent to each other formed smaller clusters of high partial correlation around the main diagonal. For gray matter volume, such patterns also appeared but with lower density and inter-cluster partial correlation. When considering the associations between different imaging modalities, we obtained a consistent pattern of significant positive intra-regional conditional dependency for the pairs amyloid-β deposition and metabolism with a mean partial correlation of $\rho = 0.27$, and between metabolism and gray matter volume (mean $\rho = 0.23$). For the pair amyloid-β deposition and gray matter volume these associations were substantially lower, that means that only 13 of 54 possible edges were detected with a mean of $\rho = 0.08$.

When estimating separate models for each diagnostic group separately, the graph structures were similar to each other with a mean Jaccard similarity of $j = 0.31$. However, the estimated regional dependency differed between groups due to the different degrees of severity of amyloid pathology, glucose metabolism, and atrophy. Regarding the partial correlation weights, we obtained an average cosine similarity of $cos = 0.92$ and Pearson coefficient of $r = 0.74$. The graphs differed in their density, leading to significant alterations of the clustering coefficient, characteristic path length, and small-world coefficient (Fig 1). In concordance with recent studies assessing other imaging modalities (cortical thinning, mean diffusivity, resting-state fMRI connectivity), we also observed a biphasic trajectory of the graph measures. This means that the clustering coefficient and small world coefficient initially increases when comparing early MCI and CN participants (Fig 1). When Alzheimer's disease progresses, i.e. in the late MCI and dementia groups, both measures decrease again, with late MCI being approximately on the same level as CN participants (Fig 1). The characteristic path length showed a similar pattern across groups, but with inverted directionality and decreasing intensity, i.e. group differences were highest in amyloid-β deposition, intermediate for glucose metabolism, and lowest for gray matter volume (Fig 1).

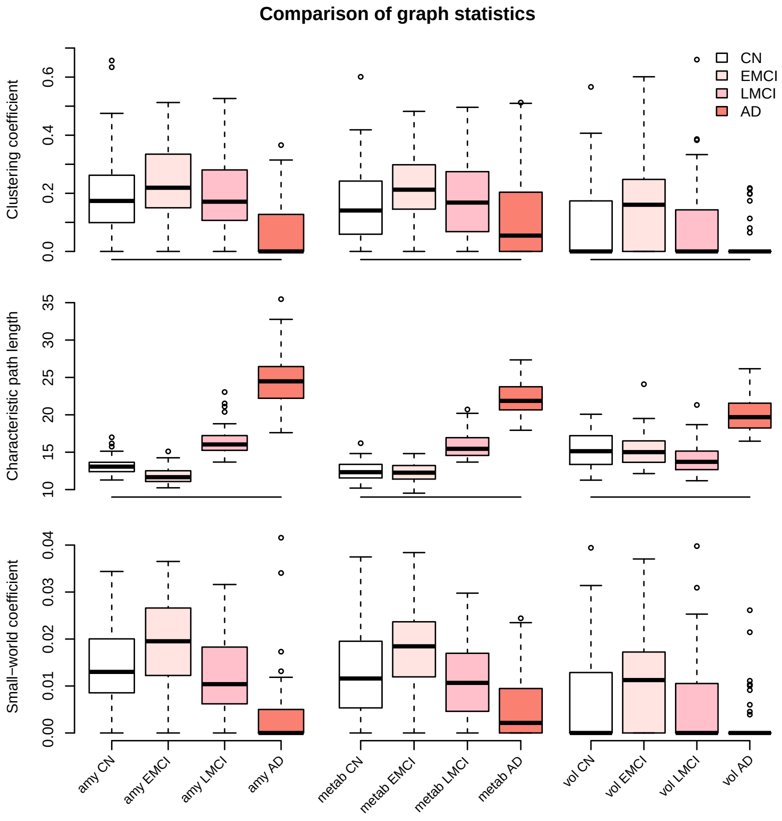


FIGURE 1. **Comparison of graph statistics for the partial correlation matrices stratified by diagnostic group and image modality.** The distribution of the weighted clustering coefficient, characteristic weighted path length, and small-world coefficient for individual brain regions is shown.

References

Mohammadi, A. and Wit, E. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, **10(1)**, 109 – 138.

Mohammadi, A. and Wit, E. (2016). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Journal of Statistical Software.*

Dyrba, M., Grothea, M., Mohammadi, A. et al (2017). Comparison of different hypotheses regarding the spread of Alzheimers disease using Markov random fields and multimodal imaging. *Journal of Alzheimer Disease.*

Dimensionality reduction through clustering of variables and canonical correlation.

Muñoz-Pichardo, J.M.[1], Pino-Mejías, R.[1], Cubiles de la Vega, M.D.[1]

[1] Departament of Statistic and Operations Research, University of Seville, Spain

E-mail for correspondence: `juanm@us.es`

Abstract: A strategy of dimensional reduction of large data sets is proposed. The procedure is based on combination of Cluster Analysis of Variables (CAV) and the Canonical Correlation Analysis (CCA) to determine synthetic variables that are representative of the clusters.

Keywords: Dimensionality reduction; Clustering; Canonical correlation.

1 Introduction

Currently, it is very common to have large data sets and consider the goal of extracting relevant information without previously determining *target* variables. That is, the objective is to discover the hidden relationships that generate the data and detect clusters of cases and/or variables that are masked by the dimensions of the data set. In this cases, dimensionality reduction techniques (especially, Principal Components PCA) and CAV are applied.

Abraham and Inouye (2014) argue that PCA is a widely used technique to detect population structures and potential outliers. However, in large data sets, the dimension causes difficulty or the impossibility of applying PCA due to computational and time consuming problems.

Vigneau and Qannari (2003) proposed grouping the variables into homogeneous groups and associating each cluster with a latent variable so that: first, the clusters must be homogeneous groups according to some criterion of homogeneity based on the correlation between variables; and second, the variables of each cluster must be closely related to the associated latent variable. The authors propose the first principal component of each

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

group as latent variable. This strategy was implemented in statistical software environment R by Saracco *et al.*(2010) and Chavent *et al.*(2012), with a procedure that mixes quantitative and qualitative attributes, applying PCMIX (Kiers, 1991).
ACP bases the dimensionality reduction in the creation of synthetic variables that maximize the explained variability, but does not consider the correlation structure. This can lead to some loss of information. To avoid this problem we propose the application of CCA, trying to conjugate the variability explained and the correlation structure.

2 Synthetic variables definition

We propose the application of CCA in the selection of synthetic variables that are representative of the clusters. This proposal aims to conjugate the amount of variability explained and the correlation structure of the data. Specifically, after determining the clusters of variables ($G_k : k = 1, \dots, K$), the basic idea to define the representative variable of the group G_k is

- Consider the groups $G_{k(1)}$ y $G_{k(2)}$ that, with their union in the agglomerative procedure of CAV, have formed the group G_k. Determine the canonical variables associated with the first and second canonical correlations between both groups of variables: $W_{1,k(1)}$ and $W_{2,k(1)}$ of $G_{k(1)}$ and $W_{1,k(2)}$ and $W_{2,k(2)}$ of $G_{k(2)}$.
- Consider as representative variables of G_k the canonical variables $W_{1,k(1)}$ and $W_{2,k(2)}$.

This procedure focuses on the correlation structure (not just variability). So, two synthetic variables are selected from each final cluster, defined as linear combinations of the original variables. It is necessary to specify this selection of variables if one of the groups has a unique original variable.

- **Case 1**: The two subgroups have a unique original variable. Assume without loss of generality that $G_{k(1)} = \{X_1\}$ and $G_{k(2)} = \{X_2\}$. Consider as representative variables of $G_k = G_{k(1)} \cup G_{k(2)}$ the original variables: $W_{1,k(1)} = X_1$ and $W_{2,k(2)} = X_2$
- **Case 2**: One of the subgroups has a unique original variable. Without loss of generality, $G_{k(1)} = \{X_1 \dots X_q\}$ and $G_{k(2)} = \{X_{q+1}\}$. Consider as representative variables X_{p+1} and the linear combination of the original variables $\{X_1 \dots X_q\}$ that has maximum variance (among all linear combinations) with the constraint that this new variable is uncorrelated with X_{q+1}.

Case 2 induces an optimization problem similar to that associated with PCA. Let $\underline{X}_{k(1)}$ be the vector of variables included in $G_{k(1)}$ and $X_{k(2)}$ be the variable included in $G_{k(2)}$. So, the representative variables are $W_{2,k(2)} = X_{k(2)}$ and $W_{1,k(1)} = \widehat{\gamma}^{\mathsf{T}}\underline{X}_{k(1)}$ such that

$$\widehat{\gamma} = \arg\sup\left\{\mathbf{c}^{\mathsf{T}}\widehat{\mathbf{\Sigma}}_{k(1)}\mathbf{c} \ : \ \mathbf{c}\in\mathbb{R}^{q-1} \ , \ \mathbf{c}^{\mathsf{T}}\mathbf{c} = 1 \ , \ \mathbf{c}^{\mathsf{T}}\widehat{\sigma}_{k(1),k(2)} = 0\right\} \tag{1}$$

being $\widehat{\mathbf{\Sigma}}_{k(1)}$ the sample covariance matrix of $\underline{X}_{k(1)}$ and $\widehat{\sigma}_{k(1),k(2)}$ the sample covariance vector between $X_{k(2)}$ and each variable in $\underline{X}_{k(1)}$.
Let $\mathbf{V}$ be a $(q\times(q-1))-$matrix such that its $(q-1)$ columns and σ forms an orthonormal basis of $\mathbb{R}^q$, with σ the normalized vector of $\widehat{\sigma}_{k(1),k(2)}$. That is, the matrix $[\mathbf{V} \quad \sigma]$ is orthonormal. Considering the matrix $\widehat{\mathbf{\Psi}}= \mathbf{V}^{\mathsf{T}}\widehat{\mathbf{\Sigma}}_{k(1)}\mathbf{V}$, the optimization problem (1) is equivalent to

$$\sup\left\{\mathbf{u}^{\mathsf{T}}\widehat{\mathbf{\Psi}}\mathbf{u} \ : \ \mathbf{u}\in\mathbb{R}^q \ , \ \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1\right\} \tag{2}$$

The solution of (2) is: $\mathbf{u} = \widehat{\mathbf{e}}_1$, the unit eigenvector associated with the largest eigenvalue $(\widehat{\psi}_1)$ of $\widehat{\mathbf{\Psi}}$, and the value of this supreme is $\widehat{\psi}_1$. So,

$$\widehat{\gamma}= \mathbf{V}\widehat{\mathbf{e}}_1, \quad \text{and} \quad \widehat{Var}(\widehat{\gamma}^{\mathsf{T}}\underline{X}_{k(1)}) = \widehat{\psi}_1$$

In general, if $\#[G_{k(1)}] > 1$ and $\#[G_{k(2)}] > 1$, let $\underline{X}_{k(1)}$ and $\underline{X}_{k(2)}$ the random vectors of $G_{k(1)}$ and $G_{k(2)}$ respectively. The first two sample canonical variables are considered

$$\begin{array}{llll} W_{1,k(1)} = \widehat{\alpha}^{\mathsf{T}}_{1,k(1)}\underline{X}_{k(1)} & \text{and} & W_{2,k(1)} = \widehat{\alpha}^{\mathsf{T}}_{2,k(1)}\underline{X}_{k(1)} & \text{of } G_{k(1)} \\ W_{1,k(2)} = \widehat{\beta}^{\mathsf{T}}_{1,k(2)}\underline{X}_{k(2)} & \text{and} & W_{2,k(2)} = \widehat{\beta}^{\mathsf{T}}_{2,k(2)}\underline{X}_{k(2)} & \text{of } G_{k(2)} \end{array}$$

where

$$\begin{array}{lcl} \widehat{\alpha}_{1,k(1)} \text{ and } \widehat{\alpha}_{2,k(1)} & : & \text{eigenvectors of } \widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}^{-1}_{k(1)} \ \widehat{\mathbf{\Sigma}}_{k(1),k(2)}\widehat{\mathbf{\Sigma}}^{-1}_{k(2)} \ \widehat{\mathbf{\Sigma}}^{\mathsf{T}}_{k(1),k(2)} \\ & & \text{associated with its two largest eigenvalues } (\varphi_1 \geq \varphi_2) \\ \widehat{\beta}_{1,k(2)} \text{ and } \widehat{\beta}_{2,k(2)} & : & \text{eigenvectors of } \widehat{\mathbf{G}} = \widehat{\mathbf{\Sigma}}^{-1}_{k(2)}\widehat{\mathbf{\Sigma}}^{\mathsf{T}}_{k(1),k(2)}\widehat{\mathbf{\Sigma}}^{-1}_{k(1)}\widehat{\mathbf{\Sigma}}_{k(1),k(2)} \\ & & \text{associated with its two largest eigenvalues } (\varphi_1 \geq \varphi_2) \end{array}$$

The synthetic variables are rescaled so that all have unit sample variance.

3 Homogeneity criterion and clustering procedure

Let $G = \{X_1 \ldots X_q\}$ be a cluster of $q > 1$ variables. Let W_1 and W_2 be the two synthetic variables associated with this group.
The homogeneity index of G is defined as

$$H(G) = \sum_{j=1}^{q}\frac{1}{2}\left[\rho^2(W_1,X_j) \ + \rho^2(W_2,X_j)\right]$$

Therefore, it is a measure of the internal correlation of the cluster, based on the linear association between the synthetic and the original variables. In the case of a partition with K clusters, $\mathcal{P} = \{G_1 \dots G_k \dots G_K\}$, the homogeneity index is defined as

$$H(\mathcal{P}) = \sum_{k=1}^{K} H(G_k).$$

Therefore, $H(G_k) \leq card(G_k) \ \ \forall k \Rightarrow H(\mathcal{P}) \leq p.$
Once the homogeneity index has been defined, the development proposed by Chavent *et al.* (2012) can be applied. The goal is to find a partition that maximizes this index. A classical hierarchical procedure is proposed, based on the dissimilarity:

$$d(C_1, C_2) = H(C_1) + H(C_2) - H(C_1 \cup C_2)$$

To evaluate the stability of the nested partitions, it is possible to use the procedure proposed by Chavent *et al.* (2012), based on the adjusted Rand index (Rand, 1971).

4 An example

In order to illustrate the proposed technique, we consider a simulated data set.The simulation procedure is similar to that used by Chen and Vigneau (2016). For $n = 100$ and $p = 16$, we consider a structure of variables with three groups, with sizes 10, 8 and 8. First, we have generated n cases of a vector $Z \sim N_6(0, \Phi)$ with

$$\Phi = \begin{pmatrix} 1.0 & 0.3 & 0 & 0 & 0 & 0 \\ 0.3 & 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0 & 0.3 & 1 \end{pmatrix}.$$

Second, let ω_{1j} and ω_{2j} be random numbers of a discrete uniform distribution $\{-1, 1\}$, δ_j a random number of a uniform distribution $U[2, 3]$ and γ_j a random number of a uniform distribution $U[0.5, 1.5]$.

$$\begin{aligned}
X_j = &\ \omega_{1j}\, \delta_j\, Z_1 + \omega_{2j}\, \gamma_j\, Z_2 + \varepsilon_j \quad j = 1, \dots, 5 \\
X_j = &\ \omega_{1j}\, \gamma_j\, Z_1 + \omega_{2j}\, \delta_j\, Z_2 + \varepsilon_j \quad j = 6, \dots, 10 \\
X_j = &\ \omega_{1j}\, \delta_j\, Z_3 + \omega_{2j}\, \gamma_j\, Z_4 + \varepsilon_j \quad j = 11, \dots, 14 \\
X_j = &\ \omega_{1j}\, \gamma_j\, Z_3 + \omega_{2j}\, \delta_j\, Z_4 + \varepsilon_j \quad j = 15, \dots, 18 \\
X_j = &\ \omega_{1j}\, \delta_j\, Z_5 + \omega_{2j}\, \gamma_j\, Z_6 + \varepsilon_j \quad j = 19, \dots, 22 \\
X_j = &\ \omega_{1j}\, \gamma_j\, Z_5 + \omega_{2j}\, \delta_j\, Z_6 + \varepsilon_j \quad j = 23, \dots, 26
\end{aligned}$$

with $\{\varepsilon_j,\ j = 1, \ldots, 26\}$ i.i.d. $N(0, 1)$. Figure 1 shows the sample correlation matrix of random vector $\underline{X} = [X_1 \ldots X_{26}]^\intercal$.

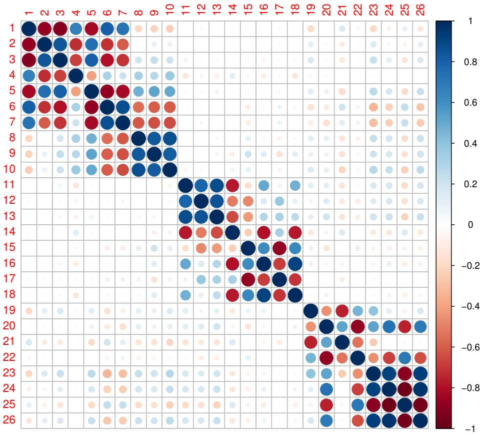


FIGURE 1. Correlation plot of simulated data set (Wei and Simko, 2017)

Finally, we applied the proposed technique, obtaining the dendrogram in Figure 2. In this figure, the height represents the change produced in the homogeneity index due to clustering of the variables. Cutting into three groups, we can see how the result of the technique reflects the process of data generation. The synthetic variables representative of the clusters are defined through the coefficients collected in Table 1.

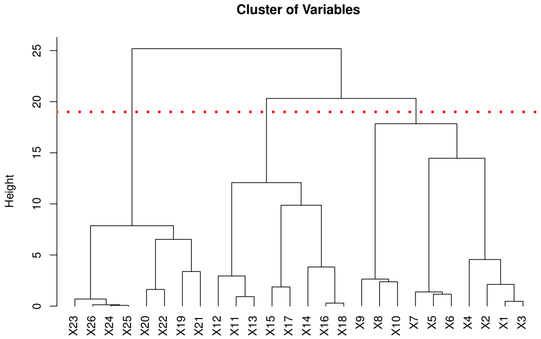


FIGURE 2. Cluster Dendrogram.

TABLE 1. Synthetic variables of the clusters and homogeneity index.

Cluster 1	Cluster 2	Cluster 3
W_{11}	W_{21}	W_{31}
X_1: -0.4648	X_{11}: 0.1480	X_{19}: 0.2556
X_2: 0.5091	X_{12}: 0.0241	X_{20}: 0.0198
X_3: -0.2544	X_{13}: 0.2007	X_{21}: -0.1415
X_4: 0.5934		X_{22}: 0.0904
X_5: -0.2629		
X_6: 0.0958		
X_7: 0.1512		
W_{12}	W_{22}	W_{32}
X_8: 0.0910	X_{14}: 0.0383	X_{23}: -0.0298
X_9: 0.1358	X_{15}: 0.1592	X_{24}: -0.0714
X_{10}: 0.2268	X_{16}: 0.0789	X_{25}: 0.0374
	X_{17}: -0.0159	X_{26}: -0.0594
	X_{18}: 0.1177	
$H(C_1) = 4.1951$	$H(C_2) = 7.1149$	$H(C_3) = 7.3129$
Homogeneity index of the partition: $H = 18.6230$		

References

Abraham, G., and Inouye, M. (2014). Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLOS ONE*, 9(4) e93766,1-5.

Chavent, M., Kuentz, V., Liquet, B., Saracco, J. (2012). ClustOfVar: An R package for the clustering of variables. *J. Stat. Software*, 50(13), 1–16.

Chen, M. and Vigneau. E. (2016). Supervised clustering of variables. *Adv. Data Anal. Classif.* 10:85–101.

Kiers, H. (1991). Simple structure in comp. analysis techniques for mixtures of qualitative and quant. variables. *Psychometrika*, 56:197–212.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. of the American Statistical Assoc.*, 66 (336), 846–850

Saracco, J., Chavent, M., and Kuentz, V. (2010). Clustering of categorical variables around latent variables. *Working Papers GRETHA* 2010-02.

Vigneau, E., and Qannari, E. (2003). Clustering of variables around latent components. *Comm. Statist.: Simulation & Comp.*, 32(4):1131-1150.

Wei, T. and Simko, W. (2017). R package ”corrplot”: Visualization of a Correlation Matrix (Vers. 0.84). https://github.com/taiyun/corrplot

A Waring regression model for overdispersed count data

Olmo-Jiménez, M. J.[1], Rodríguez-Avi, J.[1], Cueva-López, V.[1], Conde-Sánchez, A.[1], Martínez-Rodríguez, A.M.[1]

[1] Department of Statistics and Operations Research, University of Jaén, Spain

E-mail for correspondence: `mjolmo@ujaen.es`

Abstract: A regression model for overdispersed count data based on a biparametric case of the Univariate Generalized Waring (UGW) distribution, called the Biparametric Generalized Waring (BW) distribution is developed. The BW inherits the main properties of the UGW distribution. Besides the fact that the BW has fewer parameters than the UGW, the main advantage of the former is that the identification problem of the UGW first two parameters disappears. These characteristics make the BW distribution of interest as an underlying distribution in a regression model for overdispersed count data. So, we first recall the definition and the main properties of the BW distribution. Secondly, we describe the regression model based on the BW distribution and the estimation method for its parameters. Finally, we show some application examples to illustrate the utility of the proposed regression model compared with other usual regression models for overdispersed count data, such as those based on the negative binomial, generalized Poisson and UGW distributions.

Keywords: Count data; Overdispersion; Regression models; Waring distribution.

1 The conditional BW distribution

1.1 Definition

The biparametric univariate Waring (BW) distribution with parameters $\alpha, \rho > 0$ was developed by Rodríguez-Avi *et al.* (2018). It is a count data distribution of infinite range generated by the Gaussian hypergeometric function. Thus, if X follows a $BW(\alpha, \rho)$, its probability mass function (pmf) is

$$f(x) = \frac{\Gamma(\alpha+\rho)^2}{\Gamma(\alpha)^2\Gamma(\rho)}\frac{\Gamma(\alpha+x)^2}{\Gamma(2\alpha+\rho+x)}\frac{1}{x!}, \quad x = 0, 1, \ldots$$

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

This distribution is a particular case of the Univariate Generalized Waring (UGW) distribution (Irwing, 1975a,b,c). Specifically, a $BW(\alpha, \rho)$ is a $UGW(\alpha, \alpha, \rho)$.

1.2 Properties

The main properties of the BW distribution are inherited from the UGW distribution.
The mean, μ, and the variance, σ^2, of the BW distribution have the following explicit expressions:

$$\mu = \frac{\alpha^2}{\rho - 1}, \quad \sigma^2 = \frac{\alpha^2(\alpha + \rho - 1)^2}{(\rho - 1)^2(\rho - 2)}, \tag{1}$$

which exist if $\rho > 1$ and $\rho > 2$, respectively. In general, the $m - th$ raw moment exists if $\rho > m$.
The BW distribution can be obtained as the following two-step mixture:

1. $X|\lambda \sim P(\lambda)$ with pmf

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

2. $\lambda|\alpha, \theta \sim Gamma(\alpha, \theta)$ with density function

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\theta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\theta}, \quad \lambda > 0.$$

Then, $X|\alpha, \theta \sim NB(\alpha, \theta)$ with pmf

$$f(x) = \frac{1}{x!}\frac{\Gamma(x+\alpha)}{\Gamma(\alpha)}\left(\frac{1}{1+\theta}\right)^\alpha\left(\frac{\theta}{1+\theta}\right)^x, \quad x = 0, 1, \ldots$$

3. $\theta|\alpha, \rho \sim BetaII(\alpha, \rho)$ with density function

$$f(\theta) = \frac{\Gamma(\alpha+\rho)}{\Gamma(\alpha)\Gamma(\rho)}\theta^{\alpha-1}(1+\theta)^{-(\alpha+\rho)}, \quad \theta > 0.$$

Since the BW is a Poisson mixture, it is always overdispersed.
In addition, the variance of the BW model can be split into three components

$$\sigma^2 = \frac{\alpha^2}{\rho - 1} + \frac{\alpha^2(\alpha+1)}{(\rho-1)(\rho-2)} + \frac{\alpha^3(\alpha+\rho-1)}{(\rho-1)^2(\rho-2)},$$

where the first term of this decomposition represents the variability due to randomness and comes from the underlying Poisson model. The other two terms refer to the variability that is not due to randomness but it is explained by the presence of liability and proneness, respectively.

2 Model specification

Let Y be the response variable of a count model so that $Y|\mathbf{x}$ follows a $BW(\alpha_{\mathbf{x}}, \rho_{\mathbf{x}})$, where $\mathbf{x} = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_k \end{pmatrix}$ is the vector of covariates. Then, taking into account the expression of the mean of the model given in (1) and considering the effect that the covariates have on the mean, that is

$$\mu_{\mathbf{x}} = e^{\mathbf{x}'\beta}$$

where $\beta = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_k \end{pmatrix}$ is the parameter vector, different regression models based on the BW distribution can be generated by linking its parameters with the covariates. Specifically,

Model 1. We may consider that $\rho_{\mathbf{x}}$ does not depend on the covariates but $\alpha_{\mathbf{x}}$ does through the mean, that is, $\rho_{\mathbf{x}} = \rho$ is a fixed parameter whereas $\alpha_{\mathbf{x}} = +\sqrt{\mu_{\mathbf{x}}(\rho - 1)}$. From now on we refer to this model as $BWRM_1$. To guarantee the existence of the mean and the computation of $\alpha_{\mathbf{x}}$ it is necessary to impose that $\rho > 1$.

Model 2. On the other hand, we also may consider that $\alpha_{\mathbf{x}}$ does not depend on the covariates but $\rho_{\mathbf{x}}$ does through the expression of the mean, that is $\alpha_{\mathbf{x}} = \alpha$ and $\rho_{\mathbf{x}} = \alpha^2/\mu_{\mathbf{x}} + 1$. We refer to this model as $BWRM_2$. Now it is not necessary to impose any restriction on ρ because it is directly greater than 1, but α must be positive.

3 Model estimation

The estimation of the regression coefficients $\beta_0, \ldots, \beta_k$ and the parameters α or ρ, according to the case, is carried out maximizing the log-likelihood function by numerical methods. Thus, if $\mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix}$ is a sample of size n the log-likelihood function is given by

$$\begin{aligned} \ln L(\alpha, \rho|\mathbf{y}) = &\sum_{i=1}^{n} [2\ln\Gamma(\alpha + y_i) - \ln\Gamma(2\alpha + \rho + y_i)] \\ &- 2n\ln\Gamma(\alpha) + 2n\ln\Gamma(\alpha + \rho) - n\ln\Gamma(\rho). \end{aligned}$$

As we are modelling μ in function of the covariates, in each step of the optimization processs we must replace the corresponding parameter α or ρ by its expression in terms of the mean.

4 Practical application

To illustrate the performance of the proposed regression model, we have carried out several fits to real data and we have compared them with those provided by the regression models based on the negative binomial (NB), generalized Poisson (GP) and UGW distributions. These models have been fitted using R (R, 2018). Specifically, we have used the `glm.nb` function of the MASS package for the NB regression model, the `vglm` function of the pscl package for the GP regression model (Consul and Famoye, 1992) and the `gw` function of the GWRM package for the UGW regression model (Rodríguez-Avi *et al.*, 2009). Regarding the $BWRM_1$ and $BWRM_2$, we have implemented our own fitting functions using the `optim` function of the STATS package.

From these examples we conclude that the $BWRM$ can provide more accurate fits (using the Akaike information criterion) than other usual regression models as well as an interesting partition of the variance in terms of the covariates. In addition, when the $BWRM$ provides similar results than the $GWRM$, the standard errors of the regression coefficients are usually smaller.

References

Consul, P.C. and Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics*, **21**(1), 89 – 109.

Irwing, J.O. (1975a). The generalized Waring distribution. Part I. *Journal of the Royal Statistical Society, A*, **138**, 18 – 31.

Irwing, J.O. (1975b). The generalized Waring distribution. Part II. *Journal of the Royal Statistical Society, A*, **138**, 204 – 227.

Irwing, J.O. (1975c). The generalized Waring distribution. Part III. *Journal of the Royal Statistical Society, A*, **138**, 374 – 378.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J., Olmo-Jiménez, M.J. and Martínez-Rodríguez, A.M. (2009). A generalized Waring regression model for count data. *Computational Statistics and Data Analysis*, **53**, 3717 - 3725.

Rodríguez-Avi, J., Olmo-Jiménez, M.J. and Cueva-López, V. (2018). An over- and underdispersed extension of the generalized Waring distribution. *Submitted to Statistical Modelling.*

Empirical evidence of avalanche danger patterns on recreational avalanche danger accidents reported in Tyrol within 2010–2013.

Christian Pfeifer[1]

[1] Institut für Statistik, Universität Innsbruck, Universitätsstraße 15, A–6020 Innsbruck

Abstract: In this case study we study the effects of avalanche danger patterns on avalanche danger in a multiple response scenario.

Keywords: Avalanche Danger Patterns, Multiple Response Variables, Decision Strategy.

1 Introduction

In recent years backcountry skiing has become very popular. Unfortunately, there are quite a number of avalanche accidents which cause about 20 fatalities in Austria every year (Pfeifer et al. (2018)). However, efforts have been made in order to prevent backcountry avalanche accidents, see for example Munter (1997). Since 2010, the Tyrolean avalanche service is publishing special information for backcountry skiers every day which they call 'danger patterns' (Mair and Nairz (2012)) such as:

1. deep persistent weak layer
2. gliding avalanche
3. rain
4. cold following warm/warm following cold
5. snowfall after a long period of cold
6. cold, loose snow and wind
7. snow-poor zones in snow-rich surrounding

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

8. surface hoar blanketed with snow
9. graupel blanketed with snow
10. springtime scenario.

However, the authors Mair and Nairz did not give any empirical evidence of the effects of their danger patterns on avalanche danger.
The data collection design of the danger patterns (DP) is done in the manner of multiple response questionaries: At most 3 patterns are recorded every day - at least one first order DP (the most important DP) and if appropriate a second order or 3rd order DP with decreasing importance.

There is some few literature concerning models with multiple response variables – see Agresti and Liu (2001), Bilder et al. (2000) (2004) (2009) or Loughin and Scherer (1998), for example. But up to now, we could not find examples in the literature taking multiple responses with unequal importance into account.

2 Data and Statistical Models

In this approach, we are going to explore the effects of these danger patterns on the number of recreational avalanche accidents using accident data in Tyrol within 2010–2013 (winter period without 'spring condition', number of cases: 288). In order to take the skiers frequency into account we introduce weather/skiing conditions and weekend Yes/No as covariates into the loglinear model (number of daily accidents as dependent variable) – see also Pfeifer (2009).
In the following approach we compare 2 types of loglinear models,

$$\log(\mu_{\mathbf{t}}) = \mathtt{GM_x} + \mathtt{WOENDE} + \mathtt{TOURV} \tag{1}$$

where $\mathtt{GM_x}$ denote danger patterns, $\mathtt{TOURV}$ skiing conditions and $\mathtt{WOENDE}$ an indicator for during week/weekend.
In case of Model 1 $\mathtt{GM_x}$ equal to 0/1 denote no/at least one occurrence of danger pattern $\mathtt{x}$ at a specific observed day. In case of Model 2 $\mathtt{GM_x}$ equal to 0 denotes no, equal to 1 denotes a first order (or a primary) DP and equal to 2 denotes a second or third order (a secondary or low-level) DP, which is a simple approach taking the order of importance into account.
Table 1 and Table 2 show the results restricted to the effects of the DPs for the 2 different types.

3 Results and Discussion

As we can see, most of the danger patterns do not show significant results if we consider effects on the number of avalanches accidents - with the

TABLE 1. Results of Model Type 1.

GM_x	Estimate	Std. Error	z value	$Pr(>\|z\|)$
GM1	0.128	0.320	0.400	0.689
GM2	0.010	0.198	0.051	0.959
GM3	-1.093	0.719	-1.520	0.129
GM4	-0.456	0.334	-1.366	0.172
GM5	1.412	0.212	6.670	**0.000**
GM6	0.019	0.233	0.081	0.935
GM7	-15.225	661.034	-0.023	0.982
GM8	0.471	0.260	1.811	0.070
GM9	0.489	0.376	1.301	0.193
GM10	-0.825	0.427	-1.931	0.053

TABLE 2. Results of Model Type 2.

	Primary danger pattern				Secondary danger pattern			
GM_x	Est.	Std. E.	z	Pr	Est.	Std. E.	z	Pr
GM1	1.010	1.011	0.999	0.318	0.069	0.334	0.208	0.835
GM2	-0.253	0.327	-0.774	0.439	0.096	0.211	0.456	0.648
GM3	-15.183	696.863	-0.022	0.983	-0.409	0.717	-0.571	0.568
GM4	-15.185	930.829	-0.016	0.987	-0.346	0.333	-1.038	0.299
GM5	1.666	0.219	7.612	**0.000**	0.341	0.527	0.646	0.518
GM6	-0.104	0.243	-0.427	0.670	0.462	0.299	1.546	0.122
GM7	-15.463	1480.000	-0.010	0.992	-15.161	740.930	-0.020	0.984
GM8	-0.222	1.009	-0.220	0.826	0.533	0.267	1.996	**0.046**
GM9	0.336	0.202	1.662	0.097	0.489	0.376	1.301	0.193
GM10	-1.356	0.717	-1.892	0.058	-0.384	0.519	-0.739	0.460

exception of the DP GM5 ‘snowfall after a long period of cold’, which is in some sense in accordance with the results of Pfeifer and Höller (2014). Additionally, we notice a significant result in case of the DP GM8 ‘surface hoar blanketed with snow’ on low-level (or as a secondary DP).

There is some discussion in the snow science community that the actual danger levels are insufficient if we consider recreational accidents only. Especially danger level ‘3–considerable’ seems to have a (too) wide range covering the highest number of accidents. As a result of the models above, we recommend to split up danger level 3 into levels with and without danger pattern GM5!

In general, further research on multiple response models with multiple response variables either in the dependent part or in the explaning part of statitical models is recommended.

References

Agresti, A. and Liu, I. (2001). Strategies for Modeling a Categorical Variable Allowing Multiple Category Choices. *Sociological Methods & Research*, **29(4)**, 403 – 434.

Bilder, C.R., Loughin, T.M. and Nettleton, D (2000). Multiple Marginal Independence Testing for Pick Any/C Variables. *Communication in Statistics - Simulation and Computation*, **29(4)**, 1285 – 1316.

Bilder, C.R. and Loughin T.M. (2004). Testing for Marginal Independence between Two Categorical Variables with Multiple Responses. *Biometrics*, **60**, 241 – 248.

Bilder, C.R. and Loughin, T.M. (2009). Modeling multiple-response categorical data from complex surveys. *The Canadian Journal of Statistics 2009*, **37(4)**, 553 – 570.

Loughin, T.M. and Scherer. P.N. (1998). Testing for Association in Contingency Tables With Multiple Column Responses. *Biometrics*, **54**, 630 – 637.

Mair, R. and Nairz, P. (2012). *Lawine. Die entscheidenden 10 Gefahrenstufen zu erkennen: Praxis-Handbuch.* Innsbruck: Tyrolia.

Munter, W. (1997). *3x3 Lawinen.* Garmisch-Partenkirchen: Pohl & Schellhammer.

Pfeifer, C. (2009). On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities. *Natural Hazards*, **48(3)**, 425 – 438.

Pfeifer, C. and Höller, P. (2014). Effects of precipitation and temperature in alpine areas on backcountry avalanche accidents reported in the western part of Austria within 1987–2009. In: *Proceedings International Workshop of Statistical Modelling 2014.* Göttingen.

Pfeifer, C., Höller, P. and Zeileis, A. (2018). Spatial and temporal analysis of fatal off-piste and backcountry avalanche accidents in Austria with a comparison of results in Switzerland, France, Italy and the US. *Natural Hazards and Earth System Sciences*, **18**, 571 – 582.

Penalised-based estimation of covariate-specific time-dependent ROC curves

María Xosé Rodríguez-Álvarez[1], Thomas Kneib[2], Vanda Inácio de Carvalho[3]

[1] BCAM - Basque Center for Applied Mathematics & IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[2] Chair of Statistics, Georg-August-Universität Göttingen, Germany
[3] School of Mathematics, University of Edinburgh, Scotland, United Kingdom

E-mail for correspondence: mxrodriguez@bcamath.org

Abstract: This work presents a novel penalised likelihood-based estimator of the cumulative-dynamic time-dependent receiver operating characteristic (ROC) curve. The proposal allows to account for the possible modifying effect of covariates on the accuracy of prognostic biomarkers. We apply our approach to the evaluation of biomarkers for early prognosis of death after discharge in patients who suffered an acute coronary syndrome.

Keywords: time-dependent ROC curve; P-splines; hazard function.

1 Introduction

The ROC curve is the measure of diagnostic accuracy most widely used for continuous biomarkers. However, in many circumstances, the aim of a study may involve prognosis rather than diagnosis. In such cases, the disease status of an individual is not a fixed characteristic but it varies with time (e.g, death and alive). To assess the accuracy of continuous biomarkers for time-dependent disease outcomes, time-dependent extensions of *Sensitivity*, *Specificity* and ROC curve have been proposed (e.g., Pepe et al. 2008). Moreover, it is well known that the accuracy of a biomarker can be affected by external information or covariates, for instance, characteristics of the patient (Pepe, 2003). In these situations, if we failure to incorporate covariate information into the ROC analysis, the marginal or pooled

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

ROC curve could lead to erroneous conclusions, and thus conditional or covariate-specific measures of accuracy are needed. This work focuses on the estimation of the conditional cumulative-dynamic time-dependent ROC curve. In contrast to previous proposals in this setting, our approach (1) allows for non-linear effects of continuous covariates on the accuracy of prognostic biomarkers, and (2) relaxes the proportional hazards assumption.

2 Notation and preliminaries

Let T denote the time to the event of interest, Y the quantitative biomarker, and $\boldsymbol{X}$ the p-variate vector of covariates we are interested in. The conditional or covariate-specific time-dependent cumulative *Sensitivity* (Se) and dynamic *Specificity* (Sp) are defined as

$$\begin{aligned} Se^{\mathbb{C}}(v, t \mid \boldsymbol{x}) &= \Pr[Y > v \mid T \leq t, \boldsymbol{X} = \boldsymbol{x}], \\ Sp^{\mathbb{D}}(v, t \mid \boldsymbol{x}) &= \Pr[Y \leq v \mid T > t, \boldsymbol{X} = \boldsymbol{x}]. \end{aligned}$$

Thus, the conditional cumulative-dynamic time-dependent ROC curve is

$$\mathrm{ROC}_{t,\boldsymbol{x}}^{\mathbb{C}/\mathbb{D}}(p) = Se^{\mathbb{C}}\left((1 - Sp^{\mathbb{D}})^{-1}(p, t \mid \boldsymbol{x}), t \mid \boldsymbol{x}\right) \text{ with } p \in (0,1).$$

Note that with the cumulative *Sensitivity* and dynamic *Specificity* interest lies in evaluating the discriminatory capacity of the biomarker Y in distinguishing those individuals – with a covariate vector value $\boldsymbol{x}$ – that will experience the event of interest prior to time t (cases) from those with the event after t (controls). Note also that a possibly different ROC curve, and therefore discriminatory capacity, can be obtained for each covariate vector value $\boldsymbol{x}$ and each time point t.

It can be easily shown that the above expressions can be expressed as

$$Se^{\mathbb{C}}(v, t \mid \boldsymbol{x}) = \frac{\Pr[Y > v, T \leq t \mid \boldsymbol{X} = \boldsymbol{x}]}{\Pr[T \leq t \mid \boldsymbol{X} = \boldsymbol{x}]} = \frac{\int_{v}^{\infty} (1 - S(t \mid y, \boldsymbol{x}))\, dF(y \mid \boldsymbol{x})}{\int_{-\infty}^{\infty} (1 - S(t \mid y, \boldsymbol{x}))\, dF(y \mid \boldsymbol{x})}, \tag{1}$$

$$Sp^{\mathbb{D}}(v, t \mid \boldsymbol{x}) = \frac{\Pr[Y \leq v, T > t \mid \boldsymbol{X} = \boldsymbol{x}]}{\Pr[T > t \mid \boldsymbol{X} = \boldsymbol{x}]} = \frac{\int_{-\infty}^{v} S(t \mid y, \boldsymbol{x})\, dF(y \mid \boldsymbol{x})}{\int_{-\infty}^{\infty} S(t \mid y, \boldsymbol{x})\, dF(y \mid \boldsymbol{x})}. \tag{2}$$

where

$$S(t \mid y, \boldsymbol{x}) = P(T > t \mid Y = y, \boldsymbol{X} = \boldsymbol{x}) \text{ and } F(y \mid \boldsymbol{x}) = P(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}).$$

3 Penalised-based estimator

Expressions (1) and (2) make it clear that in order to estimate $Se^{\mathbb{C}}$ and $Sp^{\mathbb{D}}$ we simply need an estimator of $S\left(t \mid y, \boldsymbol{x}\right)$ and $F\left(y \mid \boldsymbol{x}\right)$. For estimating $S\left(t \mid y, \boldsymbol{x}\right)$, we assume a regression-type model for the conditional hazard function $\lambda\left(t \mid y, \boldsymbol{x}\right)$, i.e.,

$$\begin{aligned}\lambda\left(t|y, \boldsymbol{x}\right) = \exp\Bigg(&\alpha_0 + h_t\left(t\right) + h_y\left(y\right) + \sum_{a=1}^{A} f_a\left(\boldsymbol{x}_a\right) \\ &+ f_{y,t}\left(y,t\right) + \sum_{b=1}^{B} f_b\left(t, \boldsymbol{x}_b\right) + \sum_{c=1}^{C} f_c\left(y, \boldsymbol{x}_c\right)\Bigg),\end{aligned} \tag{3}$$

where $\boldsymbol{x}_a$, $\boldsymbol{x}_b$ and $\boldsymbol{x}_c$ denote subsets of covariates, and $h_{\{.\}}$ and $f_{\{.\}}$ define generic representations of different types of covariates and effects (linear or parametric, smooth, etc). Note that the inclusion of functions $f_{y,t}$ and f_b allow relaxing the proportional hazards assumption. Estimation is based on the piecewise exponential model (Friedman, 1982; Kauermann, 2005). Via a data augmentation strategy, the piecewise exponential approach allows a (penalised) Poisson-maximum likelihood estimation scheme for model (3) in the presence of censored observations. In addition, it also allows using Generalised Linear Array Models (GLAM, Currie et al, 2006) to speed up computation.

In order to estimate $F\left(y \mid \boldsymbol{x}\right)$, we propose the following model

$$Y \mid \left(\boldsymbol{X} = \boldsymbol{x}\right) = \beta_0 + \sum_{v=1}^{V} f_v\left(\boldsymbol{x}_v\right) + \varepsilon, \tag{4}$$

with $\boldsymbol{x}_v$ and f_v as defined before. We assume that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ and ε is independent of $\boldsymbol{X}$. Thus, $F\left(y \mid \boldsymbol{x}\right) = H\left(y - \beta_0 - \sum_{v=1}^{V} f_v\left(\boldsymbol{x}_v\right)\right)$, where $H(u) = \Pr(\varepsilon \leq u)$.

For the specification of the (multidimensional) smooth functions involved in models (3) and (4), use is made of penalised splines (P-splines, Eilers and Marx, 1996, 2003), in combination with (tensor-product of) B-spline basis functions. In addition, each smooth function is decomposed into a penalised and an unpenalised component (see. e.g., Currie and Durban, 2002). This decomposition presents several attractive features: (a) redundant components can be easily identified; and (b) generalised linear mixed models estimation techniques can be used. In this work, estimation of models (3) and (4) is done by means of the method described in Rodríguez-Álvarez et al. (2018).

4 Application

The Global Registry of Acute Coronary Events (GRACE) scoring system is a well-known risk score (biomarker) for early prognosis of death after discharge in patients who suffered from acute coronary syndrome (ACS). In the construction of the GRACE scoring system, the left ventricular ejection fraction (LVEF), a very well-established prognosticator of mortality in the ACS scenario, was not included, mainly due to presence of missing values. Thus, the death risk estimates from the GRACE risk score might be misleading because the LVEF is conspicuous by its absence in the construction of the GRACE. In order to check this hypothesis, we applied the proposal presented in this work with the aim of evaluating the possible effect of the LVEF on the prognostic value of the GRACE risk score.

The study population consists of 3488 consecutive patients admitted due to ACS at the Hospital Clínico de Santiago de Compostela, Spain. The event of interest was all-cause mortality during follow-up (81% censorship). Figure 1 shows the estimated time-dependent area under the ROC curve (AUC) for the GRACE score adjusted by LVEF at $t = 6$, $t = 12$ and $t = 18$ months after discharge. As can be observed, there is a clear effect of the LVEF on the prognostic value of the GRACE risk score. The red lines represent the estimated marginal/pooled time-dependent AUC, i.e., the time-dependent AUC obtained when pooling the data without regard the LVEF values. These results highlight that not accounting for the possible modifying effect of the LVEF on the prognostic value of the GRACE risk score would yield to optimistic results.

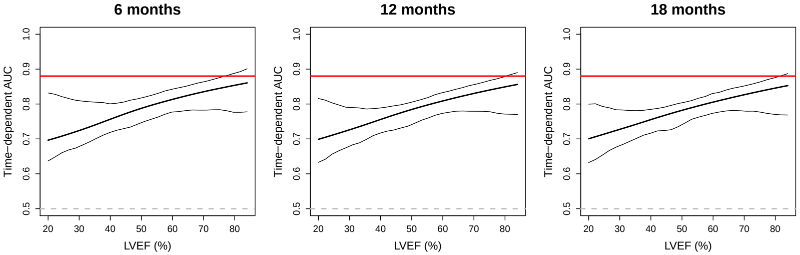


FIGURE 1. Estimated time-dependent AUCs for the GRACE score adjusted by LVEF(%) at $t = 6$, $t = 12$ and $t = 18$ months after discharge (solid black lines). The red lines represent the marginal/pooled time-dependent AUC.

Acknowledgments: This research was supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through project MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym “AFTERAM”.

References

Currie, I., and Durban. M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, **4**, 333 – 349.

Currie, I., 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*, **68**, 259 – 280.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 121.

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.

Friedman, M. (1982). Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, **10**, 101 – 113.

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

Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.

Pepe, M.S., Zheng, Y., Jin, Y., Huang, Y., Parikh C.R., and Levy, W.C. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, **14**, 86 – 113.

Rodríguez-Álvarez, M.X., Durban, M., Lee, D.-J., and Eilers, P.H.C. (2018). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *ArXiv: https://arxiv.org/abs/1801.07278*

New Methods for Generating Dynamic Reference Ranges for Blood Biomarkers With Applications in Prostate Cancer

Davood Roshan[1], John Ferguson[2], Francis J. Sullivan[3], John Newell[1]

[1] School of Mathematics, Statistics, and Applied Mathematics, National University of Ireland, Galway (NUI Galway)
[2] HRB Clinical Research Facility, NUI Galway, Ireland
[3] Prostate Cancer Institute, NUI Galway, & Galway Clinic, Ireland

E-mail for correspondence: D.roshansangachin1@nuigalway.ie

Abstract: Prostate Cancer (PC) represents the most common malignancy affecting men in western countries and Ireland has the highest rate of PC in Europe. Prostate Specific Antigen (PSA), which is a glycoprotein produced by the prostate gland, is usually increased in patients with PC. Therefore, PSA plays an essential role in the detection of patients with PC. Generally, the normal value for PSA has been considered as 4.0 ng/mL or lower. However, several other factors including age, urinary tract infection and prostatitis can cause an increase in the PSA levels. For this reason, age-related PSA reference ranges are used in most hospitals for the screening of PC. Further investigation is required to detect the presence of PC for a man whose age-related PSA is above the normal range. Prostate biopsy is the definitive diagnostic test required to establish a diagnosis of PC. However, using normal reference ranges to screen for PC can often lead to unnecessary biopsy with an increased cost and potential complications including infection, bleeding and etc. To overcome this drawback, we instead suggest generating personalized dynamic ranges for PSA and for other clinical biomarkers using Bayesian approaches and streaming algorithms. In contrast with normal ranges which in a sense weight all subjects in a certain age bracket of the population equally, dynamic reference ranges are tailored to the measurements observed on one subject, and as a result are more accurate in determining meaningful changes in biomarker trajectory. The MCMC method is applied to generate dynamic ranges from the posterior predictive distribution for the Bayesian approach while a recently proposed approximate EM algorithm for streaming data is modified to produce computationally efficient dynamic ranges for large streaming datasets.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Keywords: Reference Ranges; Personalized Dynamic Ranges; Bayesian Approach; Streaming Algorithm; Prostate Cancer.

1 Introduction

Biological markers (Biomarkers) are characteristics that represent a normal or abnormal biological process in the diagnosis of a disease or a condition. They play a key role in understanding the underlying pathogenesis of disease and extending our knowledge of normal, healthy physiology. Determining the trend and changes of biomarkers preceding a clinical stage will not only enable early detections but also facilitate the required therapeutic trials. A reference interval, generated from a cross-sectional analysis of healthy individuals free of the disease of interest, is typically used when interpreting a set of biomarker test results for a particular patient. Reference ranges often are defined as an inter-percentile range. Depending on the knowledge of the biomarker's underlying distribution, either parametric or non-parametric inter-percentile intervals can be selected. In the parametric method, the required percentiles will be calculated according to the underlying distribution of the biomarker by estimating the population parameters (e.g. mean, standard deviation). However, for the non-parametric method, where no distribution is assumed for the biomarkers, the 95% reference interval will be defined as the middle 95% of the biomarker values. A confidence interval around the percentiles is necessary for measuring the uncertainty of the interval. These 'static' normal ranges not only incorporate sampling error and uncertainty due to the sampling process but also may not be reflective of a particular individual in the population with longitudinal follow up. Therefore, when biomarkers are collected longitudinally for subjects, dynamic reference ranges which adapt to account for between and within subject variabilities are needed for effective diagnosis. This is especially useful when the within individual variability is much less than between individual variability.

In this study generating such personalized dynamic ranges for clinical biomarkers will be discussed using Bayesian approaches and streaming algorithms. The performance of the different models was assessed through simulation study under different scenarios. Also, the proposed models were evaluated on a cohort of men with longitudinally recorded PSA measurements, and used to define personalized risk zones for PC diagnosis.

2 Methodology

Bayesian approaches have the capacity to intelligently combine information from the general population with the measurements for a given individual to construct a personalized reference range. Using information gleaned from

the population allows the construction of critical ranges for the first measurement of the individual. This range is then adapted as more observations have been gathered on the same subject. Finally, these critical ranges will be predominantly based on the individual measurements as the number of biomarker values becomes very large for the individual. For example let $[x_{i1}, ..., x_{in}]$ shows a series of biomarker values for a specific individual over n follow up times. Therefore, a new measurement is considered as abnormal if it falls beyond the $\frac{\alpha}{2}*100\%$ and $(1-\frac{\alpha}{2})*100\%$ quantiles of $P(X_{i,new}|data)$; where data refers to both population and subject information up to, but not including the new observation.
The Bayesian approach for the computation of proposed ranges is computationally intensive. For example, in large data streaming problems, there is a need for a framework which is computationally efficient. This can be achievable using an approximate EM algorithm which is very quick for large data. Ippel and etc. (2016) proposed an approximation of the EM algorithm to fit a random intercept model to large streaming datasets. Unlike the EM algorithm which uses all the data to update the model parameters, their new method only uses a single data point, some summary statistics on the individual level, and the previous estimates of the model parameters, to update the model parameters. Their approach is then adapted in our study to define the dynamic range from the distribution of $X_{ij}|X_{i1}, ..., X_{i(j-1)}$ when the model parameters are estimated using both the population information and the previous records of the individual.

3 Results:

The performance of the three different approaches in detecting the abnormal observations in a sample of biomarker values was assessed by measuring the area under the ROC curve (AUC) through a simulation study. Different scenarios were considered by varying the sample size (n), individual replications (n_i), and the ratio of within subject variability to between subject variability ($r_1 = \frac{var(\sigma_i^2)}{\tau^2}, r2 = \frac{E(\sigma_i^2)}{\tau^2}$), where σ_i^2 and τ^2 representing the within subject variability for subject i and between subject variablility, respectively.
The simulation findings show that the overall performance of the Bayesian approach is the best for all situation. However, when within subject variabilities are large relative to between subject variability, all three approaches resulted in a similar performance suggesting the normal range is as good as the other two approaches (figure 1). Additionally, as can be seen in figure 2, when both sample size and individual's replication increases, the mean AUC for the EM algorithm approaches the mean AUC for the Bayesian approach with both methods significantly better than the static reference ranges.

The application of the proposed models to detection of PC using longitudinal records of PSA biomarker is displayed in figure 3. As can be seen the patient is diagnosed with PC at the seventh follow up time using both the Bayesian and EM algorithm, while the reference range was unable to detect his cancer as all of the PSA test results are within the static reference range.

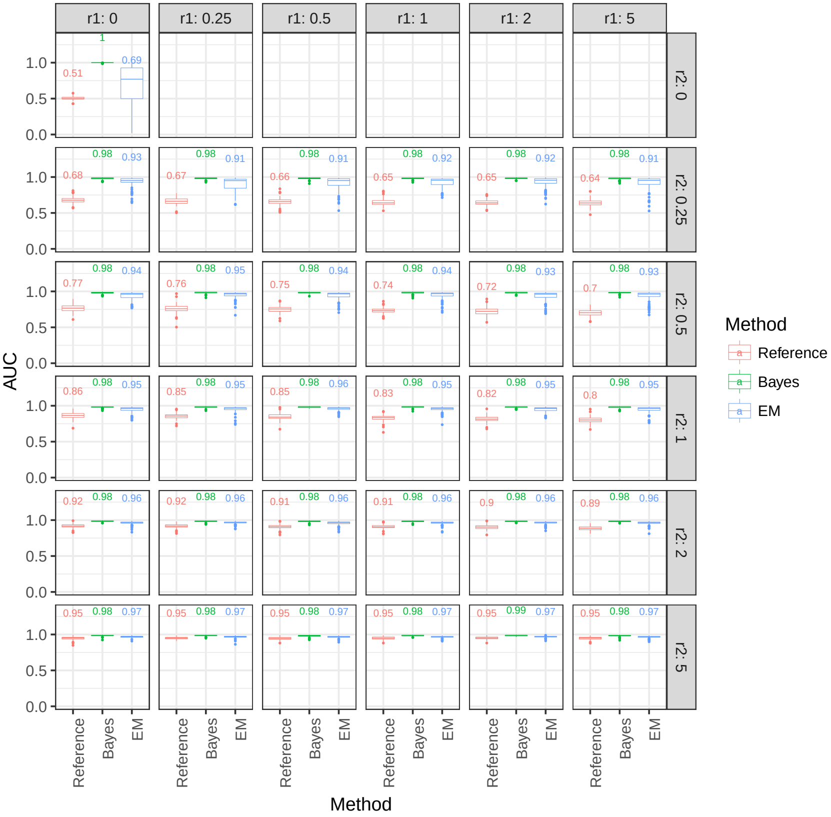


FIGURE 1. The AUC distribution of the three approaches (Reference Range, Bayesian, and Streaming EM algorithm) for detection of outliers based on different combination of $r_1 \& r_2$.

4 Conclusion:

This study introduces the idea of dynamic ranges for a longitudinal continuous response through the Bayesian model and the streaming EM algorithm.

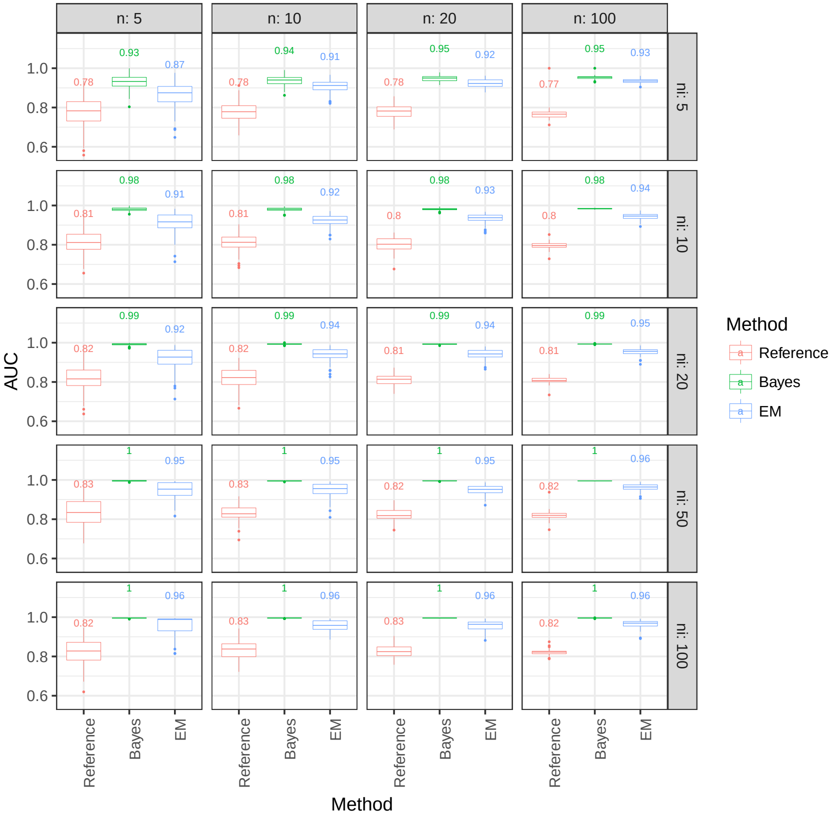


FIGURE 2. The AUC distribution of the three approaches (Reference Range, Bayesian, and Streaming EM algorithm) for detection of outliers based on different combination of $n\&n_i$

The generation of personalized dynamic ranges for clinical biomarkers helps researchers and physicians make more reliable decisions in terms of what can be considered as the normal physiology of an individual. Additionally, the models can be easily extended to allow the incorporation of other covariates and factors (e.g. age, gender) that likely affect these ranges.

Acknowledgments: This research is supported by Prostate Cancer Institute, NUIG and Irish Research Council.

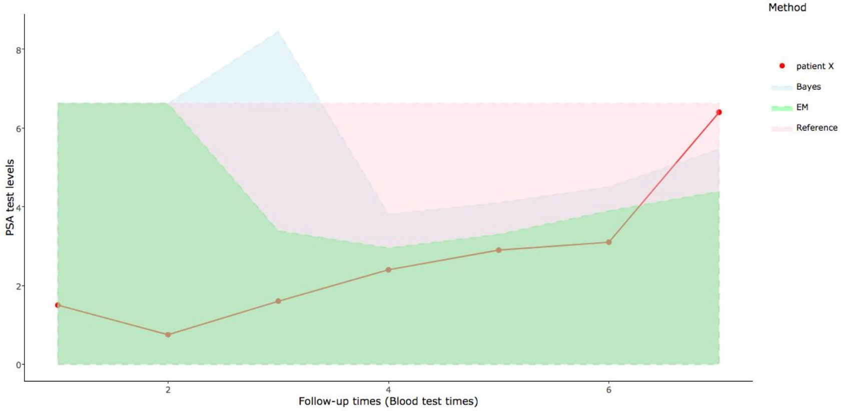


FIGURE 3. Personalized Dynamic Range for a patient diagnosed with PC based on the three different approaches.

References

Ippel, L., Maurits Clemens Kaptein, and Jeroen K. Vermunt (2016). *Estimating random-intercept models on data streams.* Computational Statistics & Data Analysis **104**, 169 – 182.

Irish Cancer Society (2012). *understanding The PSA Test* . Ireland.

Solberg, H. E. (1987). *Approved Recommendation (1987).* J. Clin. Chem. Clin. Biochem **25.9**, 645 – 656.

Sottas, Pierre-Edouard, et al. (2006). *Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio.* Biostatistics **8.2**, 285 – 296.

Network reconstruction for gene-phenotype relationships - a case study in *Populus nigra*

Sabine K. Schnabel[1], M. João Paulo[1], Georgios Bartzis[1], Joost Keurentjes[2] and Fred van Eeuwijk[1]

[1] Biometris, Wageningen University & Research, The Netherlands
[2] Laboratory of Genetics, Wageningen University & Research, The Netherlands

E-mail for correspondence: `sabine.schnabel@wur.nl`

Abstract: Data from a drought experiment in *Populus nigra* is used to develop a model-based network for the relationships between gene expression and phenotypic traits. We give an overview about the complex experiment as well as the pre-processing steps. Ultimately we propose two different procedures for reconstruction a gene-phenotype network for time-course data.

Keywords: networks; regression ; graphical lasso.

1 Setting the scene - description of the data

The data that forms the basis for the case study originates from a large experiment in *Populus nigra*: three genotypes of this tree species were planted in the greenhouse. The treatment group of plants was subjected to a drought regime by reducing the soil relative extractable water to 20% over the course of 2 weeks.

In this experiment different types of data were collected. On the one hand RNA was extracted from four replicates per treatment (control and drought). We used the total reads per gene as input for further analysis. Part of the RNA extraction took place over time at five equidistant time points.

On the other hand, phenotypic traits were measured on a subset of the plants (in both conditions). Part of the measurements were repeated at five instances during the experiment. The nature of these measurements and the RNA extraction resulted in partial harvests of the plants. Other phenotypes were only determined once, either at the final harvest or at a different time point. The measurements were also done on different tissues (especially the leaves, roots and part of the stem of the trees).

We will focus on gene expression and phenotype information that was collected over five time points. For this only one genotype was selected. Additionally we will restrict ourselves for this analysis to data that is related to the leaves (such as the leaf area, cell density etc.). The discussion will briefly touch upon other available data from the experiment.

1.1 Remarks on experimental design

The whole experiment consisted of 214 pots that were distributed over two adjacent greenhouses. Plants were moved automatically for watering and weighing to a station inside the greenhouses. While the basic experimental design was a randomized complete block design including three genotypes, two treatments and ten replicates per genotype-treatment combination, the actual set-up changed during the course of the experiment due to the pots moving to the station and the intermediate harvests.

1.2 Remarks on RNAseq analysis

The RNA seq raw reads were processes in a pipeline consisting of procedures to cut the adapter, filter and trim the data (using `cutadapt` and `Prinseq`). The reads were aligned and mapped with `TopHat` and `BowTie` to the *P. trichocarpa* genome before ultimately counting the total reads per gene using `HTSeq`.

2 Model-based networks for time-course data

In this analysis we are focussing on the data that were collected over five time points during the experimental period. Our goal is to reconstruct a network to relate differentially expressed genes and phenotypes.
In order to select these genes and phenotypes we will first fit a linear model and test for significant treatment effects. Subsequently we proceed with the fitted values from this model:

$$y = ti + trt + s(ti) \times trt + \varepsilon$$

with y being the ($\log_1 0$) gene counts respectively the trait (if necessary: transformations thereof), ti is a factor variable for time, trt represents the treatment of control vs. drought as well as an interaction term $s(ti) \times trt$ that includes time a smooth spline. We perform a joint F-test for the treatment terms to select for differentially expressed genes and phenotypes. For the network reconstruction we use the fitted values from this model.

2.1 Two approaches to network reconstruction

After the pre-processing described above we explored two different approaches for the network reconstruction. On the one hand we reconstructed the gene-phenotype network in a hierarchical set-up. This set-up was inspired by Kim et al. (2014). We assume that genes affect traits and connect to other genes, and traits may also connect to other traits. The sequence of steps for this approach is depicted in Figure 1. Our data-set ultimately resulted in a network including nine leaf traits as well as a number of genes. The genes underwent a three-step selection procedure: they were differentially expressed according the model above, a subsequent cluster analysis was performed and ultimately we studied their functional annotation which resulted in 62 genes (from more than 30000 genes in the reference genome). Forward selection in a regression of the individual traits on the genes resulted in seven genes to be included in the final network (see Figure 2).

On the other hand we followed a network estimation procedure that is based on the graphical lasso. It is for large parts in parallel to the analysis described in (Bartzis et al., 2017). This network reconstruction first focusses on estimating the gene network (using the graphical lasso with StARS regularization), find association between genes and traits, select the traits that are related to the genes and reconstruct networks based on this part of the variation. Results of this procedure on the same dataset will be reported elsewhere.

3 Conclusion

This case study of a time-course experiment in *P. nigra* included a number of steps ultimately leading to a gene-phenotype network. We used two different approaches for the network reconstruction. The results look promising both for the regression-based framework in a hierarchical set-up as well for the procedure largely based on the application of the graphical lasso. They form a starting point for further biological interpretation of the underlying mechanisms.

The analysis above focussed on time-course data limited to one type of tissues. We also analyzed other data from the same experiment that included measurements from different genotypes as well as different tissues. The presented framework above can be adapted this situation, however it also includes challenges arising from the inclusion of different tissues as well as an unbalanced measurement scheme.

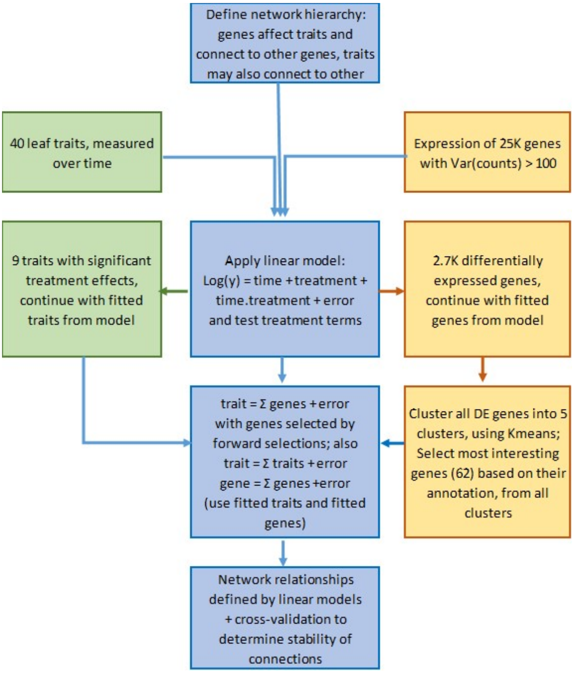

Define network hierarchy: genes affect traits and connect to other genes, traits may also connect to other
40 leaf traits, measured over time
Expression of 25K genes with Var(counts) > 100
9 traits with significant treatment effects, continue with fitted traits from model
Apply linear model: Log(y) = time + treatment + time.treatment + error and test treatment terms
2.7K differentially expressed genes, continue with fitted genes from model
trait = Σ genes + error with genes selected by forward selections; also trait = Σ traits + error gene = Σ genes + error (use fitted traits and fitted genes)
Cluster all DE genes into 5 clusters, using Kmeans; Select most interesting genes (62) based on their annotation, from all clusters
Network relationships defined by linear models + cross-validation to determine stability of connections

FIGURE 1. Flowchart for the analysis of RNAseq data and phenotypic data from a time course experiment in *P. nigra*

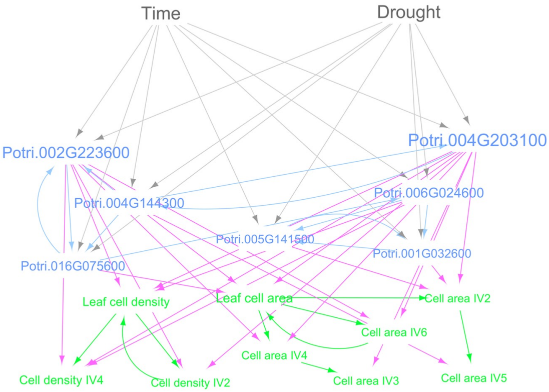


FIGURE 2. Results of a model-based network for the time-course data set

Acknowledgments: This work has been partially funded by the European Union's Seventh Framework Programme under the grant agreement FP7-311929 in the project WATBIO. (Development of improved perennial non-food biomass and bioproduct crops for water stressed environments) from 2011-2017. The data for the case study was collected within WATBIO. We would like to thank the following persons involved in this experiment: Hazel Smith, Henning Wildhagen, Cyril Douthe, David Cohen, Cyril Buré, Oliver Brendel, Didier Le Thiec, Irene Hummel, Vera Vendramin, Simone Scalabrin, Federica Cattonaro, Dennis Janz, Susanna Pollastri, Gianni Della Rocca, Matthew Haworth, Simon Rüger, Mauro Centritto, Francesco Loreto, Michele Morgante, Jaume Flexas, Marie-Beatrice Bogeat-Triboulot, Andrea Polle and Gail Taylor.

References

Bartzis, G. et al. (2017). Estimation of metabolite networks with regard to a specific covariable: applications to plant and human data. *Metabolomics*, **13**, 129 (1 – 17).

Kim, Y. et al. (2014) Mechanisms underlying robustness and tunability in a plant immune signaling network. *Cell Host Microbe*, **15**, 84 – 94.

Using Non-Systematic Data in Flood Frequency Analysis

Thomas Smith[1], Ilaria Prosdocimi[1], Thomas Kjeldsen[1]

[1] University of Bath, United Kingdom

E-mail for correspondence: T.G.Smith@bath.ac.uk

Keywords: Flood Frequency Analysis; Non-Systematic Data; Uncertainty Quantification

1 Background

Flood Frequency Analysis (FFA) aims to quantify the risk posed by floods to specific areas by, for example, estimating the size of flood that an area would expect to see on average once in 200 years. One approach towards FFA involves fitting a statistical model to the series of maximum annual river flows. In fitting such a model, hydrologists typically only use data that has been collected in the modern era through the systematic measurement of river flows. One drawback to this approach is that the duration of these records is short, being only 40-50 years long on average for the UK – far shorter than the time period for which prediction is often required.
Fortunately, there exist sources of non-systematic data which can complement the systematic record. These include flood marks carved into bridges, old photographs, and newspaper reports. Inclusion of this data into the modelling procedure can improve the quality of subsequent inference.

2 Data

This analysis concerns data collected for the river Lune at Caton in Lancashire, UK. The systematic record – available from the National River Flow Archive at http://nrfa.ceh.ac.uk/data/station/peakflow/72004 – spans 1968 to 2015 (48 years). There are an additional eight historical records presented in NERC (1975), with the earliest being in 1892, making the historical record at least 76 years long. Of particular note is the 2015 flood

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

which, at 1742m^3/s, was larger than any other flood event in both the systematic and historical records. This data is presented in Figure 1 along with the perception threshold, which is the flow rate during the historical period for which it is assumed any flow in excess would be recorded. In this case the perception threshold was set to be the flow corresponding to the smallest of the historical records.

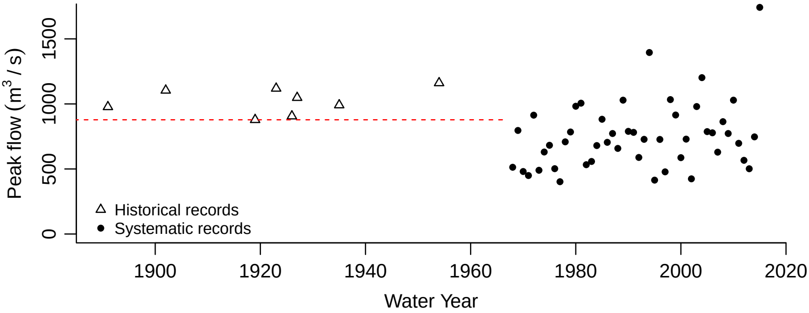


FIGURE 1. Historical and systematic data for the river Lune at Caton. The dashed line is the perception threshold.

3 Methods

For this analysis, the annual maxima were assumed to originate from a Generalised Logistic (GLO) distribution, in line with analysis of an older version of this dataset by Prosdocimi (2018), and the GLO's status as the distribution for UK river flow annual maxima preferred by the Institute of Hydrology (1999). Two separate distributions were fitted – one with, and one without, the historical records. The parameters in each case were fitted using a Maximum-Likelihood method similar to the one presented in Macdonald et al. (2014).

4 Results

The return curves – plots of the flow rate against return period – for both models are presented in Figure 2, along with associated 95% confidence intervals. It can be seen that the return curves have shifted to the right, meaning that inclusion of the historical records into the estimation procedure has resulted in a longer return period being estimated for any given flow rate. This is in line with expectations we might have from looking at

Figure 1, as it shows that there are three floods in the systematic record which are larger than any flood in the historical record.

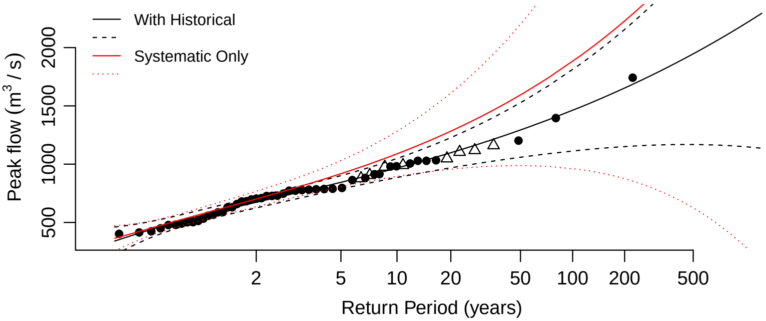


FIGURE 2. Return curves and 95% confidence intervals for the Lune at Caton for the systematic records (red), and systematic+historical records (black).

It can also be seen that the confidence intervals narrow significantly with the inclusion of historical records, demonstrating the utility of such records in practical situations where the uncertainty in a design flood estimate is required in addition to its magnitude.

5 Limitations and Future Work

One assumption in the methodology employed in this analysis is that the underlying flow-generating process is stationary, i.e. that the annual maxima in different years are identically distributed. However, the stationarity of peak river flow series is becoming an increasingly untenable assumption. Figure 1 is suggestive of non-stationarity in the series considered in this work, as the largest three floods in the 124 year record have occurred within the last 22 years. Future work should assess the extent to which non-stationarity diminishes the utility of historical records.

Another assumption made in this analysis is that the flood magnitudes are known exactly. Even for the systematic record there will be uncertainty since it's the depth of the river which is measured and then converted into a flow estimate via a rating curve. The degree of uncertainty is much more serious for historical data, however, with Hosking and Wallis (1986) suggesting errors of ±25% to be "realistic". While the error in any given historical record is difficult to assess, such work would be worthwhile since a more thorough accounting of the sources of uncertainty will be necessary to give estimates of design flood uncertainty which are more realistic than those shown in Figure 2.

References

Hosking, J.R.M. and Wallis, J.R. (1986). The Value of Historical Data in Flood Frequency Analysis. *Water Resources Research*, **22**, 1606 – 1612.

Institute of Hydrology (1999). *Flood Estimation Handbook*. Wallingford, UK: Institute of Hydrology.

Macdonald, N., Kjeldsen, T.R., Prosdocimi, I., and Sangster, H. (2014). Reassessing flood frequency for the Sussex Ouse, Lewes: the inclusion of historical flood information since AD 1650. *Natural Hazards and Earth System Sciences*, **14**, 2817 – 2828.

NERC (Natural Environment Research Council) (1975). *Flood Studies Report: Volume IV Hydrological Data*. London, UK: NERC.

Prosdocimi, I. (2018). German tanks and historical records: the estimation of the time coverage of ungauged extreme events. *Stochastic Environmental Research and Risk Assessment*, **32**, 607 – 622.

The induced smoothing for penalized quantile regression

Gianluca Sottile[1] and Vito M.R. Muggeo[1]

[1] University of Palermo, Italy

E-mail for correspondence: gianluca.sottile@unipa.it

Abstract: We discuss standard errors of parameter estimates in nonparametric quantile regression where the covariate effect is modelled flexibly via B-splines with L_1 penalty on the coefficients. The proposed approach relies on the induced smoothing paradigm aimed at replacing the non-smooth and non-monotone estimating equations with their naturally smooth counterparts. We illustrate the method analyzing a real-data.

Keywords: Quantile regression; P-spline; Induced smoothing

1 Introduction

Nonparametric quantile regression (QR; Koenker (2004) and Koenker 2005() aims to model covariate effects on the response quantiles without imposing any rigid and parametric relationship with covariates. The regression model can be written as

$$Q(\tau) = s_1(x_1) + s_2(x_2) + \ldots + z^{\mathrm{T}}\beta$$

where the smooth functions $s(\cdot)$ are expressed via low rank B-splines, and penalties are set on the corresponding coefficients to avoid overfitting. Without losing in generality we write the penalized objective to be minimized as

$$L(\beta) = \sum_i \rho_\tau(y_i - x_i^{\mathrm{T}}\beta) + \lambda \mid\mid D\beta \mid\mid_1 \tag{1}$$

where D is a penalty matrix, including possibly zero row vectors for unpenalized parameters. The smoothing parameter $\lambda > 0$ is assumed 'known' hereafter. We propose an induced smoothing (IS) approach to compute

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

standard errors for the parameter estimates. Section 2 describes the proposed algorithm, and an analysis of data concerning respiratory disease is included in Section 3.

2 Methods

The self-induced smoothing method was introduced by Brown and Wang (2005) to deal with unsmooth estimating equation. The gradient vector for the objective (1) is

$$U(\beta) = -X^{\mathrm{T}}(\tau - I(y < X\beta)) + \lambda D^{\mathrm{T}}\mathrm{sign}(D\beta). \tag{2}$$

which is clearly unsmooth. The IS aims to replace $U(\beta)$ with its smooth counterparts obtained via expectation over random perturbations weighted by the covariance matrix $\mathrm{var}(\hat{\beta}) = V$, namely $\widetilde{U}(\beta) = E_z[U(\beta + V^{1/2}z)]$ where $z \sim N(0, I_p)$. For the estimating equations (2), application of IS after some algebra leads to

$$\begin{aligned} \widetilde{U}(\beta) &= -X^{\mathrm{T}}\left\{\tau + \Phi\left(\frac{y - X\beta}{\mathrm{diag}(XVX^{\mathrm{T}})^{1/2}}\right) - 1_p\right\} + \\ &\quad +\lambda\left\{2D^T\Phi\left(\frac{D\beta}{\mathrm{diag}(DVD^{\mathrm{T}})^{1/2}}\right) - D^{\mathrm{T}}1_k\right\}, \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of a standard normal. Unlike $U(\cdot)$, $\widetilde{U}(\cdot)$ is smooth thus the derivative $\widetilde{U}'(\beta)$ exists and its computation is relatively straightforward.

Smoothness of estimating equations and relative derivatives allows to apply the usual sandwich forumula to compute the covariance matrix of estimator $\hat{\beta}$,

$$V = \widetilde{U}'^{-1}\mathcal{I}\widetilde{U}'^{-1} \tag{4}$$

where $\mathcal{I} = \tau(1-\tau)X^{\mathrm{T}}X$. Clearly, $\widetilde{U}$ requires V (see (3)) and in turn V needs $\widetilde{U}$ (see (4)), hence an iterative procedure is needed, see Brown and Wang (2005) or Cilluffo et al. (2016).

3 Application

The dataset refers to an epidemiological study carried out in 1988-1991 in the North of Italy, including 1251 males and 1316 females. The study aims to assess determinants of the inspiratory capacity (IC), a measure of lung's function, using the following nine predictors: age, height, body mass index (bmi), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma. The conditional distributions of the response IC given age and height, as suggested by Figure 2, emphasize a non-linear pattern.

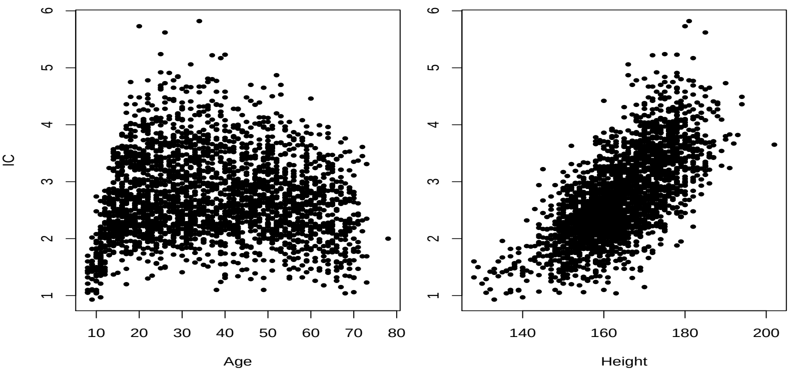


FIGURE 1. Scatterplots of inspiratory capacity versus age and height.

We use a quantile regression models (at probability values $\tau = 0.50, .90$) with smooth covariates age and heights and linear terms for the remaining covariates

$$\begin{aligned} Q(\tau) \quad &= \quad s(x_{\text{age}}) + s(x_{\text{height}}) + z_{\text{bmi}}^{\text{T}}\beta_1 + z_{\text{sex}}^{\text{T}}\beta_2 + z_{\text{smoke}}^{\text{T}}\beta_3 \\ & \quad\quad + z_{\text{occup}}^{\text{T}}\beta_4 + z_{\text{wheeze}}^{\text{T}}\beta_5 + z_{\text{asthma}}^{\text{T}}\beta_6. \end{aligned}$$

The smoothing parameters of the P-splines are selected using 10-folds cross validation. Application of the IS algorithm presented in Section 2 leads to covariance matrix based on sandwich formula (4). Furthermore, extracting the submatrices relevant to the smooth effects allows to compute the pointwise standard errors for the fitted quantile curves. Results are reported in Figure 2.

4 Conclusion

In this paper we have proposed an induced smoothing approach to compute the covariance matrix of the parameter estimates in a penalized quantile regression model. Preliminary results on a real dataset of the inspiratory capacity discussed in Section 3 seem promising. In the next version of the paper we will present further details and results from simulation experiments.

References

Brown, B.M., and Wang, Y.G. (2015). Standard errors and covariance matrices for smoothed rank estimators *Biometrika*, **92**, 149 – 158.

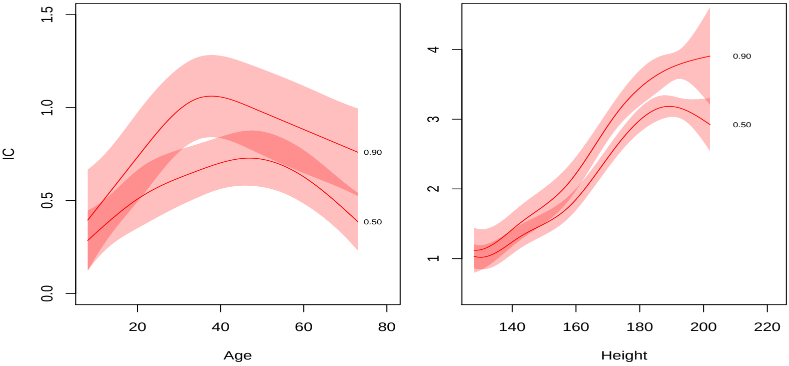


FIGURE 2. Smooth fitted quantile curves (at $\tau = 0.5$ and $\tau = 0.9$) of inspiratory capacity as a function of age and height with pointwise 95% confidence intervals based on standard errors coming from the IS approach.

Cilluffo (2015). The induced smoothed LASSO *Proceedings of the 31st IWSM*, **1**, 71 – 76.

Koenker, R. (2004). Quantile regression for longitudinal data *Journal of Multivariate Analysis*, **91**, 74 – 89.

Koenker, R. (2005). *Quantile regression*, Cambridge University Press.

Muggeo, V.M.R., Sciandra, M., Tomasello, A., Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: a practical framework with application in Ecology. *Environmental and Ecological Statistics*, **20**, 519 – 531.

A new Poisson item count technique with non-compliance

Tang, Man-Lai[1], Wu, Qin[2], Chow, Hoi Sze Daisy[3]

[1] Hang Seng Management College, Hong Kong
[2] South China Normal University, P. R. China
[3] Cheers Psychological Consultancy Services

E-mail for correspondence: mltang@hsmc.edu.hk

Abstract: Item count technique (also known as list experiement) is a popular survey method for eliciting truthful responses to sensitive questions. While item count technique may be less prone to bias than direct questioning, it may create the undesirable ceiling and/or floor effects. Although the Poisson item count technique was developed to solve the ceiling effect by replacing the list of non-sensitive questions by a single non-sensitive question with outcomes following Poisson distribution, the floor effect is not well addressed. It should be noted that all (Poisson) item count techniques rely on an impractical core assumption of no liars (i.e., compliance from the respondents). In this manuscript, we will introduce a new Poisson item count technique to measure the prevalence of non-compliance. The proposed technique allows some of the respondents not to comply with the design, which yields more accurate and reliable parameter estimate. Survey design, parameter estimation, and some results will be presented. Simulation studies are conducted to assess our method.

Keywords: Item count technique; Non-compliance; Sensitive question.

1 Design and model

Assume that the sensitive question of interest is binary (e.g., whether the respondent has ever shoplifted) and we would like to estimate the prevelance of the sensitive characteristic with non-compliance. For this purpose, let n_1 and n_2 respondents be randomly assigned to the control and treatment groups, respectively (with $n = n_1 + n_2$). All the n (i.e., $n_1 + n_2$) respondents are required to read the following non-sensitive question:
(1) How many times did you travel abroad last year?

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Besides, all the n_1 (or n_2) respondents in the control (or treatment) group need to read the following question
(2) If you were born between January and March (or April and December) and you never shoplifted (i.e., without the sensitive characteristic), the answer is 0; otherwise 1.
Finally, all respondents are required to report ONLY the sum of the answers to the two questions. For examples, respondents who were born in January and never shoplifted before should report 4 and 5 respectively if they were assigned to the control and treatment groups and travelled four times abroad last year.
Under our proposed design, the first question (i.e., number of times travelling abroad) is a counting variable with possible answers (denoted by X) being $0, 1, \ldots$ and assumed to follow the Poisson distribution with parameter being λ. In the second question, the non-sensitive question (i.e., period of being born) with binary answers (denoted by W) is assumed to be independent with the sensitive question (i.e., ever shoplifted before) with binary outcomes (denoted by Z). Here, $W = 1$ represents the respondent was born between April and December; $= 0$ otherwise, and $p = \Pr(W = 1)$ is assumed to be known. Let Z be the answer to the sensitive question (e.g., have you ever shoplifted?) with 'yes' and 'no' answers. Also, $Z = 1$ if the respondent possesses the sensitive characteristic; $= 0$ otherwise. It is clear that Z follows the Bernoulli distribution with the unknown parameter $\pi = \Pr(Z = 1)$. It is our aim to estimate π. It is noteworthy that some respondents with the embrassing characteristic might intentionally report the untruthful answer 0, with a probability θ, to demonstrate their positive image due to guilty consciousness. In order to take non-complioance into our consideration, we let U be the non-compliance variable. Obviously, U and Z are not independent. The probability of non-compliance $\theta = \Pr(U = 1)$ is also investigated, which can partly deal with the cheating behavior in the existing Poisson ICT. Furthermore, it is noticed that $\theta = \Pr(U = 1) = \Pr(U = 1|Z = 1)$, and $\Pr(U = 1|Z = 0) = 0$. Let $Q^{(i)}$ be the answer to the second question in the i-th group with $i = 1$ representing the control group; and $= 2$ representing the treatment group. Hence, we have $\Pr(Q^{(1)} = 0) = (1-p)(1-\pi)+\pi\theta$, and $\Pr(Q^{(2)} = 0) = p(1-\pi)+\pi\theta$.

2 Estimation

Suppose the observed data in the control group and treatment group are $y_1^{(1)}, \ldots, y_{n_1}^{(1)}$ and $y_1^{(2)}, \ldots, y_{n_2}^{(2)}$, respectively. Without loss of generality, we assume the first m_0 observations in the control group and the first m_1 observations in the treatment group are 0. We showed that the moment estimator for π is given by $\hat{\pi}_M = \frac{\bar{y}^{(2)}-\bar{y}^{(1)}}{1-2p}$. However, $\hat{\pi_M}$ may not lie in the interval [0, 1]. Alternatively, we propose to obtain the maximum likelihood estimate (MLE) based on the well-known EM algorithm. For this purpose, we define missing data as $Y_{\text{mis}} = \{\{x_j^{(1)}, z_j^{(1)}, u_j^{(1)}\}_{j=1}^{n_1}; \{x_j^{(2)}, z_j^{(2)}, u_j^{(2)}\}_{j=1}^{n_2}\}$. In the missing data, $\{x_j^{(1)}\}$, $\{x_j^{(2)}\}$ are the answers to the first non-sensitive (counting) question, $\{z_j^{(1)}\}$, $\{z_j^{(2)}\}$ are the answers to the sensitive question, and $\{u_j^{(1)}\}$ $\{u_j^{(2)}\}$ are the non-compliance variables in the first and second groups, respectively. After some calcualtions, we can conclude that the M step calculates the MLEs based on the complete likelihood

$$\pi = \frac{\sum_{i=1}^{n_1} z_i^{(1)} + \sum_{i=1}^{n_2} z_i^{(2)}}{n_1 + n_2}, \theta = \frac{\sum_{i=1}^{n_1} u_i^{(1)} + \sum_{i=1}^{n_2} u_i^{(2)}}{\sum_{i=1}^{n_1} z_i^{(1)} + \sum_{i=1}^{n_2} z_i^{(2)}}, \quad \text{and}$$

$$\lambda = \frac{\sum_{i=1}^{n_1} x_i^{(1)} + \sum_{i=1}^{n_2} x_i^{(2)}}{n_1 + n_2}.$$

while the E step is to find the following conditional expectation:

$$E(X_i^{(1)}|y_i^{(1)}) = \frac{y_i^{(1)}\lambda\left[(1-p)(1-\pi)+\pi\theta\right] + y_i^{(1)}(y_i^{(1)}-1)\left[p(1-\pi)+\pi(1-\theta)\right]}{\lambda\left[(1-p)(1-\pi)+\pi\theta\right] + y_i^{(1)}\left[p(1-\pi)+\pi(1-\theta)\right]},$$

$$E(X_i^{(2)}|y_i^{(2)}) = \frac{y_i^{(2)}\lambda\left[p(1-\pi)+\pi\theta\right] + y_i^{(2)}(y_i^{(2)}-1)\left[(1-p)(1-\pi)+\pi(1-\theta)\right]}{\lambda\left[p(1-\pi)+\pi\theta\right] + y_i^{(2)}\left[(1-p)(1-\pi)+\pi(1-\theta)\right]},$$

$$E(Z_i^{(1)}|y_i^{(1)}) = \frac{\pi\left[y_i^{(1)}(1-\theta)+\lambda\theta\right]}{\lambda\left[(1-p)(1-\pi)+\pi\theta\right] + y_i^{(1)}\left[p(1-\pi)+\pi(1-\theta)\right]},$$

$$E(Z_i^{(2)}|y_i^{(2)}) = \frac{\pi\left[y_i^{(2)}(1-\theta)+\lambda\theta\right]}{\lambda\left[p(1-\pi)+\pi\theta\right] + y_i^{(2)}\left[(1-p)(1-\pi)+\pi(1-\theta)\right]},$$

$$E(U_i^{(1)}|y_i^{(1)}) = \frac{\pi\lambda\theta}{\lambda\left[(1-p)(1-\pi)+\pi\theta\right] + y_i^{(1)}\left[p(1-\pi)+\pi(1-\theta)\right]}, \text{ and}$$

$$E(U_i^{(2)}|y_i^{(2)}) = \frac{\pi\lambda\theta}{\lambda\left[p(1-\pi)+\pi\theta\right] + y_i^{(2)}\left[(1-p)(1-\pi)+\pi(1-\theta)\right]}.$$

Here, the moment estimate $\hat{\pi}_M$ as an initial value for the EM algorithm. We repeat the E and M steps until the MLEs are convergent (denoted as $\hat{\pi}_{MLE}$). It is noteworthy that $\hat{\pi}_{MLE}$ is lying in the interval [0, 1].

3 Results and conclusions

To evaluate the performance of our proposed MLEs, we consider two cases for non-compliance: $\theta = 0.3$ and $\theta = 0.4$. In both cases, π are set to be (0.05, 0.1, 0.2, 0.3, 0.4), $p = 0.2$, λ is set to be 2, and $n_1 = n_2 = 1000$ for $\theta = 0.3$ and $n_1 = n_2 = 2000$ for $\theta = 0.4$. The corresponding results based on 1000 repetitions are reported in Tables 1 and 2, respectively. According to Tables 1 and 2, we observe that our MLE for π appears to be a consistent estimate, especially when π is bounded away from 0. The performance of the MLEs improves when sample size increase. In conclusion, our MLEs are reliable estimators. Future work include (i) confidence interval construction for π; (2) hypothesis testing for θ; and (3) reliable estimation methods for small sample designs and/or rare sensitive proportions.

TABLE 1. Mean of estimates based on 1000 repetitions when $\lambda = 0.2$, $\theta = 0.3$ and $n_1 = n_2 = 1000$.

	Mean of estimates		
π	$\hat{\pi}_M$	$\hat{\theta}_M$	$\hat{\lambda}_M$
0.05	0.0879	0.3904	1.9935
0.10	0.1246	0.3604	1.9961
0.20	0.2031	0.3447	2.0026
0.30	0.3018	0.3168	2.0053
0.40	0.4021	0.3097	2.0038

TABLE 2. Mean of estimates based on 1000 repetitions when $\lambda = 0.2$, $\theta = 0.4$ and $n_1 = n_2 = 2000$.

	Mean of estimates		
π	$\hat{\pi}_M$	$\hat{\theta}_M$	$\hat{\lambda}_M$
0.05	0.0705	0.4559	1.9967
0.10	0.1099	0.4381	1.9976
0.20	0.2029	0.4236	2.0002
0.30	0.3011	0.4119	1.9988
0.40	0.4017	0.4071	2.0013

Acknowledgments: The work of the first author was supported by the Research Fund of the Project UGC/FDS14/P06/17.

References

Miller, J.D. (1984). *A New Survey Technique for Studying Deviant Behavior*. Ph.D. thesis, The George Washington University.

Tian, G. L., Tang, M. L., Wu, Q., and Liu, Y. (2014). Poisson and negative binomial item count techniques for surveys with sensitive question. *Statistical Methods in Medical Research*, **26(2)**, 931 – 947.

Modelling of Solar Flare Events Using Extreme Value Theory

Thomai Tsiftsi[1]

[1] Centro de Ciencias Matemáticas, Unidad Michoacán, Universidad Nacional Autónoma de México, México

E-mail for correspondence: ttsiftsi@matmor.unam.mx

Abstract: Solar flares and other space weather events can be harmful for life and infrastructure on earth or in near-earth orbit, therefore modelling such extreme phenomena is vitally important to estimate the frequency and probability of their occurrence. We employ extreme value theory (EVT) to model extreme solar flare events, analysing X-ray flux strengths from NOAA/SWPC [NOAA (2017)]. The return levels for Carrington or Halloween like events are calculated, estimating similar events happening every 110 and 38 years respectively.

Keywords: Solar flares; Extreme value theory; Generalised Pareto distribution.

1 Introduction

Solar flares can be hazardous to human activities and the scientific community is faced with a number of such phenomena of great variability [Riley et al. (2017)]. A probabilistic analysis of these events is a top priority for the space weather agenda as they can adversely affect geopositional systems and telecommunications [Lanzerotti (2007)] and cause severe power cuts. A report provided by the National Research Council [National Research Council (2008)] states that a Carrington-like event [Carrington (1859)] is considered to be a “trillion-dollar” event that could destroy any national power grid and cause disruption in the power supply for more than a year.

To forecast events of interest to human activity we study the tail of the distribution of solar flares which describes extreme events like the Carrington event of 1859 or the “Halloween” storm of 2003. It has previously been assumed that the tail of the distribution of flare strengths x follows a power law [Riley (2012), Lu and Hamilton (1991)] or a lognormal distribution

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

[Riley et al. (2017)]; however it is believed that a power law distribution overpredicts extreme events and their occurrence [Parrott (2015)].

Here we employ Extreme Value Theory (EVT) to estimate the probability of extreme solar flares. Our main result is that a Carrington-like event (45 ± 5 Wm^{-2}) [Cliver and Dietrich (2013)] is expected approximately once every 110 years and a Halloween-like event (X35 ± 5 Wm^{-2}) [Cliver and Dietrich (2013)] approximately once every 38 years. These are in good agreement with the frequencies of occurrence in the next solar cycle as predicted by the National Oceanic and Atmospheric Administration (NOAA).

2 Extreme Value Theory (EVT)

Extreme value theory (EVT) utilises asymptotic analyses for the foundation of models of stochastic processes of unusually large or small intensity events [Coles (2001)]. The possible asymptotic distributions are universal, independent of parent (or true) distributions describing the full process and thus reduce the necessity of making a priori assumptions.

Here we use the threshold excesses approach where *all* observations greater than some high threshold are deemed extreme. In this approach, for X_1, X_2, $...X_n$ a sequence of i.i.d random variables and u a suitably large threshold parameter, the distribution of $Y = X - u$ conditional on $X > u$ that models the tail of the data is a Generalised Pareto Distribution (GPD) [Coles (2001), Leadbetter et al. (1983)]:

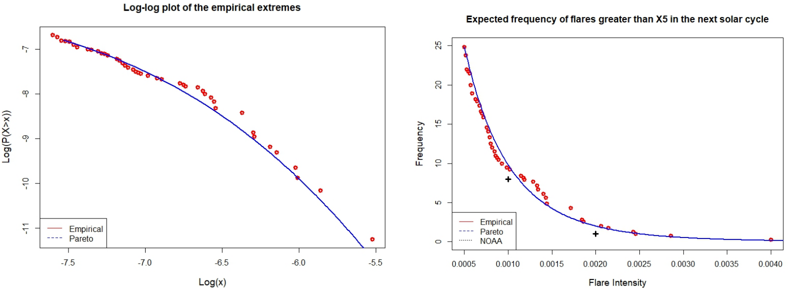
$$H(y) = \mathbb{P}(Y \leqslant y | Y > 0) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{1}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, with ξ, $\tilde{\sigma}$ the shape and scale parameter of the distribution respectively. Both parameters in this work are estimated using Maximum Likelihood Estimation (MLE). The shape parameter determines the tail's qualitative behaviour with $\xi > 0$ yielding the Pareto CDF ("heavy" tail), $\xi < 0$ yielding the Beta CDF (bounded tail) and the limiting case $\xi \to 0$ giving the exponential CDF ("light" tail).

3 Data and results

The data used in this study were X-ray fluxes spanning a period of 43 years from November 1975 to October 2017 extracted from the SWPC/NOAA website and we analysed the peak X-ray flux of each solar flare event as measured by the Geostationary Operational Environmental Satellite (GOES) spacecrafts. To get true X-ray flux measurements the data had to be divided by 0.7 to undo a scaling applied by NOAA for consistency and long term continuity [Machol and Viereck (2016)]. It is important to note here that consecutive solar flare events can be dependent as the same process can generate events of smaller intensity but still of sufficiently high severity to be considered as extreme. Such temporal dependence is often treated by the use of the conventional method of "Peaks-Over-Threshold" (POT); however it has been shown in [Fawcett and Walshaw (2007)] that this technique can lead to systematic and substantial bias in the parameter and return levels estimates. Instead, Fawcett and Walshaw suggest the use of *all* threshold exceedances. To reduce the dependence in our analysis we use only the peak X-ray flux of each event, thus cutting out any background noise or events that have been recorded multiple times and focusing on the extremes that have been proven to be the most destructive for instruments and earth-based activities. To support the claim of independence of these flux peaks we note that the partial autocorrelation function at lag 1 equals 0.049. Secondly, the estimated extremal index of the data which quantifies dependence of recorded events – as estimated by the runs declustering method [Gilleland and Katz (2016)] – was found to be $\theta \approx 1$, which verifies that EVT can safely be applied. Based upon residual life plots and stability of MLE estimates we found that an appropriate threshold ensuring minimal bias was $u = 5 \times 10^{-4}\text{Wm}^{-2}$, an X5 flare, giving a total of 93 exceedances i.e. 93 flares are greater than X5 in our dataset.

Numerical maximisation provides MLE and standard errors for the distribution parameters as: $\hat{\sigma} = 5 \times 10^{-4} \pm 0.2 \times 10^{-7}$ and $\hat{\xi} = 0.12 \pm 0.09$. The estimate of ξ suggests an unbounded distribution ($\xi > 0$) and the evidence is reasonably strong as the 95% profile likelihood confidence interval (C.I.) is almost exclusively in the positive domain with $\hat{\xi} \in (-0.017, 0.3589)$. The validity of the model is shown in Figure 1. The log-log plot indicates that our model predicts extreme events accurately whilst the expected frequencies are seen to match NOAA's estimates fairly well.


Log-log plot of the empirical extremes

Expected frequency of flares greater than X5 in the next solar cycle

FIGURE 1. Empirical and predicted probabilities of threshold exceedances on a log-log scale. The second plot shows the expected frequency of extreme events which closely match NOAA's predictions.

It has previously been suggested that solar seasonality affects the frequency and the extremity of solar flare events. In our work, seasonality in the data was identified via the discrete Fourier transform and was subtracted from the data. The threshold excesses analysis was then repeated; however it was found that the effects of seasonality were negligible with respect to *extreme events*, as the estimates of the distribution parameters were compatible in both cases within confidence intervals.

The model can further be used to estimate the strengths of extreme flares expected in a given period – this is provided by so-called **return levels**. The N-year return level z_N is exceeded by the annual maximum in any particular year with probability $1/N$ and from (1) is easily verified to be

$$z_N = u + \frac{\sigma}{\xi}\left[(N n_y \zeta_u)^{\xi} - 1\right], \tag{2}$$

where n_y is the number of observations per year, $\zeta_u = \mathbb{P}(X > u)$ which is estimated empirically, and ξ is the estimated shape parameter. Using the return levels we can estimate the expected waiting time for Carrington-like events and their 95% confidence intervals. These can be found in Table 1 and are illustrated in Figure 2 (note that caution is required for return levels corresponding to extreme extrapolation). The return levels' estimates in Table 1 are provided with their confidence intervals which were constructed via the method of profile likelihood. Although these are wider than the normal approximation confidence intervals, they are advisable for return levels as they account for the severe asymmetry of the likelihood surface very often observed and thus afford better accuracy.

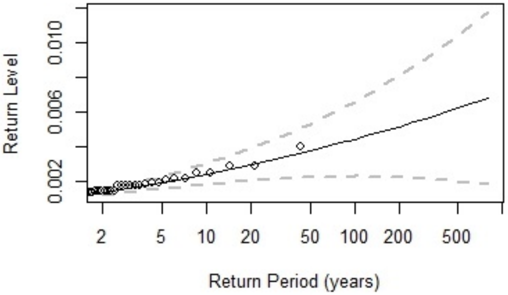


FIGURE 2. Return levels of threshold exceedances. Note an X5 flare, say, means a flux of $5 \times 10^{-4}\mathrm{Wm}^{-2}$. A "Halloween" event (X35) is a one-in-38 years event and a Carrington event (X45) is a one-in-110 years event.

According to the above result the GOES saturation level of X17 (at which the GOES system shuts down and stops recording any events and data as they are destructive for its operational system) is expected to be exceeded on average once in the next 3.5 years. A Carrington-like event (X45) is a

TABLE 1. Estimates and 95% C.I. of several return levels.

Return level	**Estimate**	**C.I.**
3.5-year	**Saturation (X17)**	**(X14.5, X20.5)**
11-year	**X24.5**	**(X20.5, X35.5)**
20-year	**X29.5**	**(X23.5, X47.5)**
38-year	**Halloween (X35)**	**(X26.5, X65)**
50-year	**X37.5**	**(X27.5, X74)**
100-year	**X44**	**(X30.5, X103)**
110-year	**Carrington (X45)**	**(X31, X108)**
150-year	**X50**	**(X36, X125)**

once in 110-year event and a “Halloween-like” event (X35) is a one in 38-years event. The probabilities of these latter events happening in the next decade are 9% and 23.8% respectively. The former is a good improvement when compared to the 12% estimate provided by [Riley (2012)].

4 Conclusions

We argue that EVT is a more rigours framework for analysing extreme events that provides a better basis for extrapolation to levels not yet observed; we have showcased that our predictions are consistent with the ones provided by NOAA. A novelty of our work is the study of solar seasonality – we have found that its effect is negligible with respect to extreme events.

Our model fits the empirical data very well and we provide “worst-case scenario” results (upper bounds of confidence intervals) for some important return levels. It is predicted that the next saturation event of the GOES system will be observed in the next 3.5 years whereas more severe events such as the Carrington and the Halloween are expected to appear once every 110 and 38 years respectively. The probability of these events happening in the next decade is 9% and 23.8% respectively which refines the predictions provided in [Riley (2012)]. The estimates of return levels are provided with more realistic confidence intervals which are evaluated using the appropriate profile likelihood functions which account for the severe asymmetry of the likelihood surface; these are important and their upper bounds should be reported carefully when it comes to protecting life as well as telecommunications and power grids to reduce or avoid the devastating damage that extreme solar flares can cause.

EVT is an excellent tool to describe the tail of the distribution of solar flares and it is expected that the estimates of the GPD will approach the true values as more data are collected. The return level predictions we report in this work can inform the preparation of earth and near-earth based devices to handle worse case scenarios that likely to be encountered.

References

Carrington, R.C. (1859). Description of a singular appearance seen in the sun on September 1, 1859. *Monthly Notices of the Royal Astronomical Society*, **20**, 13 – 15.

Cliver, E.W., and Dietrich W.F. (2013). The 1859 space weather event revisited: limits of extreme activity. *J. Space Weather Space Clim*, **3**, A31.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* London: Springer - Verlag.

Fawcett, L., and Walshaw, D. (2007). Improved estimation for temporally clustered extremes. *Environmetric*, **18(2)**, 173 – 188.

Gilleland, E., and Katz, R. (2016). extremes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, **72(8)**, 1 – 39.

Lanzerotti, L.J. (2007). Space weather effects on communications. *Space Weather – Physics and Effects.* Berlin, Heidelberg: Springer Berlin Heidelberg, 247 – 268.

Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer Series in Statistics, London: Springer - Verlag.

Lu, E.T., and Hamilton, R.J. (1991). Avalanches and the distribution of solar flares. *The astrophysical journal*, **380**, L89 – L92.

Machol, J. and Viereck, R. (2016). GOES X-ray sensor measurements. *www.ngdc.noaa.gov/stp/satellite/goes/doc/GOES-XRS-readme.pdf*

National Research Council (2008). Severe Space Weather Events – Understanding Societal and Economic Impacts: A Workshop Report. In: Washington, DC: The National Academies Press. Available at: *http://books.nap.edu/catalog.php?record id=12507.*

NOAA (2017). NOAA's X-ray fluxes database. Available at: *http://www.swpc.noaa.gov/noaa-scales-explanation.*

Parrott, S. (2015). A second look at on the probability of occurrence of extreme space weather events. Available at: *http://math.umb.edu/ sp/2ndlook.pdf.*

Riley, P. (2012). On the probability of occurrence of extreme space weather events. *Space Weather*, **10(2)**, 1 – 12.

Riley, P., Baker, D., Liu, Y.D., Verronen, P., Singer, H. and Güdel, M. (2017). Extreme space weather events: From cradle to grave. *Space Science Reviews*, **214(1)**.

Distributional regression for demand forecasting in e-grocery — a case study

Matthias Ulrich[1], Hermann Jahnke[1], Roland Langrock[1], Robert Pesch[2], Robin Senge[2]

[1] Department of Business Administration and Economics, Bielefeld University, Germany
[2] inovex GmbH, Karlsruhe, Germany

E-mail for correspondence: `matthias.ulrich@uni-bielefeld.de`

Abstract: In traditional brick-and-mortar retailing, information on customer demand typically results from point-of-sale data. These data are censored, and hence biased, due to stock-outs affecting the individual purchase. In contrast, e-retailing allows for the observation of customer preferences before stock-out information becomes known to the buyer and, therefore, yields uncensored demand data. Moreover, in e-grocery the customer selects a future delivery time slot so that future demand is partly known to the retailer at the replenishment decision time.

Considering data from a German e-grocery retailer, in this case study we discuss demand forecasting in e-grocery, making use of the corresponding new types of data that are not available in traditional retailing. Since underage and overage costs are usually asymmetric, we seek a suitable model for the entire demand distribution, rather than point forecasts only, to minimize the costs. Thus, we propose the application of Generalized Additive Models for Location, Scale and Shape (GAMLSS), which allow a flexible selection of distributions for the demand, and also a flexible modeling of covariate effects on any of the distributional parameters. As benchmark models we consider linear regression, random forests, quantile regression and quantile regression forests. The models are evaluated by comparing their out-of-sample forecasting error for varying levels of asymmetry in the costs. For each stock keeping unit (SKU) that we consider, we find that models from the GAMLSS class outperform the benchmark models.

Keywords: Distributional regression; demand forecasting; GAMLSS, e-grocery.

1 Exploratory analysis of the e-grocery data

In the following, we explore the e-grocery data available in order to motivate the statistical models considered below. To illustrate some of the key patterns, Figure 1 displays the relationship between selected explanatory variables (features) and the response variable, realized demand, for the stock keeping unit (SKU) grapes within the demand period September 2015 to August 2017.

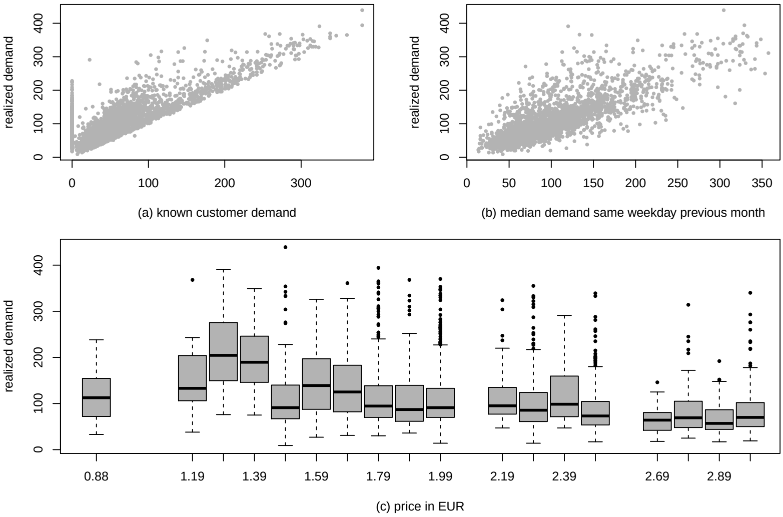


FIGURE 1. Exploratory data analysis for the e-grocery case study data.

We find, *inter alia*, the following patterns:

1. *Nonlinearity*

   Figure 1(a) shows the relationship between demand known at the time of the replenishment decision and realized demand. Realized demand equals or exceeds known demand for each observation. For relatively low known demand, i.e. below 150 units, the size of additional demand occurring during the replenishment period is relatively high compared to situations where known demand is already high, i.e. above 150 units. This indicates that the functional relationship between realized demand and known demand is nonlinear.

2. *Heteroscedasticity*

   Figure 1(b) relates the realized demand to the median demand of the same weekday in the previous month, and shows that the variance in demand increases with increasing values of this feature.

3. *Skewness*

   Figure 1(c) shows positive skewness in the distribution of the realized demand, with the degree of skewness varying across different prices. The upper whisker and the 0.75 quantile are farther from the median than the 0.25 quantile and the lower whisker. This asymmetry in the distributional shape increases with increasing price.

2 Distributional regression

2.1 Business problem

Inventory exceeding or falling short of customer demand generally causes asymmetric monetary consequences. The value of these consequences depends on SKU-specific criteria such as price, margin, customer expectation, storage capacities, and depreciation. In retailing, underage cost are often assumed to be higher than overage cost.

Our e-grocery retailer offers a significant number of perishable SKUs in the product categories fruits and meat. The associated overage costs of such SKUs include additional spoilage cost. Shelf life of these SKUs is restricted by best-before dates. We hence assume that the customer demand and the sales period are identical for the SKUs analyzed in the case study. In other words, excess inventory cannot be sold in the following demand period and thus generates spoilage.

Classical measures of forecasting accuracy are the Mean Average Error (MAE) and the Mean Average Percentage Error (MAPE). With respect to the underlying business problem of our e-grocery retailer, both metrics are inadequate because they do not consider economic consequences in their evaluation of the estimation quality. The MAE measure does not account for the asymmetry of overage and underage cost. The MAPE punishes relative deviations of slow moving and fast moving products equally.

To capture the asymmetric economic impact of absolute forecasting errors for each demand period t, we introduce the total cost C_t resulting from any potential mismatch between inventory level and realized demand. In demand period t, each excess unit of inventory generates a cost of h, while each unit that we fall short of customer demand generates a cost of b. Furthermore, we use D_t to denote the stochastic customer demand. We then aim at minimizing the expected total cost,

$$E[C_t(y_t)] = hE(y_t - D_t)^+ + bE(D_t - y_t)^+,$$

with respect to the inventory level at the beginning of the demand period, y_t. The optimal y_t defines the corresponding replenishment order quantity of the retailer for period t. For single and independent demand periods, the newsvendor problem provides the solution to the optimization problem above (Zipkin, 2000). Specifically, we suppose that the values for b and

h are defined via assessment of the retailer. The ratio $b/(b+h)$ equals the optimal demand quantile given b and h. It can be interpreted as the inventory service level selected by the retailer. The optimal inventory level is then obtained as

$$y_t^* = \underset{y_t}{\operatorname{argmin}}\, E[C_t(y_t)] = F_t^{-1}\big(b/(b+h)\big), \tag{1}$$

where F_t is the (true) cumulative distribution function of the demand distribution in period t. In practice, the optimal solution to the newsvendor problem given in (1) is not available since the c.d.f. F_t describing the stochastic demand is unknown. However, we can use data collected before time t to statistically model realized demand as a function of features (e.g. known demand at the time of the replenishment order), and subsequently predict demand at time t using $\hat{F}_t$ as obtained under the model.

2.2 Feature engineering

For all models, the demand distribution F_t is estimated using features. Feature engineering describes the process of generating suitable features from data. Both the general pattern of the demand distribution as well as any time series effects are taken into account by considering historic demand quantiles (5%, 50% and 95%), then building corresponding features using data from a) the previous quarter, b) the previous month, and c) the previous two weeks. After feature engineering, the data set contains 12 features, including also price and known demand as extracted directly from the raw data.

2.3 GAMLSS

Given the complex patterns found in the data, we propose to use Generalized Additive Models for Location, Scale and Shape (GAMLSS), as they allow a flexible selection of distributions for the demand, and also a flexible modeling of covariate effects on any of the distributional parameters (Rigby and Stasinopoulos, 2005). For our case study, we implemented the normal (NO), gamma (GA), Poisson (PO) and negative binomial (NBI) distributions as these are established in inventory management (see for example Silver and Peterson, 1985, and Ramaekers and Janssens, 2008). We implement a P-spline smoother to account for potential nonlinear relationships between features and realized demand.

Feature selection using component-wise gradient boosting, as described in Hofner et al. (2016), in our case study did not improve the out-of-sample forecast accuracy, such that we eventually trained all models using the complete feature set.

2.4 Benchmark models

Based on the existing literature on distributional regression (e.g. Koenker and Hallock, 2001, Meinshausen, 2006), the benchmark models we consider are linear regression (LM), random forests (RF), quantile regression (QR), and quantile regression forests (QRF). We select the same distributions for random forests that we applied also for GAMLSS (i.e. normal, gamma, Poisson and negative binomial).

2.5 Model training and forecasting

Our data set contains data for the period September 2015 to August 2017 from six different e-grocery fulfillment center. We split the data into a training data set (September 2015 to August 2016) and a validation data set (September 2016 to August 2017). In the training process, we move forward in time for model training and forecasting. For example, we train the year August 2016 to July 2017 to forecast August 2017.

3 Results

For each demand period t and each of the models considered, we obtain an estimate $\hat{F}_t$ for the demand distribution F_t, which we apply to derive y_t for any given demand quantile. In the validation period September 2016 to August 2017, we then calculate the total costs that occur under the y_t obtained:

$$C_t(y_t) = h(y_t - d_t)^+ + b(d_t - y_t)^+.$$

We specify the costs b and h such that the ratio $b/(b+h)$ equals the demand quantile at which we want to compare the performance of the various models considered. As an example, we specify $b = 4$ and $h = 1$ when evaluating the models for the target quantile 0.8. For each demand quantile selected, we estimate the absolute total costs by the average costs of all demand periods for the six e-grocery fulfillment center. As the absolute total costs of SKUs depend on overall demand, we standardize the costs by calculating the ratio to realized demand.

Figure 2 showcases example results for the SKU grapes, displaying relative total costs as a function of the demand quantiles considered. Note that the demand quantiles 0.8 to 0.99 are the most relevant quantiles in retail practice. For clarity of presentation, the plot shows only the best-performing models.

For each SKU that we consider, we find that models from the GAMLSS class outperform the benchmark models, with the Poisson distribution yielding the overall lowest out-of-sample costs. The superiority of the Poisson distribution increases with an increase in the selected demand quantile of the demand distribution. In addition, we find that for all models the total cost is a convex function of the demand distribution's quantile.

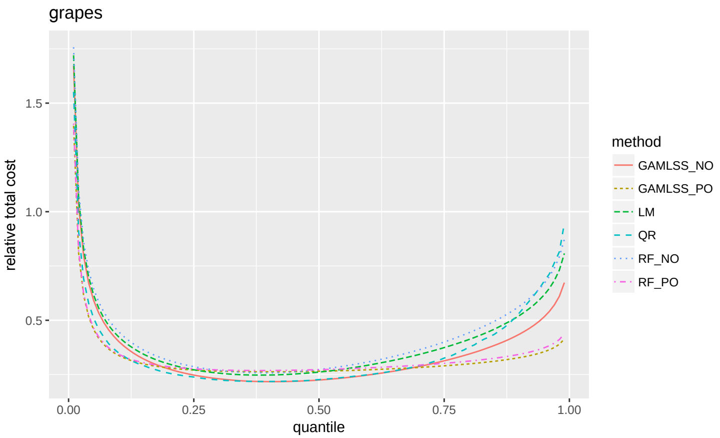


FIGURE 2. Relative total costs for demand quantiles 0.01 to 0.99.

References

Hofner, B., Mayr, A., and Schmid, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, **74**.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, **15**, 143 – 156.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, **7**, 983 – 999.

Rigby, B. and Stasinopoulos, M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507 – 554.

Ramaekers, K. and Janssens, G.K. (2008). On the choice of a demand distribution for inventory management models. *European Journal of Industrial Engineering*, **2**, 479 – 491.

Silver, A. and Peterson, R. (1985). *Decision Systems for Inventory Management*, New York: Wiley.

Zipkin, P. H. (2000). *Foundations of Inventory Management.* Boston: McGraw-Hill.

Testing for Statistical Interaction in Cox Regression Model in the Case of Non-Proportional Hazards in One of the Covariates Involved in the Interaction Effect

Kristina P. Vatcheva[1]

[1] School of Mathematical and Statistical Sciences, The University of Texas Rio Grande Valley, Brownsville, TX, USA

E-mail for correspondence: kristina.vatcheva@utrgv.edu

Abstract: We investigated the diagnostic of statistical interaction in the case of non-proportionality in hazards in one of the covariates involved in the interaction effect during the Cox regression model development process. We generated right-censored survival data using simulations with different scenarios that involved different values for the coefficient of the interaction term and the time-dependent term in Cox regression models. We evaluated and compared the empirical power of the local chi-square test for regression coefficient of statistical interaction in the different simulation scenarios for the models with and without proportionality in hazards assumption satisfied. The results of the analysis of the simulated data suggested that in the incorrectly specified Cox regression model due to a non-proportionality in hazards in a covariate, the identification of a statistical interaction with the same covariate, in some cases, required more statistical power. We recommend that the evaluation of the interaction effect in Cox regression model to be performed before and after testing and if necessary correcting for proportionality in hazards.

Keywords: Cox regression model; Proportionality in hazards; Simulation; Statistical interaction; Effect modification.

1 Introduction

Regression analysis is a widely used powerful tool in epidemiological and medical research to investigate associations between a specific exposure and an outcome. Correctly specified regression models can provide reliable parameter estimates for the regression coefficient of each of the variables in the model. This might have a direct impact as to how a researcher interprets the data, answers the study questions, and in some instances, makes important public health decisions (Greenland et al., 1994). Misspecified non-additive models with no interaction term result in biased regression coefficient estimates and ultimately erroneous interpretations (Vatcheva et al., 2016). Common methods for assessing statistical interactions are testing the product term of two or more variables included in the regression model or conducting the log-likelihood ratio test for the nested models, with and without interaction term. The Cox proportional hazards regression is a commonly used semi-parametric statistical method in epidemiological and medical research when analyzing time-to-event data for investigating the association between the survival time of patients and one or more predictor variables (Cox, 1972). Performing a proportional hazards regression analysis of survival data in applied settings requires a number of critical decisions and steps. Some of the major steps are: selection of potential predictors and confounders and fitting the main effect model, examination of the scale of the continuous variables; testing for statistical interaction and fitting the preliminary model; and selection of the final model after performing checking for adherence to the key model assumptions, diagnostics for influential observations, and testing for overall goodness-of-fit (Harrel et al., 1996; Hosmer et al., 2008). A key assumption of the Cox proportional hazards regression model is that the hazard ratio is constant over time for each of the covariates included in the model. The aim of this study was to investigate the practice of evaluation of interaction effect in the case of non-proportionality in hazards in one of the covariates involved in the interaction effect during the Cox regression model development process using simulated epidemiological data.

2 Statistical Methods

2.1 Data Generation

We generated right-censored survival data from a fully specified Cox regression model $h\,(t|x) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \ \beta_4 x_2 log(t))$, where $h(t|x)$ is the hazard rate at time t for an individual with risk vector x; $h_0(t)$ is the baseline hazard; x_1 and x_2 are two predictor variables that had an interactive effect and a violation in the proportional hazard assumptions in variable x_2; and β_1, β_2, β_3, and β_4 are the regression parameters. We considered the following simulation scenarios where we varied the magnitude of interaction effect and the magnitude of deviation from proportionality in hazards:

$$
\begin{aligned}
h(t|x) &= h_0(t)exp(2x_1 + x_2 + 0.5x_1x_2 + x_2log(t)), && (1)\\
h(t|x) &= h_0(t)exp(2x_1 + x_2 + 1x_1x_2 + x_2log(t)), && (2)\\
h(t|x) &= h_0(t)exp(2x_1 + x_2 + 1.5x_1x_2 + x_2log(t)), && (3)\\
h(t|x) &= h_0(t)exp(2x_1 + x_2 + 1.5x_1x_2 + 1.5x_2log(t)). && (4)
\end{aligned}
$$

Based on previous studies (Vatheva et al., 2015; Vatcheva et al., 2016), the pre-specified continuous variable x_1 was generated from normal distribution with mean 6.4 and variance 2.25 and the pre-specified binary variable x_2 was generated from Bernoulli distribution with probability of success $p = 0.5$. Baseline hazard rate was chosen so that more than 30% of the subjects experience the event (Vatcheva et al., 2016). For each of the scenarios 1000 datasets with sample sizes of 600 were generated. All simulations were conducted with Stata 12 using survsim module (Crowther et al., 2011).

2.2 Data Analysis

By using the simulated data under each of the simulation scenarios, Cox regression models with an interaction (correct model) and without an interaction term (incorrect model) were fitted. In each of the cases we fitted two models: with non-proportionality in hazards in variable x_2 and with stratification by variable x_2 to correct for non-proportionality in hazards in variable x_2. First, we obtained the vector of the p-values corresponding to the local chi-square test statistic of the coefficient estimates of the product term x_1x_2 in the interactive models across the 1000 repetitions. We calculated the percentage of the p-values that were less than the priory defined probability of Type I error $\alpha = 0.05$. This percentage was our empirical power, which is the empirical probability to reject an incorrect null hypothesis that the coefficient estimates of the product term x_1x_2 is zero. We evaluated and compared the power of the test for statistical interaction x_1x_2 in the different simulation scenarios between the models with and without proportionality in hazards in variable x_2. In addition, we performed and evaluated the power of Therneu-Grambsuch non-proportionality test (Therneau et al., 2000) for both correct and misspecified models due to exclusion of the interaction term and non-proportionality in hazards using Stata 15.1 phtest command. All other statistical analyses were performed using SAS 9.4. All statistical testing were two-sided and performed at significance level $\alpha = 0.05$.

3 Results

The results from the analysis of the simulated data for the empirical power to detect a significant interaction effect at a significance level $\alpha = 0.05$

based on various simulation scenarios are shown in Figure 1. The empirical power was compared to the recommended 80% power (Cohen, 1988), which is commonly used in study design for detecting an effect when there is an effect to be detected. Recall that we had generated our data using fully specified Cox regression model in a way that there is non-zero interaction effect between variables x_1 and x_2. In the case of non-proportionality in hazards in variable x_2 and simulation scenario (3) with a regression coefficient of the interaction term greater than the magnitude of the regression coefficient of the time-dependent term, the empirical power of the local chi-square test for the coefficient estimates of the product term x_1x_2 was 21.3% (Figure 1, scenario(3)). To further investigate this effect we simulated additional data with Model 3 by increasing the sample size and the event rate. When the sample size was increased from 600 to 2000, the empirical power slightly increased to 45% (Figure 1, scenario (3a)). It is well known that the power of Cox regression model is driven by the number of the events, rather than the number of subject. When the event rate was increased to 77% the empirical power increased to 68.2% (Figure 1, scenario (3b)). In simulation scenario (4) where the magnitude of the generated coefficient of time-dependent term was increased to 1.5, the empirical power of the local chi-square test for the coefficient estimates of the product term x_1x_2 was 90.2%.

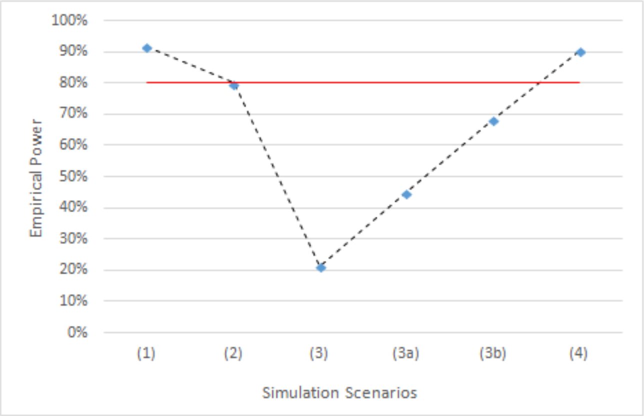


FIGURE 1. Empirical power of detecting non-zero interaction effect in various simulation scenarios.

Table 1 presents the results from the power analysis of Therneu-Grambsuch non-proportionality test. In all of the simulation scenarios the test performed better before the inclusion of the interaction term in the models. Despite that the Therneu-Grambsuch non-proportionality test may yield false positive when the model is misspecified (Therneau et al., 2000; Keele L., 2010), the test had more than 80% power to detect the non-proportionality in hazards in variable x_2 in the misspecified models fitted in simulation scenarios (1), (2) and (4)(Table 1). The inclusion of the interaction term in the models reduced the power of the test for non-proportionality in hazards in variable x_2 to as high as 50% (Table 1). After conducting Cox proportional hazard regression with stratification by variable x_2 to correct for non-proportionality in hazards in variable x_2, the power of detecting non-zero interaction effect in all simulation scenarios was 100%. Therneu-Grambsuch test accurately detected that the non-proportionality in hazards in variable x_2 was corrected.

TABLE 1. Empirical power (%) of Therneu-Grambsuch non-proportionality test in various simulation scenarios.

	Model 1		Model 2	
Variable	Misspecified	Correct	Misspecified	Correct
x_1	4	13.5	3	13.7
x_2	81.9	28.9	97.6	37.2
x_1x_2		21.1		26.5
Global Test	92.7	60.7	95.3	54.2

	Model 3		Model 4	
Variable	Misspecified	Correct	Misspecified	Correct
x_1	7.6	10.6	3.5	24.5
x_2	46.4	35.9	99.4	48.8
x_1x_2		28.9		36.1
Global Test	65.2	51.1	99.2	76.6

4 Conclusions

The results of the analysis of the simulated data suggested that in the incorrectly specified Cox regression model due to non-proportionality in hazards in a covariate, the identification of a statistical interaction with the same covariate, in some cases, required more statistical power. This findings were in the case when the magnitude of the interaction term coefficient was greater than the magnitude of coefficient of the term creating non-proportionality. Previous research proposing Therneu-Grambsuch non-proportionality test as a diagnostic strategy for Cox models (Keele, L., 2010) reported that correcting the model due to interactions and other model misspecifications must be done prior testing for non-proportional hazards. We recommend that during the Cox regression model development process, the statistical testing for interaction effect need to be performed before and after testing and correcting for the proportional hazards assumption.

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Routledge Academic.

Cox, D.R. (1972). Regression models and life-tables with discussion. *J Royal Stat Society Ser B*, **34**, 187 – 220.

Crowther, M.J., Lambert, P.C (2012). Simulating complex survival data. *The Stata Journal*, **12**, 674 – 687.

Greenland, S., Maldonado, G. (1994). The interpretation of multiplicative-model parameters as standardized parameters. *StatMed*, **13**, 989 – 999.

Harrell, F. E., Lee, K.L., Mark, D.B (1996). Tutorial in biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361 – 387.

Hosmer, D.W., Lameshow, S., May, S. (2008). *Applied survival analysis : regression modeling of time-to-event data.* New York, NY: Wiley-Interscience.

Keele, L. (2010). Proportionally difficult: Testing for nonproportional hazards in Cox models *Political Analysis*, **18**, 189 – 205.

Therneau, T.M., Grambsch, P.M. (2000). *Modeling survival data: Extending the Cox model.* New York: Springer-Verlag.

Vatcheva, K.P., Fisher-Hoch, S.P., Rahbar, M., Lee, M., Olvera, R. (2015). Association of total and differential white blood cell counts to development of type 2 diabetes in Mexican Americans in Cameron county Hispanic cohort. *Diabetes Res Open J*, **1(4)**, 103.

Vatcheva, K.P., Lee, M., McCormick, J.B., Rahbar, M.H. (2016). The effect of ignoring statistical interactions in regression analyses conducted in epidemiologic studies: an example with survival analysis using Cox proportional hazards regression model. *Epidemiology (Sunnyvale)*, **6**, 216.

Nonparametric spatial clustering using spatio-temporal data

Ashwini Venkatasubramaniam[123], Ludger Evers[2], Konstantinos Ampountolas[13]

[1] Urban Big Data Centre, University of Glasgow, United Kingdom,
[2] School of Mathematics and Statistics, University of Glasgow, United Kingdom,
[3] School of Engineering, University of Glasgow, United Kingdom

E-mail for correspondence: a.venkatasubramaniam.1@research.gla.ac.uk

Abstract: This paper proposes a nonparametric Bayesian clustering model that seeks to identify spatially contiguous clusters using data recorded over space and time. This approach utilises a modified non-sequential distance dependent Chinese restaurant process (ddCRP) to model dependencies arising from both space and network connectivity in an undirected graph and we also define a spatio-temporal precision matrix to fully account for spatial and temporal constraints within individual clusters. The method employs a Metropolis within Gibbs sampler to fully explore all possible partition structures and the developed algorithm is illustrated by an application to house prices recorded for non-overlapping areal units in England from 1995 to 2016.

Keywords: Bayesian; Clustering; Network; Nonparametric

1 Introduction

Spatial clustering methods applied to spatio-temporal data are employed to identify spatially contiguous homogeneous regions and serve as an important exploratory tool towards understanding location based differences over time. We propose a novel nonparametric Bayesian clustering algorithm that is capable of determining spatially connected clusters using non-exchangeable data. This holistic approach determines the number of clusters in a data-driven manner from observed data and incorporates spatial and temporal dependencies within individual identified clusters. This clustering approach is primarily motivated by temporal data for different locations in space such that there is a unique observation for every space and time combination.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Method

Let a group of objects be arranged as an undirected graph such that neighbouring objects are connected and a time series of observations are available for each object. Examples include sensors spread over space (such as occupancy observations recorded over six hours by sensors in a road network) or areal unit data (such as average house prices recorded over twenty years for each local authority in England). A cluster is composed of a set of connected objects that have no links to objects in the rest of the undirected graph. In a flexible Bayesian clustering approach, a prior enforces a distinct partitioning of the graph and we use a modified non-sequential distance dependent Chinese restaurant process (ddCRP) to account for non-exchangeable data. The classical *Chinese restaurant process* (CRP) suggests a generative process where a new customer i that enters the restaurant is allocated to an existing table depending on the number of customers already present and is allocated to a new table with probability α. The CRP is described using 'culinary' metaphors and customers seated at a table in a restaurant correspond to objects allocated to a cluster in a graph. An alternative generative model for the classical CRP is based on customers who choose to sit with another customer rather than at a table. The classical CRP is obtained if a customer chooses an already seated customer as their 'friend' with probability proportional to 1 and chooses themselves as their own friend with probability proportional to α. This is a special case known as the *sequential* CRP since customers cannot choose future customers as their friends. In a sequential CRP, a new cluster is formed when a customer chooses no friend and so the parameter α allows for effective control of the number of clusters. We will see that this does not hold for a *non-sequential* CRP in which the parameter α only poses limited control over the number of clusters. In a *distance dependent Chinese restaurant process* (ddCRP) [Blei and Frazier, 2011], the probability of a customer choosing another customer as their friend depends on the distance between them. If the objects are arranged on a neighbourhood graph, a suitable model is to restrict customers to choose only one of their neighbours as a friend, effectively making the distance binary. Figure 1 shows two toy examples of such CRPs. Figure 1(a) shows a sequential process where customers can only choose neighbours with a lower index (such as $\mathit{4} \rightsquigarrow \mathit{3}$) or themselves as their friend. Figure 1(b) shows a non-sequential process, in which customers can also choose neighbours with a higher index. As one can see from the figure, new clusters in a non-sequential ddCRP are formed by either self links or redundant links which create cycles (such as a cycle formed by $\mathit{3} \rightsquigarrow \mathit{4}$ and $\mathit{4} \rightsquigarrow \mathit{3}$). Unless the parameter α also controls the creation of cycles, it will not be effective at controlling the number of clusters. This is especially problematic in the neighbourhood-based model, where customers have a very limited choice of potential friends so cycles are very likely. We thus propose a modification of the ddCRP that assigns a probability of α not only to a self link but also to redundant links that lead to a cycle.

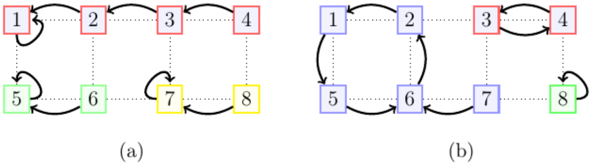


FIGURE 1. Group of objects arranged as a network

The distribution of the ith customer assignment c_i being equivalent to customer j is defined as:

$$p(c_i = j) \propto \begin{cases} \alpha \text{ if } i = j \text{ or } j \rightsquigarrow i \text{ or a cycle is formed} \\ h(d_{ij}) = 1 \text{ if } i \sim j \\ h(d_{ij}) = 0 \text{ if } i \nsim j \end{cases} \tag{1}$$

The probability of the partition structure (composed of three clusters) being formed within the network in Figure 1(b) is $\alpha^{n_L} = \alpha^3$, where n_L is the number of clusters formed by cycles and self links.

The likelihood function defines a product over the probabilities of observations at identified clusters and we assume that the observations recorded over time for each object follows a Gaussian distribution. In order to be able to fully account for the temporal and spatial dependencies within individual identified clusters, we define a spatio-temporal precision matrix using a conditional auto-regressive (CAR) model over space and a first order auto-regressive (AR-1) model over time. The presence of a unique observation for every space-time combination enables the use of Kronecker product tricks to improve computational efficiency. A cluster structure of the network based on observed data is found by posterior inference and a Metropolis within Gibbs sampler enables the space of all possible partitions to be explored by joining and breaking up clusters in the network. A general Gibbs sampler for the sequential ddCRP is described by Blei and Frazier (2011) and the general case where cycles are possible is introduced by Socher (2011). We propose a sampler for the non-sequential ddCRP to accommodate spatial and network dependencies imposed by the structure of the graph and where new clusters are formed by both cycles and self links. Unlike samplers for a traditional CRP, the sampler for a non-sequential ddCRP can efficiently move multiple objects in or out of clusters in a single step because all linked objects in the network need to be taken into account. Assuming that the cluster specific parameters can be integrated out, the sampler for cluster allocation is defined such that

$$P(c_i = j \mid c_{-i}) \propto \begin{cases} p(c_i = j)\frac{p(\mathbf{y}_A,\mathbf{y}_B)}{p(\mathbf{y}_A)p(\mathbf{y}_B)} \text{ if link } i \rightsquigarrow j \text{ joins clusters A and B} \\ p(c_i = j) \text{ otherwise} \end{cases}$$

3 Results

The developed clustering method is applied to average house price data for small areas in England recorded by the Office for National Statistics (ONS) from 1995 to 2016 for middle layer super output areas (MSOAs). The algorithm is implemented to cluster house prices aggregated at different unit levels including MSOA, the local authority, counties and regions and Figure 2 represents a cluster structure across local authority units (n = 326) in England. The clustering method identifies spatially contiguous clusters that represent distinct temporal patterns of prices at different local authority units across the network.

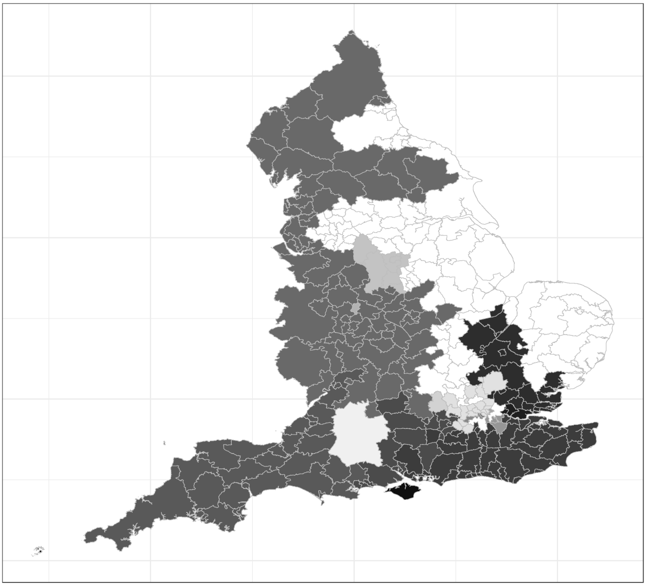


FIGURE 2. Cluster structure to describe differences in average house prices.

In future work we seek to extend the algorithm to generate dynamic clusters that change in shape over time and to incorporate data from multiple sources in order to identify meaningful relationships.

References

Blei, D.M. and Frazier, P.I. (2011) *Distance dependent Chinese restaurant processes*. Journal of Machine Learning Research, 12, 2461–2488.

Socher, R., Maas, A., and Manning, C. (2011). *Spectral Chinese restaurant processes: Nonparametric clustering based on similarities*, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 1476–1484.

Investigating patterns in macroinvertebrate communities using mixed models

Massimo Ventrucci[1], Gemma Burgazzi[2], Daniela Cocchi[1], Alex Laini[2]

[1] University of Bologna, Italy
[2] University of Parma, Italy

E-mail for correspondence: massimo.ventrucci@unibo.it

Abstract: We focus on an ecological study where the aim is to investigate small scale spatial processes within macroinvertebrate communities. We use mixed models in a Bayesian framework and discuss the use of PC priors in this setting. Preliminary results show presence of correlation within survey campaigns, suggesting possible small scale interactions between members of the communities.

Keywords: mixed models; unobserved heterogeneity; PC priors, INLA.

1 Introduction

The distribution of organisms in natural communities is often spatially structured and shows high degree of variability (Laini et al., 2014). Investigating factors affecting these communities is an important topic in community ecology and proper statistical tools are needed to disentangle the effects of abiotic factors from those of biotic interactions. The motivating example for this work regards the analysis of small scale spatial processes within macroinvertebrate communities. Data were collected in six sampling campaigns carried out in three different streams, tributaries of the Po River (Northern Italy): Nure Stream, Parma Stream and Enza Stream. For each river a sampling area was selected and sampled twice, once in summer and once in winter. The spatial design for each station included fifty random points aligned along several transects, see an example in Figure 1. At each point, abundance of macroinvertebrates (response) and abiotic factors such as flow velocity (V), water depth (P) and benthic organic matter (BOM) were recorded.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

In recent years, the linear mixed model (LMM) framework has grown a lot of attention for the analysis of ecological survey data (Zuur et al., 2009). The reason lies in interpretability of the model components. The fixed part includes the effects of observed abiotic factors (environmental covariates), while the random effects account for sources of heterogeneity driven by unobserved factors. Unobserved heterogeneity is usually interpreted as either missing covariates or evidence for biotic processes taking place among members of a given community.

In a Bayesian hierarchical framework for mixed models (Fong et al., 2010), a major issue regards the choice of the prior for the precision (i.e. inverse variance) of the random effects. Prior information on the scale of a precision parameter is typically not available. To address this issue, Simpson et al. (2017) proposed penalized complexity (PC) priors, which are defined on the scale of the distance from a base model and then transferred to the scale of the original parameter. For instance, for a Gaussian random effect with zero mean and precision τ a natural base model is obtained as $\tau \to \infty$, which corresponds to no random effect. Following Simpson et al. (2017), the PC prior for τ is a Gumbel type 2 distribution.

In this work, a LMM is used to understand the nature of residual structure remaining after accounting for environmental factors. To this aim, we will apply PC priors for correlation matrices discussed in section 6 of Simpson et al. (2017).

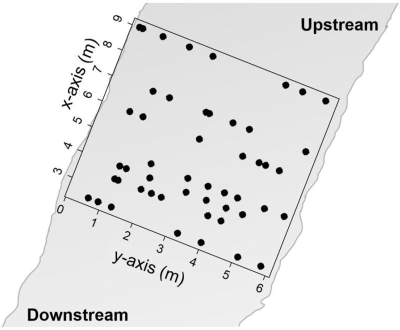


FIGURE 1. Scheme of the sampling design: black dots represent the 50 points inside the grid, with positions varying depending on sampling campaign.

2 Modelling unobserved heterogeneity with mixed models and PC priors

Let $y_{i,j}$ be the log transformed abundance observed at replicate $i = 1 : 50$ from campaign $j = 1 : 6$, we assume the model

$$y_{i,j} = \boldsymbol{x}_{i,j}^{\mathsf{T}} \boldsymbol{\beta} + \epsilon_{i,j} \quad ; \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau_\epsilon^{-1} \boldsymbol{R}^{-1}). \tag{1}$$

The environmental covariates $\boldsymbol{x}$ include the abiotic factors (V, P, BOM) observed during each campaign. The residuals $\boldsymbol{\epsilon}$ in (1) have correlation matrix $\boldsymbol{R}^{-1}$ and precision τ_ϵ. Any structure in the residuals can then be modelled by specifying a suitable form for $\boldsymbol{R}$. We use two alternatives for $\boldsymbol{R}$ that correspond to two different assumptions on $\boldsymbol{\epsilon}$: the first assumes the residuals are independent (i.e. $\boldsymbol{R} = \boldsymbol{I}$; `iid case`), while the second states the residuals are independent between campaigns and exchangeable within campaigns with correlation ρ (`exch case`). In the latter case, $\boldsymbol{R} = \boldsymbol{I}_6 \otimes \boldsymbol{C}$, where $\otimes$ indicates the Kronecker product and $\boldsymbol{C}$ is a matrix containing 1 in the diagonal and ρ out of diagonal; note, this is also referred to as compound symmetry and corresponds to the mixed model $\boldsymbol{x}_{i,j}^{\mathsf{T}}\boldsymbol{\beta} + b_j + \epsilon_{i,j}$, where b_j is a random effect for campaign and ϵ_{ij} are independent residuals. In the `iid case` the prior $\pi(\boldsymbol{\epsilon})$ depends on the hyperparameter τ_ϵ, while in the `exch case` it depends on both τ_ϵ and ρ. Regarding τ_ϵ, we use the Gumbel PC prior in both cases. Following Simpson et al. (2017) section 6, we use the PC prior for the correlation parameter ρ defined on the distance from an iid base model (i.e. $\rho = 0$). To complete the prior specification we need to set the degree of penalty - applied to $\pi(\boldsymbol{\epsilon})$ - for deviating from the iid base model $\pi(\boldsymbol{\epsilon}|\rho = 0)$. This can be done through an intuitive user-defined scaling approach: for instance, setting the prior median for ρ as equal to 0.5, which means that 0.5 prior probability mass is assigned to $\rho < 0.5$. This seems a sensible approach in general, as we will hardly have any precise prior guess about ρ in practice.
Model (1) implementing the two priors was fitted in `R-INLA` (Rue et al. 2009). The posterior summaries (mean, lower and upper quantiles) for the main model parameters are in Table 1. Environmental covariates show significant effect on the abundance of macroinvertebrate and are broadly the same in the two cases. Checking whether data shows any evidence of correlation within campaigns is important, as that may suggest the presence of intra-community processes driven by biotic factors. The hyperparameter ρ is concentrated at around 0.42 (with credible interval from 0.26 to 0.58), meaning that there is evidence for within campaign residual correlation. We analyzed posterior credible intervals for the residuals $\boldsymbol{\epsilon}$ across campaigns and they do not show any clear pattern if the exchangeable prior is used, whereas they look much more structured under the iid prior (figures not shown here).

3 Concluding remarks

Mixed models represent a valuable toolbox for statistical modelling of ecological survey data. The use of PC priors in a mixed model setting can lead to advantages in terms of avoiding overfitting models and invariance over reparametrization. The preliminary results discussed here show possible occurrence of biotic processes driving macroinvertebrate communities in the study area. Future work will investigate spatial structure in the residuals, modelling small-scale variations along the transects. This would be a further step into characterizing the nature of the biotic processes driving natural communities.

TABLE 1. Posterior summaries for model (1), using iid and exchangeable prior. In the last line, ρ is the within campaign correlation.

	iid case			exch case		
	0.025q	mean	0.975q	0.025q	mean	0.975q
(Intercept)	4.09	4.31	4.52	3.52	4.35	5.18
V	0.13	0.28	0.43	0.12	0.26	0.39
P	-0.40	-0.27	-0.13	-0.34	-0.22	-0.09
BOM	0.37	0.57	0.77	0.32	0.50	0.68
τ_ϵ	1.08	1.27	1.49	0.72	0.97	1.30
ρ				0.26	0.42	0.58

Acknowledgments: This work is supported by the PRIN 2015 grant 20154X8K23 (EPHASTAT).

References

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics (Oxford, England)*, **11(3)**, 397 – 412.

Laini, A., Vorti, A., Bolpagni, R., and Viaroli P. (2014). Small-scale variability of benthic macroinvertebrates distribution and its effects on biological monitoring. *Annales de Limnologie - International Journal of Limnology*, **50(3)**, 211 – 216.

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(2)**, 319 – 392.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sorbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, **32(1)**, 1 – 28.

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., and Smith, G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R.* Springer.

Inference for multiplicative model combination using score matching

Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, via Tomadini 30/a, I-33100 Udine, Italy

E-mail for correspondence: paolo.vidoni@uniud.it

Abstract: This paper concerns multiplicative model combination, which is a way of combining probability models using a weighted multiplication and a subsequent normalization. In particular, we focus on density estimation problems and we define a density estimator, based on a suitable model combination, using a new boosting-type algorithm with the Hyvärinen score as loss function. Finally, a simple application to the estimation of the precision matrix of a multivariate Gaussian model is presented.

Keywords: Boosting; Density estimation; Hyvärinen's divergence; Multiplicative mixture model

1 Introduction

The main focus of the paper is density estimation and, in particular, the aim is to estimate an unknown density function using a suitable combination of basic density functions, which might correspond to simple probability models describing particular features of the interest random phenomenon. The problem of combining density functions, also termed model pooling or combination of experts, is considered quite often in the machine learning and in the econometric literature (see, for example, Hinton, 2002, and Geweke and Amisano, 2011).
Let us consider a continuous random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$, with $K \geq 1$, having an unknown density function $f(\boldsymbol{z})$, $\boldsymbol{z} \in \mathbf{R}^K$, and a set $\mathcal{P} = \{p_j(\boldsymbol{z}; \theta_j),\, j = 1, \ldots, J\}$ containing $J \geq 1$ plausible density functions for $\boldsymbol{Z}$, where θ_j is a vector including the parameters of the j-th model. Let us assume that a sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, with $n \geq 1$, is available from $\boldsymbol{Z}$. The aim is to use the information given by the observed data in order to define a

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

combination of the models in $\mathcal{P}$ to be considered as a useful surrogate for the true density $f(\boldsymbol{z})$. In particular, we consider multiplicative combinations of densities defined as

$$f_p(\boldsymbol{z}; w, \theta) = c(w, \theta)^{-1} \prod_{j=1}^{J} p_j(\boldsymbol{z}; \theta_j)^{w_j}, \tag{1}$$

with $w = (w_1, \ldots, w_J)$ a J-dimensional vector of non-negative weights, $\theta = (\theta_1, \ldots, \theta_J)$ and $c(w, \theta) = \int_{\mathbf{R}^K} \prod_{j=1}^{J} p_j(\boldsymbol{z}; \theta_j)^{w_j} d\boldsymbol{z}$ the normalizing constant, supposed to be finite. We emphasize that in many applications the computation of the normalizing constant $c(w, \theta)$ could be intractable or very computationally demanding and this makes infeasible the use of likelihood-based approaches for making inference on w (and possibly on θ).

2 A boosting-type algorithm for density estimation

The objective is to find a multiplicative density combination (1) to be considered as a suitable estimator for the unknown density function $f(\boldsymbol{z})$ or, equivalently, to find an estimator for the unknown vector of weights w. We assume that the density functions are twice differentiable and, to simplify the presentation, that the model parameters θ_j, $j = 1, \ldots, J$, are known, and then omitted in the notation. The inferential procedure relies on the following divergence, introduced by Hyvärinen (2005),

$$H(f, f_p; w) = \int_{\mathbf{R}^K} ||\nabla \log f(\boldsymbol{z}) - \nabla f_p(\boldsymbol{z}; w)||^2 f(\boldsymbol{z})\, d\boldsymbol{z},$$

where $\nabla g(\boldsymbol{z}) = (\partial g(\boldsymbol{z})/\partial z_1, \cdots, \partial g(\boldsymbol{z})/\partial z_K)$ is the gradient and $|| \cdot ||$ the Euclidean norm. This divergence is non-negative, it vanishes only when $f \equiv f_p$ and, since it involves the gradient of the log-densities, it can be computed without the knowledge of the normalizing constants of $f(\boldsymbol{z})$ and $f_p(\boldsymbol{z}; w)$. Since it matches the scores, with respect to the vector $\boldsymbol{z}$, it is also referred to as the *score matching loss.* Under suitable regularity assumptions, Hyvärinen (2005) proved that minimizing $H(f, f_p; w)$ is equivalent to minimizing the expected Hyvärinen score

$$S_H(f, f_p; w) = \int_{\mathbf{R}^K} \left[2 \bigtriangleup \log f_p(\boldsymbol{z}; w) + ||\nabla \log f_p(\boldsymbol{z}; w)||^2\right] f(\boldsymbol{z})\, d\boldsymbol{z}, \tag{2}$$

where $\bigtriangleup g(\boldsymbol{z}) = \sum_{k=1}^{K} \partial^2 g(\boldsymbol{z})/\partial z_k^2$ is the Laplacian. This result holds for continuous random vectors with support $\mathbf{R}^K$. Extensions to the case of continuous, non-negative random vectors and to some particular discrete random vectors may be found in Hyvärinen (2007).

The integral in (2) admits an empirical version, based on an average over the observed sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, given by

$$\widehat{S}_H(f, f_p; w) = \frac{1}{n} \sum_{i=1}^{n} \left[2 \bigtriangleup \log f_p(\boldsymbol{z}_i; w) + ||\nabla \log f_p(\boldsymbol{z}_i; w)||^2\right] \tag{3}$$

and $\hat{w} = \min_{w \in \mathbf{R}_+^J} \widehat{S}_H(f, f_p; w)$ defines the score matching estimator for w. In order to solve this optimization problem, we consider a gradient boosting algorithm (see, for example, Friedman, 2001), which can be viewed as a simple variant of the coordinate descent method. More precisely, $\hat{w}$ is obtained by means of an iterative procedure where, chosen an initial value $\hat{w}^{(0)} \in \Omega$, we repeat the updating step $\hat{w}^{(r)} = \hat{w}^{(r-1)} + \alpha^{(r)}\, d^{(r)}$, $r = 1, 2, \ldots$, until a stopping criterion is satisfied. The vector $d^{(r)} \in \mathbf{R}^J$ indicates the search direction and, for coordinate descent methods, it corresponds to a vector e_h with a one in position $h \in \{1, \ldots, J\}$ and zero in all other positions. The coordinate h for descent corresponds to that one giving the largest component of the gradient vector in absolute value. Thus, at each step, only the weight of the selected component density $p_h(\boldsymbol{z})$ is modified in order to provide the maximal reduction in the loss function. Furthermore, $\alpha^{(r)}$ specifies the step sized and it can be obtained by linear search. Alternatively, a suitable constant step size may be defined: usually a fixed quantity with a small absolute value and the sign chosen in order to satisfy the descent condition.

This boosting-type algorithm is very simple and, although more advanced algorithms could be considered in order to achieve better convergence results, it can be surely convenient whenever the multiplicative mixture model has a large number of components. Thus, it defines a regularization procedure useful when the score matching approach is applied in high-dimensional problems. Moreover, it can be readily extended to the case where the component model parameters $\theta_1, \ldots, \theta_J$ are unknown (Vidoni, 2017).

3 Estimation of Gaussian precision matrices

We apply the boosting-type algorithm based on the Hyvärinen score for estimating the precision matrix of a multivariate Gaussian distribution. The precision matrix is defined as the inverse of the covariance matrix and, when its dimension is large, the estimation problem could be challenging. It is well-known that the precision matrix defines the conditional dependence structure of Gaussian graph models and Gaussian Markov random fields.

Let $\boldsymbol{Z}$ be a K-dimensional random vector following a multivariate Gaussian distribution with a null mean vector $\mu = 0$ and a non-singular covariance matrix $\Sigma = (\sigma_{rs})$. Given a sample $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$, $i = 1, \ldots, n$, the aim is to estimate the precision matrix $Q = \Sigma^{-1} = (q_{rs})$. To this end we consider a multiplicative mixture of K-variate Gaussian densities with a null mean vector and a suitable symmetric precision matrix $Q_j = Q_j(\theta) = (q_{j,rs}(\theta))$, $j = 1, \ldots, J$. Thus, using well-known properties of the Gaussian distribution, we can conclude that the multiplicative mixture density (1) corresponds to a Gaussian density with precision matrix $Q_H = \sum_{j=1}^J w_j Q_j$.

In this framework, we consider as objective function the empirical average (3) given by

$$\widehat{S}_H(f, f_p; \theta, w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left\{\sum_{j=1}^{J} w_j \sum_{s=1}^{K} z_{is} q_{j,ks}(\theta)\right\}^2 - 2\sum_{j=1}^{J} w_j q_{j,kk}(\theta)\right].$$

Using the algorithm presented in Section 2, we find the estimates $\hat{w}$, $\hat{\theta}$ and then the estimate $\hat{Q}_H = \sum_{j=1}^{J} \hat{w}_j Q_j(\hat{\theta})$ for the unknown precision matrix Q, provided that it is symmetric and positive-definite. Note that $\hat{Q}_H$ is defined as a linear combination of a set of simple precision matrices $\hat{Q}_j = Q_j(\hat{\theta})$, $j = 1, \ldots, J$, giving a partial description of the conditional covariance structure of the random vector $\boldsymbol{Z}$.

The choice of this system of matrices Q_j, $j = 1, \ldots, J$, is crucial for the effectiveness of the inferential procedure. For example, if we know that the true Q is a band matrix, namely it is a sparse matrix with non-zero entries confined to a diagonal band with unknown dimension, we may consider the following $J = K$ component matrices

$$Q_j = \begin{pmatrix} q_{j,11} & 0 & \cdots & q_{j,1j} & 0 & \cdots & 0 \\ 0 & q_{j,22} & 0 & \ddots & q_{j,2(j+1)} & \ddots & \vdots \\ \vdots & 0 & q_{j,33} & \ddots & \ddots & q_{j,3(j+2)} & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ q_{j,j1} & 0 & \ddots & \ddots & q_{j,jj} & \ddots & 0 \\ \vdots & q_{j,(j+1)2} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \ddots & 0 & \cdots & 0 & q_{j,KK} \end{pmatrix},$$

$j = 1, \ldots, J$, having the main diagonal and only two equal non-null, symmetric diagonals in K different positions. Whenever the conditional covariance does not follows a band structure an alternative, more general system of component precision matrices can be considered.

In order to guarantee identifiability in the objective function, we assume that $q_{1,rr} = \theta_r$ and $q_{j;rr} = 0$, $j = 2, \ldots, J$, $r = 1, \ldots, K$; furthermore, all the non-null, off-diagonal elements of the matrices are considered as equal to 1. Thus, matrix Q_1 (with a fixed weight $w_1 = 1$) is a diagonal matrix defining the conditional precision of each marginal component of vector Z, whereas the remaining matrices Q_j, $j = 2, \ldots, J$, specify the presence of some specific non-null conditional correlations, whose values are defined by the corresponding weights w_j, $j = 2, \ldots, J$. For example, the precision matrix obtained from the system of component precision matrices outlined before corresponds to

$$Q_H = \sum_{j=1}^{J} w_j Q_j = \begin{pmatrix} \theta_1 & w_2 & w_3 & \cdots & w_K \\ w_2 & \theta_2 & w_2 & \ddots & w_{K-1} \\ w_3 & w_2 & \theta_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & w_2 \\ w_K & w_{K-1} & \cdots & w_2 & \theta_K \end{pmatrix}.$$

Although, with the assumptions stated before, the component matrices Q_j, $j = 2, \ldots, J$, present a null main diagonal, and they are expected to be singular, the optimization procedure can be applied in the same way, giving a useful estimated precision matrix $\hat{Q}_H$.

A simulation study concerning the simple situation of band matrices is presented and it shows that the boosting algorithm produces an estimator for Q having better accuracy then the graphical lasso estimator (Mazumder and Hastie, 2012) and the estimator obtained as the inverse of the sample covariance matrix. We consider 100 simulated samples of dimension $n = 100, 200, 500$ from a K-variate Gaussian distribution with $K = 20$, having a null mean vector and a band precision matrix Q with non-null entries $q_{kk} = 2$, $k = 1, \ldots, K$, $q_{k(k-1)} = q_{(k-1)k} = -0.5$, $k = 2, \ldots, K$, $q_{k(k-2)} = q_{(k-2)k} = 0.4$, $k = 3, \ldots, K$, $q_{k(k-3)} = q_{(k-3)k} = -0.3$, $k = 4, \ldots, K$, and $q_{k(k-4)} = q_{(k-4)k} = 0.2$, $k = 5, \ldots, K$. We aim at comparing the empirical properties of the following estimators: $\hat{Q}_{H1}$, based on the boosting-type algorithm assuming a band structure for the component precision matrices, $\hat{Q}_{H2}$, based on the boosting-type algorithm assuming a more general structure for the component precision matrices (Vidoni, 2017), $\hat{Q}_{GL}$, based on the graphical lasso algorithm proposed by Mazumder and Hastie (2012) and S^{-1}, corresponding to the inverse of the sample covariance matrix.

We compare the alternative methods in terms of the Kullback-Leibler loss between the true and the estimated Gaussian densities

$$\mathrm{KL} = \mathrm{tr}(Q^{-1}\hat{Q}) - \log(|Q^{-1}\hat{Q}|) - K,$$

where $\mathrm{tr}(\cdot)$ and $|\cdot|$ indicate, respectively, the trace and the determinant of a matrix. Quantity KL measures how close the estimated $\hat{Q}$ is to the true Q and lower values indicate a better estimate, with $\mathrm{KL} = 0$ if $\hat{Q} = Q$.

The sample estimates of KL, with the associate standard errors, are presented in Table 1. We underline that the estimators based on the boosting-type algorithm and the Hyvärinen's divergence exhibit a good performance. In particular, $\hat{Q}_{H1}$ achieves definitely the best results, whereas the behavior of $\hat{Q}_{H2}$ is quite similar to that of the lasso-type estimator $\hat{Q}_{GL}$ in all the experimental situations, excluding the case with $n = 100$. This can be explained by recalling that $\hat{Q}_{H2}$ is defined without assuming a particular structure for the system of component precision matrices and then it corresponds to the most general and less powerful estimator of this family.

Finally, the inverse of the sample covariance matrix performs, as expected, very poorly, in particular for $n = 100, 200$.

n	$\hat{Q}_{H1}$	$\hat{Q}_{H2}$	$\hat{Q}_{GL}$	S^{-1}
100	0.405	1.775	1.303	3.067
	(0.009)	(0.025)	(0.010)	(0.043)
200	0.228	0.791	0.837	1.238
	(0.006)	(0.010)	(0.008)	(0.014)
500	0.123	0.363	0.360	0.447
	(0.003)	(0.004)	(0.004)	(0.005)

TABLE 1. Estimated Kullback-Leibler loss and standard errors (in brackets) for the boosting-type estimator with band component precision matrices $\hat{Q}_{H1}$, the boosting-type estimator based on general precision matrices $\hat{Q}_{H2}$, the graphical lasso estimator $\hat{Q}_{GL}$ and the inverse of the sample covariance matrix S^{-1}. Simulated samples of dimension $m = 100, 200, 500$ from a K-variate Gaussian distribution with $K = 20$, having a null mean vector and a band precision matrix.

Acknowledgments: This research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003 and by the University of Udine under the PRID2017 research grant.

References

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29**, 1189 – 1232.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, **164**, 130 – 141.

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, **14**, 1771 – 1800.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695 – 709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499 – 2512.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: new insights and alternatives. *Electronic Journal of Statistics*, **6**, 2125 – 2149.

Vidoni, P. (2017). Boosting multiplicative model combination. *Preprint.*

An R package for Determining Groups in Multiple Survival Curves

Nora M. Villanueva[1], Marta Sestelo[2], Luís Meira-Machado[3]

[1] Dep. Statistics and O.R., University of Vigo, Spain.
[2] SiDOR Research Group and CINBIO, University of Vigo, Spain.
[3] Centre of Molecular and Environmental Biology & Department of Mathematics and Applications, University of Minho, Portugal.

E-mail for correspondence: `nmvillanueva@uvigo.es`

Abstract: Survival analysis includes a wide variety of methods for analyzing time-to-event data. One basic but important goal in survival analysis is the comparison of survival curves between groups. Several nonparametric methods have been proposed in the literature to test for the equality of survival curves for censored data. When the null hypothesis of equality of curves is rejected, leading to the clear conclusion that at least one curve is different, it can be interesting to ascertain whether curves can be grouped or if all these curves are different from each other. We present the R `clustcurv` package which allows determining groups with an automatic selection of their number. The applicability of the proposed method is illustrated using real data.

Keywords: Log-rank Test; Multiple Survival Curves; Number of Groups; Survival Analysis

1 Introduction

Survival analysis includes a wide variety of methods for analyzing time-to-event data. One basic but important goal in survival analysis is the comparison of survival curves between groups. For example, in an observational survival study, one may be interested in comparing survival between individuals from different age groups, different genders, racial/ethnic groups, geographic localization, etc.

Several nonparametric methods have been proposed in the literature to test for the equality of survival curves for censored data. The log-rank or Mantel-Haenszel test (Mantel, 1966) is the most well-known and widely used to test the null hypothesis of no difference in survival between two or more independent groups. An alternative test that is often used is the Peto & Peto (1972) modification of the Gehan-Wilcoxon test (Gehan, 1965). Though the aforementioned methods can be used to compare multiple survival curves, methods that can be used to determine groups among a series of survival curves are not available, to the best of our knowledge. When the log-rank test (or its analogous) is used to compare three or more survival curves at once, the test reports a single p-value testing the null hypothesis that all the samples come from populations with identical survival. If the null hypothesis of equality of curves is rejected, then, this leads to the clear conclusion that at least one curve is different. However, these methods cannot be used to ascertain whether groups of curves can be performed or if all these curves are different from each other.

One naïve approach would be to perform pairwise comparisons. However, this approach would lead to a large number of comparisons (e.g. 7 groups would lead to 21 pairwise comparisons). One could make it but without the possibility of determining groups with similar survival curves. This can be achieved with the `pairwise_survdiff` of the package `survminer` (Kassambara and Kosinski, 2017) which calculates pairwise comparisons between group levels with corrections for multiple testing. Results for such a test can tell us that all combinations are different, or just one pair. However, as it was mentioned, when the number of curves increases so does the difficult of interpretation.

According to this, the paper introduces `clustcurv`, a software application for R which allows determining groups with an automatic selection of their number based on k-means or k-medians algorithms (Villanueva et al., submitted). It describes the capabilities of the package using a real dataset.

2 The clustcurv package in practice

To illustrate our method we will use one real dataset. It comes from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer (Moertel et al., 1990). This data set is freely available as part of the R package `condSURV` (Meira-Machado and Sestelo, 2016). From the total of 929 patients, 452 died. For each individual, an indicator of his/her final vital status (censored or not), the survival time (time to death) from the entry of the patient in the study (in days), and a covariate including the number of lymph nodes with detectable cancer (grouped from 1 to ≥ 10 in the dataset colonCSm) were used.

```
> devtools::install_github("noramvillanueva/clustcurv")
> library(clustcurv); library(condSURV)
> head(colonCSm)[1:2, ]
  time status nodes
1 1521      1     5
```

```
2 3087       0      1
```

The estimated survival curves after splitting the data according to the number of nodes are shown in Figure 1 (upper panel). When we confront with a dataset like this, with a categorical variable with a high number of levels, maybe a good approximation could be to establish groups with the same risk or survival probability. The unique option until now could be to use first the log-rank test and then, if the result of the application of this test is statistically significant, do a post hoc analysis like a pairwise comparison. The p-value of the log-rank test is < 0.01 and the interpretation of the resulting p-values of the pairwise comparison (not shown) becomes a problem.

```
> survdiff(Surv(time, status) ~ factor(nodes), data = colonCSm)
Call:
survdiff(formula = Surv(time, status) ~ factor(nodes),
+ data = colonCSm)
                   N Observed Expected (O-E)^2/E (O-E)^2/V
factor(nodes)=1  274       94   151.93   22.0901   33.9249
factor(nodes)=2  194       74   102.87    8.1022   10.5979
factor(nodes)=3  125       61    62.56    0.0387    0.0453
factor(nodes)=4   84       43    38.26    0.5868    0.6434
factor(nodes)=5   46       34    17.06   16.8249   17.5428
factor(nodes)=6   43       27    16.43    6.8027    7.0736
factor(nodes)=7   38       25    15.41    5.9636    6.1880
factor(nodes)=8   23       18     7.22   16.0875   16.3765
factor(nodes)=9   20       14     8.05    4.3931    4.4795
factor(nodes)=10  62       49    19.21   46.2239   48.6066
Chisq= 129  on 9 degrees of freedom, p= 0

> survminer::pairwise_survdiff(Surv(time, status) ~ nodes,
+ data = colonCSm, p.adjust.method = "BH")
```

To solve it, we applied the proposed procedure. For a significance level of 0.05 and using the Cramér-von Mises type statistic, the null hypothesis $H_0(1)$ is rejected (p-value of < 0.01) while the null hypothesis $H_0(2)$ is accepted (p-value of 0.19). The assignment of the curves to the two groups can be observed in Figure 1.

```
> res <- clustcurv_surv(time = colonCSm$time,
+ status = colonCSm$status, fac = colonCSm$nodes,
+ algorithm = "kmeans", nboot = 500, cluster = TRUE,
+ seed = 300716)
Checking 1 cluster...
Checking 2 clusters...
Finally, there are 2 clusters.
```

```
> autoplot(res, groups_by_colour = TRUE, xlab = "Time (in days)")
```

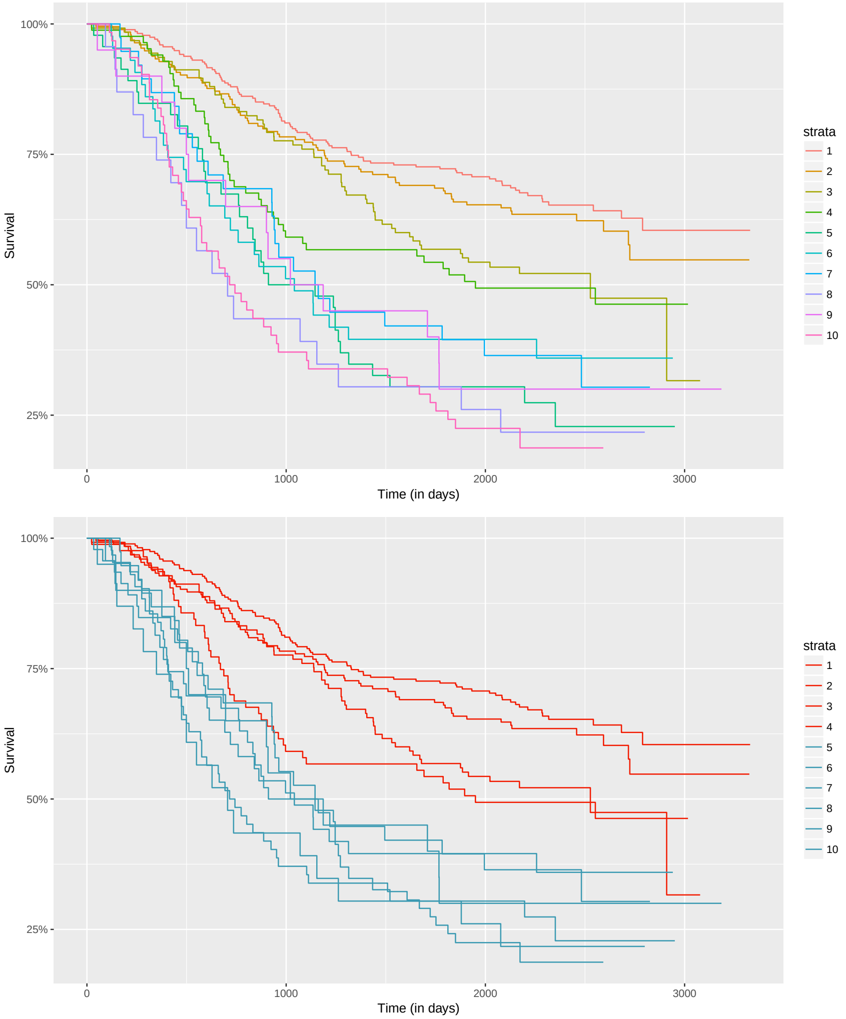


FIGURE 1. Estimated survival curves for each of the levels of the variable "nodes" using the Kaplan-Meier estimator. A specific color is assigned for each curve according to the group to which it belongs (in this case two groups, K = 2).

References

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203 – 223.

Kassambara, A. and Kosinski, M. (2017). survminer: Drawing Survival Curves using 'ggplot2'. *R package version 0.3.1*, version 0.3.1.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163 – 170.

Meira-Machado, L., Sestelo, M. (2016). condSURV: An R Package for the Estimation of the Conditional Survival Function for Ordered Multivariate Failure Time Data. *The R Journal.* , **8(2)**:460–473.

Moertel, C.G., Fleming T.R., Macdonald J.S., et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine.* **322(6)**:352–358.

Peto, R. and Peto, J. (1972).Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, **135**, 185 – 206.

Villanueva, N. M., Sestelo, M. and Meira-Machado, L. (2018). A method for determining groups in multiple survival curves. *Statistics in Medicine*, submitted.

Frailty Models for Cardiac Allograft Vasculopathy Data

Wenyu Wang[1], Ardo van den Hout[1]

[1] Department of Statistical Science, University College London

E-mail for correspondence: `w.wang.16@ucl.ac.uk`

Abstract: Frailty models are getting more and more popular in survival analysis. These models have two advantages comparing with the fixed-effects model: One is showing the effect of individual-specific and cluster-specific parameters, another is preserving the Markov assumption in survival analysis. Here we fit a frailty model to the cardiac allograft vasulopathy data, the results illustrate that it is better than the fixed-effects model. Furthermore, we can explore different distributions for the frailty for different transitions.

Keywords: Markov model; multi-state model; survival.

1 Introduction

The multi-state model describes a process where individuals move among a series of states over time. It is increasingly popular in a wide range of applications in biostatistics. For instance, breast cancer (Putter et al. 2006), HIV (Gentleman et al. 1994), ageing (Rickayzen and Walsh 2002). Generally, if death is one of the state, the multi-state model can be seen as an extention of survival analysis.

In most studies, covariate effects are fixed effects. These effects do not take into account unobserved heterogeneity with respect to individual or group level effects. In a multi-state model, there may be unobserved heterogeneity with respect to the rate of moving from one state to another. In survival analysis, such an effect on a rate is called a frailty. For example, different moving rate of each individual between transitions as the individual-effect frailty, and different groups of individuals with respect to hospital frailty as the group-effect frailty.

Another contribution of adding frailty parameters in multi-state model is to avoid violating the Markov assumption, which implies that future

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

states are only determined by the current states. It is known the transition hazard may be affected by the duration in previous states, implying that the future not only depends on current states but also the past. For example, individuals who have been longer in disease states, are more likely to move to death; see Putter and van Houwelingen (2015) for more details. It can be addressed with fitting a frailty model, where frailties represents the duration in former states.

In this study, we fit a frailty model to the cardiac allograft vasculopathy data to explore the role of random-effect parameters in multi-state models. Cardiac allograft vasculopathy (CAV) is a kind of disease, which limits survival for cardiac transplant recipients. Sharples et al. (2003) defined it by three living states, which are the grades of CAV at each time. Figure 1 shows the multi-state process. State 1 to 3 are defined by no CAV, moderate CAV, severe CAV, respectively. State 4 is an absorbing state representing dead. Here we use the data to define a progressive process defined by the history of observed states, which results in transitions (1,2), (1,4), (2,3), (2,4), (3,4).

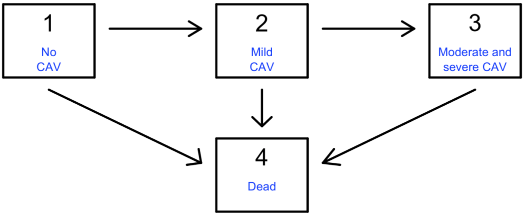


FIGURE 1. Transitions in the four-state model for cardiac allograft vasculopathy (CAV).

2 Method

2.1 Regression Model

For the fixed-effects model, the hazard function can be defined by regression. For transition (r, s), the hazard function is given by

$$h_{rs}(t|\boldsymbol{x}) = h_{rs.0}(t)\exp(\boldsymbol{\beta}_{rs}^{\top}\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x}$ is the vector of covariates, $\boldsymbol{\beta}_{rs}$ is a parameter vector, $h_{rs.0}(t)$ is the baseline hazard.

For the random-effects model, the hazard function for individual i can be defined by adding a frailty variable to equation (1).

$$h_{rs}(t|b_{rs.i}, \boldsymbol{x}) = h_{rs.0}(t)\exp(\boldsymbol{\beta}_{rs}^{\top}\boldsymbol{x} + b_{rs.i}), \tag{2}$$

where $b_{rs.i}$ is the frailty variable. Note that $b_{rs.i}$ can be changed to b_c for a cluster-specific random effect. Here we discuss the normal distribution: $b_{rs.i} \sim N(0, \sigma_{rs}^2)$. In this study, the hazard regressions are defined with different covariates in different transitions, which are displayed in Table 1. In this table, t, $bage$, $dage$ are represent years followed-up, baseline age and doner age of individuals, respectively.

TABLE 1. The hazard function in each transition for CAV data

Transition	Hazard Function, where $b_i \sim N(0, \sigma^2)$
$(1, 2)$	$h_{12}(t\|b_i, \boldsymbol{x}) = \exp(\beta_{12.0} + \beta_{12.1}t + \beta_{12.2}bage + \beta_{12.3}dage + b_i)$
$(1, 4)$	$h_{14}(t\|\boldsymbol{x}) = \exp(\beta_{14.0} + \beta_{14.2}bage + \beta_{14.3}dage)$
$(2, 3)$	$h_{23}(t\|b_i, \boldsymbol{x}) = \exp(\beta_{23.0} + \beta_{23.3}dage + b_i)$
$(2, 4)$	$h_{24}(t\|\boldsymbol{x}) = \exp(\beta_{24.0} + \beta_{24.3}dage)$
$(3, 4)$	$h_{34}(t\|\boldsymbol{x}) = \exp(\beta_{34.0} + \beta_{34.3}dage)$

2.2 Likelihood Function

Estimating the model parameters can be undertaken by maximazing the log-likelihood function. Under the Markov assumption, y_{ij} is the state for individual i at time t_{ij}, the likelihood function of frailty model for individual i is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) &= P(Y_J = y_J, \ldots, Y_2 = y_2 | Y_1 = y_1, \boldsymbol{\theta}, \boldsymbol{x}) \\ &= \int_{\Omega_{b_i}} P(Y_J = y_J, \ldots, Y_2 = y_2 | Y_1 = y_1, \theta, x, b_i) f(b_i) db_i \\ &= \begin{cases} \int_{\Omega_{b_i}} (\prod_{j=2}^{J} P(Y_j = y_j | Y_1 = y_1, \theta, x, b_i)) f(b_i) db_i, \\ \text{where } y_J \in \{1, 2, 3\} \\ \int_{\Omega_{b_i}} (\prod_{j=2}^{J-1} P(Y_j = y_j | Y_1 = y_1, \theta, x, b_i)) \times \\ \quad (\sum_{s=1}^{3} P(Y_J = s | Y_{J-1} = y_{J-1}, \theta, x) h_{s4}(t_{J-1}|\theta, x)) f(b_i) db_i, \\ \text{where } y_J \text{ is death} \end{cases} \end{aligned}$$

Thus, the likelihood function of all N individuals is given by multiplying contributions:

$$L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=1}^{N} L_i(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}).$$

Maximizing the likelihood function can be undertaken through using a general-purpose optimisation. Here we use the optimisation in software R.

3 Data Analysis

In this study, we fit both the fixed-effect model and the frailty model mentioned above, AIC of these two models are 3472.6 and 3470.4, respectively. It represents that adding a frailty as a random-effect intercept in the hazard model for transitions $(1, 2)$ and $(2, 3)$ will lead to an improvement of the loglikelihood function. The estimation of the frailty variance is 0.539 with a standard error 0.216. This illustrates that is worthwhile to distinguish movers from stayers.

4 Conclusion

The aim of this study is to discuss whether the frailty model is a bit better than the fixed-effects model. It is clear that the frailty model we fit there is better. In the future, we can explore more types of frailty models, e.g., with different distributions, for different transitions, and bivariate frailty models. Here we give an example of the frailty which follows a one-parameter gamma distribution: $B_{rs.i} \sim Gamma(\phi_{rs})$. The hazard function for individual i can be defined as $h_{rs}(t|B_{rs.i}, \boldsymbol{x}) = h_{rs.0}(t) B_{rs.i} \exp(\boldsymbol{\beta}_{rs}^{\top} \boldsymbol{x})$, where $B_{rs.i}$ is the frailty variable.

References

Gentleman, R. C., Lawless, J. F., Lindsey, J. C., & Yan, P. (1994). Multistate Markov models for analysing incomplete disease history data with illustrations for hiv disease. *Statistics in Medicine*, **13(8)**, 805 – 821.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R., & van de Velde, C. J. (2006). Estimation and prediction in a multistate model for breast cancer. *Biometrical Journal*, **48(3)**, 366 – 380.

Putter, H., & van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical Methods in Medical Research*, **24(6)**, 675 – 692.

Rickayzen, B. D., & Walsh, D. E. (2002). A multi-state model of disability for the United Kingdom: implications for future need for long-term care for the elderly. *British Actuarial Journal*, **8(2)**, 341 – 393.

Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J., and Large, S. R. (2003). Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, **76**, 679 – 682.

Spatiotemporal statistical downscaling for the fusion of in-lake and remote sensing data

Craig Wilkie[1], Claire Miller[1], Marian Scott[1], Stefan Simis[2], Steve Groom[2], Peter Hunter[3], Evangelos Spyrakos[3], Andrew Tyler[3]

[1] University of Glasgow, UK
[2] Plymouth Marine Laboratory, UK
[3] University of Stirling, UK

E-mail for correspondence: `craig.wilkie@glasgow.ac.uk`

Abstract: This paper addresses the problem of fusing data from in-lake monitoring programmes with remote sensing data, through statistical downscaling. A Bayesian hierarchical model is developed, in order to fuse the in-lake and remote sensing data using spatially-varying coefficients. The model is applied to an example dataset of log(chlorophyll-a) data for Lake Erie, one of the Great Lakes of North America.

Keywords: Bayesian hierarchical model; Statistical downscaling; Data fusion; Chlorophyll-a.

1 Introduction and background

This work is motivated by the problem of fusing data from in-lake monitoring programmes with remote sensing data, which have impressive spatial and temporal coverage but require calibration with the in-lake data to ensure accuracy. This presents a problem of change-of-support between the point-scale in-lake data and the grid-cell-scale remote sensing data.

In-lake data have been traditionally used extensively to enable water quality investigators to understand lake health. They are assumed to be accurate within measurement error, since they are obtained from water samples that are taken directly from the lake surface and then analysed in a laboratory. However, these data are expensive to collect in terms of both time and money and so are often sparse in both space and time, with a small

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

number of sampling locations across each lake. They therefore provide little information on the spatial patterns in water quality. Remote sensing data have become much more commonly available in recent years, due to the increased availability of data from Earth-facing satellite monitoring programmes. These data provide spatially comprehensive information on water quality parameters.

In this paper, data for log(chlorophyll-a), an important indicator of lake water quality, are considered. The example used is Lake Erie, one of the Great Lakes of North America, which has suffered from poor water quality in the past and is therefore of interest to regulatory bodies and local communities. The in-lake data are available for 20 locations over 20 months, collected by the US Environmental Protection Agency and made available in the LIMNADES database (https://www.limnades.org/home.psp). These data are collected at several time points within each month and are temporally aggregated onto the monthly scale, before analysis. The remotely-sensed data are available over the same time period, but with a much better spatial coverage, with grid cells of up to 300 metres in dimension, with 351,041 grid cells covering the lake, on a monthly-averaged time-scale. These European Space Agency Medium Resolution Imaging Spectrometer data were produced through the GloboLakes project and are available at https://globolakes.eofrom.space/.

The remotely-sensed data and the in-lake data for August 2007 are shown in Figure 1 below.

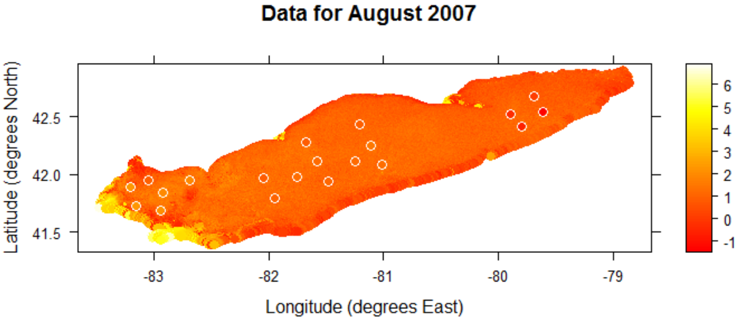


FIGURE 1. Remote sensing data for August 2007, with the in-lake data overlaid and surrounded by white circles.

This paper presents a spatiotemporal development of the model of Wilkie et al. (2015), with an application to a spatially-larger dataset. The model is based upon the approach of Gelfand et al. (2003), which was developed into a statistical downscaling model by Berrocal et al. (2010) for air quality data.

2 Methodology

A Bayesian hierarchical model is proposed for the fusion of remote sensing and in-lake data. The model allows for the n_j in-lake sampling locations to differ for each time point j (for $j = 1, \dots, t$). For the vector of response data $\mathbf{y}_j$, i.e. the vector of in-lake data collected at the n_j sampling locations at time j, and the vector of remote sensing data $\mathbf{x}_j$ recorded for the n_j grid cells containing these in-lake sampling locations, the model is written as follows:

$$\mathbf{y}_j \sim \mathrm{N}_{n_j}(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma^2_\varepsilon \mathbf{I}_{n_j}),$$

where the vectors of intercepts and slope coefficients $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are allowed to be smoothly spatially-varying and so are given the following multivariate-Normal prior distributions:

$$\boldsymbol{\alpha}_j \sim \mathrm{N}_{n_j}(\mathbf{0}, \sigma^2_\alpha \exp(-\phi_\alpha \mathbf{D}_j)) \text{ and } \boldsymbol{\beta}_j \sim \mathrm{N}_{n_j}(\mathbf{1}, \sigma^2_\beta \exp(-\phi_\beta \mathbf{D}_j)),$$

where σ^2_α and σ^2_β are the spatial variance parameters and ϕ_α and ϕ_β are the spatial decay parameters, controlling how fast the correlations in the intercept and slope parameters decrease to zero as the distance between the in-lake sampling locations increases. These parameters are shared over time, which helps to improve their estimation. The matrix $\mathbf{D}_j$ is the $n_j \times n_j$ matrix of distances between the in-lake sampling locations for time j. Finally, the remaining prior and hyperprior distributions must be specified. The spatial variance parameters and error variance parameter are given the following distributions:

$$\sigma^2_\alpha \sim \text{Inv-Gamma}(2,1),\ \sigma^2_\beta \sim \text{Inv-Gamma}(2,1) \text{ and } \sigma^2_\epsilon \sim \text{Inv-Gamma}(2,1),$$

following the example of Sahu et al. (2006). As noted by Sahu et al. (2006), the spatial decay parameters are not easy to identify and so a grid search is performed.

All full conditional posterior distributions can de derived. Therefore, the model is fitted using Gibbs sampling. To provide predictions over the lake surface, Delaunay triangulation, constrained by the lake edges, is carried out to ensure the optimal spatial coverage of the prediction locations.

The temporal aspect of the data is made use of through the sharing of information over time, with the error variance and spatial variance parameters being estimated from the data for all timepoints.

3 Example for Lake Erie

Using the example dataset for Lake Erie, the model is fitted using the R packages `Rcpp` and `RcppArmadillo`, with predictions made at 1000 locations for each of the 20 months in the dataset. These locations are defined by a Delaunay triangulation that is constrained by points along the lake

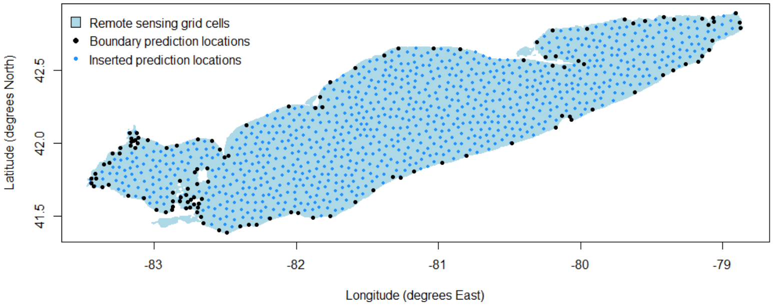

□ Remote sensing grid cells
● Boundary prediction locations
● Inserted prediction locations
Latitude (degrees North)
41.5
42.0
42.5
-83
-82
-81
-80
-79
Longitude (degrees East)

FIGURE 2. Remote sensing grid cells with prediction points overlaid, as obtained from a Delaunay triangulation constrained by the boundary points.

edges, using the R package `RTriangle`. The constraining points and the resulting inserted points are shown in Figure 2.

The model is run for 2 chains of 10,000 iterations each, with every tenth iteration saved, after a burn-in period of 100 iterations. Trace and density plots, such as the examples for the prediction at prediction location 1 for August 2007 ($\tilde{y}_{1,\text{Aug 2007}}$) shown in Figure 3, provide no evidence against the assumption that the MCMC chains have converged to their posterior distributions.

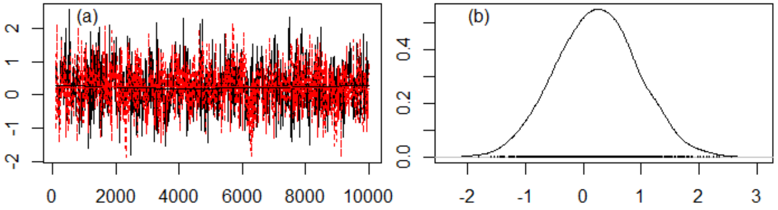


FIGURE 3. (a) Trace plot of the MCMC iterations for $\tilde{y}_{1,\text{Aug 2007}}$; (b) Plot of the posterior density for $\tilde{y}_{1,\text{Aug 2007}}$.

The resulting predictions for August 2007 are shown in Figure 4(a) and their corresponding standard errors are given in Figure 4(b). These predictions illustrate the utility of the model for calibrating the remotely-sensed data using the in-lake data, while retaining the important spatial patterns of the remote sensing data. Figure 4(a) shows the adjustments to the remote sensing image of Figure 1 as a result of the fusion with the in-lake data. In the example shown here, the model predictions show that the northeast of the lake has lower values of log(chlorophyll-a) in August 2007, while the southwest of the lake has higher values. Figure 4(b) shows that the standard errors are lowest closest to the in-lake data locations for this

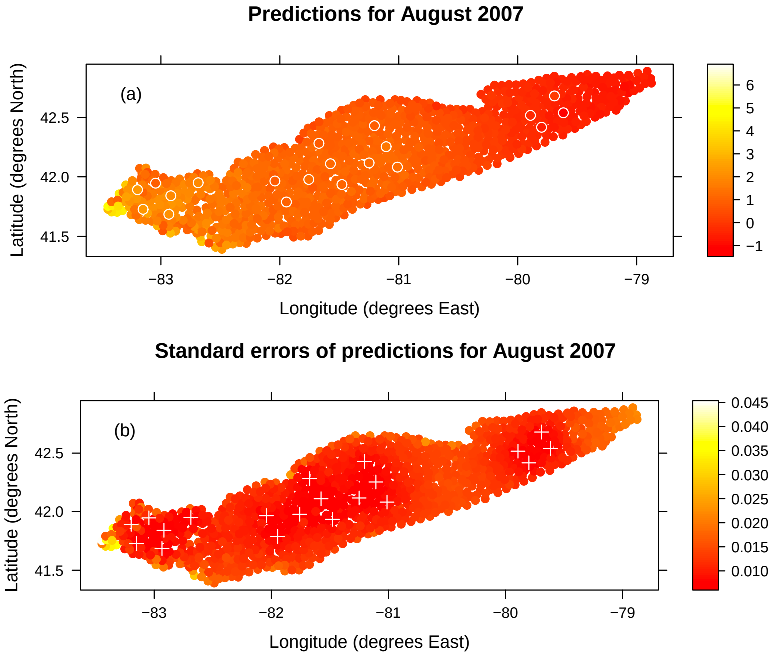


FIGURE 4. (a) Predictions for August 2007, with the in-lake data overlaid and surrounded by white circles; (b) Standard errors of predictions for August 2007, with the in-lake data locations marked by white crosses.

month, as expected. The standard errors are small in comparison to the variation across the lake, providing evidence of a true pattern across the lake surface. The resulting spatial maps, such as the example shown in Figure 4(a), would be useful for water quality investigators to identify parts of the lake of particular interest for further study.

4 Conclusions

The model described in this work enables the fusion of data from in-lake monitoring schemes, which are limited spatially and temporally, with extensive remotely-sensed data with good spatial and temporal resolution. The model makes use of data from multiple available timepoints in order to improve the estimation of the spatial variance parameters and the error variance parameter. Predictions can be made at any point location for which corresponding remotely-sensed data are available, i.e. any location within a remote sensing grid cell. Delaunay triangulation is used to optimise the spatial coverage of the prediction locations, in order to gain a better

understanding of the state of the health of the lake without increasing the computational complexity of the model.
Future work focusses on dealing with the temporal change of support, which can be accomplished through treating the data for each in-lake location and remote-sensing grid-cell as observations of smooth functions over time.

Acknowledgments: CW acknowledges the support of the School of Mathematics and Statistics, University of Glasgow, for funding the PhD during which this work was carried out. CM and MS were partly funded for this work through the NERC GloboLakes project (NE/J022810/1).

References

Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176 – 197.

Gelfand, A.E., Kim, H.-J., Sirmans, C.F., and Banerjee, S. (2003). Spatial modelling with spatially-varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387 – 396.

Sahu, S.K., Gelfand, A.E., and Holland, D.M. (2006). Spatio-temporal modelling of fine particulate matter. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 61 – 86.

Wilkie, C.J., Scott, E.M., Miller, C., Tyler, A.N., Hunter, P.D., and Spyrakos, E. (2015). Data fusion of remote-sensing and in-lake chlorophyll$_\mathrm{a}$ data using statistical downscaling. *Procedia Environmental Sciences*, **26**, 123 – 126.

Testing for zero–modification relative to a negative–binomial distribution.

Paul Wilson[1]

[1] School of Mathematics and Computer Science, University of Wolverhampton, WV1 1LY, United Kingdom

E-mail for correspondence: pauljwilson@wlv.ac.uk

Abstract: Wilson and Einbeck (2016, 2018) propose a test for zero–modification relative to a stated model that uses the observed number of zeros as a test statistic, focusing on a Poisson model. We extend the focus to a negative–binomial model with fixed size parameter and show that excellent attainment rates and power are achieved. The extension of the test to negative binomial models where both parameters are estimated is also discussed.

Keywords: zero-modification, geometric model, negative–binomial model.

1 Introduction

The concept of the zero–inflation and zero–deflation relative to a given statistical model is now firmly established in the statistical literature. The terms zero–inflation and zero–deflation have sometimes been combined towards zero–modification, meaning that there are either too few or too many zeros in the data, relative to the specified count data model. Various tests for zero–modification already exist: the likelihood ratio test, score (Rao) and Wald tests. While these tests are all viable, they rely upon asymptotic results and hence implicitly on large samples, and their test statistics do not, in their standard form, transparently distinguish between zero–inflation and zero–deflation. Wilson and Einbeck (2016, 2018) proposed a new test for zero–modification that relates more directly to the character of zero–modification than the other tests: The test employs the number of observed zeros, n_0, in the data as the test statistic, and tests whether this number is consistent with the non zero–modified model, G. This is achieved by referencing the value of n_0 to the appropriate Poisson-Binomial distribution (Chen and Liu, 1997). The author believes that this feature of the test will be very attractive to statistical practitioners. Wilson and Einbeck (2016, 2018) focus on the case where G is a Poisson model, and show that the attainment and power of the test compare extremely favourably with that of other tests of zero–modification. In common with the score test, the test of Wilson and Einbeck (2016, 2018) does not require the zero–modified model to be fitted, and unlike all other tests (except the "normal distribution" version of the Wald test) the test statistic of the observed number of zeros directly indicates the direction of the zero–modification if it occurs.

1.1 The Estimation of the Mean Parameter

Wilson and Einbeck (2016, 2018) show that for testing for zero–modification relative to a Poisson model the maximum likelihood estimator of μ, $\hat{\mu}_W = \sum_{i=1}^{n} y_i/n$ is, for a given value of n_0, a precise but biased estimator of μ, and the estimator $\hat{\mu}_T$ obtained from the mean of the positive observations is an apparently unbiased estimator of μ which is however imprecise. It is shown that the use of $\hat{\mu}_W$ and $\hat{\mu}_T$ in the proposed test results in under–attainment and over-attainment of the nominal significance rate respectively. It is also shown that the "hybrid estimator", $\hat{\mu}_H$ of μ:

$$\hat{\mu}_H = h\hat{\mu}_W + (1-h)\hat{\mu}_T \tag{1}$$

where $h = 2/3$ results in excellent attainment rates and power.

2 Zero–Modified Negative Binomial Models

In this paper we shift the focus to testing for zero–modification relative to the negative binomial (type II) model with *fixed* size parameter, α, the mean parameter μ_i (possibly) depending on covariates, i.e.

$$NB(\mu_i) = \frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(y+1\right)} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y} \frac{1}{(1+\alpha\mu_i)^{\frac{1}{\alpha}}} \tag{2}$$

Note that for this parameterization:

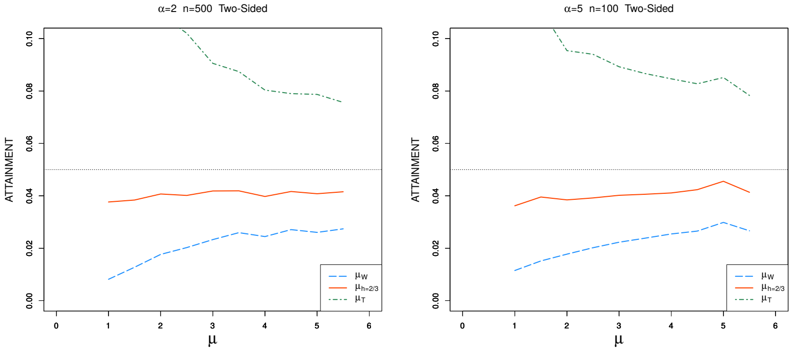
$$\mu = \mu_i \qquad\qquad \alpha^2 = \mu_i(1+\mu_i\alpha) \tag{3}$$

where $\alpha = 1$ corresponds to a geometric distribution.

Figure 1 illustrates the attainment rate when the test of Wilson and Einbeck (2016, 2018) is applied to data drawn from negative binomial distributions with $\alpha = 2$, $n = 100$ and $\alpha = 5$, $n = 500$; clearly neither $\hat{\mu}_W$, $\hat{\mu}_{H=2/3}$ nor μ_T result in accurate attainment rates, (as is the case with all sample sizes and values of α.)

Figure 2 illustrates that excellent attainment rates and power are obtained when the test of Wilson and Einbeck (2016, 2018) with the estimator $\mu_{H=0.55}$ is used to estimate the mean parameter. The attainment rates and

FIGURE 1. Observed Attainment



powers of the likelihood ratio test are also plotted for comparative purposes, as is apparent these are nearly identical with those of the proposed test.

3 Non-fixed size parameter

With the exception of geometric models, it is extremely rare in practice to fit NB models with fixed size parameters. It is encouraging that letting $h = 0.55$ in Equation (1) results in a test with excellent power and attainment, it may be shown that if the value of α is estimated from the observed data using maximum likelihood methods, the attainment of the resulting test is poor. This would however appear to be due not to problems with the proposed test as such, but to issues concerning the estimation of the size parameter of negative binomial distributions. Figure 3 shows the rates and attainments achieved when "fixed α" $NB(2, \alpha)$ models, with $\alpha = 1.6$, 1.8, 2.0, 2.2 and 2.4 are fitted to 100 data drawn from $NB(2.0, 2.0)$ data, clearly the proposed test is sensitive to imprecise estimation of the size parameter.

Figure 4 illustrates the observed distribution of the maximum likelihood estimates of the size parameters obtained when samples of 100 data are drawn from $NB(2, 2)$ data. Whilst it would appear that the estimates from the whole samples are unbiased, they are clearly imprecise, whereas those from the truncated samples are biased, and are extremely imprecise, this begs the question as to the suitability of maximum likelihood estimation of the size parameter of a negative binomial distribution. Various authors have investigated the estimation of the parameters of negative binomial distributions, the reader is referred to Section 5.8 of Johnson, Kotz and Kemp (1992) for a discussion of these.

FIGURE 2. Left: observed attainment under various values of h; right: observed power for $\hat{\mu}_H = 0.55\mu_W + 0.45\mu_T$. ($\omega$ = zero–modification parameter)

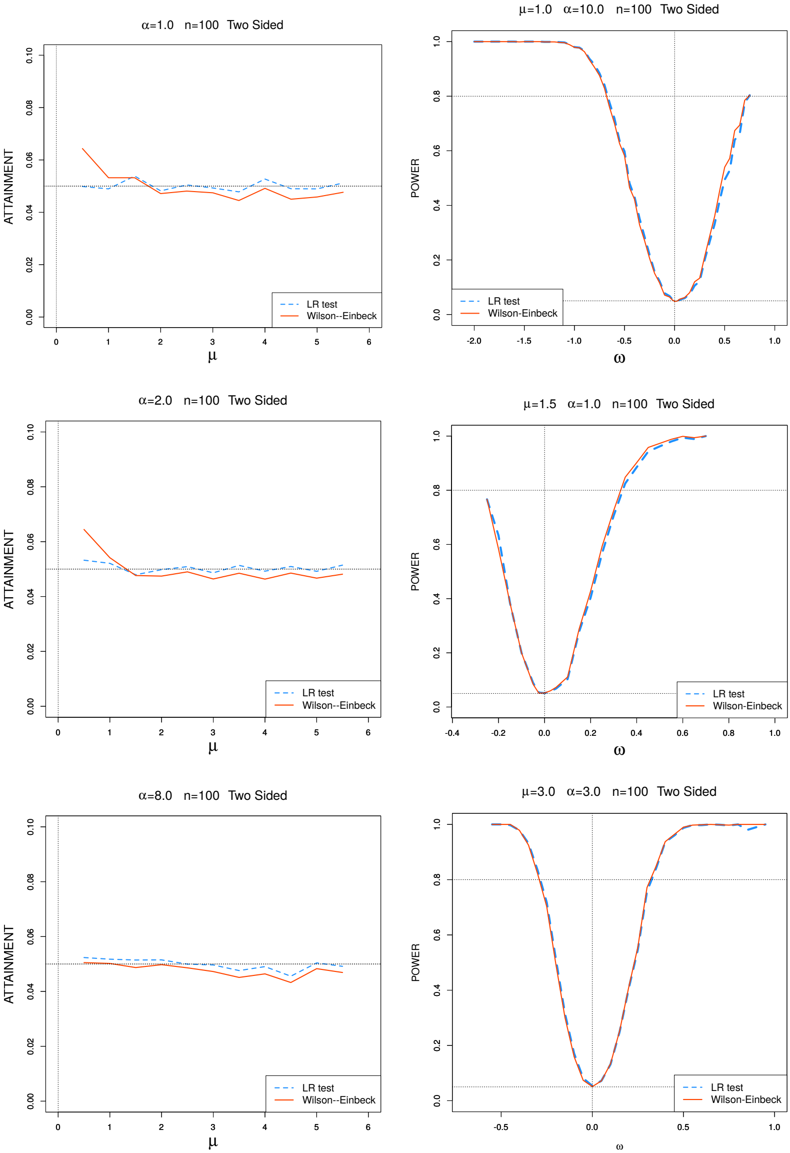


FIGURE 3. Observed attainment under correct ($\alpha = 2.0$) and "wrong" size parameter estimates.

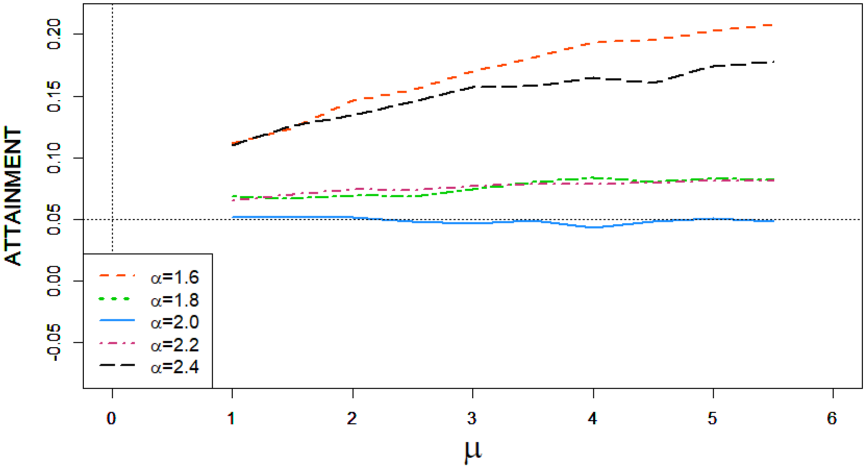
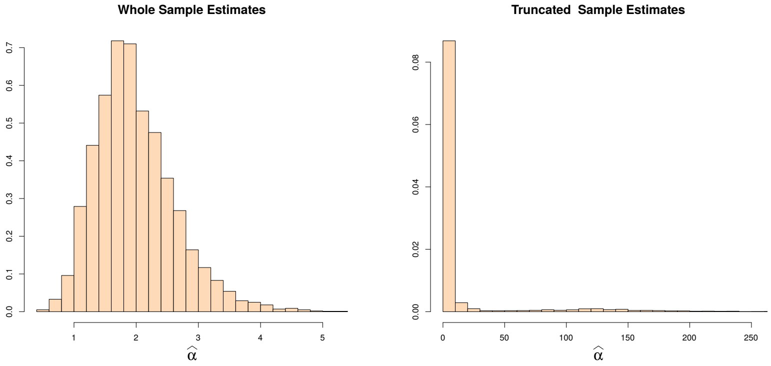


FIGURE 4. MLE of α obtained from whole and truncated samples from a $NB(2,2)$ distribution.



4 Conclusion

It is clear that the test proposed by Wilson and Einbeck (2016, 2018) has excellent attainment rate and power when used as a test of zero–modification relative to a negative binomial (type II) distribution with fixed size parameter (α) if a hybrid estimator $0.55\hat{\mu}_W + 0.45\hat{\mu}_T$ is employed. The situation for non-fixed α is not clear, and would appear to be dependent on the development of methods of obtaining precise estimates of the size parameter.

References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Johnson, N.L., Kotz, S. and Kemp A.W (1997). Univariate Discrete Distributions, (2nd Ed.), Wiley, New York, 1992.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero–modification in count data regression. In: Dupuy, J. and Josse, J. (Eds). Proc's of the 31st IWSM, Rennes, France, Vol 1, pages 325 – 330.

Wilson, P. and Einbeck, J. (2018) A new and intuitive test for zero–modification. *Statistical Modelling* Available online:

http://journals.sagepub.com/doi/abs/10.1177/1471082X18762277

Weight choice for penalized composite quantile regression and for model averaging

Jing Zhou[1], Gerda Claeskens[1], Daumantas Bloznelis[2]

[1] ORStat and Leuven Statistics Research Center, KU Leuven, Belgium
[2] Inland Norway University of Applied Sciences, Elverum, Norway

E-mail for correspondence: `jing.zhou@kuleuven.be`

Abstract: For composite estimation, weights are needed to combine the different loss functions; likewise, for model averaging, weights are used to average the different estimators. We investigate the choice of weights for both types of estimators in the setting of quantile regression. We pay particular attention to the high-dimensional case where due to the regularization, different expressions may be used to define so-called optimal weights.

Keywords: Quantile regression; Penalized estimation; Composite estimation; Weight choice.

1 Definitions of composite and model averaged quantile estimators

Let us consider a linear model $Y = X\beta + \epsilon$ where the response $Y \in \mathbb{R}^{n\times 1}$, the design matrix $X \in \mathbb{R}^{n\times p}$ and the coefficient vector $\beta \in \mathbb{R}^{p\times 1}$. The error $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ with ϵ_j's independent and identically distributed; and F and f denote respectively the cumulative distribution and the density function of ϵ_j.

For a single quantile $\tau \in (0, 1)$, Koenker (1984) defined the estimator of the $100\tau\%$ quantile of the response Y as

$$(\hat{b}_\tau, \hat{\beta}_\tau) = \arg\min_{b_\tau, \beta} \sum_{i=1}^{n} \rho_\tau(Y_i - b_\tau - X_i^\top \beta),$$

where $\rho_\tau(z) = \tau I\{z \geq 0\}z + (\tau - 1)I\{z < 0\}z$.

Estimating multiple quantile regressions with $1 < \tau_1 < \ldots < \tau_K < 1$, while allowing different weights $\nu = (\nu_1, \ldots, \nu_K)$, leads to the composite

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

estimator

$$(\hat{b}_{\tau_1,\mathrm{c}},\ldots,\hat{b}_{\tau_K,\mathrm{c}},\hat{\beta}_{\mathrm{c}}^\top(\nu)) = \arg\min_{b_{\tau_1},\ldots,b_{\tau_K},\beta} \sum_{k=1}^{K} \nu_k \sum_{i=1}^{n} \rho_{\tau_k}(Y_i - b_{\tau_k} - X_i^\top \beta).$$

To construct the model averaged estimator, we calculate the weighted average of K quantile estimates $\hat{\beta}_{\tau_k}$ from (1), which results in the model averaged estimator

$$\hat{\beta}_{\mathrm{mod.avg}}(\omega) = \sum_{k=1}^{K} \omega_k \hat{\beta}_{\tau_k},$$

where $\omega = (\omega_1,\ldots,\omega_K)^\top$ is the weight vector, and we impose the assumption that $\sum_{k=1}^{K}\omega_k = 1$ and $\omega_k \geq 0$. In low dimensions, the limiting normal distribution of the estimators, see Koenker and Bassett (1978), can be used to define weights that minimize the asymptotic variance. It is found that both estimators have the same lower bound for the variance.

2 High dimensional composite and model averaged quantile estimators

We consider the linear model $Y = X\beta + \epsilon$ from the previous section. Additionally, we allow p, the number of columns of X, to be an exponential order of the sample size n such that $\log(p) = O(n^\delta)$ with $\delta \in (0,1)$, and the sparsity $s = O(n^{\alpha_0})$, $\alpha_0 \in (0,1)$. In this sparse high-dimensional setting, Bradic et al. (2011) considered a penalized composite quantile estimator

$$\begin{aligned}&(\hat{b}_{\tau_1,\mathrm{c,pen}},\ldots,\hat{b}_{\tau_K,\mathrm{c,pen}},\hat{\beta}_{\mathrm{c,pen}}^\top(\nu))\\ &= \arg\min_{b_{\tau_1},\ldots,b_{\tau_K},\beta}\{\sum_{k=1}^{K}\nu_k\sum_{i=1}^{n}\rho_{\tau_k}(Y_i - b_{\tau_k} - X_i^\top\beta) + n\sum_{j=1}^{p}\gamma_\lambda(|\beta_j^{(0)}|)|\beta_j|\},\end{aligned}$$

where $|\beta_j^{(0)}|$ is some initial slope estimator (e.g. the Lasso, penalized quantile estimation, etc.), and $\gamma_\lambda(\cdot)$ is the derivative of the penalty function. Here, we consider only the SCAD penalty (Fan and Li, 2001) which leads to $\gamma_\lambda(u) = \lambda[I\{u \leq \lambda\} + \max(a\lambda - x, 0)I\{x > \lambda\}/\{(a-1)\lambda\}]$.
We reorganize the design matrix as $\tilde{X} = (X_a, X_b)$ and the coefficient $\beta^\top = (\beta_a^\top, \beta_b^\top)$, where X_a consists of the columns of X having non-zero coefficients β_a, and X_b to those columns having zero coefficients β_b. Bradic et al. (2011) showed the asymptotic normality for the active set β_a with the lower bound of the variance sharing the same expression with classical one $(\tilde{f}^\top A^{-1}\tilde{f})^{-1}$, where $\tilde{f} = (f(b_{\tau_1}^*),\ldots,f(b_{\tau_K}^*))$ and $b_{\tau_k}^* = F^{-1}(\tau_k)$ the true $100\tau_k\%$ quantile of the error. The optimal weight $\nu_{\mathrm{opt}} = A^{-1}\tilde{f}$ is chosen to maximize the efficiency of the estimator $\hat{\beta}_a$.

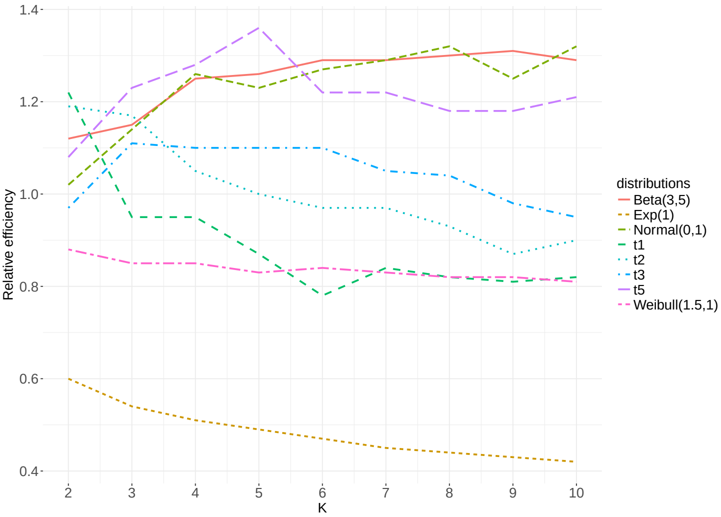


FIGURE 1. Simulated (over 1000 runs) relative efficiency of the model averaged estimator with the estimated non-negative optimal weights to the equally weighted model averaged estimator, defined as $\sum_{r=1}^{1000}\sum_{j=1}^{p}\{\hat\beta_j^r(\hat\omega_{\text{opt}}^+)-\beta_j^{\text{true}}\}^2/\sum_{r=1}^{1000}\sum_{j=1}^{p}\{\hat\beta_j^r(1_k/k)-\beta_j^{\text{true}}\}^2$.

The penalized model averaged estimator (see Bloznelis et al., 2017) is defined as

$$\hat\beta_{\text{mod.avg.pen}}(\omega)=\sum_{k=1}^{K}\omega_k\hat\beta_{\tau_k,\text{pen}}, \tag{1}$$

where

$$(\hat b_{\tau_k},\hat\beta_{\tau_k,\text{pen}})=\arg\min_{b_{\tau_k},\beta}\{\sum_{i=1}^{n}\rho_{\tau_k}(Y_i-b_{\tau_k}-X_i^\top\beta)+n\sum_{j=1}^{p}\gamma_{\lambda_k}(|\beta_j^{(0)}|)|\beta_j|\}.$$

Mimicking the same assumptions and arguments in Bradic et al. (2011), we obtain the asymptotic normality of $\hat\beta_{\text{mod.avg.pen}}(\omega)$ considering the coefficient vector corresponding to the true active set. This leads to finding a set of optimal weights that minimizes the asymptotic variance of this estimator.

3 The effect of the weight choice

Several issues lead to non-optimality of the so-called optimal weights. First, there is the issue of estimating the unknown error distribution, second, the very fact of estimating the variance to be minimized as a function of the weights, leads to weights that are different, and hence no longer optimal, as compared to when the true variance expression would be minimized. To avoid any estimation, equal weights form a simpler option. For the model averaged estimator, Figure 1 compares the simulated relative efficiency of estimators using equal weights with that of using estimated optimal weights. We observe that using the latter weights leads to higher efficiencies mostly for skewed or heavy tail distributions. For Weibull$(1.5, 1)$ and Exp(1), using non-negative optimal weights helps to improve the efficiency. For the t distribution it depends on the degrees of freedom and the number of quantiles which weight choice is preferred.

4 Compare optimal weights of the model averaged estimator

In the previous section, we worked with the expression of the limiting distribution of the estimator for the active set of coefficients. This construction ignores the selection effect of the regularization. Indeed, in practice one might not arrive at the estimated true set of active coefficients, but some truly active coefficients might have been set to zero, while other truly zero coefficients might have been estimated by a non-zero value.

An alternative approach, while taking selection into account, is to consider the robust approximate message passing algorithm (RAMP) in Bradic (2016) to construct penalized quantile estimators at K quantiles. To construct the model averaged estimator from RAMP, we inherit the same assumption in Koenker (2005) assuming the single quantile regression estimator of the slope β_τ is consistent for β_0 which is the true slope vector in the linear model $Y = X\beta + \epsilon$; the consistency holds for any fixed $\tau \in (0, 1)$. Hence we expect that for any single quantile $\tau_k, k = 1, \ldots, K$, Lemma 1 and Theorem 1 in Bradic (2016) holds; and the empirical distributions of the slopes for each quantile τ_k, as well as the achieved model averaged estimator defined in Eq.(1), converge weakly to the common probability measure f_{B_0}.

Consider the empirical mean squared error $\text{MSE} = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j^t - \beta_{0,j})^2$ at iteration t, as defined in Bradic (2016), the mean squared error of the model averaged estimator follows

$$\begin{aligned}\text{MSE}(\beta^t_{\text{mod.avg.pen}}, \beta_0) &= \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}^t_{\text{mod.avg.pen,j}} - \beta_{0,j})^2 \\ &= \frac{1}{p}\sum_{j=1}^{p}(\sum_{k=1}^{K} w_k \hat{\beta}^t_{\tau_k,j} - \beta_{0,j})^2 = w^\top \hat{\Sigma}_0 w\end{aligned}$$

where $\hat{\Sigma}_0$ is a $K \times K$ matrix with $(\hat{\Sigma}_0)_{k_1,k_2} = (\hat{\beta}^t_{\tau_{k_1}} - \beta_0)(\hat{\beta}^t_{\tau_{k_2}} - \beta_0)$ and $k_1, k_2 = 1, \ldots, K$. Following section If.1 in Rao (1973), the lower bound of the MSE of the model averaged estimator is $(\mathbf{1}_K^\top \hat{\Sigma}_0^{-1}\mathbf{1}_K)^-$ where $\mathbf{1}_K$ denotes a column vector with length K and $(\cdot)^-$ denotes a generalized inverse of a matrix. Furthermore, the lower bound is attained at $w_{\text{opt}} = \hat{\Sigma}_0^{-1}\mathbf{1}_K(\mathbf{1}_K^\top \hat{\Sigma}_0^{-1}\mathbf{1}_K)^-$; while restricting the weights to be nonnegative, we obtain an alternative optimal weight

$$w_{\text{opt}^+} = \underset{0\le w\le 1, \mathbf{1}_K w=1}{\operatorname{argmin}} \; w^\top \hat{\Sigma}_0 w. \tag{2}$$

Notice that the expression of the core covariance-like matrix $\hat{\Sigma}_0$ contains the true value of the regression coefficient which is unknown in practise. To utilize the above-mentioned lower bound of the MSE and optimal weight, we derive a consistent estimator for the covariance-like matrix. The estimator requires only estimations from each iteration of the RAMP algorithm, as well as an empirical covariance estimation. Derivation of the consistent estimator relies on the limiting Gaussian distribution of the RAMP algorithm for the large system.

The derived estimator for the (k_1, k_2)'th component of the matrix $\hat{\Sigma}_0$ is as follows (See Zhou and Claeskens, 2018)

$$\begin{aligned}(\hat{\Sigma})_{k_1,k_2} \quad = \quad & -\bar{\zeta}_{\tau_{k_1},\tau_{k_2}} + \left\langle \eta(\tilde{\beta}^t_{\tau_{k_1}}; \theta_{\tau_{k_1},t}) - \tilde{\beta}_{\tau_{k_1},t}, \eta(\tilde{\beta}^t_{\tau_{k_2}}; \theta_{\tau_{k_2},t}) - \tilde{\beta}_{\tau_{k_2},t}\right\rangle + \\ & \bar{\zeta}_{\tau_{k_1},\tau_{k_2}}\left\langle \eta'(\tilde{\beta}^t_{\tau_{k_1}}; \theta_{\tau_{k_1},t})\right\rangle + \bar{\zeta}_{\tau_{k_1},\tau_{k_2}}\left\langle \eta'(\tilde{\beta}^t_{\tau_{k_2}}; \theta_{\tau_{k_2},t})\right\rangle,\end{aligned}$$

where the $\tilde{\beta}^t_{\tau_{k_i}}$'s are the pseudo-data sequences, which are the noisy coefficient estimations, from the RAMP algorithm; $\bar{\zeta}_{\tau_{k_1},\tau_{k_2}}$ is the covariance of the $\tilde{\beta}^t_{\tau_{k_1}}$ and $\tilde{\beta}^t_{\tau_{k_2}}$, which can be estimated empirically; $\eta(\cdot;\theta)$ is the soft-thresholding function with parameter θ.

We investigate the performance of the two versions of the optimal weights, without and with selection uncertainty, by a simulation study. We choose the sample size $n = 320$, dimension $p = 500$, and the quantile components at $\tau_k = 0.3, 0.5, 0.7$. The distribution of the true coefficient is $P(\beta_0 = 1) = P(\beta_0 = 1) = 0.064$ and $P(\beta_0 = 0) = 0.872$. The design matrix $X \sim \mathcal{N}(0, 1/n)$, which is an assumption of the RAMP algorithm. Figure 2 compares the simulated MSE of the model averaged estimators using the optimal weights in Bloznelis et al. (2017) over that using the optimal weight in Eq.(2). We see that the model averaged estimator using the optimal weight in Eq.(2) outperforms that using the optimal weight in Bloznelis et al. (2017) consistently for all three error distributions.

References

Bloznelis, D., Claeskens, G. and Zhou, J. (2017). Composite versus model-averaged quantile regression. Technical report.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society, Series B*, **73(3)**, 325 – 349.

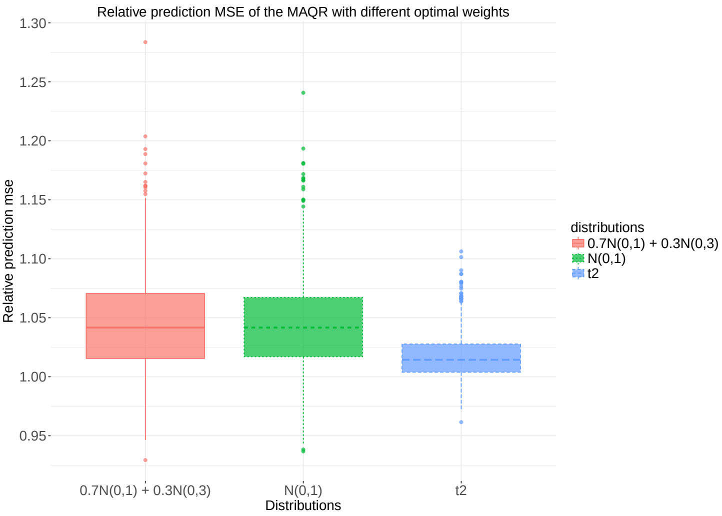


FIGURE 2. Relative prediction MSE of the model averaged estimator with two sets of optimal weights: $w_{\text{opt}+,1}$ as in Bloznelis et al. (2017) and $w_{\text{opt}+,2}$ from Eq.(2). Boxplot used 1000 simulations for each error distribution; the relative out-of-sample prediction MSE is defined as $\frac{1}{n}\sum_{i=1}^{n}\{Y_i - X_i^\top \hat\beta_{\text{mod.avg.pen}}(w_{\text{opt}+,1})\}^2 / \frac{1}{n}\sum_{i=1}^{n}\{Y_i - X_i^\top \hat\beta_{\text{mod.avg.pen}}(w_{\text{opt}+,2})\}^2$. Relative MSE larger than 1 indicates that the model averaged estimator with the optimal weight $w_{\text{opt}+,2}$ has lower prediction MSE.

Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, **10(2)**, 3894 – 3944.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96(456)**, 1348–1360.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R., and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, 43 – 61.

Rao, R. C. (1973). *Linear statistical inference and its applications*. Wiley New York.

Zhou, J. and Claeskens, G. (2018). Optimal weights of the composite and model-averaged quantile regression estimators. Technical report.

Fractional Nonconformance Based Guardbanding

Xin Zhou[1], Kondaswamy Govindaraju[1], Geoff Jones[1]

[1] Institute of Fundamental Sciences, Massey University, Palmerston North, 4442, New Zealand

E-mail for correspondence: zhouxin07@gmail.com

Abstract: In this study, we developed a guardbanded acceptance control chart plan to monitor a short run process as well as dispose the products manufactured from the process simultaneously. The proposed approach was trialled for a whole milk powder production process. Our analysis shows that the optimum guardband widths obtained under risk and cost models are consistent and comparable.

Keywords: Fractional nonconformance; Short run production; Measurement error; Acceptance control chart; Guardbanding.

1 Introduction

The Phase I and II approach of the traditional control chart does not fit for short run production environments because the process parameters must be estimated *dynamically* to monitor the in-control state of the process as well as keep the fraction nonconforming low.

Bulk product quality characteristics, such as the percentage protein or fat, are determined using analytical methods which involve considerable measurement uncertainty. *Guardbanding* is an offset technique used to compensate for measurement and sampling uncertainty and thereby reducing the risk of nonconformance to specifications. Chou and Chen (2005) developed a kernel density estimator after deconvolution of the density to adjust for the measurement error and then determined the optimal guardband limit that controls both producer's and consumer's risks. Williams and Hawkins (1993) proposed a cost model taking into account measurement error and then adjusting the guardband width.

Acceptance control chart is a hybrid approach to both control charting and limiting the proportion nonconforming of products exceeding specifi-

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

cations. An introduction to acceptance control charts and further discussion are in Duncan (1986) and Wu (1998). The traditional acceptance control methodology requires an indifference zone to be set, and it does not allow for measurement errors. As a result, these traditional acceptance control chart procedures cannot be used for bulk products.
Much of the current literature on acceptance control charting and guardbanding assumes that the process characteristic is normal distributed. No unified method is available in the literature to deal with short run process monitoring, acceptance control chart and guardbanding. Govindaraju and Jones (2015) proposed a probabilistic measure for quantifying the nonconformance after adjusting for measurement uncertainty when the underlying measurement error distribution is known. This fractional nonconformance (FNC) statistic was initially applied for acceptance sampling inspection, and was further implemented for short run process monitoring by Zhou et al. (2017). Our current research is focussed on the use of fractional nonconformance principles for guardbanded acceptance control charting under both normal and beta processes involving measurement errors.

2 Guardbanded acceptance control chart plan

We develop a guardbanded acceptance control chart plan to monitor the short run process as well as dispose the products manufactured from the process simultaneously. Under the acceptance control chart plan, the production run is accepted only if the process is under control and the overall nonconformance level is low. Zhou et al. (2017) found that average FNC ($\hat{p_{A_j}}$) is more sensitive than individual FNC ($\hat{p_{I_j}}$) in detection of a sudden shift in the process. Let $X \sim N(\mu_X, \sigma_X^2)$, $Z \sim N(0, \sigma_Z^2)$ and $Y \sim N(\mu_X, \sigma_X^2 + \sigma_Z^2)$, assuming that there is no instrument bias, where X, Y, Z are the true measurement, the apparent measurement and the measurement error respectively. For a random sample with apparent sample measurements $(y_i, y_2, ..., y_n)$, $\hat{p_{I_j}}$ and $\hat{p_{A_j}}$ can be calculated using Equations 1 and 2 when σ_z is known and then used to evaluate the nonconformance level and monitor the process respectively.

$$\hat{p_{I_j}} = P(x_j > \text{USL}) = P(z < y_j - \text{USL}) = \Phi\left(\frac{y_j - \text{USL}}{\sigma_z}\right) \tag{1}$$

$$\hat{p}_{A_j} = \frac{1}{t}\sum_{j=1}^{t} \Phi\left(\frac{y_j - \text{USL}}{\sigma_z}\right) \tag{2}$$

The control chart rule for process monitor as well as the rule for product acceptance must be met at the same time for greater assurance to the consumer. Therefore, the probability of acceptance of a batch with size n under the acceptance control chart plan is defined as below:

$$P_a = \Pr\left(\{\sum_{j=1}^{n} \hat{p_{I_j}} \leq nAc\} \cap \{\hat{p_{A_j}} < \mathrm{UCL}_A\}\right) \tag{3}$$

where Ac and UCL_A are the fractional acceptance number and the upper control limit of average fractional nonconformance control chart respectively. For $j = 1, 2, ..., n$, the event $\{\hat{p_{A_j}} < \mathrm{UCL}_A\}$ ensures that the short-run production process is in-control while the event $\{\sum_{j=1}^{n} \hat{p_{I_j}} \leq nAc\}$ ensures that the lot or product quality does not exceed the set nonconformance requirement.

The overall Type I error of the proposed acceptance control chart plan (α_a) can be split into Type I errors of acceptance sampling plan (α_l) and process control plan (α_p), which are determined by the selection of Ac and UCL_A respectively. The guardband coefficient g, which defined as the ratio of guardband limit (UGL) and specification limit (USL), controls both α_l and α_p. When g decreases from 1, a tightened limit UGL is adopted instead of USL, and as a result, both $\hat{p_{I_j}}$ and $\hat{p_{A_j}}$ will increase. In other words, a small g will increase the probability of Type I error for both lot acceptance and process control. Hence, for a given α_a, several combinations of α_l and α_p are possible; so is the case with the combinations of g, Ac and UCL_A. This research is focussed on finding the optimum guardbanded acceptance control chart parameters controlling both producer's and consumer's risks.

2.1 Risk model

In the *risk model*, the producer's risk is fixed at a certain level, say $\alpha_a = 5\%$, and the optimum guardband is the one which results in the minimum consumer's risk. Under both normal process model $Y \sim N(0.25, 0.01)$ and beta process model $Y \sim \mathrm{Beta}(500, 1500)$, we found that the optimum guardband selection does not rely on the process distribution too much when the process mean and SD are matched, see Figure 1. In other words, the optimum g obtained under normal model is robust for independent processes assuming process variation σ_Y^2 is known. Although this assumption is validated on empirical grounds for compositional material production processes, it may not be legitimate for discrete item production processes. Hence, we further investigated the properties of guardbanded acceptance control chart plan for unknown dispersion parameters. Even though we expected a tighter guardband width to compensate for the higher sampling-related uncertainty, we found that the optimum guardband width for SD-unknown process is not as stringent as the SD-known process due to the producer's risk constraint, as shown in Figure 2.

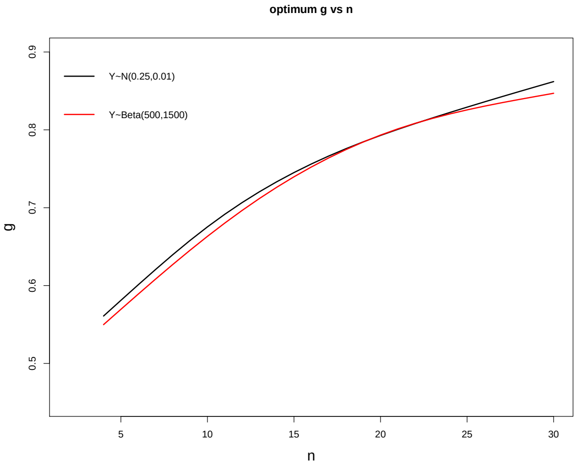


FIGURE 1. Optimum guardband $Y \sim N(0.25, 0.01)$ vs $Y \sim \text{Beta}(500, 1500)$

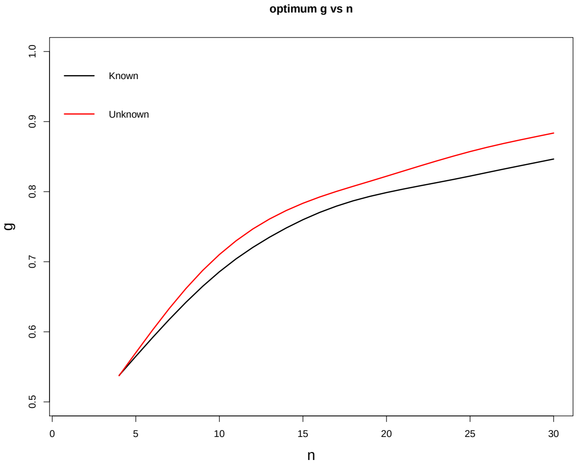


FIGURE 2. Optimum guardband SD-known vs SD-unknown

2.2 Cost model

A simple *cost model* was also studied to determine the optimum guardband. We considered different cost ratios ($c = 10, 50, 100$ and 500) of Type II and Type I errors and found that the optimum guardband is primarily a function of the cost ratio. The higher the cost ratio is, the tighter the guardband width should be. The optimum guardband becomes constant for high cost ratios particularly when the production length is 25 or more, see Figure 3. The consumer's risk was found to be high when the production length is small. The guardbanding approach can reduce the consumer's risk significantly for small production lengths. When the production length reaches a certain limit such as 25, the consumer's risk tends to be stable. Hence, the guardbanding approach is preferable for consumer protection when the production length is shorter, say below 50.

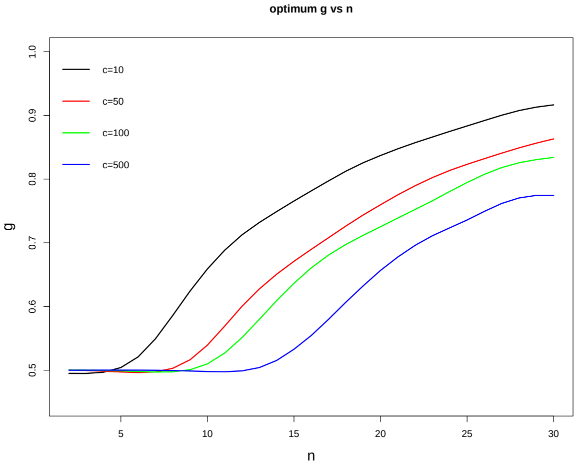


FIGURE 3. Optimum guardband for different cost ratios

3 Data analysis

The proposed guardbanded acceptance control chart plan was motivated by dairy industry problems. The newly developed guardbanded acceptance control chart plan was trialled using the whole milk powder production data. We examined 24 batches of data with length between 4 and 20 samples. Quality characteristics of the whole milk powder, including moisture, protein, fat and P:SNF were analyzed. We found that the guardbanded approach also works well in very short process. In other words, the guardbanded acceptance control chart plan can be applied to both grand and sublots formed. Our analysis also showed that the optimum guardband widths obtained in risk and cost models are consistent and comparable. The implementation of the proposed guardbanded acceptance control chart plan was done dynamically using a Shiny *app* hosted at https://zhouxin07.shinyapps.io/guardbanding/.

Acknowledgments: This research is supported by the Primary Growth Partnership (PGP) scheme sponsored by Fonterra Cooperative Group and the New Zealand Government. Thanks to Roger Kissling of Fonterra for his guidance and suggestions.

References

Chou, Y. M., and Chen, K. S. (2005). Determination of Optimal Measurement Guardbands. *Quality Technology & Quantitative Management*, **2(1)**, 65 – 75.

Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 5th edn. Richard D Irwin Inc, Homewood, IL.

Govindaraju, K. and Jones, G. (2015). Fractional acceptance numbers for lot quality assurance. In *Frontiers in Statistical Quality Control*, **11**, 271 – 286. Springer.

Williams, R. H. and Hawkins, C. F. (1993). The economics of guardband placement. In *24th IEEE International Test Conference*, 218 - 225.

Wu, Z. (1998). An adaptive acceptance control chart for tool wear. *International Journal of Production Research*, **36(6)**, 1571 – 1586.

Zhou, X., Govindaraju, K., and Jones, G. (2017). Monitoring fractional nonconformance for short-run production. *Quality Engineering*, Published online: 21 Sep 2017. https://doi.org/10.1080/08982112.2017.1360499.

Author Index