SORT 36 (2) July-December 2012, 115-138

# On developing ridge regression parameters: a graphical investigation

Gisela Muniz[1], B. M. Golam Kibria[2], Kristofer Mansson[3] and Ghazi Shukur[3,4]

## Abstract

In this paper we review some existing and propose some new estimators for estimating the ridge parameter. All in all 19 different estimators have been studied. The investigation has been carried out using Monte Carlo simulations. A large number of different models have been investigated where the variance of the random error, the number of variables included in the model, the correlations among the explanatory variables, the sample size and the unknown coefficient vector were varied. For each model we have performed 2000 replications and presented the results both in term of figures and tables. Based on the simulation study, we found that increasing the number of correlated variable, the variance of the random error and increasing the correlation between the independent variables have negative effect on the mean squared error. When the sample size increases the mean squared error decreases even when the correlation between the independent variables and the variance of the random error are large. In all situations, the proposed estimators have smaller mean squared error than the ordinary least squares and other existing estimators.

*MSC:* Primary 62J07, Secondary 62F10.

*Keywords:* Linear model, LSE, MSE, Monte Carlo simulations, multicollinearity, ridge regression.

## 1. Introduction

In most empirical studies practitioners are often concerned about the specification of the models under consideration, especially with regards to problems associated with the residuals, with the aim of assessing white noise errors to judge whether the model is well specified. Model misspecification can be due to omission of one or several relevant

[1,2] Department of Mathematics and Statistics, Florida International University, Miami, Florida, USA.

[3] Department of Economics, Finance and Statistics, Jonkoping University, Sweden.

[4] Department of Economics and Statistics, Centre for Labour Market and Discrimination Studies, Linnaeus University, Sweden.

Received: May 2011

Accepted: March 2012

variables, inclusion of unnecessary variables, wrong functional form, misspecified dynamics, autocorrelation, heteroscedasicity, etc. It is recommended that some diagnostic tests should be conducted in order to assure the whiteness of the model under consideration, otherwise the estimated results can be inefficient, biased or inconsistent.

However, there are other problems that also might influence the results, e.g. multicollinearity. This problem occurs in situations when the explanatory variables are highly inter-correlated. Empirically the problem of multicollinearity can be observed, for example in cement production, when the amount of different compounds in the clinkers is regressed on the heat evolved of cement (see the classical Portland cement data used in Muniz and Kibria, 2009). Another examples in economics could be the effect of different amenities that attract visitors and tourists on the gross regional product. A final example is when the usage of different types of vehicles on the number of injured people in traffic is analyzed. In these situations it becomes difficult to disentangle the separate effects of each of the explanatory variables on the response variable. As a result, estimated parameters can be wrongly insignificant or have (unexpectedly) wrong signs. Note that multicollinearity is more a problem with the data than with the model itself, and hence this kind of problem can not be identified by residual analysis. As a result, a common deficiency in many applied studies is the absence of paying serious attention to this problem. Indeed, although model misspecification is an important area in statistical modelling, multicollinearity is an important issue too.

The history of multicollinearity dates back at least to the paper by Frisch (1934) who introduced the concept to denote a situation where the variables dealt with are subject to two or more relations. One way to deal with this problem is called ridge regression, first introduced by Hoerl and Kennard (1970a,b). At this stage, the main interest lies in finding a value of the ridge parameter, say $k$, such that the reduction in the variance term of the slope parameter is greater than the increase in its squared bias. The authors proved that there is a nonzero value of such ridge parameter for which the mean squared error (MSE) for the slope parameter using ridge regression is smaller than the variance of the ordinary least squares (OLS) estimator of the respective parameter. Many authors thereafter worked in this area of research and developed and proposed different estimates for the ridge regression parameter. To mention a few, McDonald and Galarneau (1975), Lawless and Wang (1976), Saleh and Kibria (1996), Haq and Kibria (1996), Kibria (2003), Khalaf and Shukur (2005) and Alkhamisi, Khalaf and Shukur (2006). In Kibria (2003) and Alkhamisi *et al.* (2006), the authors used simulation techniques to study the properties of some new proposed estimators and compared their properties with some popular existing estimators. Under certain conditions, they found that the MSEs of some of the new proposed estimators are smaller than the corresponding MSE of the OLS estimator and other existing estimators. Recently, Muniz and Kibria (2009) developed five new ridge parameters based on Kibria (2003) and Khalaf and Shukur (2005) in models with two explanatory variables. They found the new parameters outperform the previous ones in term of smaller MSEs.

In this paper we aim to extend the study by Muniz and Kibria (2009) by developing nine more new ridge parameters and to extend the models by including more explanatory variables. We also study models with four explanatory variables that are more realistic in empirical work than models with only two variables. Proceeding in this manner, it is possible to investigate the effect of the additional included variables on the MSEs.

The paper is organized as follows: In Section 2 we present the model we analyze, and give the formal definition of the ridge regression parameters used in this study. In Section 3, the design of our Monte Carlo experiment together with the factors that can affect the small sample properties of these proposed parameters are introduced. In Section 4 we describe the results concerning the various parameters in term of MSE. The conclusions of the paper are presented in Section 5.

## 2. Methodology

In this section we present the proposed ridge regression estimators. This includes a brief background on the methods suggested by Hoerl and Kennard (1970a), and that developed by Khalaf and Shukur (2005), Alkhamisi and Shukur (2008), Alkhamisi *et al.* (2006) and Muniz and Kibria (2009). Moreover, we present the new ridge parameter, (denoted by $K_{AS}$), together with the other five new versions.

### 2.1. Notations and some preliminaries

The multiple linear regression model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ observed matrix of the regressors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors.

The ordinary least squares estimator (OLS) of the regression coefficients $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}, \tag{2.2}$$

Suppose, there exists an orthogonal matrix $\boldsymbol{D}$ such that $\boldsymbol{D}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{D} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ are the eigenvalues of the matrix $\boldsymbol{C} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. The orthogonal (canonical form) version of the multiple regression model (2.1) is

$$\boldsymbol{Y} = \boldsymbol{X}^{*}\boldsymbol{\alpha} + \boldsymbol{e}$$

where $\boldsymbol{X}^* = \boldsymbol{X}\,\boldsymbol{D}$ and $\boldsymbol{\alpha} = \boldsymbol{D}^{\mathsf{T}}\,\boldsymbol{\beta}$. In case the matrix $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is ill-conditioned however (in the sense of there is a near-linear dependency among the columns of the matrix) the OLS estimator of $\boldsymbol{\beta}$ has a large variance, and multicollinearity is said to be present. Ridge regression replaces $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ with $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + k\boldsymbol{I}$, $(k > 0)$. Then the generalized ridge regression estimator of $\boldsymbol{\alpha}$ are given as follows:

$$\hat{\boldsymbol{\alpha}}(k) = \left(\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^* + \boldsymbol{K}\right)^{-1}\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{Y} \tag{2.3}$$

where $\boldsymbol{K} = \operatorname{diag}(k_1, k_2, \ldots, k_p)$, $k_i > 0$ and $\hat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}^{-1}\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{Y}$ is the OLS estimator of $\boldsymbol{\alpha}$.

According to Hoerl and Kennard (1970a) the value of $k_i$ which minimizes the $\operatorname{MSE}(\hat{\boldsymbol{\alpha}}(\boldsymbol{K}))$ is

$$k_i = \frac{\sigma^2}{\alpha_i^2}, \tag{2.4}$$

where $\sigma^2$ represents the error variance of the multiple regression model, and $\alpha_i$ is the $i^{th}$ element of $\alpha$.

## 2.2. Proposed estimators

In this section, we review some already available ridge estimators and propose some new ones.

### 2.2.1. Estimators based on Hoerl and Kennard (1970)

Hocking, Speed and Lynn (1976) showed that for known optimal $k_i$, the generalized ridge regression estimator is superior to all other estimators within the class of biased estimators they considered. Nevertheless, the optimal value of $k_i$ fully depends on the unknown $\sigma^2$ and $\alpha_i$, which must be estimated from the observed data. Hoerl and Kennard (1970) suggested to replace $\sigma^2$ and $\alpha_i^2$ by their corresponding unbiased estimators in (2.4). That is,

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \tag{2.5}$$

where $\hat{\sigma}^2$ is the residual mean square estimate, which is an unbiased estimator of $\sigma^2$ and $\hat{\alpha}_i$ is the $i^{th}$ element of $\hat{\alpha}$, which is an unbiased estimator of $\alpha$.

Hoerl and Kennard (1970) suggested $k$ to be

$$k_{HK1} = \hat{k}_{HK} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\max}^2} \tag{2.6}$$

where $\hat{\alpha}_{\max}$ is the maximum element of $\hat{\alpha}$. Now, when $\sigma^2$ and $\alpha$ are known then $\hat{k}_{HK}$ will give smaller MSE than the OLS.

Hoerl *et al.* (1975), proposed a different estimator of $k$ by taking the harmonic mean of $\hat{k}_i$. That is

$$k_{HK2} = \hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \alpha_i^2} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\alpha}}^{\mathsf{T}}\hat{\boldsymbol{\alpha}}} \tag{2.7}$$

### *2.2.2. Estimators based on Kibria (2003)*

Kibria (2003) proposed some new estimators based on the *generalized ridge regression* approach. They are as follows:

By using the geometric mean of $\hat{k}_i$, which produces the following estimator

$$k_{K1} = \hat{k}_{GM} = \frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p} \hat{\alpha}_i^2\right)^{\frac{1}{p}}} \tag{2.8}$$

By using the median of $\hat{k}_i$, which produces the following estimator for $p \geq 3$

$$k_{K2} = \hat{k}_{MED} = \text{median}\left\{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right\}, \qquad i = 1,2,\ldots,p \tag{2.9}$$

### *2.2.3. Estimators based on Khalaf and Shukur (2005)*

Khalaf and Shukur (2005) suggested a new method to estimate the ridge parameter $k$, as a modification of $k_{HK1}$ as

$$k_{S1} = \hat{k}_{KS} = \frac{t_{\max}\hat{\sigma}^2}{(n-p)\,\hat{\sigma}^2 + t_{\max}\hat{\alpha}_{\max}^2} \tag{2.10}$$

where $t_{\max}$ is the maximum eigenvalue of the matrix $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$.

Following Kibria (2003) and Khalaf and Shukur (2005), Alkhamisi *et al.* (2006) proposed the following estimators for $k$:

$$k_{S2} = \hat{k}_{arith}^{KS} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{t_i\hat{\sigma}^2}{(n-p)\,\hat{\sigma}^2 + t_i\hat{\alpha}_i^2}\right), \qquad i = 1,2,\ldots,p \tag{2.11}$$

$$k_{S3} = \hat{k}_{\max}^{KS} = \max\left(\frac{t_i\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_i\hat{\alpha}_i^2}\right), \qquad i = 1,2,\ldots,p \tag{2.12}$$

$$k_{S4} = \hat{k}^{KS}_{md} = \text{median}\left(\frac{t_i\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_i\hat{\alpha}_i^2}\right). \tag{2.13}$$

### *2.2.4. Some proposed new estimators*

Following Kibria (2003), Khalaf and Shukur (2005), Alkhamisi *et al.* (2006) and Alkhamisi and Shukur (2008), we proposed the following estimators. First, following Kibria (2003) and Khalaf and Shukur (2005), we propose the following estimator

$$k_{KM1} = \hat{k}^{KS}_{gm} = \left(\prod_{i=1}^{p}\frac{t_i\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_i\hat{\alpha}_i^2}\right)^{\frac{1}{p}} \tag{2.14}$$

In Muniz and Kibria (2009) some ridge parameters that are functions of the optimal value shown in equation (2.4) was proposed. These functions of the optimal value used firstly different quantiles such as the median and max just as in Khalaf and Shukur (2005) and secondly the square root transformation that was shown to work well in Alkhamisi and Shukur (2008). These ridge parameters correspond to:

$$k_{KM2} = \max\left(\frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}}\right) \tag{2.15}$$

$$k_{KM3} = \max\left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}\right) \tag{2.16}$$

$$k_{KM4} = \left(\prod_{i=1}^{p}\frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}}\right)^{\frac{1}{p}} \tag{2.17}$$

$$k_{KM5} = \left(\prod_{i=1}^{p}\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}\right)^{\frac{1}{p}} \tag{2.18}$$

$$k_{KM6} = \text{median}\left(\frac{1}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}}\right) \tag{2.19}$$

$$k_{KM7} = \text{median}\left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}}\right) \tag{2.20}$$

Furthermore, we will propose some new estimators of the ridge parameter. These also use different quantiles and the square root transformation proposed in Khalaf and Shukur (2005) and Alkhamisi and Shukur (2008) respectively. However, the base of the different functions is no longer the optimal value but a modification proposed by Khalaf and Shukur (2005). This modification, which in general leads to larger values of the ridge parameters than those derived from the optimal value, was shown to work well in the simulation study conducted in that paper:

$$k_{KM8} = \max\left(\frac{1}{\sqrt{\frac{t_{\max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2+t_{\max}\hat{\alpha}_i^2}}}\right) \tag{2.21}$$

$$k_{KM9} = \max\left(\sqrt{\frac{t_{\max}\hat{\sigma}^2}{(n-p)\,\hat{\sigma}^2+t_{\max}\hat{\alpha}_i^2}}\right) \tag{2.22}$$

$$k_{KM10} = \left(\prod_{i=1}^{p}\frac{1}{\sqrt{\frac{t_{\max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2+t_{\max}\hat{\alpha}_i^2}}}\right)^{\frac{1}{p}} \tag{2.23}$$

$$k_{KM11} = \left(\prod_{i=1}^{p}\sqrt{\frac{t_{\max}\hat{\sigma}^2}{(n-p)\,\hat{\sigma}^2+t_{\max}\hat{\alpha}_i^2}}\right)^{\frac{1}{p}} \tag{2.24}$$

$$k_{KM12} = \text{median}\left(\frac{1}{\sqrt{\frac{t_{\max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2+t_{\max}\hat{\alpha}_i^2}}}\right) \tag{2.25}$$

Note that the new proposed estimators: $k_{S1}$ (in 2.10), $k_{S2}$ (in 2.11), $k_{KM3}$ (in 2.16), $k_{KM7}$ (in 2.20), $k_{KM8}$ (in 2.21), $k_{KM9}$ (in 2.22), $k_{KM10}$ (in 2.23), $k_{KM11}$ (in 2.24) and finally $k_{KM12}$ (in 2.25) were not investigated in Muniz and Kibria (2009).

## 3. The Monte Carlo design

The aim of this paper is to compare the performance of our new proposed estimators and the other estimators together with the OLS. Since a theoretical comparison is not possible, a simulation study has been conducted using S-plus 8.0 software. The design of a good simulation study is dependent on (i) what factors are expected to affect the properties of the estimators under investigation and (ii) what criteria are being used to judge the results. Since ridge estimators are supposed to have smaller MSE compared to OLS, the MSE will be used to measure the goodness of an estimator, while the first question will be treated briefly.

Since the degree of collinearity among the explanatory variable is of central importance, we followed Muniz and Kibria (2009) in generating the explanatory variable using the following device:

$$x_{ij} = \left(1-\gamma^2\right)^{1/2} z_{ij} + \gamma z_{ip}, \qquad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,p \tag{3.26}$$

where $\gamma^2$ represents the correlation between the explanatory variables, and $z_{ij}$ are independent standard pseudo-random numbers. The $n$ observations for the dependent variable are then determined by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \qquad i = 1,2,\ldots,n \tag{3.27}$$

where $e_i$ are i.i.d. N $\left(0,\sigma^2\right)$ pseudo-random numbers, and $\beta_0$ is taken to be zero without loss of generality.

### 3.1. Factors that vary in the Monte Carlo simulations

Since our primary interest lies in the performance of our proposed estimators according to the strength of the multcollinearity, we used different degrees of correlation between the variables and let $\gamma$ = 0.7, 0.8 and 0.9. We also want to see the effect of the sample sizes on the performance of the estimators. Therefore, in this study, we considered $n =$ 10, 20, 30, 40, 50 and 100 which will cover models with small, medium and large sample sizes. The number of the explanatory variables is also of great importance since the bad impact of the collinearity on the MSE might be stronger when more variables in the model are correlated. We hence generated models with $p = 2$ and 4 explanatory variables. To see whether the magnitude of the error variance has a significant effect of the performances of the proposed estimators, we used different values of the error standard deviations $\sigma =$ 0.01, 0.5, 1, 3, and 5. For each set of explanatory variables we considered the coefficient vector that corresponded to the largest eigenvalue of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ matrix subject to the constraint that $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} = 1$. Newhouse and Oman (1971) stated that if the mean squared error (MSE) is a function of $\boldsymbol{\beta}$, $\sigma^2$, and $k$, and if the explanatory variables are fixed, then the MSE is minimized when we choose this coefficient vector.

For given values of $n, p, \boldsymbol{\beta}, \gamma$, and $\sigma$, the set of explanatory variables was generated. Then the experiment was repeated 2000 times and the average mean squared error was calculated for all 15 estimators.

## 4. Results and discussion

In this section we present the results of our Monte Carlo experiment concerning the MSEs of the different proposed estimators compare to the OLS. A conventional way to report the results of a Monte Carlo experiment is to tabulate the values of these MSEs under different conditions. When determining the manner of presentation, some account has to be taken to the results obtained. Our original intention was to start by presenting results for all the main effects in term of tables. However, since the results are too extensive, presenting the results in term of tables will make it difficult to follow the main findings. We hence present our most important findings in the form of figures that summarize most of the results with respect to the different features under investigation. More exact results of the simulated MSEs for the 15 estimators are provided in the appendix (all results are not included in tables, however, but are available on request from the authors). Simulated MSEs for fixed $n$, $p$ and $\gamma$ and different values of $\sigma$ are presented in Table A.1, for fixed $n$, $p$ and $\sigma$ and different values of $\gamma$ are presented in Table A.2, for fixed $p$, $\gamma$ and $\sigma$ and different values of $n$ are presented in Table A.3.

### 4.1. Performance as a function of $\sigma$

In Table A.1 we have provided the MSEs of the estimators as a function of the variance of the errors($\sigma$). When the value of $\sigma$ increases, the MSE of the estimators also increases. For all values of $\sigma$, the ridge regression estimators have smaller MSE than the OLS. However, the performance of the proposed estimators $k_{KM4}$, $k_{KM5}$, $k_{KM8}$, $k_{KM10}$, $k_{KM12}$, and $k_{K1}$, $k_{K2}$ is better than the performance of the rest of the analysed estimators. This behaviour was almost constant for any sample size and number of variables considered. However, when the standard deviation is large, i.e. $(\sigma = 5)$, the new $k_{KM8}$, $k_{KM12}$ outperform all the other estimators in term of producing less MSE.

For given $\gamma = 0.70$ and $n = 10$, the performance of estimators as a function of the standard deviation of the errors for $p = 2$ and $p = 4$ are provided in Figures 1 and 2 respectively. From these figures we observe that as the standard deviation increases, the MSE also increases. The same is true when shifting from models with 2 variables to those with 4, especially for the OLS, $k_{HK,}$ $k_{S1,}$ $k_{S2,}$ (see Figure 2).

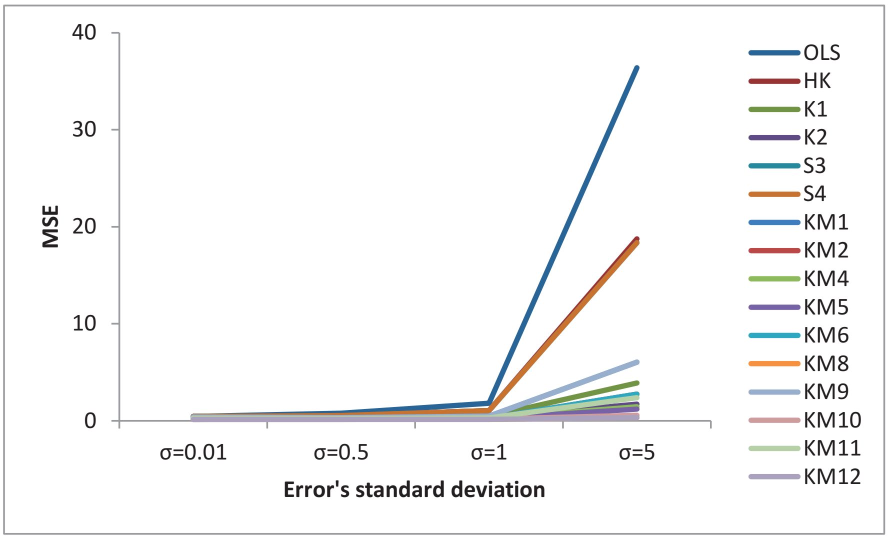

***Figure 1:*** *Performance of the estimators as a function of $\sigma$ when $p = 2$.*

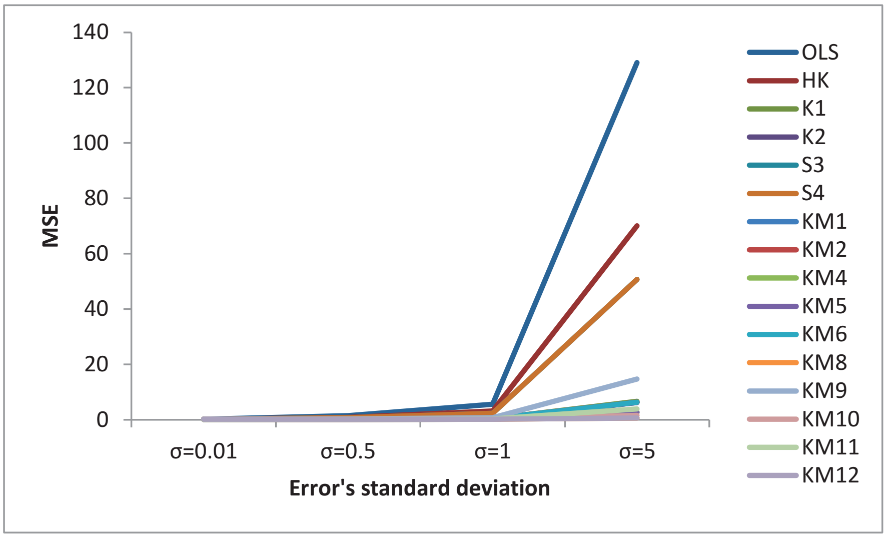

***Figure 2:*** *Performance of the estimators as a function of $\sigma$ when $p = 4$.*

## 4.2. Performance as a function of $\gamma$

In Table A.2 we have provided the MSEs of the estimators as a function of the correlation between the explanatory variables. For smaller sigma ($\sigma = 0.01$) the change in the correlation between the explanatory variables had almost no effect on the MSEs. In all situations they remained almost the same for any sample size or number of parameters,

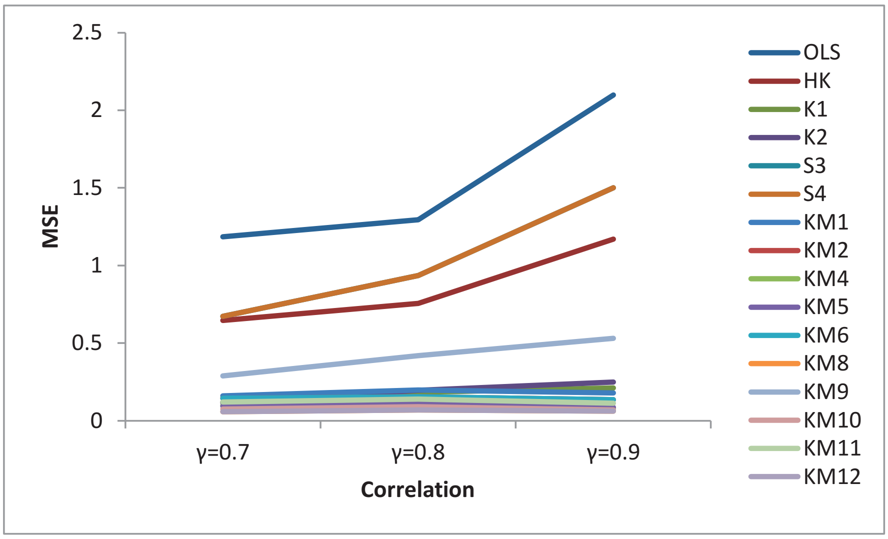

***Figure 3:*** *Performance of the estimators as a function of* $\gamma$ *when* $n = 20$.

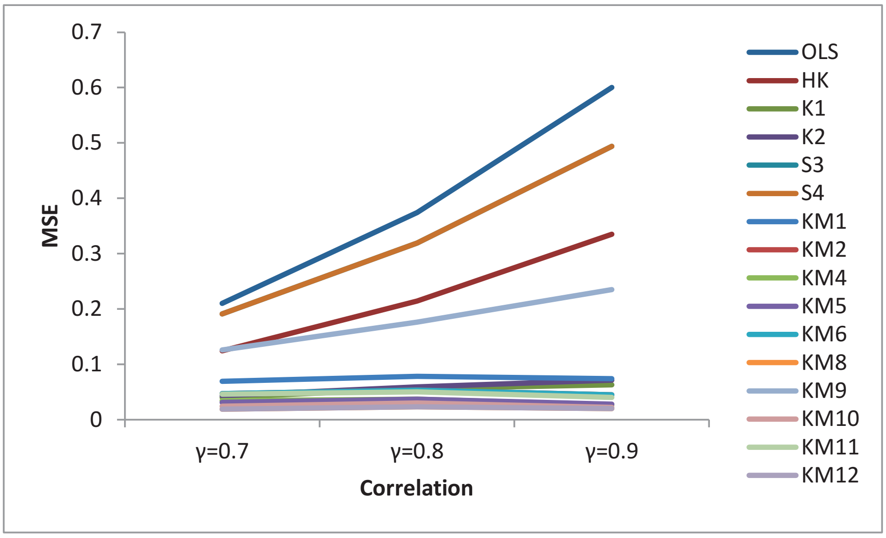

***Figure 4:*** *Performance of the estimators as a function of* $\gamma$ *when* $n = 50$.

and their MSEs are very small. When $\sigma$ increases, the higher correlation between the independent variables resulted in an increase of the MSE of the $k$-estimators. In general, $k_{KM4}$, $k_{KM5}$, $k_{KM8}$, $k_{KM10}$, $k_{KM12}$ and $k_{K1}$, $k_{K2}$ performed better than others.

For given $\sigma = 1$ and $p = 4$ the performance of estimators as a function of the correlation between the explanatory variables for $n = 20$ and $n = 50$ is provided in Figures 3 and 4 respectively. From these figures we observe that as correlation

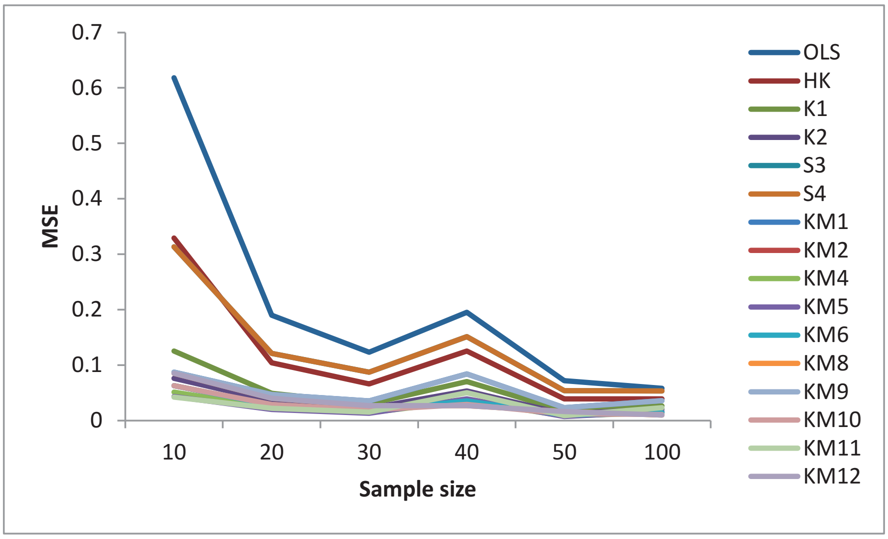

***Figure 5:*** *Performance of the estimators as a function of n when* $\sigma = 0.5$.

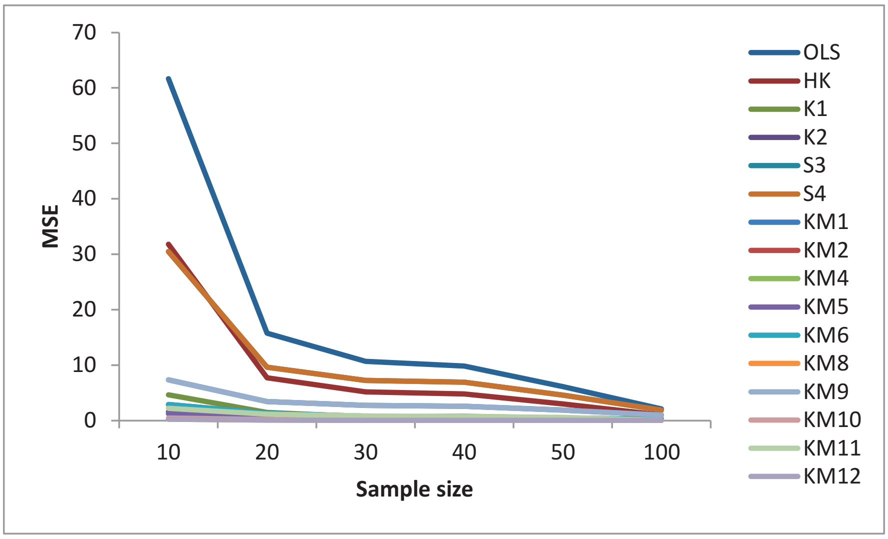

***Figure 6:*** *Performance of the estimators as a function of n when* $\sigma = 5$.

increases, the MSE also increases. However, the MSE decreases when the number of observations increases from 20 to 50. All of the ridge estimators have smaller MSE compared with OLS and they are very close to one another.

### 4.3. Performance as a function of $n$

In Table A.3 we have provided the MSEs of the estimators as a function of the sample size. We observed that, in general, when the sample size increases the MSE decreases, or remains the same. Even for the large values of $\gamma$ and $\sigma$, if we increase the sample size the MSE of estimators decreases. Again in this situation, as $n$ and $p$ increase the performance of $k_{KM4}$, $k_{KM5}$, $k_{KM8}$, $k_{KM10}$, $k_{KM12}$, and $k_{K1}$, $k_{K2}$ is better than the rest of the $k$ estimators.

For given $\gamma = 0.90$ and $p = 2$, the performance of the estimators as a function of the sample size for $\sigma = 0.5$ and $\sigma = 5$ is provided in Figures 5 and 6 respectively. From these figures, we observe that as the sample size increases, the MSE decreases. Except for a few situations, this pattern was constant for all of the estimators. Note the huge increase in the MSE when shifting from $\sigma = 0.5$ to $\sigma = 5$.

## 5. Concluding remarks

In this paper we have reviewed and proposed some new estimators for estimating the ridge parameter $k$. The new proposed estimators are defined based on the work of Kibria (2003), Khalaf and Shukur (2005) and Alkhamisi *et al.* (2006). The performance of the estimators depends on the variance of the random error $(\sigma)$, the correlations among the explanatory variables $(\gamma)$, the sample size $(n)$ and the unknown coefficient vector $\boldsymbol{\beta}$. Based on the simulation study, some conclusions can be drawn. However, these restrictions may be restricted to the set of experimental conditions that are investigated. We used the MSE criterion to measure the performance of the estimators. The increase of the number of correlated variables, of $\sigma$ and of the correlation between the independent variables have a negative effect in the MSE, in the sense that it also increases. When the sample size increases the MSE decreases, even when the correlation between the independent variables and $\sigma$ are large. In all situations, the proposed estimators have smaller MSE than the ordinary least squared estimators. Five of them, $k_{KM4}$, $k_{KM5}$, $k_{KM8}$, $k_{KM10}$, $k_{KM12}$, and the $k_{K1}$, $k_{K2}$ performed better than the rest in the sense of smaller MSE. Finally, it appears that the proposed estimators $k_{KM4}$, $k_{KM5}$, $k_{KM8}$, $k_{KM10}$, $k_{KM12}$ are useful and may be recommended to practitioners. The $k_{KM8}$ and $k_{KM12}$ estimators are particularly also recommended when working with models with large residual variances since they outperform all the others in such cases.

## Acknowledgement

The idea of this work initiated while the last two authors were on a research exchange leave at the Department of Mathematics and Statistics, Florida International University (FIU), Miami, USA, in April 2011. The authors acknowledge the excellent research facilities provided by FIU.

## References

Alkhamisi, M., Khalaf, G. and Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics-Theory and Methods*, 35, 2005–2020.

Alkhamisi, M. and Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics-Theory and Methods*, 37(4), 544–564.

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72, 77–91.

Frisch, R. (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems*, Publication 5 (Oslo: University Institute of Economics, 1934).

Galton, Sir Francis. (1885). Regression towards mediocrity in heredity stature. *Journal of Anthropological Institute*, 15, 246–263.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76, 131–139.

Hocking, R. R., Speed, F. M. and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, 18, 425–438.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, 12, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: application to non-orthogonal problems, *Technometrics*, 12, 69–82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulation. *Communications in Statistics*, 4, 105–123.

Khalaf, G. and Shukur, G. (2005). Choosing ridge parameters for regression problems. *Communications in Statistics-Theory and Methods*, 34, 1177–1182.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators *Communications in Statistics-Theory and Methods*, 32, 419–435.

Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics A*, 5, 307–323.

McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407–416.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*, Third Edition, John Wiley, New York.

Muniz, G. and Kibria, B. M. G. (2009). On some ridge regression estimators: an empirical comparisons. *Communications in Statistics-Simulation and Computation*, 38(3), 621–630.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, second edition. Duxbury. Belmont, CA.

Newhouse, J. P. and Oman, S. D. (1971). *An Evaluation of Ridge Estimators*. Rand Corporation, P- 716-PR.

Singh, S. and Tracy, D. S. (1999). Ridge-regression using scrambled responses. *Metrika*, 147–157.

***Table A.1:*** *Simulated MSE for fixed $n$, $p$, and $\gamma$ and different values of $\sigma$.*

$n = 10, p = 2, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.427 | 0.426 | 0.426 | 0.426 | 0.426 | 0.426 | 0.389 | 0.125 | 0.129 | 0.272 | 0.125 | 0.125 | 0.389 | 0.129 | 0.272 | 0.125 |
| 0.5 | 0.765 | 0.516 | 0.336 | 0.269 | 0.517 | 0.517 | 0.304 | 0.172 | 0.171 | 0.189 | 0.172 | 0.136 | 0.304 | 0.149 | 0.222 | 0.136 |
| 1 | 1.799 | 1.051 | 0.438 | 0.297 | 1.042 | 1.042 | 0.479 | 0.270 | 0.223 | 0.215 | 0.270 | 0.143 | 0.479 | 0.163 | 0.291 | 0.143 |
| 5 | 36.39 | 18.75 | 3.898 | 1.701 | 18.36 | 18.36 | 6.025 | 2.733 | 1.446 | 1.204 | 2.733 | 0.317 | 6.025 | 0.547 | 2.373 | 0.317 |

$n = 10, p = 2, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.057 | 0.121 | 0.094 | 0.025 | 0.121 | 0.121 | 0.057 | 0.095 | 0.025 | 0.120 |
| 0.5 | 0.438 | 0.244 | 0.117 | 0.084 | 0.243 | 0.243 | 0.087 | 0.065 | 0.053 | 0.047 | 0.065 | 0.084 | 0.087 | 0.063 | 0.048 | 0.084 |
| 1 | 1.608 | 0.866 | 0.254 | 0.133 | 0.840 | 0.840 | 0.271 | 0.121 | 0.084 | 0.082 | 0.121 | 0.089 | 0.271 | 0.075 | 0.115 | 0.089 |
| 5 | 38.59 | 20.17 | 4.069 | 1.577 | 20.19 | 20.19 | 6.265 | 2.712 | 1.384 | 1.138 | 2.712 | 0.271 | 6.265 | 0.477 | 3.325 | 0.271 |

$n = 10, p = 2, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM 11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.052 | 0.120 | 0.094 | 0.020 | 0.120 | 0.052 | 0.094 | 0.020 | 0.121 | 0.052. |
| 0.5 | 0.618 | 0.329 | 0.125 | 0.076 | 0.313 | 0.313 | 0.087 | 0.063 | 0.051 | 0.043 | 0.063 | 0.085 | 0.087 | 0.063 | 0.042 | 0.085 |
| 1 | 2.445 | 1.267 | 0.296 | 0.137 | 1.191 | 1.191 | 0.301 | 0.133 | 0.195 | 0.081 | 0.133 | 0.094 | 0.301 | 0.074 | 0.109 | 0.090 |
| 5 | 61.65 | 31.81 | 4.630 | 1.542 | 30.49 | 30.49 | 7.311 | 2.890 | 1.224 | 1.222 | 2.890 | 0.256 | 7.311 | 0.438 | 2.203 | 0.256 |

$n = 20, p = 2, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.035 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.030 | 0.032 | 0.032 | 0.020 | 0.032 | 0.032 | 0.030 | 0.032 | 0.020 | 0.032 |
| 0.5 | 0.090 | 0.061 | 0.045 | 0.041 | 0.072 | 0.072 | 0.043 | 0.029 | 0.025 | 0.025 | 0.029 | 0.038 | 0.043 | 0.029 | 0.028 | 0.038 |
| 1 | 0.264 | 0.152 | 0.080 | 0.064 | 0.209 | 0.209 | 0.119 | 0.059 | 0.047 | 0.043 | 0.059 | 0.040 | 0.119 | 0.035 | 0.068 | 0.040 |
| 5 | 5.529 | 2.823 | 1.010 | 0.645 | 4.299 | 4.299 | 2.393 | 1.198 | 0.853 | 0.524 | 0.198 | 0.116 | 2.393 | 0.246 | 1.293 | 0.116 |

$n = 20, p = 2, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.189 | 0.189 | 0.189 | 0.189 | 0.189 | 0.189 | 0.174 | 0.055 | 0.057 | 0.123 | 0.055 | 0.055 | 0.174 | 0.057 | 0.123 | 0.057 |
| 0.5 | 0.334 | 0.225 | 0.149 | 0.20 | 0.255 | 0.255 | 0.156 | 0.077 | 0.077 | 0.086 | 0.077 | 0.05 | 0.156 | 0.065 | 0.107 | 0.059 |
| 1 | 0.776 | 0.444 | 0.182 | 0.127 | 0.535 | 0.535 | 0.265 | 0.117 | 0.099 | 0.098 | 0.117 | 0.061 | 0.265 | 0.071 | 0.146 | 0.061 |
| 5 | 13.77 | 6.66 | 1.312 | 0.619 | 8.96 | 8.96 | 3.500 | 1.265 | 0.703 | 0.521 | 1.265 | 0.113 | 3.500 | 0.217 | 1.280 | 0.113 |

$n = 20, p = 2, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.025 | 0.032 | 0.032 | 0.011 | 0.032 | 0.032 | 0.025 | 0.032 | 0.011 | 0.032 |
| 0.5 | 0.190 | 0.104 | 0.049 | 0.035 | 0.121 | 0.121 | 0.047 | 0.028 | 0.023 | 0.020 | 0.028 | 0.040 | 0.047 | 0.029 | 0.022 | 0.040 |
| 1 | 0.666 | 0.340 | 0.094 | 0.056 | 0.419 | 0.419 | 0.155 | 0.050 | 0.037 | 0.037 | 0.050 | 0.042 | 0.155 | 0.034 | 0.058 | 0.042 |
| 5 | 15.76 | 7.692 | 1.441 | 0.599 | 9.601 | 9.601 | 3.440 | 1.303 | 0.635 | 0.457 | 1.303 | 0.093 | 3.440 | 0.176 | 1.141 | 0.093 |

$n = 50, p = 2, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.011 | 0.012 | 0.013 | 0.008 | 0.012 | 0.012 | 0.011 | 0.011 | 0.008 | 0.012 |
| 0.5 | 0.027 | 0.020 | 0.016 | 0.015 | 0.025 | 0.025 | 0.018 | 0.011 | 0.010 | 0.010 | 0.011 | 0.016 | 0.018 | 0.012 | 0.016 | 0.016 |
| 1 | 0.076 | 0.046 | 0.027 | 0.023 | 0.071 | 0.071 | 0.049 | 0.023 | 0.020 | 0.017 | 0.023 | 0.016 | 0.049 | 0.014 | 0.030 | 0.016 |
| 5 | 1.561 | 0.783 | 0.323 | 0.231 | 1.444 | 1.444 | 0.994 | 0.434 | 0.329 | 0.204 | 0.434 | 0.029 | 0.994 | 0.075 | 0.578 | 0.029 |

$n = 50, p = 2, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.073 | 0.071 | 0.071 | 0.071 | 0.071 | 0.071 | 0.068 | 0.020 | 0.022 | 0.054 | 0.020 | 0.020 | 0.068 | 0.022 | 0.054 | 0.020 |
| 0.5 | 0.096 | 0.070 | 0.053 | 0.047 | 0.089 | 0.089 | 0.067 | 0.033 | 0.032 | 0.035 | 0.033 | 0.022 | 0.067 | 0.025 | 0.048 | 0.022 |
| 1 | 0.172 | 0.107 | 0.061 | 0.049 | 0.156 | 0.156 | 0.106 | 0.050 | 0.043 | 0.039 | 0.050 | 0.022 | 0.106 | 0.026 | 0.067 | 0.022 |
| 5 | 2.381 | 1.184 | 0.388 | 0.250 | 2.099 | 2.099 | 1.246 | 0.467 | 0.322 | 0.217 | 0.467 | 0.035 | 1.246 | 0.081 | 0.616 | 0.035 |

$n = 50, p = 2, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.009 | 0.020 | 0.015 | 0.004 | 0.012 | 0.012 | 0.009 | 0.015 | 0.004 | 0.012 |
| 0.5 | 0.072 | 0.039 | 0.017 | 0.012 | 0.054 | 0.054 | 0.023 | 0.010 | 0.008 | 0.007 | 0.010 | 0.016 | 0.023 | 0.012 | 0.009 | 0.016 |
| 1 | 0.248 | 0.123 | 0.033 | 0.019 | 0.189 | 0.189 | 0.079 | 0.020 | 0.013 | 0.013 | 0.020 | 0.016 | 0.079 | 0.012 | 0.025 | 0.016 |
| 5 | 6.132 | 2.978 | 0.464 | 0.199 | 4.572 | 4.572 | 1.858 | 0.452 | 0.225 | 0.169 | 0.452 | 0.026 | 1.858 | 0.052 | 0.539 | 0.026 |

$n = 100, p = 2, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.008 | 0.004 | 0.009 | 0.009 | 0.005 | 0.008 | 0.004 | 0.009 |
| 0.5 | 0.014 | 0.010 | 0.008 | 0.007 | 0.013 | 0.013 | 0.010 | 0.005 | 0.005 | 0.005 | 0.005 | 0.008 | 0.010 | 0.006 | 0.006 | 0.008 |
| 1 | 0.039 | 0.023 | 0.013 | 0.011 | 0.038 | 0.038 | 0.028 | 0.011 | 0.009 | 0.008 | 0.011 | 0.008 | 0.028 | 0.007 | 0.017 | 0.008 |
| 5 | 0.812 | 0.412 | 0.163 | 0.113 | 0.779 | 0.779 | 0.572 | 0.221 | 0.165 | 0.102 | 0.221 | 0.012 | 0.572 | 0.031 | 0.326 | 0.012 |

$n = 100, p = 2, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.007 | 0.003 | 0.009 | 0.009 | 0.005 | 0.007 | 0.003 | 0.009 |
| 0.5 | 0.016 | 0.010 | 0.008 | 0.007 | 0.015 | 0.015 | 0.010 | 0.005 | 0.004 | 0.004 | 0.005 | 0.008 | 0.010 | 0.006 | 0.006 | 0.008 |
| 1 | 0.048 | 0.027 | 0.013 | 0.011 | 0.045 | 0.045 | 0.031 | 0.010 | 0.008 | 0.008 | 0.010 | 0.008 | 0.031 | 0.006 | 0.016 | 0.008 |
| 5 | 1.026 | 0.518 | 0.178 | 0.114 | 0.968 | 0.968 | 0.653 | 0.212 | 0.150 | 0.096 | 0.212 | 0.011 | 0.653 | 0.028 | 0.033 | 0.011 |

$n = 100, p = 2, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.036 | 0.034 | 0.034 | 0.034 | 0.034 | 0.034 | 0.032 | 0.010 | 0.010 | 0.023 | 0.010 | 0.010 | 0.032 | 0.010 | 0.023 | 0.010 |
| 0.5 | 0.058 | 0.039 | 0.026 | 0.021 | 0.053 | 0.053 | 0.036 | 0.014 | 0.014 | 0.016 | 0.014 | 0.010 | 0.036 | 0.011 | 0.023 | 0.010 |
| 1 | 0.118 | 0.067 | 0.029 | 0.021 | 0.106 | 0.106 | 0.064 | 0.021 | 0.018 | 0.017 | 0.021 | 0.010 | 0.064 | 0.012 | 0.032 | 0.010 |
| 5 | 2.115 | 1.005 | 0.222 | 0.118 | 1.861 | 1.861 | 0.960 | 0.226 | 0.132 | 0.099 | 0.226 | 0.013 | 0.960 | 0.029 | 0.013 | 0.013 |

$n = 10, p = 4, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.184 | 0.183 | 0.183 | 0.183 | 0.183 | 0.183 | 0.071 | 0.159 | 0.103 | 0.063 | 0.159 | 0.159 | 0.133 | 0.103 | 0.063 | 0.159 |
| 0.5 | 1.482 | 0.847 | 0.190 | 0.171 | 0.664 | 0.664 | 0.133 | 0.140 | 0.104 | 0.105 | 0.140 | 0.124 | 0.234 | 0.103 | 0.111 | 0.124 |
| 1 | 5.564 | 3.088 | 0.392 | 0.359 | 2.278 | 2.278 | 0.329 | 0.327 | 0.185 | 0.188 | 0.327 | 0.139 | 0.701 | 0.151 | 0.233 | 0.139 |
| 5 | 129.0 | 70.10 | 6.667 | 6.191 | 50.63 | 50.63 | 6.225 | 6.286 | 2.659 | 2.616 | 6.286 | 0.568 | 14.72 | 1.587 | 3.912 | 0.568 |

$n = 10, p = 4, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.170 | 0.169 | 0.169 | 0.169 | 0.169 | 0.169 | 0.192 | 0.169 | 0.190 | 0.193 | 0.169 | 0.169 | 0.193 | 0.190 | 0.193 | 0.169 |
| 0.5 | 12.04 | 6.365 | 0.341 | 0.289 | 3.308 | 3.308 | 0.264 | 0.217 | 0.213 | 0.223 | 0.217 | 0.183 | 0.409 | 0.205 | 0.234 | 0.183 |
| 1 | 47.61 | 24.49 | 0.661 | 0.495 | 11.43 | 11.43 | 0.485 | 0.384 | 0.514 | 0.283 | 0.306 | 0.384 | 0.196 | 0.247 | 0.359 | 0.196 |
| 5 | 1190. | 613.5 | 10.40 | 6.984 | 292.0 | 292.0 | 7.455 | 5.690 | 2.463 | 2.868 | 5.690 | 0.582 | 21.58 | 1.507 | 4.237 | 0.582 |

$n = 10, p = 4, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.188 | 0.177 | 0.179 | 0.179 | 0.179 | 0.179 | 0.209 | 0.164 | 0.160 | 0.184 | 0.164 | 0.164 | 0.243 | 0.159 | 0.184 | 0.164 |
| 0.5 | 23.17 | 11.50 | 0.405 | 0.321 | 5.521 | 5.521 | 0.270 | 0.190 | 0.182 | 0.197 | 0.190 | 0.164 | 0.480 | 0.173 | 0.215 | 0.164 |
| 1 | 94.17 | 48.44 | 0.865 | 0.627 | 22.26 | 22.26 | 0.517 | 0.406 | 0.241 | 0.275 | 0.406 | 0.173 | 1.278 | 0.205 | 0.336 | 0.173 |
| 5 | 2313 | 1174 | 13.87 | 9.616 | 553.3 | 553.3 | 8.052 | 6.160 | 2.021 | 2.583 | 6.160 | 0.452 | 26.49 | 1.187 | 3.965 | 0.452 |

$n = 20, p = 4, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.060 | 0.060 | 0.047 | 0.054 | 0.060 | 0.060 | 0.062 | 0.047 | 0.054 | 0.060 |
| 0.5 | 0.359 | 0.223 | 0.082 | 0.079 | 0.221 | 0.221 | 0.081 | 0.067 | 0.059 | 0.060 | 0.067 | 0.052 | 0.117 | 0.053 | 0.068 | 0.052 |
| 1 | 1.185 | 0.646 | 0.121 | 0.114 | 0.673 | 0.673 | 0.160 | 0.145 | 0.097 | 0.093 | 0.145 | 0.057 | 0.288 | 0.075 | 0.121 | 0.057 |
| 5 | 28.52 | 14.95 | 1.624 | 1.495 | 15.52 | 15.52 | 2.851 | 2.599 | 1.390 | 1.208 | 2.599 | 0.209 | 5.958 | 0.785 | 1.921 | 0.209 |

$n = 20, p = 4, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.063 | 0.064 | 0.064 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 |
| 0.5 | 0.367 | 0.241 | 0.111 | 0.110 | 0.280 | 0.280 | 0.102 | 0.087 | 0.080 | 0.082 | 0.087 | 0.067 | 0.154 | 0.076 | 0.088 | 0.067 |
| 1 | 1.295 | 0.756 | 0.182 | 0.196 | 0.936 | 0.936 | 0.198 | 0.154 | 0.109 | 0.110 | 0.154 | 0.069 | 0.419 | 0.090 | 0.139 | 0.069 |
| 5 | 30.56 | 16.70 | 2.512 | 3.022 | 21.69 | 21.69 | 3.336 | 2.627 | 1.147 | 1.041 | 2.627 | 0.169 | 8.985 | 0.588 | 1.850 | 0.169 |

$n = 20, p = 4, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.0630 | 0.0620 | 0.062 | 0.062 | 0.062 | 0.062 | 0.049 | 0.061 | 0.057 | 0.022 | 0.061 | 0.061 | 0.054 | 0.057 | 0.052 | 0.061 |
| 0.5 | 0.579 | 0.344 | 0.112 | 0.113 | 0.419 | 0.419 | 0.085 | 0.070 | 0.063 | 0.064 | 0.070 | 0.059 | 0.171 | 0.060 | 0.070 | 0.059 |
| 1 | 2.099 | 1.170 | 0.210 | 0.248 | 1.500 | 1.500 | 0.179 | 0.136 | 0.083 | 0.087 | 0.136 | 0.061 | 0.530 | 0.070 | 0.112 | 0.061 |
| 5 | 51.09 | 27.72 | 3.112 | 35.62 | 35.62 | 3.155 | 2.742 | 8.514 | 0.835 | 0.820 | 2.742 | 0.129 | 11.745 | 0.416 | 1.478 | 0.129 |

$n = 50, p = 4, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.018 | 0.021 | 0.016 | 0.016 | 0.021 | 0.021 | 0.022 | 0.016 | 0.016 | 0.021 |
| 0.5 | 0.071 | 0.049 | 0.026 | 0.027 | 0.066 | 0.066 | 0.030 | 0.023 | 0.020 | 0.020 | 0.023 | 0.019 | 0.047 | 0.018 | 0.0241 | 0.019 |
| 1 | 0.210 | 0.124 | 0.040 | 0.043 | 0.191 | 0.191 | 0.069 | 0.047 | 0.034 | 0.031 | 0.047 | 0.019 | 0.126 | 0.024 | 0.046 | 0.019 |
| 5 | 4.868 | 2.686 | 0.489 | 0.552 | 4.413 | 4.413 | 1.386 | 0.969 | 0.517 | 0.408 | 0.969 | 0.042 | 2.817 | 0.225 | 0.810 | 0.042 |

$n = 50, p = 4, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.021 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.021 | 0.021 | 0.021 | 0.021 | 0.021 | 0.020 | 0.021 | 0.021 | 0.020 |
| 0.5 | 0.112 | 0.073 | 0.035 | 0.035 | 0.098 | 0.098 | 0.037 | 0.029 | 0.028 | 0.028 | 0.029 | 0.023 | 0.061 | 0.026 | 0.031 | 0.023 |
| 1 | 0.374 | 0.214 | 0.056 | 0.059 | 0.319 | 0.319 | 0.078 | 0.053 | 0.037 | 0.037 | 0.053 | 0.023 | 0.176 | 0.030 | 0.050 | 0.023 |
| 5 | 9.112 | 4.978 | 0.698 | 0.911 | 7.735 | 7.735 | 1.481 | 0.924 | 0.394 | 0.356 | 0.924 | 0.040 | 4.065 | 0.176 | 0.730 | 0.040 |

$n = 50, p = 4, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.015 | 0.021 | 0.017 | 0.015 | 0.021 | 0.021 | 0.019 | 0.017 | 0.015 | 0.021 |
| 0.5 | 0.166 | 0.099 | 0.032 | 0.033 | 0.138 | 0.138 | 0.030 | 0.021 | 0.019 | 0.019 | 0.021 | 0.020 | 0.071 | 0.018 | 0.022 | 0.020 |
| 1 | 0.600 | 0.335 | 0.063 | 0.071 | 0.494 | 0.494 | 0.074 | 0.045 | 0.027 | 0.028 | 0.045 | 0.020 | 0.235 | 0.022 | 0.040 | 0.020 |
| 5 | 14.84 | 8.068 | 0.999 | 1.504 | 12.04 | 12.04 | 1.516 | 1.023 | 0.314 | 0.301 | 1.023 | 0.035 | 5.570 | 0.139 | 0.635 | 0.035 |

$n = 100, p = 4, \gamma = 0.7$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.008 | 0.010 | 0.008 | 0.008 | 0.010 | 0.010 | 0.010 | 0.008 | 0.008 | 0.010 |
| 0.5 | 0.030 | 0.021 | 0.012 | 0.012 | 0.029 | 0.029 | 0.015 | 0.011 | 0.010 | 0.010 | 0.011 | 0.009 | 0.023 | 0.009 | 0.012 | 0.009 |
| 1 | 0.089 | 0.053 | 0.019 | 0.020 | 0.085 | 0.085 | 0.037 | 0.024 | 0.018 | 0.016 | 0.024 | 0.010 | 0.063 | 0.012 | 0.024 | 0.010 |
| 5 | 1.966 | 1.067 | 0.213 | 0.245 | 1.879 | 1.879 | 0.727 | 0.440 | 0.253 | 0.199 | 0.440 | 0.016 | 1.365 | 0.103 | 0.423 | 0.016 |

$n = 100, p = 4, \gamma = 0.8$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.009 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |
| 0.5 | 0.040 | 0.027 | 0.015 | 0.015 | 0.038 | 0.038 | 0.018 | 0.013 | 0.012 | 0.012 | 0.013 | 0.010 | 0.028 | 0.011 | 0.014 | 0.010 |
| 1 | 0.131 | 0.077 | 0.023 | 0.025 | 0.123 | 0.123 | 0.041 | 0.026 | 0.018 | 0.017 | 0.026 | 0.010 | 0.083 | 0.013 | 0.025 | 0.010 |
| 5 | 3.065 | 1.679 | 0.278 | 0.330 | 2.869 | 2.869 | 0.816 | 0.458 | 0.219 | 0.180 | 0.458 | 0.015 | 1.889 | 0.084 | 0.409 | 0.015 |

$n = 100, p = 4, \gamma = 0.9$

| $\sigma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | KM11 | KM12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |
| 0.5 | 0.069 | 0.043 | 0.018 | 0.018 | 0.063 | 0.063 | 0.019 | 0.013 | 0.012 | 0.012 | 0.013 | 0.010 | 0.040 | 0.011 | 0.014 | 0.010 |
| 1 | 0.236 | 0.132 | 0.029 | 0.035 | 0.213 | 0.213 | 0.045 | 0.026 | 0.016 | 0.016 | 0.026 | 0.010 | 0.125 | 0.013 | 0.024 | 0.010 |
| 5 | 5.757 | 3.139 | 0.420 | 0.599 | 5.195 | 5.195 | 0.913 | 0.432 | 0.165 | 0.158 | 0.432 | 0.015 | 2.967 | 0.068 | 0.369 | 0.015 |

***Table A.2:*** *Simulated MSE for fixed $n$, $p$, and $\sigma$ and different values of $\gamma$.*

$n = 10, p = 2, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.427 | 0.426 | 0.426 | 0.426 | 0.426 | 0.426 | 0.389 | 0.125 | 0.129 | 0.272 | 0.125 | 0.125 | 0.389 | 0.129 | 0.272 | 0.125 |
| 0.80 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.057 | 0.121 | 0.094 | 0.025 | 0.121 | 0.121 | 0.057 | 0.095 | 0.025 | 0.120 |
| 0.90 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.052 | 0.120 | 0.094 | 0.020 | 0.120 | 0.052 | 0.094 | 0.020 | 0.121 | 0.052. |

$n = 10, p = 2, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 1.799 | 1.051 | 0.438 | 0.297 | 1.042 | 1.042 | 0.479 | 0.270 | 0.223 | 0.215 | 0.270 | 0.143 | 0.479 | 0.163 | 0.291 | 0.143 |
| 0.80 | 1.608 | 0.866 | 0.254 | 0.133 | 0.840 | 0.840 | 0.271 | 0.121 | 0.084 | 0.082 | 0.121 | 0.089 | 0.271 | 0.075 | 0.115 | 0.089 |
| 0.90 | 2.445 | 1.267 | 0.296 | 0.137 | 1.191 | 1.191 | 0.301 | 0.133 | 0.195 | 0.081 | 0.133 | 0.094 | 0.301 | 0.074 | 0.109 | 0.090 |

$n = 10, p = 2, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 36.39 | 18.75 | 3.898 | 1.701 | 18.36 | 18.36 | 6.025 | 2.733 | 1.446 | 1.204 | 2.733 | 0.317 | 6.025 | 0.547 | 2.373 | 0.317 |
| 0.80 | 38.59 | 20.17 | 4.069 | 1.577 | 20.19 | 20.19 | 6.265 | 2.712 | 1.384 | 1.138 | 2.712 | 0.271 | 6.265 | 0.477 | 3.325 | 0.271 |
| 0.90 | 61.65 | 31.81 | 4.630 | 1.542 | 30.49 | 30.49 | 7.311 | 2.890 | 1.224 | 1.222 | 2.890 | 0.256 | 7.311 | 0.438 | 2.203 | 0.256 |

$n = 10, p = 4, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.184 | 0.183 | 0.183 | 0.183 | 0.183 | 0.183 | 0.071 | 0.159 | 0.103 | 0.063 | 0.159 | 0.159 | 0.133 | 0.103 | 0.063 | 0.159 |
| 0.80 | 0.170 | 0.169 | 0.169 | 0.169 | 0.169 | 0.169 | 0.192 | 0.169 | 0.190 | 0.193 | 0.169 | 0.169 | 0.193 | 0.190 | 0.193 | 0.169 |
| 0.90 | 0.188 | 0.177 | 0.179 | 0.179 | 0.179 | 0.179 | 0.209 | 0.164 | 0.160 | 0.184 | 0.164 | 0.164 | 0.243 | 0.159 | 0.184 | 0.164 |

$n = 10, p = 4, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 5.564 | 3.088 | 0.392 | 0.359 | 2.278 | 2.278 | 0.329 | 0.327 | 0.185 | 0.188 | 0.327 | 0.139 | 0.701 | 0.151 | 0.233 | 0.139 |
| 0.80 | 47.61 | 24.49 | 0.661 | 0.495 | 11.43 | 11.43 | 0.485 | 0.384 | 0.514 | 0.283 | 0.306 | 0.384 | 0.196 | 0.247 | 0.359 | 0.196 |
| 0.90 | 94.17 | 48.44 | 0.865 | 0.627 | 22.26 | 22.26 | 0.517 | 0.406 | 0.241 | 0.275 | 0.406 | 0.173 | 1.278 | 0.205 | 0.336 | 0.173 |

$n = 10, p = 4, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 129 | 70.10 | 6.667 | 6.191 | 50.63 | 50.63 | 6.225 | 6.286 | 2.659 | 2.616 | 6.286 | 0.568 | 14.72 | 1.587 | 3.912 | 0.568 |
| 0.80 | 1190 | 613.5 | 10.40 | 6.984 | 292 | 292 | 7.455 | 5.690 | 2.463 | 2.868 | 5.690 | 0.582 | 21.58 | 1.507 | 4.237 | 0.582 |
| 0.90 | 2313. | 1174 | 13.87 | 9.616 | 553.3 | 553.3 | 8.052 | 6.160 | 2.021 | 2.583 | 6.160 | 0.452 | 26.49 | 1.187 | 3.965 | 0.452 |

$n = 20, p = 2, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K11 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.035 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.030 | 0.032 | 0.032 | 0.020 | 0.032 | 0.032 | 0.030 | 0.032 | 0.020 | 0.032 |
| 0.80 | 0.189 | 0.189 | 0.189 | 0.189 | 0.189 | 0.189 | 0.174 | 0.055 | 0.057 | 0.123 | 0.055 | 0.055 | 0.174 | 0.057 | 0.123 | 0.057 |
| 0.90 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.025 | 0.032 | 0.032 | 0.011 | 0.032 | 0.032 | 0.025 | 0.032 | 0.011 | 0.032 |

$n = 20, p = 2, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.264 | 0.152 | 0.080 | 0.064 | 0.209 | 0.209 | 0.119 | 0.059 | 0.047 | 0.043 | 0.059 | 0.040 | 0.119 | 0.035 | 0.068 | 0.040 |
| 0.80 | 0.776 | 0.444 | 0.182 | 0.127 | 0.535 | 0.535 | 0.265 | 0.117 | 0.099 | 0.098 | 0.117 | 0.061 | 0.265 | 0.071 | 0.146 | 0.061 |
| 0.90 | 0.666 | 0.340 | 0.094 | 0.056 | 0.419 | 0.419 | 0.155 | 0.050 | 0.037 | 0.037 | 0.050 | 0.042 | 0.155 | 0.034 | 0.058 | 0.042 |

$n = 20, p = 2, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 5.529 | 2.823 | 1.010 | 0.645 | 4.299 | 4.299 | 2.393 | 1.198 | 0.853 | 0.524 | 0.198 | 0.116 | 2.393 | 0.246 | 1.293 | 0.116 |
| 0.80 | 13.77 | 6.66 | 1.312 | 0.619 | 8.96 | 8.96 | 3.500 | 1.265 | 0.703 | 0.521 | 1.265 | 0.113 | 3.500 | 0.217 | 1.280 | 0.113 |
| 0.90 | 15.76 | 7.692 | 1.441 | 0.599 | 9.601 | 9.601 | 3.440 | 1.303 | 0.635 | 0.457 | 1.303 | 0.093 | 3.440 | 0.176 | 1.141 | 0.093 |

$n = 20, p = 4, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.060 | 0.060 | 0.047 | 0.054 | 0.060 | 0.060 | 0.062 | 0.047 | 0.054 | 0.060 |
| 0.80 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.063 | 0.064 | 0.064 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 |
| 0.90 | 0.063 | 0.062 | 0.062 | 0.062 | 0.062 | 0.062 | 0.049 | 0.061 | 0.057 | 0.022 | 0.061 | 0.061 | 0.054 | 0.057 | 0.052 | 0.061 |

$n = 20, p = 4, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 1.185 | 0.646 | 0.121 | 0.114 | 0.673 | 0.673 | 0.160 | 0.145 | 0.097 | 0.093 | 0.145 | 0.057 | 0.288 | 0.075 | 0.121 | 0.057 |
| 0.80 | 1.295 | 0.756 | 0.182 | 0.196 | 0.936 | 0.936 | 0.198 | 0.154 | 0.109 | 0.110 | 0.154 | 0.069 | 0.419 | 0.090 | 0.139 | 0.069 |
| 0.90 | 2.099 | 1.170 | 0.210 | 0.248 | 1.500 | 1.500 | 0.179 | 0.136 | 0.083 | 0.087 | 0.136 | 0.061 | 0.530 | 0.070 | 0.112 | 0.061 |

$n = 20, p = 4, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 28.52 | 14.95 | 1.624 | 1.495 | 15.52 | 15.52 | 2.851 | 2.599 | 1.390 | 1.208 | 2.599 | 0.209 | 5.958 | 0.785 | 1.921 | 0.209 |
| 0.80 | 30.56 | 16.70 | 2.512 | 3.022 | 21.69 | 21.69 | 3.336 | 2.627 | 1.147 | 1.041 | 2.627 | 0.169 | 8.985 | 0.588 | 1.850 | 0.169 |
| 0.90 | 51.09 | 27.72 | 3.112 | 35.62 | 35.62 | 3.155 | 2.742 | 8.514 | 0.835 | 0.820 | 2.742 | 0.129 | 11.745 | 0.416 | 1.478 | 0.129 |

$n = 50, p = 2, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.011 | 0.012 | 0.013 | 0.008 | 0.012 | 0.012 | 0.011 | 0.011 | 0.008 | 0.012 |
| 0.80 | 0.073 | 0.071 | 0.071 | 0.071 | 0.071 | 0.071 | 0.068 | 0.020 | 0.022 | 0.054 | 0.020 | 0.020 | 0.068 | 0.022 | 0.054 | 0.020 |
| 0.90 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.009 | 0.020 | 0.015 | 0.004 | 0.012 | 0.012 | 0.009 | 0.015 | 0.004 | 0.012 |

$n = 50, p = 2, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.076 | 0.046 | 0.027 | 0.023 | 0.071 | 0.071 | 0.049 | 0.023 | 0.020 | 0.017 | 0.023 | 0.016 | 0.049 | 0.014 | 0.030 | 0.016 |
| 0.80 | 0.172 | 0.107 | 0.061 | 0.049 | 0.156 | 0.156 | 0.106 | 0.050 | 0.043 | 0.039 | 0.050 | 0.022 | 0.106 | 0.026 | 0.067 | 0.022 |
| 0.90 | 0.248 | 0.123 | 0.033 | 0.019 | 0.189 | 0.189 | 0.079 | 0.020 | 0.013 | 0.013 | 0.020 | 0.016 | 0.079 | 0.012 | 0.025 | 0.016 |

$n = 50, p = 2, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 1.561 | 0.783 | 0.323 | 0.231 | 1.444 | 1.444 | 0.994 | 0.434 | 0.329 | 0.204 | 0.434 | 0.029 | 0.994 | 0.075 | 0.578 | 0.029 |
| 0.80 | 2.381 | 1.184 | 0.388 | 0.250 | 2.099 | 2.099 | 1.246 | 0.467 | 0.322 | 0.217 | 0.467 | 0.035 | 1.246 | 0.081 | 0.616 | 0.035 |
| 0.90 | 6.132 | 2.978 | 0.464 | 0.199 | 4.572 | 4.572 | 1.858 | 0.452 | 0.225 | 0.169 | 0.452 | 0.026 | 1.858 | 0.052 | 0.539 | 0.026 |

$n = 50, p = 4, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.018 | 0.021 | 0.016 | 0.016 | 0.021 | 0.021 | 0.022 | 0.016 | 0.016 | 0.021 |
| 0.80 | 0.021 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.021 | 0.021 | 0.021 | 0.021 | 0.021 | 0.020 | 0.021 | 0.021 | 0.020 |
| 0.90 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.015 | 0.021 | 0.017 | 0.015 | 0.021 | 0.021 | 0.019 | 0.017 | 0.015 | 0.021 |

$n = 50, p = 4, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.210 | 0.124 | 0.040 | 0.043 | 0.191 | 0.191 | 0.069 | 0.047 | 0.034 | 0.031 | 0.047 | 0.019 | 0.126 | 0.024 | 0.046 | 0.019 |
| 0.80 | 0.374 | 0.214 | 0.056 | 0.059 | 0.319 | 0.319 | 0.078 | 0.053 | 0.037 | 0.037 | 0.053 | 0.023 | 0.176 | 0.030 | 0.050 | 0.023 |
| 0.90 | 0.600 | 0.335 | 0.063 | 0.071 | 0.494 | 0.494 | 0.074 | 0.045 | 0.027 | 0.028 | 0.045 | 0.020 | 0.235 | 0.022 | 0.040 | 0.020 |

$n = 50, p = 4, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 4.868 | 2.686 | 0.489 | 0.552 | 4.413 | 4.413 | 1.386 | 0.969 | 0.517 | 0.408 | 0.969 | 0.042 | 2.817 | 0.225 | 0.810 | 0.042 |
| 0.80 | 9.112 | 4.978 | 0.698 | 0.911 | 7.735 | 7.735 | 1.481 | 0.924 | 0.394 | 0.356 | 0.924 | 0.040 | 4.065 | 0.176 | 0.730 | 0.040 |
| 0.90 | 14.84 | 8.068 | 0.999 | 1.504 | 12.04 | 12.04 | 1.516 | 1.023 | 0.314 | 0.301 | 1.023 | 0.035 | 5.570 | 0.139 | 0.635 | 0.035 |

$n = 100, p = 2, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.008 | 0.004 | 0.009 | 0.009 | 0.005 | 0.008 | 0.004 | 0.009 |
| 0.80 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.007 | 0.003 | 0.009 | 0.009 | 0.005 | 0.007 | 0.003 | 0.009 |
| 0.90 | 0.036 | 0.034 | 0.034 | 0.034 | 0.034 | 0.034 | 0.032 | 0.010 | 0.010 | 0.023 | 0.010 | 0.010 | 0.032 | 0.010 | 0.023 | 0.010 |

$n = 100, p = 2, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.039 | 0.023 | 0.013 | 0.011 | 0.038 | 0.038 | 0.028 | 0.011 | 0.009 | 0.008 | 0.011 | 0.008 | 0.028 | 0.007 | 0.017 | 0.008 |
| 0.80 | 0.048 | 0.027 | 0.013 | 0.011 | 0.045 | 0.045 | 0.031 | 0.010 | 0.008 | 0.008 | 0.010 | 0.008 | 0.031 | 0.006 | 0.016 | 0.008 |
| 0.90 | 0.118 | 0.067 | 0.029 | 0.021 | 0.106 | 0.106 | 0.064 | 0.021 | 0.018 | 0.017 | 0.021 | 0.010 | 0.064 | 0.012 | 0.032 | 0.010 |

$n = 100, p = 2, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.812 | 0.412 | 0.163 | 0.113 | 0.779 | 0.779 | 0.572 | 0.221 | 0.165 | 0.102 | 0.221 | 0.012 | 0.572 | 0.031 | 0.326 | 0.012 |
| 0.80 | 1.026 | 0.518 | 0.178 | 0.114 | 0.968 | 0.968 | 0.653 | 0.212 | 0.150 | 0.096 | 0.212 | 0.011 | 0.653 | 0.028 | 0.033 | 0.011 |
| 0.90 | 2.115 | 1.005 | 0.222 | 0.118 | 1.861 | 1.861 | 0.960 | 0.226 | 0.132 | 0.099 | 0.226 | 0.013 | 0.960 | 0.029 | 0.013 | 0.013 |

$n = 100, p = 4, \sigma = 0.01$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.008 | 0.010 | 0.008 | 0.008 | 0.010 | 0.010 | 0.010 | 0.008 | 0.008 | 0.010 |
| 0.80 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.009 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |
| 0.90 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |

$n = 100, p = 4, \sigma = 1$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 0.089 | 0.053 | 0.019 | 0.020 | 0.085 | 0.085 | 0.037 | 0.024 | 0.018 | 0.016 | 0.024 | 0.010 | 0.063 | 0.012 | 0.024 | 0.010 |
| 0.80 | 0.131 | 0.077 | 0.023 | 0.025 | 0.123 | 0.123 | 0.041 | 0.026 | 0.018 | 0.017 | 0.026 | 0.010 | 0.083 | 0.013 | 0.025 | 0.010 |
| 0.90 | 0.236 | 0.132 | 0.029 | 0.035 | 0.213 | 0.213 | 0.045 | 0.026 | 0.016 | 0.016 | 0.026 | 0.010 | 0.125 | 0.013 | 0.024 | 0.010 |

$n = 100, p = 4, \sigma = 5$

| $\gamma$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.70 | 1.966 | 1.067 | 0.213 | 0.245 | 1.879 | 1.879 | 0.727 | 0.440 | 0.253 | 0.199 | 0.440 | 0.016 | 1.365 | 0.103 | 0.423 | 0.016 |
| 0.80 | 3.065 | 1.679 | 0.278 | 0.330 | 2.869 | 2.869 | 0.816 | 0.458 | 0.219 | 0.180 | 0.458 | 0.015 | 1.889 | 0.084 | 0.409 | 0.015 |
| 0.90 | 5.757 | 3.139 | 0.420 | 0.599 | 5.195 | 5.195 | 0.913 | 0.432 | 0.165 | 0.158 | 0.432 | 0.015 | 2.967 | 0.068 | 0.369 | 0.015 |

***Table A.3:*** *Simulated MSE for fixed $p$, $\gamma$ and $\sigma$ and different values of n.*

$p = 2, \gamma = 0.7, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.427 | 0.426 | 0.426 | 0.426 | 0.426 | 0.426 | 0.389 | 0.125 | 0.129 | 0.272 | 0.125 | 0.125 | 0.389 | 0.129 | 0.272 | 0.125 |
| 20 | 0.035 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.030 | 0.032 | 0.032 | 0.020 | 0.032 | 0.032 | 0.030 | 0.032 | 0.020 | 0.032 |
| 30 | 0.122 | 0.121 | 0.121 | 0.121 | 0.121 | 0.121 | 0.117 | 0.035 | 0.037 | 0.093 | 0.035 | 0.035 | 0.117 | 0.037 | 0.093 | 0.035 |
| 40 | 0.089 | 0.089 | 0.089 | 0.089 | 0.089 | 0.089 | 0.086 | 0.026 | 0.028 | 0.071 | 0.026 | 0.026 | 0.086 | 0.028 | 0.071 | 0.026 |
| 50 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.011 | 0.012 | 0.013 | 0.008 | 0.012 | 0.012 | 0.011 | 0.011 | 0.008 | 0.012 |
| 100 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.008 | 0.004 | 0.009 | 0.009 | 0.005 | 0.008 | 0.004 | 0.009 |

$p = 2, \gamma = 0.7, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.765 | 0.516 | 0.336 | 0.269 | 0.517 | 0.517 | 0.304 | 0.172 | 0.171 | 0.189 | 0.172 | 0.136 | 0.304 | 0.149 | 0.222 | 0.136 |
| 20 | 0.090 | 0.061 | 0.045 | 0.041 | 0.072 | 0.072 | 0.043 | 0.029 | 0.025 | 0.025 | 0.029 | 0.038 | 0.043 | 0.029 | 0.028 | 0.038 |
| 30 | 0.161 | 0.118 | 0.089 | 0.078 | 0.145 | 0.145 | 0.107 | 0.056 | 0.055 | 0.060 | 0.056 | 0.038 | 0.107 | 0.044 | 0.079 | 0.038 |
| 40 | 0.114 | 0.086 | 0.067 | 0.059 | 0.107 | 0.107 | 0.083 | 0.043 | 0.042 | 0.046 | 0.043 | 0.028 | 0.083 | 0.032 | 0.063 | 0.028 |
| 50 | 0.027 | 0.020 | 0.016 | 0.015 | 0.025 | 0.025 | 0.018 | 0.011 | 0.010 | 0.010 | 0.011 | 0.016 | 0.018 | 0.012 | 0.016 | 0.016 |
| 100 | 0.014 | 0.010 | 0.008 | 0.007 | 0.013 | 0.013 | 0.010 | 0.005 | 0.005 | 0.005 | 0.005 | 0.008 | 0.010 | 0.006 | 0.006 | 0.008 |

$p = 2, \gamma = 0.7, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1.799 | 1.051 | 0.438 | 0.297 | 1.042 | 1.042 | 0.479 | 0.270 | 0.223 | 0.215 | 0.270 | 0.143 | 0.479 | 0.163 | 0.291 | 0.143 |
| 20 | 0.264 | 0.152 | 0.080 | 0.064 | 0.209 | 0.209 | 0.119 | 0.059 | 0.047 | 0.043 | 0.059 | 0.040 | 0.119 | 0.035 | 0.068 | 0.040 |
| 30 | 0.279 | 0.171 | 0.098 | 0.080 | 0.240 | 0.240 | 0.158 | 0.083 | 0.073 | 0.066 | 0.083 | 0.039 | 0.158 | 0.047 | 0.104 | 0.039 |
| 40 | 0.194 | 0.124 | 0.073 | 0.059 | 0.177 | 0.177 | 0.126 | 0.066 | 0.058 | 0.051 | 0.066 | 0.029 | 0.126 | 0.035 | 0.084 | 0.029 |
| 50 | 0.076 | 0.046 | 0.027 | 0.023 | 0.071 | 0.071 | 0.049 | 0.023 | 0.020 | 0.017 | 0.023 | 0.016 | 0.049 | 0.014 | 0.030 | 0.016 |
| 100 | 0.039 | 0.023 | 0.013 | 0.011 | 0.038 | 0.038 | 0.028 | 0.011 | 0.009 | 0.008 | 0.011 | 0.008 | 0.028 | 0.007 | 0.017 | 0.008 |

$p = 2, \gamma = 0.7, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 36.39 | 18.75 | 3.898 | 1.701 | 18.36 | 18.36 | 6.025 | 2.733 | 1.446 | 1.204 | 2.733 | 0.317 | 6.025 | 0.547 | 2.373 | 0.317 |
| 20 | 5.529 | 2.823 | 1.010 | 0.645 | 4.299 | 4.299 | 2.393 | 1.198 | 0.853 | 0.524 | 0.198 | 0.116 | 2.393 | 0.246 | 1.293 | 0.116 |
| 30 | 4.417 | 2.277 | 0.757 | 0.465 | 3.654 | 3.654 | 2.041 | 0.795 | 0.553 | 0.402 | 0.795 | 0.072 | 2.041 | 0.160 | 1.022 | 0.072 |
| 40 | 2.685 | 1.378 | 0.534 | 0.358 | 2.373 | 2.373 | 1.470 | 0.587 | 0.426 | 0.295 | 0.587 | 0.049 | 1.470 | 0.114 | 0.786 | 0.049 |
| 50 | 1.561 | 0.783 | 0.323 | 0.231 | 1.444 | 1.444 | 0.994 | 0.434 | 0.329 | 0.204 | 0.434 | 0.029 | 0.994 | 0.075 | 0.578 | 0.029 |
| 100 | 0.812 | 0.412 | 0.163 | 0.113 | 0.779 | 0.779 | 0.572 | 0.221 | 0.165 | 0.102 | 0.221 | 0.012 | 0.572 | 0.031 | 0.326 | 0.012 |

$p = 2, \gamma = 0.8, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.057 | 0.121 | 0.094 | 0.025 | 0.121 | 0.121 | 0.057 | 0.095 | 0.025 | 0.120 |
| 20 | 0.035 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.030 | 0.032 | 0.032 | 0.020 | 0.032 | 0.032 | 0.030 | 0.032 | 0.020 | 0.032 |
| 30 | 0.021 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.018 | 0.034 | 0.027 | 0.009 | 0.034 | 0.034 | 0.034 | 0.027 | 0.009 | 0.034 |
| 40 | 0.089 | 0.089 | 0.089 | 0.089 | 0.089 | 0.089 | 0.086 | 0.026 | 0.028 | 0.071 | 0.026 | 0.026 | 0.086 | 0.028 | 0.071 | 0.026 |
| 50 | 0.073 | 0.071 | 0.071 | 0.071 | 0.071 | 0.071 | 0.068 | 0.020 | 0.022 | 0.054 | 0.020 | 0.020 | 0.068 | 0.022 | 0.054 | 0.020 |
| 100 | 0.010 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.009 | 0.007 | 0.003 | 0.009 | 0.009 | 0.005 | 0.007 | 0.003 | 0.009 |

$p = 2, \gamma = 0.8, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.438 | 0.244 | 0.117 | 0.084 | 0.243 | 0.243 | 0.087 | 0.065 | 0.053 | 0.047 | 0.065 | 0.084 | 0.087 | 0.063 | 0.048 | 0.084 |
| 20 | 0.334 | 0.225 | 0.149 | 0.20 | 0.255 | 0.255 | 0.156 | 0.077 | 0.077 | 0.086 | 0.077 | 0.05 | 0.156 | 0.065 | 0.107 | 0.059 |
| 30 | 0.084 | 0.051 | 0.031 | 0.025 | 0.065 | 0.065 | 0.032 | 0.017 | 0.014 | 0.013 | 0.017 | 0.026 | 0.032 | 0.019 | 0.016 | 0.026 |
| 40 | 0.045 | 0.029 | 0.020 | 0.018 | 0.039 | 0.039 | 0.023 | 0.013 | 0.011 | 0.011 | 0.013 | 0.019 | 0.023 | 0.014 | 0.013 | 0.019 |
| 50 | 0.096 | 0.070 | 0.053 | 0.047 | 0.089 | 0.089 | 0.067 | 0.033 | 0.032 | 0.035 | 0.033 | 0.022 | 0.067 | 0.025 | 0.048 | 0.022 |
| 100 | 0.016 | 0.010 | 0.008 | 0.007 | 0.015 | 0.015 | 0.010 | 0.005 | 0.004 | 0.004 | 0.005 | 0.008 | 0.010 | 0.006 | 0.006 | 0.008 |

$p = 2, \gamma = 0.8, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1.608 | 0.866 | 0.254 | 0.133 | 0.840 | 0.840 | 0.271 | 0.121 | 0.084 | 0.082 | 0.121 | 0.089 | 0.271 | 0.075 | 0.115 | 0.089 |
| 20 | 0.776 | 0.444 | 0.182 | 0.127 | 0.535 | 0.535 | 0.265 | 0.117 | 0.099 | 0.098 | 0.117 | 0.061 | 0.265 | 0.071 | 0.146 | 0.061 |
| 30 | 0.273 | 0.143 | 0.053 | 0.037 | 0.207 | 0.207 | 0.099 | 0.038 | 0.027 | 0.025 | 0.038 | 0.027 | 0.099 | 0.043 | 0.027 | 0.099 |
| 40 | 0.140 | 0.076 | 0.036 | 0.027 | 0.120 | 0.120 | 0.068 | 0.028 | 0.021 | 0.019 | 0.028 | 0.020 | 0.068 | 0.016 | 0.034 | 0.020 |
| 50 | 0.172 | 0.107 | 0.061 | 0.049 | 0.156 | 0.156 | 0.106 | 0.050 | 0.043 | 0.039 | 0.050 | 0.022 | 0.106 | 0.026 | 0.067 | 0.022 |
| 100 | 0.048 | 0.027 | 0.013 | 0.011 | 0.045 | 0.045 | 0.031 | 0.010 | 0.008 | 0.008 | 0.010 | 0.008 | 0.031 | 0.006 | 0.016 | 0.008 |

$p = 2, \gamma = 0.8, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 38.59 | 20.17 | 4.069 | 1.577 | 20.19 | 20.19 | 6.265 | 2.712 | 1.384 | 1.138 | 2.712 | 0.271 | 6.265 | 0.477 | 3.325 | 0.271 |
| 20 | 13.77 | 6.66 | 1.312 | 0.619 | 8.96 | 8.96 | 3.500 | 1.265 | 0.703 | 0.521 | 1.265 | 0.113 | 3.500 | 0.217 | 1.280 | 0.113 |
| 30 | 6.236 | 3.051 | 0.788 | 0.430 | 4.653 | 4.653 | 2.168 | 0.762 | 0.469 | 0.329 | 0.762 | 0.058 | 2.168 | 0.124 | 0.898 | 0.058 |
| 40 | 3.133 | 1.549 | 0.480 | 0.294 | 2.660 | 2.660 | 1.490 | 0.559 | 0.371 | 0.252 | 0.559 | 0.038 | 1.490 | 0.08 | 0.711 | 0.038 |
| 50 | 2.381 | 1.184 | 0.388 | 0.250 | 2.099 | 2.099 | 1.246 | 0.467 | 0.322 | 0.217 | 0.467 | 0.035 | 1.246 | 0.081 | 0.616 | 0.035 |
| 100 | 1.026 | 0.518 | 0.178 | 0.114 | 0.968 | 0.968 | 0.653 | 0.212 | 0.150 | 0.096 | 0.212 | 0.011 | 0.653 | 0.028 | 0.033 | 0.011 |

$p = 2, \gamma = 0.9, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.073 | 0.052 | 0.120 | 0.094 | 0.020 | 0.120 | 0.052 | 0.094 | 0.020 | 0.121 | 0.052. |
| 20 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.032 | 0.025 | 0.032 | 0.032 | 0.011 | 0.032 | 0.032 | 0.025 | 0.032 | 0.011 | 0.032 |
| 30 | 0.021 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.016 | 0.034 | 0.027 | 0.007 | 0.034 | 0.034 | 0.016 | 0.027 | 0.007 | 0.034 |
| 40 | 0.090 | 0.089 | 0.089 | 0.089 | 0.089 | 0.089 | 0.079 | 0.026 | 0.027 | 0.053 | 0.026 | 0.026 | 0.079 | 0.027 | 0.054 | 0.026 |
| 50 | 0.013 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 | 0.009 | 0.020 | 0.015 | 0.004 | 0.012 | 0.012 | 0.009 | 0.015 | 0.004 | 0.012 |
| 100 | 0.036 | 0.034 | 0.034 | 0.034 | 0.034 | 0.034 | 0.032 | 0.010 | 0.010 | 0.023 | 0.010 | 0.010 | 0.032 | 0.010 | 0.023 | 0.010 |

$p = 2, \gamma = 0.9, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.618 | 0.329 | 0.125 | 0.076 | 0.313 | 0.313 | 0.087 | 0.063 | 0.051 | 0.043 | 0.063 | 0.085 | 0.087 | 0.063 | 0.042 | 0.085 |
| 20 | 0.190 | 0.104 | 0.049 | 0.035 | 0.121 | 0.121 | 0.047 | 0.028 | 0.023 | 0.020 | 0.028 | 0.040 | 0.047 | 0.029 | 0.022 | 0.040 |
| 30 | 0.123 | 0.066 | 0.030 | 0.022 | 0.087 | 0.087 | 0.035 | 0.017 | 0.014 | 0.013 | 0.017 | 0.027 | 0.035 | 0.019 | 0.015 | 0.027 |
| 40 | 0.195 | 0.125 | 0.070 | 0.053 | 0.151 | 0.151 | 0.084 | 0.035 | 0.035 | 0.039 | 0.035 | 0.027 | 0.084 | 0.029 | 0.050 | 0.027 |
| 50 | 0.072 | 0.039 | 0.017 | 0.012 | 0.054 | 0.054 | 0.023 | 0.010 | 0.008 | 0.007 | 0.010 | 0.016 | 0.023 | 0.012 | 0.009 | 0.016 |
| 100 | 0.058 | 0.039 | 0.026 | 0.021 | 0.053 | 0.053 | 0.036 | 0.014 | 0.014 | 0.016 | 0.014 | 0.010 | 0.036 | 0.011 | 0.023 | 0.010 |

$p = 2, \gamma = 0.9, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 2.445 | 1.267 | 0.296 | 0.137 | 1.191 | 1.191 | 0.301 | 0.133 | 0.195 | 0.081 | 0.133 | 0.094 | 0.301 | 0.074 | 0.109 | 0.090 |
| 20 | 0.666 | 0.340 | 0.094 | 0.056 | 0.419 | 0.419 | 0.155 | 0.050 | 0.037 | 0.037 | 0.050 | 0.042 | 0.155 | 0.034 | 0.058 | 0.042 |
| 30 | 0.449 | 0.230 | 0.061 | 0.034 | 0.304 | 0.304 | 0.116 | 0.034 | 0.023 | 0.023 | 0.034 | 0.027 | 0.116 | 0.022 | 0.039 | 0.027 |
| 40 | 0.499 | 0.275 | 0.093 | 0.057 | 0.358 | 0.358 | 0.160 | 0.052 | 0.044 | 0.045 | 0.052 | 0.027 | 0.160 | 0.031 | 0.071 | 0.027 |
| 50 | 0.248 | 0.123 | 0.033 | 0.019 | 0.189 | 0.189 | 0.079 | 0.020 | 0.013 | 0.013 | 0.020 | 0.016 | 0.079 | 0.012 | 0.025 | 0.016 |
| 100 | 0.118 | 0.067 | 0.029 | 0.021 | 0.106 | 0.106 | 0.064 | 0.021 | 0.018 | 0.017 | 0.021 | 0.010 | 0.064 | 0.012 | 0.032 | 0.010 |

$p = 2, \gamma = 0.9, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 61.65 | 31.81 | 4.630 | 1.542 | 30.49 | 30.49 | 7.311 | 2.890 | 1.224 | 1.222 | 2.890 | 0.256 | 7.311 | 0.438 | 2.203 | 0.256 |
| 20 | 15.76 | 7.692 | 1.441 | 0.599 | 9.601 | 9.601 | 3.440 | 1.303 | 0.635 | 0.457 | 1.303 | 0.093 | 3.440 | 0.176 | 1.141 | 0.093 |
| 30 | 10.65 | 5.177 | 0.686 | 0.382 | 7.235 | 7.235 | 2.718 | 0.754 | 0.385 | 0.301 | 0.754 | 0.052 | 2.718 | 0.101 | 0.826 | 0.052 |
| 40 | 9.842 | 4.804 | 0.761 | 0.301 | 6.910 | 6.910 | 2.565 | 0.538 | 0.288 | 0.256 | 0.538 | 0.042 | 2.565 | 0.085 | 0.698 | 0.042 |
| 50 | 6.132 | 2.978 | 0.464 | 0.199 | 4.572 | 4.572 | 1.858 | 0.452 | 0.225 | 0.169 | 0.452 | 0.026 | 1.858 | 0.052 | 0.539 | 0.026 |
| 100 | 2.115 | 1.005 | 0.222 | 0.118 | 1.861 | 1.861 | 0.960 | 0.226 | 0.132 | 0.099 | 0.226 | 0.013 | 0.960 | 0.029 | 0.013 | 0.013 |

$p = 4, \gamma = 0.7, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.184 | 0.183 | 0.183 | 0.183 | 0.183 | 0.183 | 0.071 | 0.159 | 0.103 | 0.063 | 0.159 | 0.159 | 0.133 | 0.103 | 0.063 | 0.159 |
| 20 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.060 | 0.060 | 0.047 | 0.054 | 0.060 | 0.060 | 0.062 | 0.047 | 0.054 | 0.060 |
| 30 | 0.379 | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.031 | 0.037 | 0.033 | 0.030 | 0.037 | 0.037 | 0.036 | 0.033 | 0.030 | 0.037 |
| 40 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 | 0.015 | 0.025 | 0.016 | 0.012 | 0.025 | 0.025 | 0.022 | 0.016 | 0.012 | 0.025 |
| 50 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.018 | 0.021 | 0.016 | 0.016 | 0.021 | 0.021 | 0.022 | 0.016 | 0.016 | 0.021 |
| 100 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.008 | 0.010 | 0.008 | 0.008 | 0.010 | 0.010 | 0.010 | 0.008 | 0.008 | 0.010 |

$p = 4, \gamma = 0.7, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1.482 | 0.847 | 0.190 | 0.171 | 0.664 | 0.664 | 0.133 | 0.140 | 0.104 | 0.105 | 0.140 | 0.124 | 0.234 | 0.103 | 0.111 | 0.124 |
| 20 | 0.359 | 0.223 | 0.082 | 0.079 | 0.221 | 0.221 | 0.081 | 0.067 | 0.059 | 0.060 | 0.067 | 0.052 | 0.117 | 0.053 | 0.068 | 0.052 |
| 30 | 0.143 | 0.097 | 0.051 | 0.051 | 0.123 | 0.123 | 0.052 | 0.043 | 0.038 | 0.039 | 0.043 | 0.035 | 0.084 | 0.036 | 0.043 | 0.035 |
| 40 | 0.089 | 0.057 | 0.027 | 0.027 | 0.078 | 0.078 | 0.028 | 0.022 | 0.018 | 0.018 | 0.022 | 0.021 | 0.049 | 0.016 | 0.021 | 0.021 |
| 50 | 0.071 | 0.049 | 0.026 | 0.027 | 0.066 | 0.066 | 0.030 | 0.023 | 0.020 | 0.020 | 0.023 | 0.019 | 0.047 | 0.018 | 0.024 | 0.019 |
| 100 | 0.030 | 0.021 | 0.012 | 0.012 | 0.029 | 0.029 | 0.015 | 0.011 | 0.010 | 0.010 | 0.011 | 0.009 | 0.023 | 0.009 | 0.012 | 0.009 |

$p = 4, \gamma = 0.7, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 5.564 | 3.088 | 0.392 | 0.359 | 2.278 | 2.278 | 0.329 | 0.327 | 0.185 | 0.188 | 0.327 | 0.139 | 0.701 | 0.151 | 0.233 | 0.139 |
| 20 | 1.185 | 0.646 | 0.121 | 0.114 | 0.673 | 0.673 | 0.160 | 0.145 | 0.097 | 0.093 | 0.145 | 0.057 | 0.288 | 0.075 | 0.121 | 0.057 |
| 30 | 0.455 | 0.266 | 0.083 | 0.087 | 0.383 | 0.383 | 0.119 | 0.091 | 0.063 | 0.060 | 0.091 | 0.037 | 0.225 | 0.048 | 0.083 | 0.037 |
| 40 | 0.291 | 0.167 | 0.047 | 0.050 | 0.253 | 0.253 | 0.077 | 0.055 | 0.036 | 0.033 | 0.055 | 0.022 | 0.150 | 0.025 | 0.050 | 0.022 |
| 50 | 0.210 | 0.124 | 0.040 | 0.043 | 0.191 | 0.191 | 0.069 | 0.047 | 0.034 | 0.031 | 0.047 | 0.019 | 0.126 | 0.024 | 0.046 | 0.019 |
| 100 | 0.089 | 0.053 | 0.019 | 0.020 | 0.085 | 0.085 | 0.037 | 0.024 | 0.018 | 0.016 | 0.024 | 0.010 | 0.063 | 0.012 | 0.024 | 0.010 |

$p = 4, \gamma = 0.7, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 129.0 | 70.10 | 6.667 | 6.191 | 50.63 | 50.63 | 6.225 | 6.286 | 2.659 | 2.616 | 6.286 | 0.568 | 14.72 | 1.587 | 3.912 | 0.568 |
| 20 | 28.52 | 14.95 | 1.624 | 1.495 | 15.52 | 15.52 | 2.851 | 2.599 | 1.390 | 1.208 | 2.599 | 0.209 | 5.958 | 0.785 | 1.921 | 0.209 |
| 30 | 10.56 | 5.809 | 1.060 | 1.178 | 8.725 | 8.725 | 2.214 | 1.668 | 0.849 | 0.709 | 1.668 | 0.096 | 4.821 | 0.410 | 1.304 | 0.096 |
| 40 | 6.643 | 3.564 | 0.639 | 0.723 | 5.760 | 5.760 | 1.634 | 1.154 | 0.629 | 0.518 | 1.154 | 0.056 | 3.380 | 0.292 | 0.978 | 0.056 |
| 50 | 4.868 | 2.686 | 0.489 | 0.552 | 4.413 | 4.413 | 1.386 | 0.969 | 0.517 | 0.408 | 0.969 | 0.042 | 2.817 | 0.225 | 0.810 | 0.042 |
| 100 | 1.966 | 1.067 | 0.213 | 0.245 | 1.879 | 1.879 | 0.727 | 0.440 | 0.253 | 0.199 | 0.440 | 0.016 | 1.365 | 0.103 | 0.423 | 0.016 |

$p = 4, \gamma = 0.8, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.170 | 0.169 | 0.169 | 0.169 | 0.169 | 0.169 | 0.192 | 0.169 | 0.190 | 0.193 | 0.169 | 0.169 | 0.193 | 0.190 | 0.193 | 0.169 |
| 20 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.064 | 0.063 | 0.064 | 0.064 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 |
| 30 | 0.040 | 0.039 | 0.039 | 0.039 | 0.039 | 0.039 | 0.036 | 0.038 | 0.037 | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.037 | 0.038 |
| 40 | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 | 0.016 | 0.026 | 0.017 | 0.013 | 0.026 | 0.026 | 0.023 | 0.017 | 0.013 | 0.026 |
| 50 | 0.021 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.021 | 0.021 | 0.021 | 0.021 | 0.021 | 0.020 | 0.021 | 0.021 | 0.020 |
| 100 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.009 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |

$p = 4, \gamma = 0.8, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 12.04 | 6.365 | 0.341 | 0.289 | 3.308 | 3.308 | 0.264 | 0.217 | 0.213 | 0.223 | 0.217 | 0.183 | 0.409 | 0.205 | 0.234 | 0.183 |
| 20 | 0.367 | 0.241 | 0.111 | 0.110 | 0.280 | 0.280 | 0.102 | 0.087 | 0.080 | 0.082 | 0.087 | 0.067 | 0.154 | 0.076 | 0.088 | 0.067 |
| 30 | 0.184 | 0.121 | 0.059 | 0.059 | 0.154 | 0.154 | 0.059 | 0.047 | 0.044 | 0.044 | 0.047 | 0.038 | 0.094 | 0.041 | 0.049 | 0.038 |
| 40 | 0.113 | 0.071 | 0.029 | 0.029 | 0.097 | 0.097 | 0.030 | 0.023 | 0.018 | 0.019 | 0.023 | 0.024 | 0.058 | 0.017 | 0.022 | 0.022 |
| 50 | 0.112 | 0.073 | 0.035 | 0.035 | 0.098 | 0.098 | 0.037 | 0.029 | 0.028 | 0.028 | 0.029 | 0.023 | 0.061 | 0.026 | 0.031 | 0.023 |
| 100 | 0.040 | 0.027 | 0.015 | 0.015 | 0.038 | 0.038 | 0.018 | 0.013 | 0.012 | 0.012 | 0.013 | 0.010 | 0.028 | 0.011 | 0.014 | 0.010 |

$p = 4, \gamma = 0.8, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 47.61 | 24.49 | 0.661 | 0.495 | 11.43 | 11.43 | 0.485 | 0.384 | 0.514 | 0.283 | 0.306 | 0.384 | 0.196 | 0.247 | 0.359 | 0.196 |
| 20 | 1.295 | 0.756 | 0.182 | 0.196 | 0.936 | 0.936 | 0.198 | 0.154 | 0.109 | 0.110 | 0.154 | 0.069 | 0.419 | 0.090 | 0.139 | 0.069 |
| 30 | 0.653 | 0.385 | 0.101 | 0.109 | 0.532 | 0.532 | 0.129 | 0.093 | 0.064 | 0.063 | 0.093 | 0.040 | 0.282 | 0.050 | 0.085 | 0.040 |
| 40 | 0.378 | 0.215 | 0.055 | 0.059 | 0.322 | 0.322 | 0.083 | 0.057 | 0.035 | 0.033 | 0.057 | 0.023 | 0.182 | 0.024 | 0.050 | 0.023 |
| 50 | 0.374 | 0.214 | 0.056 | 0.059 | 0.319 | 0.319 | 0.078 | 0.053 | 0.037 | 0.037 | 0.053 | 0.023 | 0.176 | 0.030 | 0.050 | 0.023 |
| 100 | 0.131 | 0.077 | 0.023 | 0.025 | 0.123 | 0.123 | 0.041 | 0.026 | 0.018 | 0.017 | 0.026 | 0.010 | 0.083 | 0.013 | 0.025 | 0.010 |

$p = 4, \gamma = 0.8, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1190 | 613 | 10.40 | 6.984 | 292.0 | 292.0 | 7.455 | 5.690 | 2.463 | 2.868 | 5.690 | 0.582 | 21.58 | 1.507 | 4.237 | 0.582 |
| 20 | 30.56 | 16.70 | 2.512 | 3.022 | 21.69 | 21.69 | 3.336 | 2.627 | 1.147 | 1.041 | 2.627 | 0.169 | 8.985 | 0.588 | 1.850 | 0.169 |
| 30 | 14.29 | 7.724 | 1.260 | 1.633 | 11.44 | 11.44 | 2.190 | 1.643 | 0.703 | 0.606 | 1.643 | 0.083 | 5.743 | 0.328 | 1.159 | 0.083 |
| 40 | 8.662 | 4.724 | 0.760 | 0.890 | 7.315 | 7.315 | 1.701 | 1.283 | 0.579 | 0.473 | 1.283 | 0.056 | 4.022 | 0.257 | 0.933 | 0.056 |
| 50 | 9.112 | 4.978 | 0.698 | 0.911 | 7.735 | 7.735 | 1.481 | 0.924 | 0.394 | 0.356 | 0.924 | 0.040 | 4.065 | 0.176 | 0.730 | 0.040 |
| 100 | 3.065 | 1.679 | 0.278 | 0.330 | 2.869 | 2.869 | 0.816 | 0.458 | 0.219 | 0.180 | 0.458 | 0.015 | 1.889 | 0.084 | 0.409 | 0.015 |

$p = 4, \gamma = 0.9, \sigma = 0.01$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.032 | 0.037 | 0.033 | 0.032 | 0.037 | 0.037 | 0.035 | 0.033 | 0.032 | 0.037 |
| 20 | 0.063 | 0.062 | 0.062 | 0.062 | 0.062 | 0.062 | 0.049 | 0.061 | 0.057 | 0.022 | 0.061 | 0.061 | 0.054 | 0.057 | 0.052 | 0.061 |
| 30 | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.038 | 0.032 | 0.037 | 0.033 | 0.032 | 0.037 | 0.037 | 0.035 | 0.033 | 0.032 | 0.037 |
| 40 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | 0.024 | 0.027 | 0.024 | 0.023 | 0.027 | 0.027 | 0.026 | 0.024 | 0.024 | 0.027 |
| 50 | 0.023 | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 | 0.015 | 0.021 | 0.017 | 0.015 | 0.021 | 0.021 | 0.019 | 0.017 | 0.015 | 0.021 |
| 100 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |

$p = 4, \gamma = 0.9, \sigma = 0.5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.308 | 0.184 | 0.065 | 0.066 | 0.247 | 0.247 | 0.053 | 0.040 | 0.036 | 0.037 | 0.040 | 0.035 | 0.118 | 0.035 | 0.041 | 0.035 |
| 20 | 0.579 | 0.344 | 0.112 | 0.113 | 0.419 | 0.419 | 0.085 | 0.070 | 0.063 | 0.064 | 0.070 | 0.059 | 0.171 | 0.060 | 0.070 | 0.059 |
| 30 | 0.308 | 0.184 | 0.065 | 0.066 | 0.247 | 0.247 | 0.053 | 0.040 | 0.036 | 0.037 | 0.040 | 0.035 | 0.118 | 0.035 | 0.041 | 0.035 |
| 40 | 0.171 | 0.105 | 0.043 | 0.044 | 0.144 | 0.144 | 0.042 | 0.030 | 0.027 | 0.028 | 0.030 | 0.026 | 0.080 | 0.025 | 0.032 | 0.026 |
| 50 | 0.166 | 0.099 | 0.032 | 0.033 | 0.138 | 0.138 | 0.030 | 0.021 | 0.019 | 0.019 | 0.021 | 0.020 | 0.071 | 0.018 | 0.022 | 0.020 |
| 100 | 0.069 | 0.043 | 0.018 | 0.018 | 0.063 | 0.063 | 0.019 | 0.013 | 0.012 | 0.012 | 0.013 | 0.010 | 0.040 | 0.011 | 0.014 | 0.010 |

$p = 4, \gamma = 0.9, \sigma = 1$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1.149 | 0.651 | 0.123 | 0.147 | 0.900 | 0.900 | 0.121 | 0.091 | 0.051 | 0.051 | 0.091 | 0.036 | 0.383 | 0.041 | 0.070 | 0.036 |
| 20 | 2.099 | 1.170 | 0.210 | 0.248 | 1.500 | 1.500 | 0.179 | 0.136 | 0.083 | 0.087 | 0.136 | 0.061 | 0.530 | 0.070 | 0.112 | 0.061 |
| 30 | 1.149 | 0.651 | 0.123 | 0.147 | 0.900 | 0.900 | 0.121 | 0.091 | 0.051 | 0.051 | 0.091 | 0.036 | 0.383 | 0.041 | 0.070 | 0.036 |
| 40 | 0.600 | 0.336 | 0.073 | 0.084 | 0.490 | 0.490 | 0.092 | 0.067 | 0.039 | 0.039 | 0.067 | 0.026 | 0.241 | 0.030 | 0.055 | 0.026 |
| 50 | 0.600 | 0.335 | 0.063 | 0.071 | 0.494 | 0.494 | 0.074 | 0.045 | 0.027 | 0.028 | 0.045 | 0.020 | 0.235 | 0.022 | 0.040 | 0.020 |
| 100 | 0.236 | 0.132 | 0.029 | 0.035 | 0.213 | 0.213 | 0.045 | 0.026 | 0.016 | 0.016 | 0.026 | 0.010 | 0.125 | 0.013 | 0.024 | 0.010 |

$p = 4, \gamma = 0.9, \sigma = 5$

| $n$ | OLS | HK | K1 | K2 | S3 | S4 | KM1 | KM2 | KM4 | KM5 | KM6 | KM8 | KM9 | KM10 | K12 | K12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 27.93 | 15.14 | 1.876 | 2.760 | 21.46 | 21.46 | 2.269 | 1.690 | 0.509 | 0.507 | 1.690 | 0.068 | 8.729 | 0.246 | 0.987 | 0.068 |
| 20 | 51.09 | 27.72 | 3.112 | 35.62 | 35.62 | 3.155 | 2.742 | 8.514 | 0.835 | 0.820 | 2.742 | 0.129 | 11.74 | 0.416 | 1.478 | 0.129 |
| 30 | 27.93 | 15.14 | 1.876 | 2.760 | 21.46 | 21.46 | 2.269 | 1.690 | 0.509 | 0.507 | 1.690 | 0.068 | 8.729 | 0.246 | 0.987 | 0.068 |
| 40 | 9.842 | 4.804 | 0.761 | 0.301 | 6.910 | 6.910 | 2.565 | 0.538 | 0.288 | 0.256 | 0.538 | 0.042 | 2.565 | 0.085 | 0.698 | 0.042 |
| 50 | 14.84 | 8.068 | 0.999 | 1.504 | 12.04 | 12.04 | 1.516 | 1.023 | 0.314 | 0.301 | 1.023 | 0.035 | 5.570 | 0.139 | 0.635 | 0.035 |
| 100 | 5.757 | 3.139 | 0.420 | 0.599 | 5.195 | 5.195 | 0.913 | 0.432 | 0.165 | 0.158 | 0.432 | 0.015 | 2.967 | 0.068 | 0.369 | 0.015 |

SORT 36 (2) July-December 2012, 139-152

# A note on the use of supply-use tables in impact analyses*

Manfred Lenzen[1] and José M. Rueda-Cantuche[2]

**Abstract**

Little attention has so far been paid to the problems inherent in interpreting the meaning of results from standard impact analyses using symmetric input-output tables. Impacts as well as drivers of these impacts must be either of the product type or of the industry type. Interestingly, since supply-use tables distinguish products and industries, they can cope with product impacts driven by changes in industries, and vice versa. This paper contributes in two ways. Firstly, the demand-driven Leontief quantity model, both for industry-by-industry as well as for product-by-product tables, is formalised on the basis of supply-use tables, thus leading to impact multipliers, both for industries and products. Secondly, we demonstrate how the supply-use formulation can improve the incorporation of disparate satellite data into input-output models, by offering both industry and product representation. Supply-use blocks can accept any mix of industry and product satellite data, as long as these are not overlapping.

*MSC:* 91B76–Environmental economics (natural resource models, harvesting, pollution, etc.)

*Keywords:* Technology assumptions, supply-use framework, multipliers.

## 1. Background

An input-output matrix of technical coefficients (**A**) generally depicts either the direct requirements of commodity $i$ needed to produce one physical unit of commodity $j$ or, alternatively, the direct inputs from industry $i$ needed to produce one physical unit of industry $j$. The former is built up with a product-by-product input-output table and the latter, with an industry-by-industry input-output table. Both are called symmetric input-

*The views expressed in this paper are those of the authors and should not be attributed to the European Commission or its services.

[1] ISA, School of Physics A28. The University of Sydney NSW 2006, Australia.
E-mail: m.lenzen@physics.usyd.edu.au

[2] European Commission, Joint Research Centre, Institute for Prospective Technological Studies (IPTS) and Pablo de Olavide University at Seville. E-mail: jmruecan@upo.es

Received: June 2011
Accepted: April 2012

output tables for having products or industries in both rows and columns; and the choice between them has only been addressed so far in detail by Rueda-Cantuche (2011).

Before the SNA-68 (UN, 1968), national statistical institutes almost exclusively constructed industry-by-industry input-output tables instead of commodity-by-commodity tables[1] and they used to set up the so-called transaction tables (ten Raa, 1994). In such tables, each element displayed the input requirements of industry *i* per unit of industry *j*'s production, as well as the final demand compartments (household and government consumption, investment and exports net of imports). Ten Raa (1994) noted that an input-output transaction table reduced the construction of a matrix of technical coefficients **A** just to a matter of dividing each element by their corresponding total output.

However, there were three different problems identified here. Firstly, products and industries cannot always be classified in the same way. Secondly, in addition to a multitude of inputs, industries may also have a multitude of outputs. Thirdly, products contained in each row and column of an industry-by-industry table are not homogeneous in terms of production (see e.g. Rainer, 1989).

To address these complications, the Systems of National Accounts proposed by the United Nations (1968, 1993), first established the concepts of use and make matrices within an input-output framework. Demand (use) and supply (make) of commodities were described by industries. This new framework provided a more accurate description of product flows and at the same time, made economists face a new problem in the construction of technical coefficients. Basically, the construction of technical coefficients was reduced to a matter of treatment of secondary products. Many establishments produce only one group of commodities, which are the primary products of the industry to which they are classified. However, some establishments produce commodities that are not among the primary products of the industry to which they belong. As a result, non-zero off-diagonal elements would appear in the make matrix. Alternative treatments of secondary products rest upon the separation of outputs and inputs associated with secondary products so that they can be added to the outputs and inputs of the industry in which the secondary product is a characteristic output. Assumptions on these inputs structures imply an **A**-matrix of technical coefficients as a function of the use and make matrices. The reader should be aware that a make matrix (industry by product) is merely the transposition of a supply matrix (product by industry) and we may use both indistinctly.

The matrix of technical coefficients has been used for economic analysis by means of the so-called Leontief quantity model and the Leontief price model, which are based on the following two equations: $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{y}$ and $\mathbf{p} = \mathbf{p}\mathbf{A} + \mathbf{v}$.

Here, **x** is a column vector of total output; **y**, a column vector of final demand; **p**, a row vector of prices; and **v**, a row vector of value-added coefficients. The standard

[1] In what follows, we will refer to the "commodity-by-commodity input-output tables" and "product-by-product input-output tables" as fully equivalent.

Leontief quantity model would be given by $\mathbf{x} = (\mathbf{I} - \mathbf{A}\ )^{-1}\mathbf{y}$, and the standard Leontief price model as $\mathbf{p}\ = \mathbf{v}(\mathbf{I} - \mathbf{A}\ )^{-1}$.

The first equation is used for national or regional economic planning; in accordance, there will be a direct effect over the output levels which will depend on the final demand variations ($\Delta\mathbf{y}$) and additional indirect effects that will be determined by the so called Leontief inverse matrix, $(\mathbf{I} - \mathbf{A})^{-1}$. The second equation can be used to assess the price effects resulting from an energy shock, which surely will bring about variations in the value-added shares of a product, to mention an example. For the time being and for the sake of clarification, we have deliberately omitted the identification of the outputs as products or industries. We will introduce this distinction later on.

Within this context, two major trade-offs were recently indentified concerning the choice of type of symmetric input-output tables to be used in input-output analyses (Rueda-Cantuche, 2011). The main limitation of these tables relates to their underlying symmetry, which implies that they must be defined as either product-by-product or industry-by-industry.

On the one hand, the Leontief quantity model, which is driven by demand for products, presents a trade-off whenever the impact analysis relates to external accounts (environment, employment, etc) pre-multiplying the Leontief inverse matrix and which are only available at industry detail. Then, either one could incorrectly assume that these external accounts reflect product detail, and employ a product–by-product input-output table in order to assess the effects of a unit change in final demand of a single product; or, alternatively, correctly take the external accounts as industry-specific information and use industry-by-industry tables. The latter practice, however, would preclude calculating effects of changes in the final demand of single products, because in an industry-by-industry table, final demand only exists as mixed bundles of goods and services produced by particular industries.

On the other hand, the Leontief price model, which is driven by industry supply, imposes trade-offs whenever the impact analysis relates to external accounts that are only available at industry detail. In this case, key questions such as the fuel price effects generated by an increase in the labour costs of the petroleum refining industry cannot really be answered by input-output price models as it may be generally thought. Either one could incorrectly assume that variations in the primary costs (labour) happen within homogenous branches of activity rather than in industries and thus, employ product-by-product tables or instead, one could correctly assume that price changes of labour costs effectively occur within entire industries and therefore, use industry-by-industry tables. In the latter case, the reported price impacts will refer to the fuel industry rather than to the fuel product itself.

Rueda-Cantuche (2011) proposed the use of supply and use tables instead of input-output tables for resolving the different trade-offs efficiently. Indeed, supply and use tables are defined and compiled at product-by-industry detail and do not require the symmetries causing the trade-offs described above. However, this author did not go beyond the mere statement and discussion of the convenience of extending the use of supply and use tables in input-output analysis. Therefore, this paper is aimed at formalizing Rueda-Cantuche's argument concerning the trade-offs that he identified.

In the next section, we will introduce the basics of the construction of symmetric input-output tables, which will be further described under a common schematic representation in Section 3. In what follows, Section 4 will generalize the calculation of impact multipliers for industries and commodities, separately. Section 5 describes the main empirical findings for the Brazilian economy in 2005 and finally, the last section will draw the main conclusions of this paper.

## 2. Introduction

Amongst other textbooks, the *United Nations Handbook on Input-Output Table Compilation* (UN 1999) distinguishes two basic technology assumptions for the construction of symmetric product-by-product input-output tables: in the *industry technology assumption*, the production recipe is unique to an industry, while products' input recipe is a weighted sum over industries' production recipes; in the *commodity technology assumption*[2], the input recipe is unique to a product, while industries' production recipes are a weighted sum over their primary and secondary outputs.

In practice, both assumptions are known to have drawbacks: Applying the commodity technology assumption can lead to negative elements during table construction, and requires the supply matrix to be square, which could lead to loss of detail in rectangular accounts.[3] The commodity technology assumption has proven to be theoretically superior while the industry technology assumption has been shown to be implausible (Kop Jansen and Ten Raa 1990). Comparative advantages of these perspectives are however not the concern of this work.

The construction of industry-by-industry input-output tables[4] requires two main assumptions stating that when product output is translated into industry output, the pattern

[2] In what follows, we will refer to the "commodity technology assumption" and the "product technology assumption" as fully equivalent.

[3] Konijn and Steenge (1995), Almon (2000), Bohlin and Widell (2006) and Smith and McDonald (2011) suggest ways of getting around the problems associated with the technology assumption. Konijn and Steenge (1995) suggest an input allocation procedure that uses activities of industries in their production of products. However, the data necessary to make this method operational are generally not available. Almon (2000) suggests a balancing algorithm that explicitly deals with cases where the subtraction of inputs of from a secondary production recipe would generate negative entries. Bohlin and Widell (2006) (extended by Smith and McDonald, 2011) apply an optimisation calculus, where they define the technology assumptions in terms of process coefficients that are both industry- and product-specific (see Ten Raa and Rueda-Cantuche 2007), and then minimize the variance of these process coefficients subject to summation rules.

[4] In this context, Yamano and Ahmad (2006) argue that the "description of the conversion (industry-technology assumption) is inaccurate where industry-by-industry tables are concerned, and is better described as *a fixed product sales structure* assumption. In other words the conversion merely assumes that the proportion of domestically produced commodity A bought by industry B from industry C is proportional to industry C's share of the total (domestic) economy production of commodity A. Put this way, it is clear that this is a far less demanding assumption than that implied by the equivalent, but differently named, "industry technology" assumption.

of sales will remain the same. This is the so-called sales structure approach that only admits two options: 1) where industry supply is independent of the products delivered (*fixed industry sales structure*), and 2) where industry supply is independent of the producing industry (*fixed product sales structure*). Employing arguments similar to those used in discussing the industry and product technology assumptions for the construction of product-by-product input-output tables (Ten Raa and Rueda-Cantuche 2007), Rueda-Cantuche and Ten Raa (2009) proved that the fixed industry sales approach is theoretically superior.

Notwithstanding the above theoretical considerations, statistical offices construct national input-output tables based on hybrid technology or combined fixed sales structure assumptions. So, in what follows, we will simply take what statistical office publish as given, and start with a formulation of their different assumptions using a supply-use framework; then, we will show how the supply-use blocks can be useful in simultaneously generating multipliers both for industries and for products (and thus solving the trade-offs caused by the symmetry of input-output tables).

We will show in the next sections how the industry technology and the fixed product sales structure assumptions can be jointly formulated in a common framework that allows carrying out impact analyses simultaneously in terms of products and industries. The same will apply for the product technology and the fixed industry sales structure assumptions.

## 3. Schematic representation of the assumptions made in the construction of input-output tables

In the *United Nations Handbook on Input-Output Table Compilation* (UN 1999) and the *Eurostat Manual of Supply, Use and Input-Output Tables* (Eurostat, 2008), there are various assumptions to be used for the construction of industry by industry or product-by-product symmetric input-output tables.[5] In the following, we will show that at least the ones referred to in the last section can be represented in one unified supply-use formulation. We will use the standard Eurostat Manual notation (UN 1999). Notice that the supply matrix, which we will denote $\mathbf{V}^\mathsf{T}$ corresponds to the transposition of the so called "make matrix".

[5] A number of authors suggest further alternatives such as mixed technology and activity technology assumptions (see Gigantes (1970); Schinnar (1978); Konijn and Steenge (1995) but for a comprehensive list, see also Ten Raa and Rueda-Cantuche (2003)) For a generalized formulation of the industry and product technology assumptions, see Ten Raa and Rueda-Cantuche (2007).

### 3.1. Industry-related assumptions

Let a single-region *supply-use transaction block* $\mathbf{T}$ be represented by:

$$\mathbf{T}=\begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \quad , \tag{1}$$

with $\mathbf{U}$ being a product-by-industry *use* matrix, showing the input $U_{ij}$ of product $i$ into industry $j$, and $\mathbf{V}$ being a industry-by-product *make* matrix, with $V_{ij}$ showing the output by industry $i$ of product $j$. This block formulation is well known in the input-output literature[6] and it is already proposed as an example by Eurostat (2008).

Let $\mathbf{T}$ satisfy the national accounting identity:

$$\begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{e}_c \\ \mathbf{e}_i \end{bmatrix} + \begin{bmatrix} \mathbf{y}_c \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{q} \\ \mathbf{g} \end{bmatrix} \quad , \tag{2}$$

where $[\mathbf{e}_c\,\mathbf{e}_i]^{\mathsf{T}}$ is the row summation vector formed by two summation sub-vectors corresponding to commodities ($\mathbf{e}_c$) and industries ($\mathbf{e}_i$), superscript $^{\mathsf{T}}$ denotes transposition, $\mathbf{y}_c$ is a vector of *final demand* of products, and $\mathbf{q}$ and $\mathbf{g}$ are vectors of *total product and industry outputs*, respectively. Equation 2 includes the product balance $\mathbf{U}\mathbf{e}_i + \mathbf{y}_c = \mathbf{q}$, and the industry balance $\mathbf{V}\mathbf{e}_c = \mathbf{g}$. Therefore, it can be transformed into:

$$\begin{aligned}
&\begin{bmatrix} \mathbf{y}_c \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{q} \\ \mathbf{g} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{e}_c \\ \mathbf{e}_i \end{bmatrix} = \left\{ \begin{bmatrix} \hat{\mathbf{q}} & 0 \\ 0 & \hat{\mathbf{g}} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \right\} \begin{bmatrix} \mathbf{e}_c \\ \mathbf{e}_i \end{bmatrix} = \\
&\left\{ \begin{bmatrix} \hat{\mathbf{q}} & 0 \\ 0 & \hat{\mathbf{g}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{q}}^{-1} & 0 \\ 0 & \hat{\mathbf{g}}^{-1} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{q}}^{-1} & 0 \\ 0 & \hat{\mathbf{g}}^{-1} \end{bmatrix} \right\} \begin{bmatrix} \hat{\mathbf{q}} & 0 \\ 0 & \hat{\mathbf{g}} \end{bmatrix} \begin{bmatrix} \mathbf{e}_c \\ \mathbf{e}_i \end{bmatrix} = \\
&\left\{ \begin{bmatrix} \mathbf{I}_c & 0 \\ 0 & \mathbf{I}_i \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{q}}^{-1} & 0 \\ 0 & \hat{\mathbf{g}}^{-1} \end{bmatrix} \right\} \begin{bmatrix} \mathbf{q} \\ \mathbf{g} \end{bmatrix} \\
&\Leftrightarrow \begin{bmatrix} \mathbf{q} \\ \mathbf{g} \end{bmatrix} = \left\{ \mathrm{I} - \begin{bmatrix} 0 & \mathbf{U} \\ \mathbf{V} & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{q}}^{-1} & 0 \\ 0 & \hat{\mathbf{g}}^{-1} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{y}_c \\ 0 \end{bmatrix} = \left\{ \mathbf{I} - \begin{bmatrix} 0 & \mathbf{B} \\ \mathbf{D} & 0 \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{y}_c \\ 0 \end{bmatrix}
\end{aligned} \quad , \tag{3}$$

where $\mathbf{D}$ and $\mathbf{B}$[7] form the *supply* and *use coefficient matrices*, $\mathbf{I}$ is an identity matrix, and the hat symbol (ˆ) denotes a diagonalised vector. $\mathbf{B} = \mathbf{U}\hat{\mathbf{g}}^{-1}$ is called the (product-by industry) *use coefficients* matrix (input structures), and $\mathbf{D} = \mathbf{V}\hat{\mathbf{q}}^{-1}$ is called the (industry-by-product) *market share* matrix.

[6] Note that the supply-use-block formulation requires the make matrix $\mathbf{V}$ to be defined as industry-by-product, and not as product-by-industry.

[7] Our B matrix is equivalent to the Z matrix in Eurostat (2008).

Using the supply-use-block formulation in Equation (3), a compound Leontief inverse can be written as:

$$\mathrm{L}_{\mathrm{I}}^{*}=\begin{bmatrix}\mathbf{I} & -\mathbf{B}\\ -\mathbf{D} & \mathbf{I}\end{bmatrix}^{-1} \quad . \tag{4}$$

Applying the partitioned inverse of Miyazawa (1968), Equation (4) can be written as:

$$\mathbf{L}_{\mathrm{I}}^{*}=\begin{bmatrix}\mathbf{I}+\mathbf{BL}_{\mathrm{I,ii}}\mathbf{D} & \mathbf{BL}_{\mathrm{I,ii}}\\ \mathbf{L}_{\mathrm{I,ii}}\mathbf{D} & \mathbf{L}_{\mathrm{I,ii}}\end{bmatrix} \quad , \tag{5}$$

where $\mathbf{L}_{\mathrm{I,ii}}=(\mathbf{I}-\mathbf{DB})^{-1}$ is precisely the Leontief inverse of the industry-by-industry type of a technical coefficient matrix constructed on the basis of the *fixed product sales structure* (see Eurostat, 2008, p. 349). Considering the series expansion of $\mathbf{L}_{\mathrm{I,ii}}=(\mathbf{I}+\mathbf{DB}+(\mathbf{DB})(\mathbf{DB})+\cdots)$, we find:

$$\begin{aligned}\mathbf{BL}_{\mathrm{I,ii}}\mathbf{D}=\mathbf{B}\left(\mathbf{I}+\mathbf{DB}+(\mathbf{DB})^{2}+\cdots\right)\mathbf{D}=\mathbf{BD}+\mathbf{B}(\mathbf{DB})\mathbf{D}+\mathbf{B}(\mathbf{DBDB})\mathbf{D}+\cdots=\\ =\mathbf{BD}+(\mathbf{BD})(\mathbf{BD})+(\mathbf{BD})(\mathbf{BD})(\mathbf{BD})+\cdots,\end{aligned}$$

which leads to:

$$\mathbf{I}+\mathbf{BL}_{\mathrm{I,ii}}\mathbf{D}=\mathbf{I}+\mathbf{BD}+(\mathbf{BD})(\mathbf{BD})+\cdots=(\mathbf{I}-\mathbf{BD})^{-1}=\mathbf{L}_{\mathrm{I,cc}},$$

and which is identical to the series expansion of the Leontief inverse of a product-by-product type technical coefficients matrix constructed with the *industry technology model* (see Eurostat, (2008), p. 349). Equation (5) can be simplified to

$$\mathbf{L}_{\mathrm{I}}^{*}=\begin{bmatrix}\mathbf{L}_{\mathrm{I,cc}} & \mathbf{L}_{\mathrm{I,cc}}\mathbf{B}\\ \mathbf{L}_{\mathrm{I,ii}}\mathbf{D} & \mathbf{L}_{\mathrm{I,ii}}\end{bmatrix}. \tag{6}$$

Regarding the off-diagonal elements, the reader may find easily that $\mathbf{L}_{\mathrm{I,ii}}\mathbf{D}=\mathbf{DL}_{\mathrm{I,cc}}$ and $\mathbf{BL}_{\mathrm{I,ii}}=\mathbf{L}_{\mathrm{I,cc}}\mathbf{B}$. The matrices of market shares $\mathbf{D}$ and of input structures $\mathbf{B}$ are clearly used to convert the resulting impacts of industries into those of products, and impacts of products into those of industries, respectively.

Hence, and this is the first result of this paper, when supply and use matrices are handled under an integrated supply-use framework, the compound Leontief inverse elegantly reproduces the product-by-product type model under the industry technology assumption and the industry-by-industry model under the fixed product sales structure assumption.

### 3.2. Product-related assumptions

Product technology assumes an input recipe that is characteristic for a certain product. Here, we use the relationships $\mathbf{V}\mathbf{e}_c = \hat{\mathbf{g}}\mathbf{e}_i = \mathbf{g}$ and $\mathbf{V}^{\mathsf{T}}\mathbf{e}_i = \mathbf{q}$, and re-write the national accounting identity in Equation (2) as:

$$\begin{aligned}&\begin{bmatrix}0 & \mathbf{U}\\ \mathbf{V} & 0\end{bmatrix}\begin{bmatrix}\mathbf{e}_c\\ \mathbf{e}_i\end{bmatrix}+\begin{bmatrix}\mathbf{y}_c\\ 0\end{bmatrix}=\begin{bmatrix}\mathbf{U}\mathbf{e}_i+\mathbf{y}_c\\ \mathbf{V}\mathbf{e}_c\end{bmatrix}=\begin{bmatrix}\mathbf{U}\mathbf{e}_i+\mathbf{y}_c\\ \hat{\mathbf{g}}\mathbf{e}_i\end{bmatrix}=\\ &\qquad=\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\begin{bmatrix}\mathbf{e}_i\\ \mathbf{e}_i\end{bmatrix}+\begin{bmatrix}\mathbf{y}_c\\ 0\end{bmatrix}=\begin{bmatrix}\mathbf{V}^{\mathsf{T}}\mathbf{e}_i\\ \mathbf{g}\end{bmatrix}=\begin{bmatrix}\mathbf{q}\\ \mathbf{g}\end{bmatrix},\end{aligned} \tag{7}$$

Then, Equation (7) can be transformed into:

$$\begin{aligned}&\begin{bmatrix}\mathbf{y}_c\\ 0\end{bmatrix}=\begin{bmatrix}\mathbf{V}^{\mathsf{T}}\mathbf{e}_i\\ \mathbf{g}\end{bmatrix}-\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\begin{bmatrix}\mathbf{e}_i\\ \mathbf{e}_i\end{bmatrix}=\left\{\begin{bmatrix}\mathbf{V}^{\mathsf{T}} & 0\\ 0 & \hat{\mathbf{g}}\end{bmatrix}-\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\right\}\begin{bmatrix}\mathbf{e}_i\\ \mathbf{e}_i\end{bmatrix}=\\ &\left\{\begin{bmatrix}(\mathbf{V}^{\mathsf{T}}) & 0\\ 0 & \hat{\mathbf{g}}\end{bmatrix}\begin{bmatrix}(\mathbf{V}^{\mathsf{T}})^{-1} & 0\\ 0 & \hat{\mathbf{g}}^{-1}\end{bmatrix}-\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\begin{bmatrix}(\mathbf{V}^{\mathsf{T}})^{-1} & 0\\ 0 & \hat{\mathbf{g}}^{-1}\end{bmatrix}\right\}\begin{bmatrix}(\mathbf{V}^{\mathsf{T}}) & 0\\ 0 & \hat{\mathbf{g}}\end{bmatrix}\begin{bmatrix}\mathbf{e}_i\\ \mathbf{e}_i\end{bmatrix}=\\ &\left\{\begin{bmatrix}\mathbf{I}_i & 0\\ 0 & \mathbf{I}_i\end{bmatrix}-\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\begin{bmatrix}(\mathbf{V}^{\mathsf{T}})^{-1} & 0\\ 0 & \hat{\mathbf{g}}^{-1}\end{bmatrix}\right\}\begin{bmatrix}\mathbf{V}^{\mathsf{T}}\mathbf{e}_i\\ \mathbf{g}\end{bmatrix}\\ &\Leftrightarrow\begin{bmatrix}\mathbf{V}^{\mathsf{T}}\mathbf{e}_i\\ \mathbf{g}\end{bmatrix}=\left\{\mathbf{I}-\begin{bmatrix}0 & \mathbf{U}\\ \hat{\mathbf{g}} & 0\end{bmatrix}\begin{bmatrix}(\mathbf{V}^{\mathsf{T}})^{-1} & 0\\ 0 & \hat{\mathbf{g}}^{-1}\end{bmatrix}\right\}^{-1}\begin{bmatrix}\mathbf{y}_c\\ 0\end{bmatrix}=\left\{\mathbf{I}-\begin{bmatrix}0 & \mathbf{B}\\ \mathbf{C}^{-1} & 0\end{bmatrix}\right\}^{-1}\begin{bmatrix}\mathbf{y}_c\\ 0\end{bmatrix},\end{aligned} \tag{8}$$

where $\mathbf{C} \;= \mathbf{V}^{\mathsf{T}}\hat{\mathbf{g}}^{-1}$ and $\mathbf{B}$ form the *supply* and *use coefficients* blocks, respectively.

Using the supply-use-block formulation in Equation (8), a new compound Leontief inverse can be written as:

$$\mathbf{L}^{*}_{\mathrm{C}}=\begin{bmatrix}\mathbf{I} & -\mathbf{B}\\ -\mathbf{C}^{-1} & \mathbf{I}\end{bmatrix}^{-1}. \tag{9}$$

Applying the partitioned inverse of Miyazawa 1968), Equation (9) can be written as:

$$\mathbf{L}^{*}_{\mathrm{C}}=\begin{bmatrix}\mathbf{I}+\mathbf{B}\mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} & \mathbf{B}\mathbf{L}_{\mathrm{C,ii}}\\ \mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} & \mathbf{L}_{\mathrm{C,ii}}\end{bmatrix}, \tag{10}$$

where $\mathbf{L}_{\mathrm{C,ii}} = \left(\mathbf{I}-\mathbf{C}^{-1}\mathbf{B}\right)^{-1}$ is the Leontief inverse of the industry-by-industry technical coefficient matrix constructed on the basis of the *fixed industry sales structure* (Eurostat, (2008), p. 349). Considering the series expansion of $\mathbf{L}_{\mathrm{C,ii}} = \left(\mathbf{I}+\mathbf{C}^{-1}\mathbf{B}+\left(\mathbf{C}^{-1}\mathbf{B}\right)\left(\mathbf{C}^{-1}\mathbf{B}\right)+\cdots\right)$, we find:

$$\begin{aligned}\mathbf{B}\mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} &= \mathbf{B}\left(\mathbf{I}+\mathbf{C}^{-1}\mathbf{B}+\left(\mathbf{C}^{-1}\mathbf{B}\right)^2+\cdots\right)\mathbf{C}^{-1} = \\ &= \mathbf{B}\mathbf{C}^{-1}+\mathbf{B}\left(\mathbf{C}^{-1}\mathbf{B}\right)\mathbf{C}^{-1}+\mathbf{B}\left(\mathbf{C}^{-1}\mathbf{B}\right)\left(\mathbf{C}^{-1}\mathbf{B}\right)\mathbf{C}^{-1}+\cdots = \\ &= \mathbf{B}\mathbf{C}^{-1}+\left(\mathbf{B}\mathbf{C}^{-1}\right)\left(\mathbf{B}\mathbf{C}^{-1}\right)+\left(\mathbf{B}\mathbf{C}^{-1}\right)\left(\mathbf{B}\mathbf{C}^{-1}\right)\left(\mathbf{B}\mathbf{C}^{-1}\right)+\cdots\end{aligned}$$

and yields:

$$\mathbf{I}+\mathbf{B}\mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} = \mathbf{I}+\mathbf{B}\mathbf{C}^{-1}+\left(\mathbf{B}\mathbf{C}^{-1}\right)\left(\mathbf{B}\mathbf{C}^{-1}\right)+\cdots = \left(\mathbf{I}-\mathbf{B}\mathbf{C}^{-1}\right)^{-1} = \mathbf{L}_{\mathrm{C,cc}},$$

which is the series expansion of the Leontief inverse of a product-by-product technical coefficient matrix using the *product technology model* (see Eurostat, (2008), p. 349). Then, Equation (10) can be reduced to:

$$\mathbf{L}_{\mathrm{C}}^{*} = \begin{bmatrix} \mathbf{L}_{\mathrm{C,cc}} & \mathbf{L}_{\mathrm{C,cc}}\mathbf{B} \\ \mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} & \mathbf{L}_{\mathrm{C,ii}} \end{bmatrix}. \tag{11}$$

It is easy to show that the off-diagonal terms transform as $\mathbf{L}_{\mathrm{C,ii}}\mathbf{C}^{-1} = \mathbf{C}^{-1}\mathbf{L}_{\mathrm{C,cc}}$ and $\mathbf{B}\mathbf{L}_{\mathrm{C,ii}} = \mathbf{L}_{\mathrm{C,cc}}\mathbf{B}$. The matrices $\mathbf{C}^{-1}$ and $\mathbf{B}$ are used to convert impacts of industries into those of products, and impacts of products into those of industries, respectively.

Hence, as the second result of this paper, when supply and use matrices are integrated in a supply-use framework, the compound Leontief inverse elegantly reproduces the product-by-product type model assuming the product technology assumption and the industry-by-industry model assuming the fixed industry sales structure assumption.

It is interesting to note that these two models provide negative elements in the resulting technical coefficient matrices, while the models dealt with in the subsection 3.1 always provide non-negative terms.

## 4. Generalized input-output calculations

It was always the intention of Leontief to combine the input-output table with external, physical information, for example in order to examine questions relating to environmental impacts or the labour market (Leontief and Ford, 1970; Leontief and Duchin, 1986). Since Leontief's work there have been numerous publications of what Miller and Blair (2009) call *generalized* input-output analyses. For example, Kagawa and Suh (2009), and Suh *et al.* (2010) (see also references therein) use make and use matrices in environmental Life Cycle Assessment. When applied to a supply-use framework, the generalised calculus elegantly reproduces industry and product multipliers in one single shot. Assume for example that external physical information $\mathbf{f}_{\mathrm{i}}$ is available only at the industry level. Invoking the industry technology assumption, as in Equation (6), multipliers can be written as:

$$\left[\begin{array}{cc} 0 & \mathbf{f}_\mathrm{i} \end{array}\right] \mathbf{L}_\mathrm{I}^* = \left[\begin{array}{cc} \mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D} & \mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii} \end{array}\right], \tag{12}$$

with $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}$ representing industries, and $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D}$ representing commodities. This feature was applied in a generalized multi-region analyses of embodied $\mathrm{CO}_2$ for Denmark and its trading partners (Lenzen *et al.* 2004). Alternatively, if physical information $\mathbf{f}_\mathrm{c}$ is available only for products, multipliers are then defined as:

$$\left[\begin{array}{cc} \mathbf{f}_\mathrm{c} & 0 \end{array}\right] \mathbf{L}_\mathrm{I}^* = \left[\begin{array}{cc} \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc} & \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc}\,\mathbf{B} \end{array}\right]. \tag{13}$$

In a study on Brazil by Wachsmann *et al.* (2009), physical information on energy consumption was generally available for industries as $\mathbf{f}_\mathrm{i}$, but for some industries, detailed commodity information was available as $\mathbf{f}_\mathrm{c}$. Hence, a vector $\left[\begin{array}{cc} \mathbf{f}_\mathrm{c} & \mathbf{f}_\mathrm{i}^* \end{array}\right]$ was constructed with $\mathbf{f}_\mathrm{i}^*$ representing the industry data $\mathbf{f}_\mathrm{i}$ and setting the industries represented in $\mathbf{f}_\mathrm{c}$ to zero. The industry and product multipliers are then:

$$\left[\begin{array}{cc} \mathbf{f}_\mathrm{c} & \mathbf{f}_\mathrm{i}^* \end{array}\right] \mathbf{L}_\mathrm{I}^* = \left[\begin{array}{cc} \mathbf{f}_\mathrm{i}^*\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D} + \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc} & \mathbf{f}_\mathrm{i}^*\,\mathbf{L}_\mathrm{I,ii} + \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc}\,\mathbf{B} \end{array}\right]. \tag{14}$$

In Equations (12–14), the matrices $\mathbf{D}$ and $\mathbf{B}$ are used to convert industry data into product data ($\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}$), and vice versa ($\mathbf{f}_\mathrm{c}$). Similar relationships can be derived for models assuming product-related assumptions.

## 5. Empirical application

Provided rectangular supply-use frameworks with more products than industries, the calculation of total energy intensities (energy multipliers) can differentiate between products and industries, and thus add value over conventional mutlipliers based on input-output tables. Take for instance, the petrol and coke refining industry, which may produce petrol, fuel oil and diesel oil, amongst other products.

In order to prove the utility of supply-use tables in impact analysis and the theoretical framework presented before, we will run two experiments aiming to compare supply-use-based with input-output-based energy multipliers for the Brazilian economy in 2005. In particular, we will first determine simultaneously industry and commodity multipliers as in (12), assuming that energy data are only available at the industry level ($\mathbf{f}_\mathrm{i}$). Second, we will discuss the difference between supply-use-based commodity multipliers and input-output-based industry multipliers when a mix of energy industry data $\mathbf{f}_\mathrm{i}$ and energy commodity data $\mathbf{f}_\mathrm{c}$ is used.

The Brazilian supply-use tables for 2005 issued by the Instituto Brasileiro de Geografia e Estatística (IBGE, 2008) distinguish 110 commodities, but only 55 industries. Commodity detail is higher than industry detail especially for agriculture, food manufacturing, and refining. Whilst energy data are not available at the high commodity detail

***Table 1:*** *Energy multipliers (in units of kt oil equivalent per million 2005 Reais–ktoe/mR\$, where 1 ktoe/mR\$ = 102 terajoules per million US\$) for Brazilian cropping and forestry industries and commodities.*

| | **Industry data only** | **Industry & commodity data** |
|---|---|---|
| **SUT industry multipliers** | $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}$ | $\mathbf{f}_\mathrm{i}^*\,\mathbf{L}_\mathrm{I,ii} + \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc}\,\mathbf{B}$ |
| Cropping and forestry | 0.1999 | 0.1595 |
| Grazing and fishing | 0.1024 | 0.1051 |
| **SUT commodity multipliers** | $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D}$ | $\mathbf{f}_\mathrm{i}^*\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D} + \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc}$ |
| Rice in the husk | 0.1960 | 0.1136 |
| Corn | 0.1721 | 0.1114 |
| Wheat | 0.1999 | 0.1140 |
| Sugar cane | 0.1981 | 0.1138 |
| Soy beans | 0.1972 | 0.1136 |
| Other crops | 0.1910 | 0.1129 |
| Manioc | 0.1948 | 0.1135 |
| Tobacco | 0.1992 | 0.1139 |
| Cotton | 0.1956 | 0.1136 |
| Citrus fruit | 0.1974 | 0.1137 |
| Coffee | 0.1970 | 0.1137 |
| Forestry products | 0.1973 | 1.3943 |

for the agriculture and food manufacturing sectors, energy data on refining distinguishes diesel oil, fuel oil, petrol, and LPG (EPE 2011).

In a first experiment, we re-classified the Brazilian raw energy data into the 55-industry classification $\mathbf{f}_\mathrm{i}$. As (12) shows, supply-use-based and input-output-based industry multipliers are the same: $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}$. However, the supply-use framework allows the simultaneous determination of commodity multipliers $\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D}$ (see (12)). Except for wheat, which is solely produced by the 'Cropping and forestry' industry, commodity multipliers are lower than industry multipliers for all crops (Table 1, industry data only column). This is because some crops are partly produced by mixed-business broadacre farms in the less energy-intensive 'Grazing and fishing' sector. Such co-product detail is only available in supply-use tables, and lost in input-output tables. The error associated with this loss of detail is 16% for corn, and 1-3% for other crops.[8]

In a second experiment, we re-classified only the raw energy data for the petroleum and coke refining sector into the 110-commodity classification $\mathbf{f}_\mathrm{c}$, and deleted the entry for petroleum and coke refining in the industry data $\mathbf{f}_\mathrm{i}^*$. As Equation (14) shows, both supply-use-based industry and commodity multipliers are now different from input-output-based industry multipliers. Once again, the supply-use framework allows the simultaneous determination of commodity multipliers. Only now, the distinction of 'Forestry products' as an energy-intensive industry becomes apparent (Table 1, industry and commodity data column). This is because wood charcoal operations that are only

[8] Relative errors are calculated as $|\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D} - \mathbf{f}_\mathrm{i}^*\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D} + \mathbf{f}_\mathrm{c}\,\mathbf{L}_\mathrm{I,cc}|/\mathbf{f}_\mathrm{i}\,\mathbf{L}_\mathrm{I,ii}\,\mathbf{D}$, in this case $|0.1721 - 0.1114|/0.1721 \approx 16\%$.

***Table 2:*** *Energy multipliers (in units of kt oil equivalent per million 2005 Reais–ktoe/mR$) for Brazilian petroleum and coke refining industries and commodities.*

| | **Industry & commodity data** | |
|---|---|---|
| | IO industry multipliers | SUT multipliers |
| Petroleum and coke refining | 0.9859 | 0.8406 |
| LPG | | 1.1012 |
| Petrol | | 0.4727 |
| Gasoalcohol | | 0.4693 |
| Fuel oil | | 1.6410 |
| Diesel oil | | 1.2405 |
| Other petroleum and coke refining products | | 0.8804 |

part of the 'Forestry products' sub-sector consume much more energy than cropping. Once, again, such detail is only available in supply-use tables, since forestry and cropping is aggregated in input-output tables. The error associated with this aggregation is in the order of 80%.

Similar errors between 70% and 90% can be observed when comparing the one input-output-based multiplier for the 'Petroleum and coke refining' industry, and the supply-use-based commodity multipliers for the six refining sub-sectors (Table 2). Here, LPG, fuel oil, and diesel oil appear more energy-intensive than petrol and gasoalcohol, which once again cannot be discerned from input-output industry multipliers.

Our results bear significant implications for real-world policy. Assume for example that the Brazilian Government debated the impact of a 90 R$/toe energy tax (about 5% on top of the price of petrol, for example) on agricultural commodities, and in turn on different food products. Such a policy question would be rather mis-informed by any analysis using only industry-specific energy data (see Table 1). Opponents of such energy taxes could base their arguments on multipliers derived from industry data, and warn that if the government went ahead with the tax, households (who consumed 46.5 bR$ of crop sector output in 2008) would be short by 90 R$/toe × 0.19 ktoe/mR$ × 46.5 bR$ = 800 mR$. However, upon using mixed industry and commodity data in a SUT framework, it would become clear that some of this tax impact would in reality affect forestry products (charcoal), and not crop-based products, and that the real adverse impact on households would be significantly lower at 90 R$/toe × 0.11 ktoe/mR$ × 46.5 bR$ = 450 mR$.

## 6. Conclusions

We believe that the unnoticed drawback underlying the use of input-output tables in impact analyses is their symmetry, in the sense that they must be defined either on a product-by-product or on an industry-by-industry basis. Rueda-Cantuche (2011) identified two major trade-offs in the calculation of impact multipliers when using symmetric input-output tables. However, the author only stated that supply-use tables would overcome this undesirable effect but without formalising his argument. This note extends Rueda-Cantuche's reasoning and shows that the use of supply-use tables in a common framework concerning product- and industry-related assumptions may overcome the undesirable limitations of symmetric input-output tables. We show that the industry technology and the fixed product sales structure assumptions can be jointly formulated in a common framework that allows us to carry out impact analyses simultaneously in terms of products and industries. The same applies for the product technology and the fixed industry sales structure assumptions. As we have proven for the empirical example of Brazilian energy multipliers, using rectangular supply-use tables has significant advantages for real-world impact analyses whenever physical satellite data (environmental, socio-economic, tourism, etc.) are available.

## Acknowledgments

This work was supported in part by the Norwegian Research Council under its project 196090/V10, and in part by the Australian Research Council under its Discovery Project DP0985522. Richard Wood provided helpful comments on an early version of this paper. The views expressed in this paper belong to the authors and should not be attributed to the European Commission or its services.

## References

Almon, C. (2000). Product-to-product tables via product-technology with no negative flows. *Economic Systems Research*, 12, 27–43.

Bohlin, L. and Widell, L. M. (2006). Estimation of commodity-by-commodity input-output matrices. *Economic Systems Research*, 18, 205–215.

EPE (2011). *Balanço Energético Nacional 2010-Matrizes consolidadas*. Internet site http://ben.epe.gov.br, Empresa de Pesquisa Energética, Ministéria de Minas e Energia, Governo Federal.

Eurostat (2008). *Eurostat Manual of Supply, Use and Input-Output Tables*. Luxembourg, Luxembourg, European Commission.

Gigantes, T. (1970). *The Representation of Technology in Input-Output Systems*. Geneva, Switzerland, North-Holland Publishing Company.

IBGE (2008). *Matriz de Insumo-Produto Brasil 2000/2005*. Contas Nacionais número 23, Internet site http://www.ibge.gov.br/home/estatistica/economia/matrizinsumo_produto/publicacao.pdf, Rio de Janeiro, Brazil, Instituto Brasileiro de Geografia e Estatística, Ministério de Planejamento e Orçamento, Fundação IBGE, Diretoria de Pesquisa, Coordenação de Contas Nacionais.

Kagawa, S. and Suh, S. (2009). *Multistage Process-Based Make-Use System*. In: S. Suh (ed.) *Handbook of Input-Output Analysis in Industrial Ecology*, 777–800.

Konijn, P. J. A. and Steenge, A. E. (1995). Compilation of input-output tables from the national accounts. *Economic Systems Research*, 7, 31–45.

Kop Jansen, P. S. M. and Ten Raa, T. (1990). The choice of model in the construction of input-output coefficients matrices. *International Economic Review*, 31, 213–227.

Lenzen, M., L.-L. Pade and Munksgaard, J. (2004). $CO_2$ multipliers in multi-region input-output models. *Economic Systems Research*, 16, 391–412.

Leontief, W. and Duchin, F. (1986). *The Future Impact of Automation on Workers*. Oxford, UK, Oxford University Press.

Leontief, W. and Ford, D. (1970). Environmental repercussions and the economic structure: an input-output approach. *Review of Economics and Statistics*, 52, 262–271.

Miller, R. E. and Blair, P. D. (2009). *Input-Output Analysis: Foundations and Extensions*. Cambridge, UK, Cambridge University Press.

Miyazawa, K. (1968). Input-output analysis and interrelational income multiplier as a matrix. *Hitotsubashi Journal of Economics*, 18, 39–58.

Rainer, N. (1989). Descriptive versus analytical make-use systems: some Austrian experiences, in: R. Miller, K. Polenske & A. Z. Rose (eds) *Frontiers of Input–Output Analysis* (New York, Oxford University Press).

Rueda-Cantuche, J. M. (2011). The choice of type of input-output table revisited: moving towards the use of supply-use tables in impact analysis. *Statistics and Operations Research Transactions*, 35, 21–38.

Rueda-Cantuche, J. M. and Ten Raa, T. (2009). The choice of model in the construction of industry coefficients matrices. *Economic Systems Research*, 21, 363–376.

Schinnar, A. P. (1978). A method for computing Leontief multipliers from rectangular input-output accounts. *Environment and Planning A*, 10, 137–143.

Smith, N. and McDonald, G. (2011). Estimation of symmetric input-output tables: An extension to Bohlin and Widell. *Economic Systems Research*, 23.

Suh, S., B. Weidema, J. H. Schmidt and Heijungs, R. (2010). Generalized make and use framework for allocation in life cycle assessment. *Journal of Industrial Ecology*, 14, 335–353.

Ten Raa, T. (1994). On the methodology of input–output analysis, *Regional Science and Urban Economics*, 24, 3–27.

Ten Raa, T. and Rueda-Cantuche, J. M. (2003). The construction of input-output coefficients matrices in an axiomatic context: some further considerations, *Economic Systems Research*, 15, 439–455.

Ten Raa, T. and Rueda-Cantuche, J. M. (2007). A generalized expression for the commodity and the industry technology models in input-output analysis. *Economic Systems Research*, 19, 99–104.

United Nations (1968). *A System of National Accounts*, Studies in Methods Series F, nr. 2, rev. 3. New York, USA, United Nations.

United Nations (1993). *Revised System of National Accounts*, Studies in Methods Series F, no. 2, rev. 4. New York, USA, United Nations.

United Nations Department for Economic and Social Affairs Statistics Division (1999). *Handbook of Input-Output Table Compilation and Analysis*. New York, USA, United Nations.

Wachsmann, U., R. Wood, M. Lenzen and Schaeffer, R. (2009). Structural decomposition of energy use in Brazil from 1970 to 1996. *Applied Energy*, 86, 578–587.

Yamano, N. and Ahmad, N. (2006). *The OECD input-output database: 2006 edition*. STI Working Paper 2006/8, DSTI/DOC(2006)8, Paris, France, Directorate for Science, Technology and Industry, Organisation for Economic Co-operation and Development.

SORT 36 (2) July-December 2012, 153-180

# The new class of Kummer beta generalized distributions

R. R. Pescim[1], G. M. Cordeiro[2], C. G. B. Demétrio[3],
E. M. M. Ortega[4] and S. Nadarajah[5]

## Abstract

Ng and Kotz (1995) introduced a distribution that provides greater flexibility to extremes. We define and study a new class of distributions called the Kummer beta generalized family to extend the normal, Weibull, gamma and Gumbel distributions, among several other well-known distributions. Some special models are discussed. The ordinary moments of any distribution in the new family can be expressed as linear functions of probability weighted moments of the baseline distribution. We examine the asymptotic distributions of the extreme values. We derive the density function of the order statistics, mean absolute deviations and entropies. We use maximum likelihood estimation to fit the distributions in the new class and illustrate its potentiality with an application to a real data set.

*MSC:* 62E15

*Keywords:* Generalized distribution, Kummer beta distribution, likelihood ratio test, moment, order statistic, Weibull distribution.

## 1. Introduction

Beta distributions are very versatile and can be used to analyze different types of data sets. Many of the finite range distributions encountered in practice can be easily

[1] ESALQ, Departamento de Ciências Exatas, Universidade de São Paulo-USP, Piracicaba, Brasil.
E-mail: rrpescim@usp.br

[2] Departamento de Estatística, Universidade Federal de Pernambuco-UFPE, Recife, Brazil.
E-mail: gausscordeiro@uol.com.br

[3] ESALQ, Departamento de Ciências Exatas, Universidade de São Paulo-USP, Piracicaba, Brazil.
E-mail: clarice@esalq.usp.br

[4] ESALQ, Departamento de Ciências Exatas, Universidade de São Paulo-USP, Piracicaba, Brazil.
E-mail: edwin@esalq.usp.br

[5] School of Mathematics, University of Manchester, Manchester, UK.
E-mail: mbbsssn2@manchester.ac.uk
Received: November 2011
Accepted: April 2012

transformed into the standard beta distribution. In econometrics, quite often the data are analyzed by using finite-range distributions. Generalized beta distributions have been widely studied in statistics and numerous authors have developed various classes of these distributions. Eugene *et al.* (2002) proposed a general class of distributions based on the logit of a beta random variable by employing two parameters whose role is to introduce skewness and to vary tail weights.

Following Eugene *et al.* (2002), who defined the beta normal (BN) distribution, Nadarajah and Kotz (2004) introduced the beta Gumbel distribution (BGu), provided expressions for the moments, examined the asymptotic distribution of the extreme order statistics and performed maximum likelihood estimation (MLE). Nadarajah and Gupta (2004) defined the beta Fréchet (BF) distribution and derived analytical shapes of the probability density and hazard rate functions. Nadarajah and Kotz (2005) proposed the beta exponential (BE) distribution, derived the moment generating function (mgf), the first four moments, and the asymptotic distribution of the extreme order statistics and discussed MLE. Most recently, Pescim *et al.* (2010) and Paranaíba *et al.* (2011) have studied important mathematical properties of the beta generalized half-normal (BGHN) and beta Burr XII (BBXII) distributions. However, those distributions do not offer flexibility to the extremes (right and left) of the probability density functions (pdfs). Therefore, they are not suitable for analyzing data sets with high degrees of asymmetry.

Ng and Kotz (1995) proposed the Kummer beta distribution on the unit interval $(0,1)$ with cumulative distribution function (cdf) and pdf given by

$$F(x) = K\int_0^x t^{a-1}\,(1-t)^{b-1}\,\exp(-ct)dt,$$

and

$$f(x) = Kx^{a-1}\,(1-x)^{b-1}\,\exp(-cx), \quad 0 < x < 1,$$

respectively, where $a > 0$, $b > 0$ and $-\infty < c < \infty$. Here,

$$K^{-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,{}_1F_1(a;a+b;-c) \tag{1}$$

and

$${}_1F_1(a;a+b;-c) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 t^{a-1}\,(1-t)^{b-1}\exp(-ct)dt = \sum_{k=0}^{\infty}\frac{(a)_k(-c)^k}{(a+b)_k\,k!}$$

is the confluent hypergeometric function (Abramowitz and Stegun, 1968), $\Gamma(\cdot)$ is the gamma function and $(d)_k = d(d+1)\ldots(d+k-1)$ denotes the ascending factorial. Independently, Gordy (1998) has also defined the Kummer beta distribution in relation

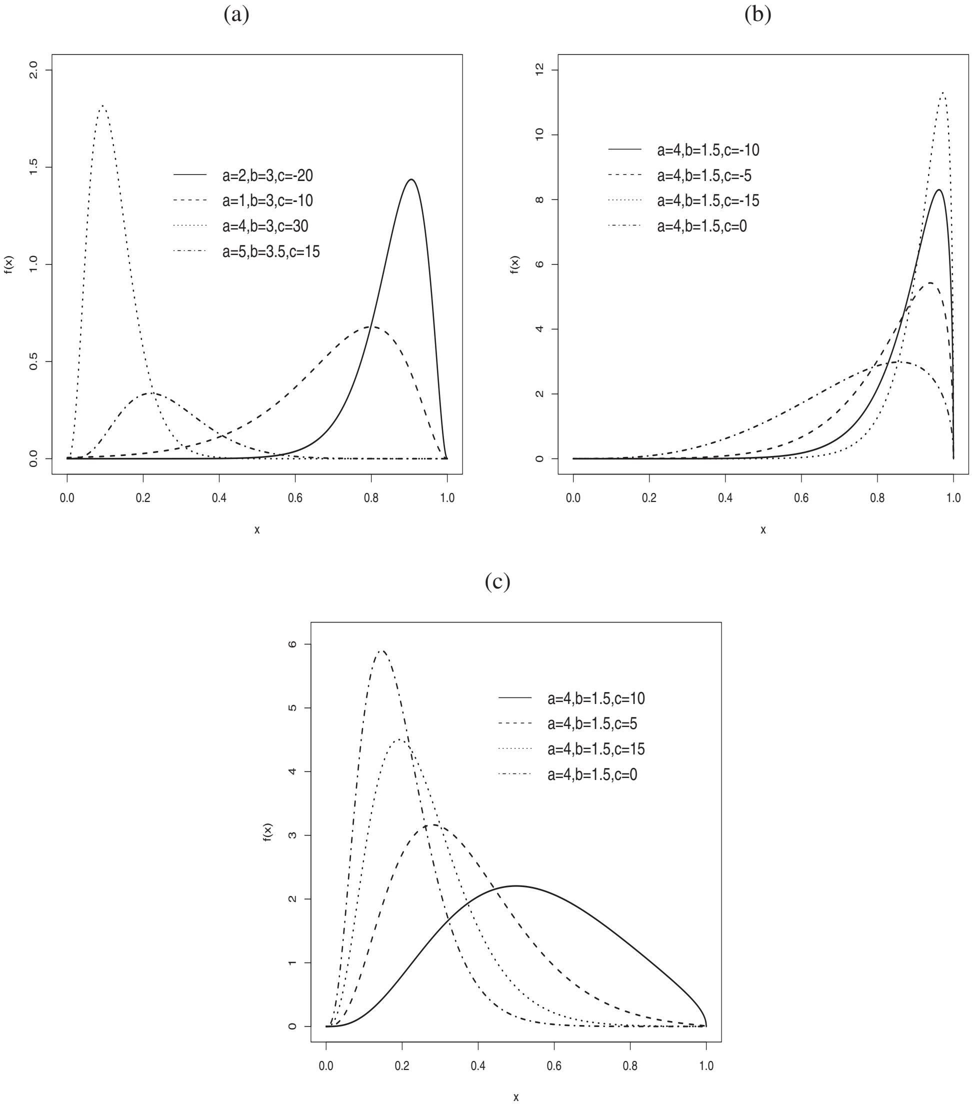

***Figure 1:*** *Plots of the Kummer beta pdf for some parameter values.*

to the problem of common value auction. This distribution is an extension of the beta distribution. It yields bimodal distributions on finite range for $a < 1$ (and certain values of the parameter $c$). Plots of the Kummer beta pdf are displayed in Figure 1 for selected parameter values.

Consider starting from a parent continuous cdf $G(x)$. A natural way of generating families of distributions from a simple parent distribution with pdf $g(x) = dG(x)/dx$ is to apply the quantile function to a family of distributions on the interval $(0,1)$. We now use the same methodology of Eugene *et al.* (2002) and Cordeiro and de Castro (2011) to construct a new class of Kummer beta generalized (KBG) distributions. From an arbitrary parent cdf $G(x)$, the KBG family of distributions is defined by

$$F(x) = K \int_0^{G(x)} t^{a-1} (1-t)^{b-1} \exp(-ct)dt, \tag{2}$$

where $a > 0$ and $b > 0$ are shape parameters introducing skewness, and thereby promoting weight variation of the tails. The parameter $-\infty < c < \infty$ "squeezes" the pdf to the left or to the right.

The pdf corresponding to (2) can be expressed as

$$f(x) = K g(x) G(x)^{a-1} \{1 - G(x)\}^{b-1} \exp\{-c\, G(x)\}, \tag{3}$$

where $K$ is defined in (1).

The KBG family of distributions defined by (3) is an alternative family of models to the class of distributions proposed by Alexander *et al.* (2012). The shape parameter $c > 0$ in Alexander *et al.* (2012) together with $a > 0$ and $b > 0$ promotes the weight variation of the tails and adds flexibility. On the other hand, the parameter $-\infty < c < \infty$ of the proposed family offers flexibility to the extremes (left and/or right) of the pdfs. Therefore, the new family of distributions is suitable for analyzing data sets with high degrees of asymmetry.

For each continuous $G$ distribution (here and henceforth "$G$" denotes the baseline distribution), we associate the KBG-$G$ distribution with three extra parameters $a$, $b$ and $c$ defined by the pdf (3). Setting $u = t/G(x)$ in equation (2), we obtain

$$\begin{aligned} F(x) &= K\, G(x)^a \int_0^1 u^{a-1} \left[1 - G(x)\, u\right]^{b-1} \exp\left[-c\, G(x)\, u\right] du \\ &= \frac{K}{a} G(x)^a \Phi_1 \left(a; 1-b; a+1; -c\, G(x); G(x)\right), \end{aligned}$$

where $\Phi_1$ is the confluent hypergeometric function of two variables defined by (Erdélyi *et al.*, 1953)

$$\Phi_1(a;b;c;x;y) = \sum_{j,m=0}^{\infty} \frac{(a)_{j+m}\,(b)_j}{(c)_{j+m}} x^j y^m$$

for $|x| < 1$ and $|y| < 1$.

Special generalized distributions can be generated as follow. The KBG-normal (KBGN) distribution is obtained by taking $G(x)$ in equation (3) to be the normal cdf. Analogously, the KBG-Weibull (KBGW), KBG-gamma (KBGGa) and KBG-Gumbel (KBGGu) distributions are obtained by taking $G(x)$ to be the cdf of the Weibull, gamma and Gumbel distributions, respectively. Hence, each new KBG-$G$ distribution can be obtained from a specified $G$ distribution. The Kummer beta distribution is clearly a basic example of the KBG distribution when $G$ is the uniform distribution on $[0,1]$. The $G$ distribution corresponds to $a = b = 1$ and $c = 0$. For $c = 0$, the KBG-$G$ distribution reduces to the beta-$G$ distribution proposed by Eugene *et al.* (2002). Further, for $b = 1$ and $c = 0$, the KBG-$G$ distribution becomes the exponentiated-$G$ distribution. One major benefit of the KBG family of distributions is its ability to fit skewed data that cannot be properly fitted by existing distributions.

We study some mathematical properties of the KBG family of distributions because it extends several widely-known distributions in the literature. The article is outlined as follows. Section 2 provides some special cases. In Section 3, we derive general expansions for the new pdf in terms of the baseline pdf $g(x)$ multiplied by a power series in $G(x)$. We can easily apply these expansions to several KBG distributions. In Section 4, we derive two simple expansions for moments of the KBG-$G$ distribution as linear functions of probability weighted moments (PWMs) of the $G$ distribution. The mean absolute deviations and Rényi entropy are determined in Sections 5 and 6, respectively. In Section 7, we provide some expansions for the pdf of the order statistics. Extreme values are obtained in Section 8. Some inferential tools are discussed in Section 9. In Section 10, we analyze a real data set using a special KBG distribution. Section 11 ends with some concluding remarks.

## 2. Special KBG generalized distributions

The KBG pdf (3) allows for greater flexibility of its tails and promotes variation of the tail weights to the extremes of the distribution. It can be widely applied in many areas of engineering and biological sciences. The pdf (3) will be most tractable when the cdf $G(x)$ and the pdf $g(x)$ have simple analytic expressions. We now define some of the many distributions which arise as special sub-models within the KBG class of distributions.

### 2.1. KBG-normal

The KBGN pdf is obtained from (3) by taking $G(\cdot)$ and $g(\cdot)$ to be the cdf and pdf of the normal distribution, $\mathrm{N}(\mu, \sigma^2)$, so that

$$f(x) = \frac{K}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{a-1}\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{b-1}\exp\left\{-c\,\Phi\left(\frac{x-\mu}{\sigma}\right)\right\},$$

where $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, $a$ and $b$ are positive shape parameters, $c \in \mathbb{R}$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. A random variable with the above pdf is denoted by $\mathrm{X} \sim \mathrm{KBGN}(a, b, c, \mu, \sigma^2)$. For $\mu = 0$ and $\sigma = 1$, we have the standard KBGN distribution. Further, the KBGN distribution with $a = 2$, $b = 1$ and $c = 0$ is the skew normal distribution with shape parameter equal to one (Azzalini, 1985).

### 2.2. KBG-Weibull

The cdf of the Weibull distribution with parameters $\beta > 0$ and $\alpha > 0$ is $G(x) = 1 - \exp\{-(\beta x)^{\alpha}\}$ for $x > 0$. Correspondingly, the KBG-Weibull (KGBW) pdf is

$$f(x) = K\,\alpha\,\beta^{\alpha}\,x^{\alpha-1}\,[1-\exp\{-(\beta x)^{\alpha}\}]^{a-1}\exp\{-c\,[1-\exp\{-(\beta x)^{\alpha}\}]-b(\beta x)^{\alpha}\},$$

where $x, a, b, \beta > 0$ and $c \in \mathbb{R}$. Let $\text{KBGW}(a,b,c,\alpha,\beta)$ denote a random variable with this pdf. For $\alpha = 1$, we obtain the KBG-exponential (KBGE) distribution. $\text{KBGW}(1,b,0,1,\beta)$ is an exponential random variable with parameter $\beta^{*} = b\,\beta$.

### 2.3. KBG-gamma

Let $Y$ be a gamma random variable with cdf $G(y) = \Gamma_{\beta y}(\alpha)/\Gamma(\alpha)$ for $y$, $\alpha$, $\beta > 0$, where $\Gamma(\cdot)$ is the gamma function and $\Gamma_{z}(\alpha) = \int_{0}^{z} t^{\alpha-1}\mathrm{e}^{-\mathrm{t}}\mathrm{dt}$ is the incomplete gamma function. The pdf of a random variable $X$, say X $\sim$ $\text{KBGGa}(a,b,c,\beta,\alpha)$, having the KBGGa distribution can be expressed as

$$f(x) = \frac{K\beta^{\alpha}x^{\alpha-1}\exp(-\beta x)}{\Gamma(\alpha)^{a+b-1}}\exp\left\{-c\,\frac{\Gamma_{\beta x}(\alpha)}{\Gamma(\alpha)}\right\}\Gamma_{\beta x}(\alpha)^{a-1}\left\{\Gamma(\alpha)-\Gamma_{\beta x}(\alpha)\right\}^{b-1}.$$

For $\alpha = 1$ and $c = 0$, we obtain the KBGE distribution. $\text{KBGGa}(1,b,0,\beta,1)$ is an exponential random variable with parameter $\beta^{*} = b\,\beta$.

### 2.4. KBG-Gumbel

The pdf and cdf of the Gumbel distribution with location parameter $\mu > 0$ and scale parameter $\sigma > 0$ are given by

$$g(x) = \sigma^{-1}\exp\left\{\frac{x-\mu}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right\}, x > 0,$$

and

$$G(x) = 1 - \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\},$$

respectively. The mean and variance are equal to $\mu - \gamma\sigma$ and $\pi^{2}\sigma^{2}/6$, respectively, where $\gamma \approx 0.57722$ is the Euler's constant. By inserting these equations in (3), we obtain a KBGGu random variable, say $\text{KBGGu}(a,b,c,\mu,\sigma)$.

Figure 2 displays some possible shapes of the four KBG pdfs. These plots show the great flexibility achieved with the new distributions.

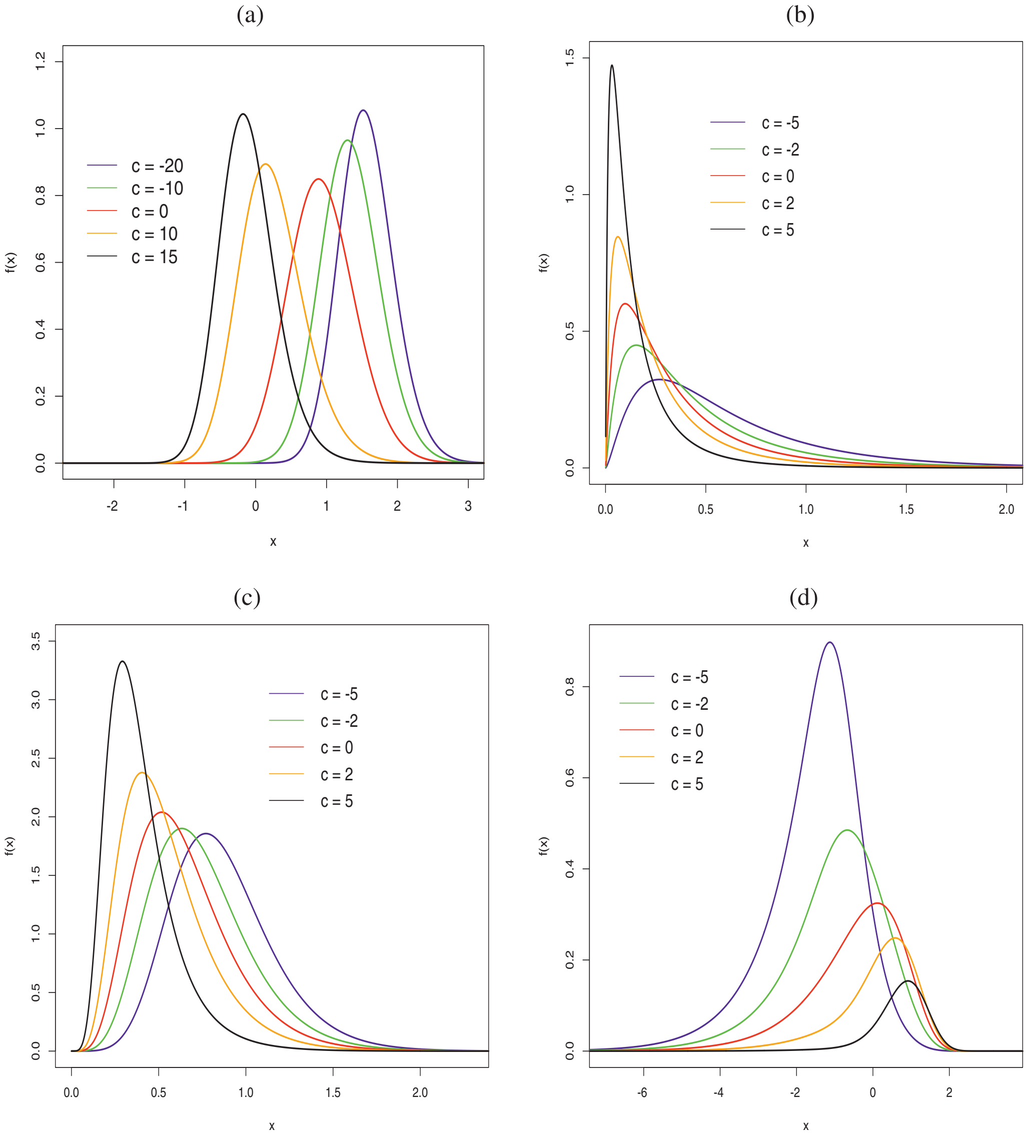

***Figure 2:*** *(a) KBGN(8,2,c,0,1), (b) KBGW(5,3,c,0.5,4), (c) KBGGa(3,1.5,c,4,2) and (d) KBGGu(0.8,1,c,0,1) pdfs (the red lines represent the beta-G pdfs).*

## 3. Expansions for pdf and cdf

The cdf $F(x)$ and pdf $f(x) = dF(x)/dx$ of the KBG-$G$ distribution are usually straightforward to compute given $G(x)$ and $g(x) = dG(x)/dx$. However, we provide expansions for these functions as infinite (or finite) weighted sums of cdf's and pdf's of exponentiated-$G$ distributions. In the next sections, based on these expansions, we obtain some structural properties of the KBG-$G$ distribution, including explicit expressions for moments, mean absolute deviations, pdf of order statistics and moments of order statistics.

Using the exponential expansion in (2), we write

$$F(x)=\sum_{i=0}^{\infty} w_i H_{a+i,b}(x), \tag{4}$$

where $w_i=[K B(a+i,b)(-c)^i]/i!$ and

$$H_{a,b}(x)=\frac{1}{B(a,b)}\int_0^{G(x)} t^{a-1}(1-t)^{b-1}dt$$

denotes the beta-$G$ cdf with positive shape parameters $a$ and $b$ (Eugene *et al.*, 2002). Equation (4) reveals that the KBG-$G$ cdf is a linear combination of beta-$G$ cdf's. This result is important. It can be used to derive properties of any KBG-$G$ distribution from those of beta-$G$ distributions.

For $b>0$ real non-integer, we have the power series representation

$$\{1-G(x)\}^{b-1}=\sum_{j=0}^{\infty}(-1)^j\binom{b-1}{j}G(x)^j, \tag{5}$$

where the binomial coefficient is defined for any real. Expanding $\exp\{-cG(x)\}$ in power series and using (5) in equation (2), the KBG-$G$ cdf can be expressed as

$$F(x)=\sum_{i,j=0}^{\infty} w_{i,j}\, G(x)^{a+i+j}, \tag{6}$$

where

$$w_{i,j}=\frac{K(-1)^{i+j}c^i}{i!\,(a+i+j)}\binom{b-1}{j}.$$

If $b$ is an integer, the index $i$ in the previous sum stops at $b-1$. If $a$ is an integer, equation (6) reveals that the KBG-$G$ pdf can be expressed as the baseline pdf multiplied by an infinite power series of its cdf.

If $a$ is a real non-integer, we can expand $G(x)^{a+i+j}$ as follows

$$G(x)^{a+i+j}=\sum_{k=0}^{\infty}(-1)^k\binom{a+i+j}{k}[1-G(x)]^k.$$

Then,

$$G(x)^{a+i+j} = \sum_{k=0}^{\infty}\sum_{r=0}^{k}(-1)^{k+r}\binom{a+i+j}{k}\binom{k}{r}G(x)^r.$$

Further, equation (2) can be rewritten as

$$F(x) = \sum_{i,j,k=0}^{\infty}\sum_{r=0}^{k} t_{i,j,k,r}\,G(x)^r, \tag{7}$$

where

$$t_{i,j,k,r} = t_{i,j,k,r}(a,b,c) = (-1)^{k+r}\binom{a+i+j}{k}\binom{k}{r}w_{i,j}$$

and $w_{i,j}$ is defined in (6). Replacing $\sum_{k=0}^{\infty}\sum_{r=0}^{k}$ by $\sum_{r=0}^{\infty}\sum_{k=r}^{\infty}$ in equation (7), we obtain

$$F(x) = \sum_{r=0}^{\infty} b_r\,G(x)^r, \tag{8}$$

where the coefficient $b_r = \sum_{i,j=0}^{\infty}\sum_{k=r}^{\infty} t_{i,j,k,r}$ denotes a sum of constants.

Expansion (8), which holds for any real non-integer $a$, expresses the KBG-$G$ cdf as an infinite weighted power series of $G$. If $b$ is an integer, the index $i$ in (7) stops at $b-1$.

We also note that the cdf of the KBG family can be expressed in terms of exponentiated-$G$ cdfs. We have

$$F(x) = \sum_{r=0}^{\infty} b_r\,V_r(x), \tag{9}$$

where $V_r = G(x)^r$ is an exponentiated-$G$ cdf (Exp-$G$ cdf for short) with power parameter $r$.

The corresponding expansions for the KBG pdf are obtained by simple differentiation of (6) for $a > 0$ integer

$$f(x) = g(x)\sum_{i,j=0}^{\infty} w_{i,j}^{*}\,G(x)^{a+i+j-1}, \tag{10}$$

where $w_{i,j}^{*} = (a+i+j)\,w_{i,j}$. Analogously, from equations (8) and (9), for $a > 0$ real non-integer, we obtain

$$f(x) = g(x)\sum_{r=0}^{\infty} b_r^{*}\,G(x)^r, \tag{11}$$

and

$$f(x)=\sum_{r=0}^{\infty} c_r\, v_{r+1}(x), \tag{12}$$

where $b_r^*=(r+1)\,b_{r+1}$ and $c_r=b_{r+1}$ for $r=0,1\ldots$, and $v_{r+1}=(r+1)\,g(x)\,G(x)^r$ denotes the Exp-$G$ pdf with parameter $r+1$. Equation (12) reveals that the KBG-$G$ pdf is a linear combination of Exp-$G$ pdfs. This result is important to derive properties of the KBG-$G$ distribution from those of the Exp-$G$ distribution.

Mathematical properties of exponentiated distributions have been studied by many authors in recent years, see Mudholkar *et al.* (1995) for exponentiated Weibull, Gupta *et al.* (1998) for exponentiated Pareto, Gupta and Kundu (2001) for exponentiated exponential and Nadarajah and Gupta (2007) for exponentiated gamma.

Equations (10)-(12) are the main results of this section. They play an important role in this paper.

## 4. Moments and generating function

### 4.1. Moments

The $s$th moment of the KBG-$G$ distribution can be expressed as an infinite weighted sum of PWMs of order $(s,q)$ of the parent $G$ distribution from equation (10) for $a$ integer and from (11) for $a$ real non-integer. We assume that $Y$ and $X$ follow the baseline $G$ and KBG-$G$ distributions, respectively. The $s$th moment of $X$ can be expressed in terms of the $(s,q)$th PWMs of $Y$, say $\tau_{s,q}=\mathrm{E}[Y^s\,G(Y)^q]$ (for $q=0,1,\ldots$), as defined by Greenwood *et al.* (1979). The moments $\tau(s,q)$ can be derived for most parent distributions.

For an integer $a$, we have

$$\mu_s'=\mathrm{E}(X^s)=\sum_{i,j=0}^{\infty} w_{i,j}^*\,\tau_{s,a+i+j-1}.$$

For a real non-integer $a$, we can write from (11)

$$\mu_s'=\sum_{r=0}^{\infty} b_r^*\,\tau_{s,r}.$$

So, we can calculate the moments of any KBG-$G$ distribution as infinite weighted sums of PWMs of the $G$ distribution.

Alternatively, we can express $\mu_s'$ from (11) in terms of the baseline quantile function $Q(u)=G^{-1}(u)$. We have

$$\mu_s' = \sum_{r=0}^{\infty} b_r^* \int x^s \, g(x) \, G(x)^r \, dx.$$

Setting $u = G(x)$ in the last equation, we obtain

$$\mu_s' = \sum_{r=0}^{\infty} b_r^* \int_0^1 u^r \, Q(u)^s \, dt.$$

Now, we express moments of KBG distributions from equation (12) in terms of moments of Exp-$G$ distributions. Let $Y_{r+1}$ have the Exp-$G$ pdf $v_{r+1} = (r+1)\, g(x)\, G(x)^r$ with power parameter $(r+1)$. As a first example, consider $G$ the Weibull distribution with scale parameter $\lambda > 0$ and shape parameter $\gamma > 0$. If $Y_{r+1}$ has the exponentiated Weibull distribution, its moments are

$$\mathrm{E}(Y^s) = \frac{(r+2)}{\lambda^s} \Gamma\left(\frac{s}{\gamma}+1\right) \sum_{i=0}^{\infty} \frac{(-r)_i}{i!\,(i+1)^{(s+\gamma)/\gamma}},$$

where $(a)_i = a(a+1)\ldots(a+i-1)$ denotes the ascending factorial. From this expectation and equation (12), the $s$th moment of the KBG-Weibull distribution is

$$\mu_s' = \lambda^{-s} \Gamma\left(\frac{s}{\gamma}+1\right) \sum_{r,i=0}^{\infty} \frac{(r+2)\, c_r\, (-r)_i}{i!\,(i+1)^{(s+\gamma)/\gamma}}.$$

For a second example, take the Gumbel distribution with cdf $G(x) = 1 - \exp\{-\exp(-\frac{x-\mu}{\sigma})\}$. The moments of $Y_{r+1}$ having the exponentiated Gumbel distribution with parameter $(r+1)$ can be obtained from Nadarajah and Kotz (2006) as

$$\mathrm{E}(Y_{r+1}^s) = (r+1) \sum_{i=0}^{s} \binom{s}{i} \mu^{s-i} (-\sigma)^i \left(\frac{\partial}{\partial p}\right)^i \left[(r+1)^{-p}\,\Gamma(p)\right]\Bigg|_{p=1}.$$

From the last equation and (12), the $s$th moment of the KBG-Gumbel (KBGGu) distribution becomes

$$\mu_s' = \sum_{r=0}^{\infty} c_r\,(r+1) \sum_{i=0}^{s} \binom{s}{i} \mu^{s-i} (-\sigma)^i \left(\frac{\partial}{\partial p}\right)^i \left[(r+1)^{-p}\,\Gamma(p)\right]\Bigg|_{p=1}.$$

Finally, as a third example, consider the standard logistic cdf $G(x) = [1+\exp(-x)]^{-1}$. We can easily obtain the $s$th moment of the KBG-logistic (KBGL) distribution as

$$\mu_s' = \sum_{r=0}^{\infty} c_r \left(\frac{\partial}{\partial t}\right)^s B\left(t+(r+1), 1-t\right)\Bigg|_{t=0}.$$

### 4.2. Generating function

Let $X \sim$ KBG-$G(a,b,c)$. We provide four representations for the mgf $M(t) = \mathrm{E}[\exp(tX)]$ of $X$. Clearly, the first one is

$$M(t) = \sum_{s=0}^{\infty} \frac{\mu_s'}{s!} t^s,$$

where $\mu_s' = \mathrm{E}(X^s)$. The second one comes from

$$\begin{aligned} M(t) &= K\,\mathrm{E}\left[\exp\left[tX - cG(X)\right] G^{a-1}(X) \left\{1 - G(X)\right\}^{b-1}\right] \\ &= K \sum_{j=0}^{\infty} (-1)^j \binom{b-1}{j} \mathrm{E}\left[\frac{\exp\left(tX - Uc\right)}{U^{-(a+j-1)}}\right], \end{aligned}$$

where $U$ is a uniform random variable on the unit interval. Note that $X$ and $U$ are not independent.

A third representation for $M(t)$ is obtained from (12)

$$M(t) = \sum_{i=0}^{\infty} c_i \, M_{i+1}(t),$$

where $M_{i+1}(t)$ is the mgf of $Y_{i+1} \sim$ Exp-$G(i+1)$. Hence, for any KBG-$G$ distribution, $M(t)$ can be immediately determined from the mgf of the $G$ distribution.

A fourth representation for $M(t)$ can be derived from (11) as

$$M(t) = \sum_{i=0}^{\infty} b_i^* \, \rho(t,i), \tag{13}$$

where $\rho(t,r) = \int_{-\infty}^{\infty} \exp(tx)\, g(x)\, G(x)^r dx$ can be expressed in terms of the baseline quantile function $Q(u)$ as

$$\rho(t,a) = \int_0^1 u^a \exp\left[t\,Q(u)\right] du. \tag{14}$$

We can obtain the mgf of several KBG distributions from equations(13) and (14). For example, the mgfs of the KBG-exponential (KBGE) (with parameter $\lambda$), KBGL and KBG-Pareto (KBGPa) (with parameter $\nu > 0$) are easily calculated as

$$M(t) = \sum_{i=0}^{\infty} b_i^* \, B\left(i+1, 1-\lambda t^{-1}\right), \;\; M(t) = \sum_{i=0}^{\infty} b_i^* \, B\left(i+1, 1-t\right),$$

and

$$M(t) = \exp(-t) \sum_{i,p=0}^{\infty} \frac{b_i^* t^p}{p!} B\left(i+1, 1-p\nu^{-1}\right),$$

respectively.

Clearly, four representations for the characteristic function (chf) $\phi(t) = E[\exp(\mathrm{i}tX)]$ of the KBG-$G$ distribution are immediately obtained from the above representations for the mgf by $\phi(t) = M(\mathrm{i}t)$, where $\mathrm{i} = \sqrt{-1}$.

## 5. Mean absolute deviations

Let $X \sim$ KBG-$G(a,b,c)$. The mean absolute deviations about the mean ($\delta_1(X)$) and about the median ($\delta_2(X)$) can be expressed as

$$\delta_1(X) = \mathrm{E}\left(\left|X - \mu_1'\right|\right) = 2\mu_1' F\left(\mu_1'\right) - 2T\left(\mu_1'\right),$$
$$\delta_2(X) = \mathrm{E}\left(|X - M|\right) = \mu_1' - 2T(M), \quad (15)$$

respectively, where $\mu_1' = \mathrm{E}(X)$, $F(\mu_1')$ comes from (2), $M = \text{Median}(X)$ denotes the median determined from the nonlinear equation $F(M) = 1/2$, and $T(z) = \int_{-\infty}^{z} x f(x) dx$. Setting $u = G(x)$ in (11) yields

$$T(z) = \sum_{r=0}^{\infty} b_r^* T_r(z), \quad (16)$$

where the integral $T_r(z)$ can be expressed in terms of $Q(u) = G^{-1}(u)$ by

$$T_r(z) = \int_0^{G(z)} u^r Q(u)\, du. \quad (17)$$

The mean absolute deviations of any KBG distribution can be computed from equations (15)-(17). For example, the mean absolute deviations of the KBGE (with parameter $\lambda$), KBGL and KBGPa (with parameter $\nu > 0$) are immediately calculated using

$$T_r(z) = \lambda^{-1} \Gamma(r+2) \sum_{j=0}^{\infty} \frac{(-1)^j \left[1 - \exp(-j\lambda z)\right]}{\Gamma(r+2-j)(j+1)!},$$

$$T_r(z) = \frac{1}{\Gamma(z)} \sum_{j=0}^{\infty} \frac{(-1)^j \Gamma(r+j+1) \left[1 - \exp(-jz)\right]}{(j+1)!}$$

and

$$T_r(z)=\sum_{j=0}^{\infty}\sum_{k=0}^{j}\frac{(-1)^j\binom{r+1}{j}\binom{j}{k}}{(1-kv)}\,z^{1-kv},$$

respectively.

An alternative representation for $T(z)$ can be derived from (12) as

$$T(z)=\int_{-\infty}^{z}x\,f(x)dx=\sum_{r=0}^{\infty}c_r\,J_{r+1}(z), \tag{18}$$

where

$$J_{r+1}(z)=\int_{-\infty}^{z}x\,v_{r+1}(x)dx. \tag{19}$$

Equation (19) is the basic quantity to compute mean absolute deviations of Exp-$G$ distributions. Hence, the KBG mean absolute deviations depend only on the quantity $J_{r+1}(z)$. So, alternative representations for $\delta_1(X)$ and $\delta_2(X)$ are

$$\delta_1(X)=2\mu_1'F\left(\mu_1'\right)-2\sum_{r=0}^{\infty}c_r\,J_{r+1}\left(\mu_1'\right)\quad\text{and}\quad\delta_2(X)=\mu_1'-2\sum_{r=0}^{\infty}c_r\,J_{r+1}(M).$$

A simple application is provided for the KBGW distribution. The exponentiated Weibull pdf with parameter $r+1$ is given by

$$v_{r+1}(x)=(r+1)\,d\,\beta^d\,x^{d-1}\exp\left\{-(\beta x)^d\right\}\left[1-\exp\left\{-(\beta x)^d\right\}\right]^r$$

for $x>0$. Then,

$$\begin{aligned}J_{r+1}(z)&=(r+1)\,d\,\beta^d\int_0^z x^d\exp\left\{-(\beta x)^d\right\}\left[1-\exp\left\{-(\beta x)^d\right\}\right]^r dx\\&=r\,d\,\beta^d\sum_{k=0}^{\infty}(-1)^k\binom{r}{k}\int_0^z x^d\exp\left[-(k+1)(\beta x)^d\right]dx.\end{aligned}$$

We calculate the last integral using the incomplete gamma function $\gamma(\alpha,x)=\int_0^x w^{\alpha-1}\,\mathrm{e}^{-\mathrm{w}}\mathrm{dw}$ for $\alpha>0$. Then,

$$J_{r+1}(z)=(r+1)\,\beta^{-1}\sum_{k=0}^{\infty}\frac{(-1)^k\binom{r}{k}}{(k+1)^{1+d^{-1}}}\,\gamma\left(1+d^{-1},(k+1)(\beta z)^d\right).$$

Equations (16) and (18) are the main results of this section. These equations can be applied to Bonferroni and Lorenz curves defined for a given probability $p$ by

$$B(p)=\frac{T(q)}{p\mu_1'} \quad\text{and}\quad L(p)=\frac{T(q)}{\mu_1'},$$

where $\mu_1'=\mathrm{E}(X)$ and $q=F^{-1}(p)$.

## 6. Entropies

An entropy is a measure of variation or uncertainty of a random variable $X$. The most popular measures of entropy are the Shannon entropy (Shannon, 1951) and the Rényi entropy.

### 6.1. Shannon entropy

The Shannon entropy (Shannon, 1951) is defined by $\mathrm{E}\{-\log[f(X)]\}$. Let $X$ has the pdf (3). We can write

$$\begin{aligned}
\mathrm{E}\{-\log[f(X)]\} &= -\log(K)-\mathrm{E}\{\log[g(X)]\}+(1-a)\mathrm{E}\{\log[G(X)]\}\\
&\quad +(1-b)\mathrm{E}\{\log[1-G(X)]\}+c\,\mathrm{E}[G(X)]\\
&= -\log K-\mathrm{E}\{\log[g(X)]\}+(a-1)\sum_{k=1}^{\infty}\frac{1}{k}\,\mathrm{E}\left\{[1-G(X)]^k\right\}\\
&\quad +(b-1)\sum_{k=1}^{\infty}\frac{1}{k}\,\mathrm{E}\left[G^k(X)\right]+c\,\mathrm{E}[G(X)]\\
&= -\log(K)-\mathrm{E}\{\log[g(X)]\}+(a-1)\sum_{k=1}^{\infty}\frac{K(a,b+k,c)}{kK(a,b,c)}\\
&\quad +(b-1)\sum_{k=1}^{\infty}\frac{K(a+k,b,c)}{kK(a,b,c)}+\frac{cK(a+1,b,c)}{K(a,b,c)}, \qquad (20)
\end{aligned}$$

where $K=K(a,b,c)$ is given by (1). The only unevaluated term in (20) is $\mathrm{E}\{\log[g(X)]\}$.

### 6.2. Rényi entropy

The Rényi entropy is given by

$$\mathfrak{I}_R(\xi)=\frac{1}{1-\xi}\log\left[\int_{-\infty}^{\infty}f^{\xi}(x)dx\right],\ \xi>0 \text{ and } \xi\neq 1.$$

The integral can be expressed as

$$\int_{-\infty}^{\infty} f^{\xi}(x)dx = K^{\xi} \int_{-\infty}^{\infty} g^{\xi}(x)\, G^{\xi(a-1)}(x)\ [1-G(x)]^{\xi(b-1)} \exp\left[-\xi\, c\, G(x)\right] dx.$$

Expanding the exponential and the binomial terms and changing variables, we obtain

$$\int_{-\infty}^{\infty} f^{\xi}(x)dx = K^{\xi} \sum_{i,j=0}^{\infty} \frac{(-1)^{i+j}(c\,\xi)^{i}}{i!} \binom{\xi(b-1)}{j} I_{i,j}(\xi), \tag{21}$$

where $I_{i,j}(\xi)$ denotes the integral

$$I_{i,j}(\xi) = \int_{0}^{1} g^{\xi-1}\left(Q(u)\right)\ u^{i+j+\xi(a-1)}\, du,$$

to be calculated for each KBG-model. For the KBGE (with parameter $\lambda$), KBGL and KBGPa (with parameter $\nu$), we obtain

$$I_{i,j}(\xi) = \lambda^{\xi-1} B\left(i+j+\xi(a-1)+1, \xi\right),\ \ I_{i,j}(\xi) = B\left(i+j+\xi\, a\,, \xi\right),$$

and

$$I_{i,j}(\xi) = \nu^{\xi-1} B\left(i+j+\xi(a-1)+1, \nu^{-1}(\xi-1)+\xi\right),$$

respectively. Equation (21) is the main result of this section.

## 7. Order statistics

Order statistics have been used in a wide range of problems, including robust statistical estimation and detection of outliers, characterization of probability distributions and goodness-of-fit tests, entropy estimation, analysis of censored samples, reliability analysis, quality control and strength of materials.

Suppose $X_1,\ldots,X_n$ is a random sample from a continuous distribution and let $X_{1:n} < \cdots < X_{i:n}$ denote the corresponding order statistics. There has been a large amount of work relating to moments of order statistics $X_{i:n}$. See Arnold *et al.* (1992), David and Nagaraja (2003) and Ahsanullah and Nevzorov (2005) for excellent accounts. It is well-known that

$$f_{i:n}(x) = \frac{f(x)}{B(i,n-i+1)} F(x)^{i-1} \left\{1-F(x)\right\}^{n-i},$$

where $B(\cdot,\cdot)$ denotes the beta function. Using the binomial expansion in the last equation, we have

$$f_{i:n}(x)=\frac{f(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}F(x)^{i+j-1}. \tag{22}$$

We now provide an expression for the pdf of KBG order statistics as a function of the baseline pdf multiplied by infinite weighted sums of powers of $G(x)$. Based on this result, we express the ordinary moments of the order statistics of any KBG-$G$ distribution as infinite weighted sums of the PWMs of the $G$ distribution.

Replacing (8) in equation (22), we have

$$F(x)^{i+j-1}=\left(\sum_{r=0}^{\infty}b_r\,u^r\right)^{i+j-1}, \tag{23}$$

where $u=G(x)$ is the baseline cdf.

We use the identity $(\sum_{k=0}^{\infty}a_k\,x^k)^n=\sum_{k=0}^{\infty}d_{k,n}\,x^k$ (see Gradshteyn and Ryzhik, 2000), where

$$d_{0,n}=a_0^n \qquad \text{and} \qquad d_{k,n}=(ka_0)^{-1}\sum_{m=1}^{k}\left[m(n+1)-k\right]a_m\,d_{k-m,n}$$

(for $k=1,2,\ldots$) in equation (23) to obtain

$$F(x)^{i+j-1}=\sum_{r=0}^{\infty}d_{r,i+j-1}\,G(x)^r, \tag{24}$$

where

$$d_{0,i+j-1}=b_0^{i+k-1} \quad \text{and} \quad d_{r,i+j-1}=(kb_r)^{-1}\sum_{m=1}^{r}\left[(i+j)m-r\right]b_m\,d_{r-m,i+j-1}.$$

For real non-integer $a$, inserting (11) and (24) into equation (22) and changing indices, we rewrite $f_{i:n}(x)$ for the KBG distribution in the form

$$f_{i:n}(x)=\frac{g(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{u,v=0}^{\infty}b_u^*\,d_{u,i+j-1}\,G(x)^{u+v}. \tag{25}$$

For an integer $a$, we obtain from equations (10), (22) and (24)

$$f_{i:n}(x)=\frac{g(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{p,q,u=0}^{\infty}w_{p,q}^*\,d_{u,i+j-1}\,G(x)^{a+p+q+u-1}. \tag{26}$$

Equations (25) and (26) immediately yield the pdf of KBG order statistics as a function of the baseline pdf multiplied by infinite weighted sums of powers of $G(x)$. Hence, the moments of KBG-$G$ order statistics can be expressed as infinite weighted sums of PWMs of the $G$ distribution. Clearly, equation (26) can be expressed as linear combinations of Exp-$G$ pdfs. So, the moments and the mgf of KBG order statistics follow immediately from linear combinations of those quantities for Exp-$G$ distributions.

## 8. Extreme values

If $\overline{X} = (X_1 + \cdots + X_n)/n$ denotes the mean of a random sample from (3), then by the usual central limit theorem $\sqrt{n}(\overline{X} - E(X))/\sqrt{Var(X)}$ approaches the standard normal distribution as $n \to \infty$ under suitable conditions. Sometimes one would be interested in the asymptotics of the extreme values $M_n = \max(X_1, \ldots, X_n)$ and $m_n = \min(X_1, \ldots, X_n)$.

Firstly, suppose that $G$ belongs to the max domain of attraction of the Gumbel extreme value distribution. Then by Leadbetter *et al.* (1987, Chapter 1), there must exist a strictly positive function, say $h(t)$, such that

$$\lim_{t\to\infty} \frac{1 - G(t + xh(t))}{1 - G(t)} = \exp(-x)$$

for every $x \in (-\infty, \infty)$. But, using L'Hopital's rule, we note that

$$\begin{aligned}
\lim_{t\to\infty} \frac{1 - F(t + xh(t))}{1 - F(t)} &= \lim_{t\to\infty} \frac{[1 + xh'(t)]f(t + xh(t))}{f(t)} \\
&= \lim_{t\to\infty} \frac{[1 + xh'(t)]g(t + xh(t))}{g(t)} \left[\frac{G(t + xh(t))}{G(t)}\right]^{a-1} \\
&\quad \times \left[\frac{1 - G(t + xh(t))}{1 - G(t)}\right]^{b-1} \exp\{cG(t) - cG(t + xh(t))\} \\
&= \exp(-bx)
\end{aligned}$$

for every $x \in (-\infty, \infty)$. So, it follows that $F$ also belongs to the max domain of attraction of the Gumbel extreme value distribution with

$$\lim_{n\to\infty} \Pr\{a_n(M_n - b_n) \le x\} = \exp\{-\exp(-bx)\}$$

for some suitable norming constants $a_n > 0$ and $b_n$.

Secondly, suppose that $G$ belongs to the max domain of attraction of the Fréchet extreme value distribution. Then by Leadbetter *et al.* (1987, Chapter 1), there must exist a $\beta > 0$ such that

$$\lim_{t\to\infty} \frac{1-G(tx)}{1-G(t)} = x^{\beta}$$

for every $x > 0$. But, using L'Hopital's rule, we note that

$$\begin{aligned}
\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} &= \lim_{t\to\infty} \frac{xf(tx)}{f(t)} \\
&= \lim_{t\to\infty} \frac{xg(tx)}{g(t)} \left[\frac{G(tx)}{G(t)}\right]^{a-1} \left[\frac{1-G(tx)}{1-G(t)}\right]^{b-1} \exp\{cG(t)-cG(tx)\} \\
&= x^{b\beta}
\end{aligned}$$

for every $x > 0$. So, it follows that $F$ also belongs to the max domain of attraction of the Fréchet extreme value distribution with

$$\lim_{n\to\infty} \Pr\{a_n(M_n - b_n) \leq x\} = \exp\left(-x^{b\beta}\right)$$

for some suitable norming constants $a_n > 0$ and $b_n$.

Thirdly, suppose that $G$ belongs to the max domain of attraction of the Weibull extreme value distribution. Then by Leadbetter *et al.* (1987, Chapter 1), there must exist a $\alpha > 0$ such that

$$\lim_{t\to-\infty} \frac{G(tx)}{G(t)} = x^{\alpha}$$

for every $x < 0$. But, using L'Hopital's rule, we note that

$$\begin{aligned}
\lim_{t\to-\infty} \frac{F(tx)}{F(t)} &= \lim_{t\to-\infty} \frac{xf(tx)}{f(t)} \\
&= \lim_{t\to\infty} \frac{xg(tx)}{g(t)} \left[\frac{G(tx)}{G(t)}\right]^{a-1} \left[\frac{1-G(tx)}{1-G(t)}\right]^{b-1} \exp\{cG(t)-cG(tx)\} \\
&= x^{a\beta}.
\end{aligned}$$

So, it follows that $F$ also belongs to the max domain of attraction of the Weibull extreme value distribution with

$$\lim_{n\to\infty} \Pr\{a_n(M_n - b_n) \leq x\} = \exp\{-(-x)^{a\alpha}\}$$

for some suitable norming constants $a_n > 0$ and $b_n$.

The same argument applies to min domains of attraction. That is, $F$ belongs to the same min domain of attraction as that of $G$.

## 9. Inference

Let $\gamma$ be the $p$-dimensional parameter vector of the baseline distribution in equations (2) and (3). We consider independent random variables $X_1,\ldots,X_n$, each $X_i$ following a KBG-$G$ distribution with parameter vector $\theta=(a,b,c,\gamma)$. The log-likelihood function, $\ell=\ell(\theta)$, for the model parameters is

$$\begin{aligned}\ell(\theta) &= n\log(K)+\sum_{i=1}^{n}\log g(x_i;\gamma)-c\sum_{i=1}^{n}G(x_i;\gamma)\\ &\quad+(a-1)\sum_{i=1}^{n}\log\{G(x_i;\gamma)\}+(b-1)\sum_{i=1}^{n}\log\{1-G(x_i;\gamma)\}.\end{aligned}\tag{27}$$

The elements of score vector are given by

$$\frac{\partial\ell(\theta)}{\partial a}=\frac{n}{K}\frac{\partial K}{\partial a}+\sum_{i=1}^{n}\log\{G(x_i;\gamma)\},$$

$$\frac{\partial\ell(\theta)}{\partial b}=\frac{n}{K}\frac{\partial K}{\partial b}+\sum_{i=1}^{n}\log\{1-G(x_i;\gamma)\},$$

$$\frac{\partial\ell(\theta)}{\partial c}=\frac{n}{K}\frac{\partial K}{\partial c}-\sum_{i=1}^{n}G(x_i;\gamma),$$

and

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\gamma_j} &=\sum_{i=1}^{n}\Bigg[\frac{1}{g(x_i;\gamma)}\frac{\partial g(x_i;\gamma)}{\partial\gamma_j}-c\frac{\partial g(x_i;\gamma)}{\partial\gamma_j}+\frac{(a-1)}{G(x_i;\gamma)}\frac{\partial G(x_i;\gamma)}{\partial\gamma_j}\\ &\qquad+\frac{(b-1)}{1-G(x_i;\gamma)}\frac{\partial G(x_i;\gamma)}{\partial\gamma_j}\Bigg]\end{aligned}$$

for $j=1,\ldots,p$, where

$$\frac{\partial K}{\partial a}=-\frac{\left\{[\psi(a)-\psi(a+b)]\,{}_1F_1(a,a+b,-c)+\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial a}\right\}}{B(a,b)\,[{}_1F_1(a,a+b,-c)]^2},$$

$$\frac{\partial K}{\partial b}=-\frac{\left\{[\psi(b)-\psi(a+b)]\,{}_1F_1(a,a+b,-c)+\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial b}\right\}}{B(a,b)\,[{}_1F_1(a,a+b,-c)]^2},$$

$$\frac{\partial K}{\partial c}=\frac{a\,{}_1F_1(a+1,a+b+1,-c)}{(a+b)B(a,b)\,{}_1F_1(a,a+b,-c)},$$

$$\begin{aligned}\frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial a} &= -\left[\psi(a)-\psi(a+b)\right]\, {}_1F_1(a,a+b,-c)\\ &\quad -\sum_{k=0}^{\infty}\frac{(a)_k(-c)^k}{k!(a+b)_k}\left[\psi(a+b+k)-\psi(a+k)\right],\end{aligned}$$

and

$$\frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial b}=\psi(a+b)\, {}_1F_1(a,a+b,-c)+\sum_{k=0}^{\infty}\frac{(a)_k(-c)^k}{k!(a+b)_k}\psi(a+b+k).$$

These partial derivatives depend on the specified baseline distribution. Numerical maximization of the log-likelihood above was accomplished by using the RS method (Rigby and Stasinopoulos, 2005) available in the R contributed gamlss package (Stasinopoulos and Rigby, 2007; R Development Core Team, 2009).

For interval estimation of each parameter in $\boldsymbol{\theta}=(a,b,c,\boldsymbol{\tau}^T)^T$, and tests of hypotheses, we require the expected information matrix. Interval estimation for the model parameters can be based on standard likelihood theory. The elements of the information matrix for (27) are given in the Appendix. Under suitable regularity conditions, the asymptotic distribution of the MLE, $\widehat{\theta}$, is multivariate normal with mean vector $\theta$ and covariance matrix estimated by $\{-\partial^2\ell(\theta)/\partial\theta\partial\theta^T\}$ at $\theta=\widehat{\theta}$. The required second derivatives were computed numerically.

Consider two nested KBG-$G$ distributions: a KBG-$G_A$ distribution with parameters $\theta_1,\ldots,\theta_r$ and maximized log-likelihood $-2\,\ell(\widehat{\theta}_A)$; and, a KBG-$G_B$ distribution containing the same parameters $\theta_1,\ldots,\theta_r$ plus additional parameters $\theta_{r+1},\ldots,\theta_p$ and maximized log-likelihood $-2\,\ell(\widehat{\theta}_B)$, the models being identical otherwise. For testing the KBG-$G_A$ distribution against the KBG-$G_B$ distribution, the likelihood ratio statistic (LR) is equal to $w=-2\,\{\ell(\widehat{\theta}_A)-\ell(\widehat{\theta}_B)\}$. It has an asymptotic $\chi^2_{p-r}$ distribution.

We compare non-nested KBG-$G$ distributions by using the Akaike information criterion given by $\mathrm{AIC}=-2\,\ell(\widehat{\theta})+2p^*$ and the Bayesian information criterion defined by $\mathrm{BIC}=-2\,\ell(\widehat{\theta})+p^*\log(\theta)$, where $p^*$ is the number of model parameters. The distribution with the smallest value for any of these criteria (among all distributions considered) is usually taken as the one that gives the best description of the data.

## 10. Application-Ball bearing fatigue data

In this section, we shall compare the fits of the KBGW, beta Weibul (BW), Birnbaum-Saunders (BS) and Weibull distributions to the data set studied by Lieblein and Zelen (1956). They described the data from fatigue endurance tests for deep-groove ball bearings. The main objective of the study was to estimate parameters in the equation relating bearing life to load. The data are a subset of $n=23$ bearing failure times for units tested at one level of stress reported by Lawless (1982). Because of the lower

***Table 1:*** *MLEs and information criteria for the ball bearing data.*

| Model | $d$ | $\beta$ | $a$ | $b$ | $c$ | AIC | BIC |
|---|---|---|---|---|---|---|---|
| KBGW | 1.5040 | 0.0456 | 15.9411 | 0.1972 | 12.7943 | 223.9 | 229.6 |
| BW | 1.5254 | 0.0435 | 3.3335 | 0.2032 | 0 | 233.7 | 238.3 |
| Weibull | 2.1018 | 0.0122 | 1 | 1 | 0 | 231.3 | 233.6 |
| | $\alpha$ | $\beta$ | - | - | - | | |
| BS | 0.5391 | 62.9794 | - | - | - | 230.2 | 232.5 |

bound on cycles (or time) to fail at zero, the distributional shape is typical of reliability data.

Firstly, in order to estimate the model parameters, we consider the MLE method discussed in Section 9. We take initial estimates of $d$ and $\beta$ as those obtained by fitting the Weibull distribution. All computations were performed using the statistical software R. Table 1 lists the MLEs of the parameters and the values of the following statistics: AIC and BIC as discussed before. The results indicate that the KBGW model has the smallest values for these statistics among all fitted models. So, it could be chosen as the most suitable model.

A comparison of the proposed distribution with some of its sub-models using LR statistics is shown in Table 2. The $p$-values indicate that the proposed model yields the best fit to the data set. In order to assess if the model is appropriate, we plot in Figure 3 the histogram of the data and the fitted KBGW, BW, Weibull and BS pdfs. We conclude that the KBGW distribution is a suitable model for the data.

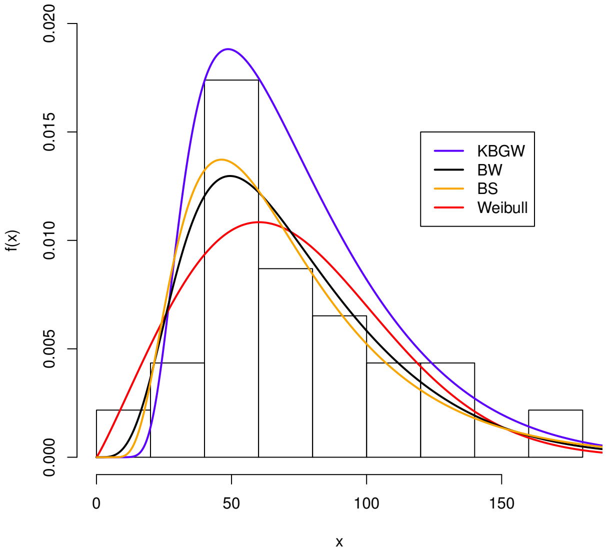

***Figure 3:*** *Fitted KBGW, BW, Weibull and BS pdfs for the ball bearing data.*

***Table 2:*** *LR statistics for the ball bearing data.*

| Model | Hypotheses | Statistic w | $p$-value |
|---|---|---|---|
| KBGW vs BW | $H_0 : c = 0$ vs $H_1 : H_0$ is false | 11.85 | 0.00057 |
| KBGW vs Weibull | $H_0 : a = b = 1$ and $c = 0$ vs $H_1 : H_0$ is false | 13.45 | 0.00375 |

Secondly, we apply formal goodness-of-fit tests in order to verify which distribution gives the best fit to the data. We consider the Cramér-Von Mises ($W^*$) and Anderson-Darling ($A^*$) statistics. In general, the smaller the values of the statistics, $W^*$ and $A^*$, the better the fit to the data. Let $H(x;\boldsymbol{\theta})$ denote a cdf, where the form of $H$ is known but $\boldsymbol{\theta}$ (a $k$-dimensional parameter vector, say) is unknown. To obtain the statistics, $W^*$ and $A^*$, we proceed as follows: (i) compute $v_i = H(x_i;\widehat{\boldsymbol{\theta}})$, where the $x_i$'s are in ascending order, $y_i = \Phi^{-1}(\cdot)$ is the standard normal quantile function and $u_i = \Phi\{(y_i - \bar{y})/s_y\}$, where $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ and $s_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$; (ii) calculate $W^2 = \sum_{i=1}^n \{u_i - (2i-1)/(2n)\}^2 + 1/(12n)$ and $A^2 = -n - n^{-1}\sum_{i=1}^n \{(2i-1)\log(u_i) + (2n+1-2i)\log(1-u_i)\}$ and (iii) modify $W^2$ into $W^* = W^2(1+0.5/n)$ and $A^*$ into $A^* = A^2(1+0.75/n+2.25/n^2)$. For further details, the reader is referred to Chen and Balakrishnan (1995). The values of the statistics, $W^*$ and $A^*$, for all fitted models are given in Table 3. Thus, according to these formal tests, the KBGW model fits the data better than other models. These results illustrate the flexibility of the KBGW distribution and the necessity for the additional shape parameters.

***Table 3:*** *Goodness-of-fit tests for the ball bearing data.*

| Model | Statistic | |
|---|---|---|
| | $W^*$ | $A^*$ |
| KBGW | 0.00507 | 0.19916 |
| BW | 0.20587 | 0.57785 |
| Weibull | 0.13587 | 0.34791 |
| BS | 0.02298 | 0.34791 |

## 11. Conclusions

Following the idea of the class of beta generalized distributions and the distribution due to Ng and Kotz (1995), we define a new family of Kummer beta generalized (KBG) distributions to extend several widely known distributions such as the normal, Weibull, gamma and Gumbel distributions. For each continuous $G$ distribution, we define the corresponding KBG-$G$ distribution using simple formulae. Some mathematical properties of the KBG distributions are readily obtained from those of the parent distributions. The moments of any KBG-$G$ distribution can be expressed explicitly in terms of infinite weighted sums of probability weighted moments (PWMs) of the $G$ distribution.

The same happens for the moments of order statistics of the KBG distributions. We discuss maximum likelihood estimation and inference on the parameters. We consider likelihood ratio statistics and goodness-of-fit tests to compare the KBG-*G* model with its baseline model. An application to real data shows the feasibility of the proposed class of models. We hope this generalization may attract wider applications in statistics.

## Acknowledgments

The authors are grateful to the referees and the Associate Editor for helpful comments and suggestions. The authors gratefully acknowledge grants from CNPq (Brazil).

## Appendix: elements of the information matrix

The elements of this matrix for (27) can be worked out as:

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial a^2}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial a}\right)-\frac{\partial^2 K}{\partial a^2}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial b\partial c}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial b}\right)\left(\frac{\partial K}{\partial c}\right)-\frac{\partial^2 K}{\partial b\partial c}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial c^2}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial c}\right)-\frac{\partial^2 K}{\partial c^2}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial a\partial b}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial a}\right)\left(\frac{\partial K}{\partial b}\right)-\frac{\partial^2 K}{\partial a\partial b}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial a\partial c}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial a}\right)\left(\frac{\partial K}{\partial c}\right)-\frac{\partial^2 K}{\partial a\partial c}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial b^2}\right)=-\frac{n}{K}\mathrm{E}\left[\frac{1}{K}\left(\frac{\partial K}{\partial b}\right)-\frac{\partial^2 K}{\partial b^2}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial a\partial\gamma_j}\right)=-\sum_{i=1}^{n}\mathrm{E}\left[\frac{1}{G(x_i;\gamma)}\frac{\partial G(x_i;\gamma)}{\partial\gamma_j}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial b\partial\gamma_j}\right)=-\sum_{i=1}^{n}\mathrm{E}\left[\frac{1}{1-G(x_i;\gamma)}\frac{\partial G(x_i;\gamma)}{\partial\gamma_j}\right],$$

$$\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial c\partial\gamma_j}\right)=\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial g\left(x_i;\gamma\right)}{\partial\gamma_j}\right],$$

$$\begin{aligned}\mathrm{E}\left(-\frac{\partial^2\ell(\theta)}{\partial\gamma_k\partial\gamma_j}\right) &= \sum_{i=1}^{n}\mathrm{E}\left[\frac{1}{g^2\left(x_i;\gamma\right)}\frac{\partial^2 g\left(x_i;\gamma\right)}{\partial\gamma_j\partial\gamma_k}\right]+c\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial^2 g\left(x_i;\gamma\right)}{\partial\gamma_j\partial\gamma_k}\right]+\\ &\quad \sum_{i=1}^{n}\mathrm{E}\left[\frac{(a-1)}{G^2\left(x_i;\gamma\right)}\frac{\partial^2 G\left(x_i;\gamma\right)}{\partial\gamma_j\partial\gamma_k}\right]+\sum_{i=1}^{n}\mathrm{E}\left[\frac{(1-b)}{\left\{1-G\left(x_i;\gamma\right)\right\}^2}\frac{\partial^2 G\left(x_i;\gamma\right)}{\partial\gamma_j\partial\gamma_k}\right]\end{aligned}$$

for $j=1,\ldots,p$, where

$$\begin{aligned}\frac{\partial^2 K}{\partial a^2} &= -\left\{\frac{[\psi'(a)-\psi'(a+b)]}{{}_1F_1(a,a+b,-c)}+\frac{[\psi(a)-\psi(a+b)]^2}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial a}\right.\\ &\quad+\frac{1}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial^2\,{}_1F_1(a,a+b,-c)}{\partial a^2}+\frac{[\psi(a)-\psi(a+b)]^2}{{}_1F_1(a,a+b,-c)}\\ &\quad+\frac{2}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial a}\\ &\quad\left.+\frac{1}{[{}_1F_1(a,a+b,-c)]^3}\left(\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial a}\right)^2\right\},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 K}{\partial b^2} &= -\left\{\frac{[\psi'(b)-\psi'(a+b)]}{{}_1F_1(a,a+b,-c)}+\frac{[\psi(b)-\psi(a+b)]^2}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial b}\right.\\ &\quad+\frac{1}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial^2\,{}_1F_1(a,a+b,-c)}{\partial b^2}+\frac{[\psi(b)-\psi(a+b)]^2}{{}_1F_1(a,a+b,-c)}\\ &\quad+\frac{2}{[{}_1F_1(a,a+b,-c)]^2}\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial b}\\ &\quad\left.+\frac{1}{[{}_1F_1(a,a+b,-c)]^3}\left(\frac{\partial\,{}_1F_1(a,a+b,-c)}{\partial b}\right)^2\right\},\end{aligned}$$

$$\frac{\partial^2 K}{\partial c^2}=-\left\{\frac{a(a+1){}_1F_1(a+2,a+b+2,-c)}{(a+b)B(a,b){}_1F_1(a,a+b,-c)}+\frac{a^2\left[{}_1F_1(a+1,a+b+1,-c)\right]^2}{(a+b)^2B(a,b)\left[{}_1F_1(a,a+b,-c)\right]^2}\right\},$$

$$\begin{aligned}
\frac{\partial^2 K}{\partial a \partial b} = & -\left\{ \frac{[\psi'(a+b)]}{{}_1F_1(a,a+b,-c)} + \frac{[\psi(b)-\psi(a+b)]}{[{}_1F_1(a,a+b,-c)]^2} \frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial a} \right. \\
& + \frac{1}{[{}_1F_1(a,a+b,-c)]^2} \frac{\partial^2\, {}_1F_1(a,a+b,-c)}{\partial a \partial b} + \left[ \frac{\psi(b)-\psi(a+b)}{{}_1F_1(a,a+b,-c)} \right]^2 \\
& + \frac{2\,[\psi(a)-\psi(a+b)]}{[{}_1F_1(a,a+b,-c)]^2} \frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial a} \\
& \left. + \frac{2}{[{}_1F_1(a,a+b,-c)]^3} \frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial b} \frac{\partial\, {}_1F_1(a,a+b,-c)}{\partial a} \right\},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 K}{\partial a \partial c} = & \; {}_1F_1(a+1,a+b+1,-c) + \frac{a}{(a+b)} + a\,[\psi'(a) - \psi'(a+b)] \\
& + \frac{a}{{}_1F_1(a,a+b,-c)} \frac{\partial\, {}_1F_1(a+1,a+b+1,-c)}{\partial a} \\
& + \frac{a}{{}_1F_1(a+1,a+b+1,-c)} \frac{\partial {}_1F_1(a+1,a+b+1,-c)}{\partial a},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 K}{\partial b \partial c} = & \; {}_1F_1(a+1,a+b+1,-c) + \frac{a}{(a+b)} + a\,[\psi'(b) - \psi'(a+b)] \\
& + \frac{a}{{}_1F_1(a,a+b,-c)} \frac{\partial\, {}_1F_1(a+1,a+b+1,-c)}{\partial b} \\
& + \frac{a}{{}_1F_1(a+1,a+b+1,-c)} \frac{\partial {}_1F_1(a+1,a+b+1,-c)}{\partial b},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2\, {}_1F_1(a,a+b,-c)}{\partial a^2} = & -\left[ \psi'(a+b) - \psi'(a) + \{\psi(a)-\psi(a+b)\}^2 \right] {}_1F_1(a,a+b,-c) \\
& - \sum_{k=0}^{\infty} \frac{(a)_k(-c)^k}{k!(a+b)_k} \,[-2\psi(a)\psi(a+k) + 2\psi(a+b)\psi(a+k) \\
& + 2\,\psi(a)\psi(a+b+k) - 2\psi(a+b)\psi(a+b+k) + \psi^2(a+k) \\
& - 2\,\psi(a+k)\psi(a+b+k) + \psi^2(a+b+k) \\
& + \psi'(a+k) - \psi'(a+b+k)]\,,
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2\, {}_1F_1(a,a+b,-c)}{\partial b^2} &= -\left[\psi'(a+b)-\psi^2(a+b)\right]\, {}_1F_1(a,a+b,-c)\\
&\quad -\sum_{k=0}^{\infty}\frac{(a)_k(-c)^k}{k!(a+b)_k}\left[-2\psi(a+b)\psi(a+b+k)\right.\\
&\quad \left.-\psi'(a+b+k)+\psi^2(a+b+k)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2\, {}_1F_1(a,a+b,-c)}{\partial a\partial b} &= \left[\psi'(a+b)-\psi^2(a+b)-\psi(a)\psi(a+b)\right]\, {}_1F_1(a,a+b,-c)\\
&\quad -\sum_{k=0}^{\infty}\frac{(a)_k(-c)^k}{k!(a+b)_k}\left[2\psi(a+b)\psi(a+b+k)-\psi^2(a+b+k)\right.\\
&\quad -\psi(a+k)\psi(a+b)+\psi(a+k)\psi(a+b+k)\\
&\quad \left.-\psi(a)\psi(a+b+k)+\psi(a+b+k)\right].
\end{aligned}
$$

## References

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of Mathematical Functions*. Dover Publications, New York.

Ahsanullah, M. and Nevzorov, V. B. (2005). *Order Statistics: Examples and Exercises*. Nova Science Publishers, Inc, Hauppauge, New York.

Alexander, C., Cordeiro, G. M., Ortega, E. M. M. and Sarabia, J. M. (2012). Generalized beta-generated distributions. *Computational Statistics and Data Analysis*, 56, 1880-1897.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*. John Wiley and Sons, New York.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 31, 171-178.

Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, 27, 154-161.

Cordeiro, G. M. and de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, 883-893.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, third edition. John Wiley and Sons, Hoboken, New Jersey.

Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F. (1953). *Higher Transcendental Functions*, volume I. McGraw-Hill Book Company, New York.

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

Gordy, M. B. (1998). Computationally convenient distributional assumptions for common-value actions. *Computational Economics*, 12, 61-78.

Gradshteyn I. S. and Ryzhik I. M. (2000). *Table of Integrals, Series, and Products*. Academic Press, San Diego.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, 15, 1049-1054.

Gupta, R. C., Gupta, P. L. and Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Communications Statistics-Theory and Methods*, 27, 887-904.

Gupta, R. D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. *Biometrical Journal*, 43, 117-130.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New York.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1987). *Extremes and Related Properties of Random Sequences and Process*. Springer Verlag, New York.

Lieblein J. and Zelen, M. (1956). Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research National Bureau of Standards*, 57, 273-316.

Mudholkar, G. S, Srivastava, D. K. and Friemer, M. (1995). The exponential Weibull family: A reanalysis of the bus-motor failure data. *Technometrics*, 37, 436-445.

Nadarajah, S. and Gupta, A. K. (2004). The beta Fréchet distribution. *Far East Journal of Theoretical Statistics*, 15, 15-24.

Nadarajah, S. and Gupta, A. K. (2007). A generalized gamma distribution with application to drought data. *Mathematics and Computer in Simulation*, 74, 1-7.

Nadarajah, S. and Kotz, S. (2004). The beta Gumbel distribution. *Mathematical Problems in Engineering*, 10, 323-332.

Nadarajah, S. and Kotz, S. (2005). The beta exponential distribution. *Reliability Engineering and System Safety*, 91, 689-697.

Nadarajah, S. and Kotz, S. (2006). The exponentiated type distribution. *Acta Applicandae Mathematicae*, 92, 97-111.

Ng, K. W. and Kotz, S. (1995). Kummer-gamma and Kummer-beta univariate and multivariate distributions. *Research Report*, 84, Department of Statistics, The University of Hong Kong, Hong Kong.

Paranaíba, P. F., Ortega, E. M. M., Cordeiro, G. M. and Pescim, R. R. (2011). The beta Burr XII distribution with application to lifetime data. *Computational Statistics and Data Analysis*, 55, 1118-1136.

Pescim, R. R., Demétrio, C. G. B., Cordeiro, G. M., Ortega, E. M. M. and Urbano, M. R. (2010). The beta generalized half-normal distribution. *Computation Statistics and Data Analysis*, 54, 945-957.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54, 507-554.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 30, 50-64.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location, scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 1-46.

# Hurdle negative binomial regression model with right censored count data

Seyed Ehsan Saffari[1], Robiah Adnan[2] and William Greene[3]

**Abstract**

A Poisson model typically is assumed for count data. In many cases because of many zeros in the response variable, the mean is not equal to the variance value of the dependent variable. Therefore, the Poisson model is no longer suitable for this kind of data. Thus, we suggest using a hurdle negative binomial regression model to overcome the problem of overdispersion. Furthermore, the response variable in such cases is censored for some values. In this paper, a censored hurdle negative binomial regression model is introduced on count data with many zeros. The estimation of regression parameters using maximum likelihood is discussed and the goodness-of-fit for the regression model is examined.

*MSC:* 62J12

*Keywords:* Hurdle negative binomial regression, censored data, maximum likelihood method, simulation.

## 1. Introduction

Commonly, for the modelling of counts such as the number of reported insurance claims, the starting point is the Poisson distribution:

$$f_{Y_i}(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{1}$$

where covariates are included in the model by the parameter $\lambda_i = \exp(x_i^\mathsf{T}\beta)$ where $x_i$ is a vector of explanatory variables (Dionne and Vanasse, 1989). The Poisson distribution

[1,2] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.

[3] Department of Economics, Stern School of Business, New York University 44 West 4th St., New York, NY, 10012, USA.

Received: December 2011
Accepted: May 2012

is equidispersed since its mean and variance are both equal to $\lambda_i$. Because the Poisson distribution has some severe drawbacks that limit its use, other distributions can be used, such as hurdle models (Boucher *et al.*, 2007).

Mullahy (1986) has first discussed hurdle count data models. Hurdle models allow for a systematic difference in the statistical process governing individuals (observations) below the hurdle and individuals above the hurdle. In particular, a hurdle model is mixed by a binary outcome of the count being below or above the hurdle (the selection variable), with a truncated model for outcomes above the hurdle. That is why hurdle models sometimes are also called two-part models.

The most important usage of a hurdle count data model is the hurdle at zero. The hurdle at zero formulation can account for excess zeros. It means that this model can be used in situations where there are many zeros at the response variable. In this case, the hurdle at zero defines a probability $(\Pr(Y=0))$ that is the first part of the two part-models.

The hurdle model is flexible and can handle both under- and overdispersion problem. A generalized hurdle model is introduced by Gurmu (1998) for the analysis of overdispersed or underdispersed count data. Greene (2005) has discussed about the comparison between hurdle and zero-inflated models as two part-models. Some researchers have discussed the applications of hurdle models, such as Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997). A hurdle model to the annual number of recreational boating trips by a family is discussed by Gurmu and Trivedi (1996). Dalrymple, Hudson and Ford (2003) applied three mixture models including a hurdle model and argued its application in the incidence of sudden infant death syndrome (SIDS). Boucher, Denuit and Guillen (2007) compared generalized heterogeneous, zero-inflated, hurdle, and compound frequency models for the annual number of claims reported to the insurer. Saffari, Adnan and Greene (2011) argued the overdispersion problem on count data using a right truncated Poisson regression model.

Suppose that $g_1(0)$ is the probability value when the value for response variable is zero and that $g_2(k), k=1,2,\ldots$ is a probability function when the response variable is a positive integer. Therefore, the probability function of the hurdle-at-zero model is given by:

$$p(Y_i=k)=\begin{cases} g_1(0), & k=0, \\ (1-g_1(0))\,g_2(k), & k=1,2,\ldots \end{cases} \tag{2}$$

Mullahy (1986) discussed the hurdle-at-zero model and he believes that both parts of the hurdle model are based on probability functions for nonnegative integers such as $f_1$ and $f_2$. In terms of the general model above, let $g_1(0)=f_1(0)$ and $g_2(k)=f_2(k)/(1-f_2(0))$. In the case of $g_2$, normalization is required because $f_2$ has support over the nonnegative integers $(k=0,1,\ldots)$ whereas the support of $g_2$ must be over the positive integers $(k=1,2,\ldots)$. This means that we need to truncate the probability function $f_2$. However, this is a theoretical concept, i.e., truncation on $f_2$ does not mean that there is truncation of the population here. All we need to do is to work with a distribution with positive support, and the second part of a hurdle model can use a displaced distribution or any distribution with positive support as well.

Under the Mullahy (1986) assumptions, the probability distribution of the hurdle-at-zero model is given by

$$\begin{aligned} f(Y=0) &= f_1(0) \\ f(Y=k) &= \frac{1-f_1(0)}{1-f_2(0)} f_2(k) = \theta f_2(k), \quad k=1,2,\dots \end{aligned}$$

where $f_2$ is referred to as *parent*-process. The numerator of $\theta$ presents the probability of crossing the hurdle and the denominator gives a normalization that accounts for the (purely technical) truncation of $f_2$. It follows that if $f_1 = f_2$ or, equivalently, $\theta = 1$ then the hurdle model collapses to the parent model. The expected value of the hurdle model is given by

$$E(Y) = \theta \sum_{k=1}^{\infty} k f_2(k) \tag{3}$$

and the difference between this expected value and the expected value of the parent model is the factor $\theta$. In addition, the variance value of the hurdle model is given by

$$Var(Y) = \theta \sum_{k=1}^{\infty} k^2 f_2(k) - \left[\theta \sum_{k=1}^{\infty} k f_2(k)\right]^2 \tag{4}$$

If $\theta$ exceeds 1, it means that the probability of crossing the hurdle is greater than the sum of the probabilities of positive outcomes in the parent model. Therefore, increasing the expected value of the hurdle model is related to the expected value of the parent model. Alternatively, if $\theta$ is less than 1 (that is the usual case in an application with excess zeros), it means that the probability of not crossing the hurdle is greater than the probability of a zero in the parent model, thus decreasing the expected value of the hurdle model relatively to the expected value of the parent model. Therefore, this model gives a new explanation of excess zeros as being a characteristic of the mean function rather than a characteristic of the variance function. The mean function of the hurdle model provides additional nonlinearities relative to the standard model in order to account for the corner solution outcome, much as in other corner solution models, and this is just like as how a Tobit model works.

Consequently, the model can be overdispersed and that depends on the value of the parent processes. To overcome overdispersion, we would like to cut the values of the response variable that are very big. In statistics, this is called truncation and because we want to truncate the values that are bigger than a constant, it is called a right truncation.

There are many options to choose the processes $f_1$ and $f_2$. Some of the most popular hurdle model choices are nested models where $f_1$ and $f_2$ come from the same distribution, such as the Poisson distribution (Mullahy, 1986) or the Negative Binomial (Pohlmeier and Ulrich, 1995). However, non-nested models (Grootendorst (1995), (Gurmu, 1998), or Winkelmann (2003)) can also be used. These models do not nest with a standard count distributions such as the Poisson or the NB types, but are overlapping with suggested model by Vuong (1989) since models can be equivalent for certain parameter restrictions.

Maximum likelihood is used to estimate the parameters. The log-likelihood function of a hurdle model can be expressed as:

$$ll = \sum_{i=1}^{n} I_{\{y_i=0\}} \log f_1(0;\theta_i) + I_{\{y_i>0\}} \log(1 - f_1(0;\theta_i)) + \sum_{i=1}^{n} I_{\{y_i>0\}} \log \frac{f_2(y_i;\theta_i)}{1 - f_2(0;\theta_i)} \quad (5)$$

The interesting aspect of the hurdle model is to estimate the parameters by two separate steps. In fact, we can estimate the zero-part parameters by using MLE on the first part of the likelihood function while the other parameters only use the second part, only composed with non-zero elements. We have used SAS code to implement this algorithm and this characteristic of the model helps us to save computer time in the estimation (Chou and Steenhard, 2009).

In this article, the main objective is to explain how we can use hurdle negative binomial regression model in right censored data. In Section 2, the hurdle negative binomial regression model is defined and the likelihood function of hurdle negative binomial regression model in right censored data is formulated. In Section 3, the parameter estimation is discussed using maximum likelihood. In Section 4, the goodness-of-fit for the regression model is examined and a test statistic for examining the dispersion of regression model in right censored data is proposed. An example is conducted for a censored hurdle negative binomial regression model in terms of the parameter estimation, standard errors and goodness-of-fit statistic in Section 5.

## 2. The model

Let $Y_i (i = 1, 2, \ldots, n)$ be a nonnegative integer-valued random variable and suppose $Y_i = 0$ is observed with a frequency significantly higher than can be modeled by the usual model. We consider a hurdle negative binomial regression model in which the response variable $Y_i (i = 1, \ldots, n)$ has the distribution

$$\Pr(Y_i = y_i) = \begin{cases} w_0, & y_i = 0, \\ (1 - w_0) \dfrac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \dfrac{(1 + \alpha\mu_i)^{-\alpha^{-1} - y_i} \alpha^{y_i} \mu_i^{y_i}}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}}, & y_i > 0, \end{cases} \quad (6)$$

or

$$\Pr(Y_i = y_i) = \begin{cases} w_0, & y_i = 0, \\ (1-w_0)\dfrac{g}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}}, & y_i > 0, \end{cases} \tag{7}$$

where

$$g = g(y_i;\mu_i,\alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}(1+\alpha\mu_i)^{-\alpha^{-1}-y_i}\alpha^{y_i}\mu_i^{y_i} \tag{8}$$

where $\alpha(\geq 0)$ is a dispersion parameter that is assumed not to depend on covariates. In addition, we suppose $0 < w_0 < 1$ and $w_0 = w_0(z_i)$ satisfy

$$\operatorname{logit}(w_0) = \log(\frac{w_0}{1-w_0}) = \sum_{j=1}^{m} z_{ij}\delta_j \tag{9}$$

where $(z_{i1} = 1, z_{i2}, \ldots, z_{im})$ is the $i$-th row of covariate matrix $Z$ and $(\delta_1, \delta_2, \ldots, \delta_m)$ is an unknown $m$-dimensional column vector of parameters. In this set up, the non-negative function $w_0$ is modeled via logit link function. This function is linear and other appropriate link functions that allow $w_0$ being negative may be used. In addition, there is interest in capturing any systematic variation in $\mu_i$, the value of $\mu_i$ is most commonly placed within a loglinear model

$$\log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j \tag{10}$$

and $\beta_j$'s are the independent variables in the regression model and $m$ is the number of these independent variables. Furthermore, in this paper we suppose that $w_0$ and $\beta_j$ are not related.

The value of response variable, $Y_i$, for some observations in a data set, may be censored. If censoring occurs for the $i$th observation, we have $Y_i \geq y_i$ (right censoring). However, if no censoring occurs, we know that $Y_i = y_i$. Thus, we can define an indicator variable $d_i$ as

$$d_i = \begin{cases} 1 & \text{if } Y_i \geq y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We can now write

$$\Pr(Y_i \geq y_i) = \sum_{j=y_i}^{\infty} \Pr(Y_i = j) = 1 - \sum_{j=0}^{y_i-1} \Pr(Y_i = j) \tag{12}$$

Therefore, the log-likelihood function of the censored hurdle regression model can be written as

$$\log L(\theta_i; y_i) = \sum_{i=1}^{n} \left\{ (1-d_i)\left[ I_{y_i=0} \log f(0;\theta_i) + I_{y_i>0} f(y_i;\theta_i) \right] + d_i \log \left( \sum_{j=y_i}^{\infty} \Pr(Y_i = j) \right) \right\} \tag{13}$$

We now obtain the log-likelihood function for the hurdle negative binomial regression model, we have

$$\begin{aligned} LL \quad &= \sum_{i=1}^{n} \Big\{ (1- \quad d_i) \Big[ I_{y_i=0} \log w_0 + I_{y_i>0} \Big\{ \log(1-w_0) + \log g - \log\Big(1-(1+\alpha\mu_i)^{-\alpha^{-1}}\Big) \Big\} \Big] \\ &\qquad + d_i \log \sum_{j=y_i}^{\infty} \Pr(Y_i = j) \Big\} \end{aligned} \tag{14}$$

## 3. Parameter estimation

In this section we estimate the parameters by maximum likelihood. By taking the partial derivatives of the likelihood function and setting them equal to zero, the likelihood equation for estimating the parameters is obtained. Thus we obtain

$$\begin{aligned} \frac{\partial LL}{\partial \beta_r} =& \sum_{i=1}^{k} \Bigg\{ (1-d_i) I_{y_i>0} \left[ \frac{g'_{\mu_i}}{g} - \frac{(1+\alpha\mu_i)^{-\alpha^{-1}-1}}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} \right] \mu_i x_{ir} \\ &+ \frac{d_i}{\sum_{j=y_i}^{\infty} \Pr(Y_i = j)} \frac{\partial \sum_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \beta_r} \Bigg\} = 0 \\ \frac{\partial LL}{\partial \alpha} =& \sum_{i=1}^{k} \Bigg\{ (1-d_i) I_{y_i>0} \Bigg[ \frac{g'_\alpha}{g} \\ &- \frac{\alpha^{-1}\mu_i/(1+\alpha\mu_i) - \alpha^{-2}\log(1+\alpha\mu_i)}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} (1+\alpha\mu_i)^{-\alpha^{-1}} \Bigg] \\ &+ \frac{d_i}{\sum_{j=y_i}^{\infty} \Pr(Y_i = j)} \frac{\partial \sum_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \alpha} \Bigg\} = 0 \\ \frac{\partial LL}{\partial \delta_s} =& \sum_{i=1}^{n} (1-d_i) \left[ I_{y_i=0}(1-w_0) - I_{y_i>0} w_0 \right] z_{is} = 0 \end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \sum\limits_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \beta_r} &= \sum_{j=y_i}^{\infty} (1-w_0) \frac{g'_{\mu_i}(1-(1+\alpha\mu_i)^{-\alpha^{-1}}) - (1+\alpha\mu_i)^{-\alpha^{-1}-1} g}{(1-(1+\alpha\mu_i)^{-\alpha^{-1}})^2} \mu_i x_{ir} \\
\frac{\partial \sum\limits_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \alpha} &= \sum_{j=y_i}^{\infty} (1-w_0) \left[ \frac{g'_\alpha}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} + \frac{\log(1+\alpha\mu_i)^{\alpha^{-2}} - \frac{\alpha^{-1}\mu_i}{1+\alpha\mu_i}}{\left(1-(1+\alpha\mu_i)^{-\alpha^{-1}}\right)^2} g(1+\alpha\mu_i)^{-\alpha^{-1}} \right] \\
g'_\alpha &= \left( \frac{\Gamma'(y_i+\alpha^{-1})}{\Gamma(y_i+\alpha^{-1})} - \frac{\Gamma'(\alpha^{-1})}{\Gamma(\alpha^{-1})} + \alpha^{-2}\log(1+\alpha\mu_i) - \frac{\alpha^{-1}+y_i}{1+\alpha\mu_i}\mu_i + \alpha^{-1} y_i \right) g \\
g'_\mu &= \frac{y_i - \mu_i}{\mu_i(1+\alpha\mu_i)} g
\end{aligned}$$

Since these partial derivative equations cannot be further simplified, we have applied the Newton-Raphson method with ridging as the optimization algorithm, using code in SAS given in the Appendix. Furthermore, the Convergence of the algorithm does not necessarily mean that a global maximum has been found, it just means that the convergence criteria have been achieved, and thus it can be a local maximum.

## 4. Goodness-of-fit statistics

For count regression models, a measure of goodness of fit may be based on the deviance statistic $D$ defined as

$$D = -2\left[\log L(\hat{\theta}_i;\hat{\mu}_i) - \log L(\hat{\theta}_i;y_i)\right] \tag{15}$$

where $\log L(\hat{\theta}_i;\hat{\mu}_i)$ and $\log L(\hat{\theta}_i;y_i)$ are the model's likelihood evaluated respectively under $\hat{\mu}_i$ and $y_i$. The log-likelihood functions are given in equation (7).

For an adequate model, the asymptotic distribution of the deviance statistic $D$ is chi-square distribution with $n-k-1$ degrees of freedom. Therefore, if the value for the deviance statistic $D$ is close to the degrees of freedom, the model may be considered as adequate. When we have many regression models for a given data set, the regression model with the smallest value of the deviance statistic $D$ is usually chosen as the best model for describing the given data.

In many data sets, the $\mu_i$'s may not be reasonably large and so the deviance statistic $D$ may not be suitable. Thus, the log-likelihood statistic $\log(\hat{\theta}_i;y_i)$ can be used as an alternative statistic to compare the different models. Models with the largest log-likelihood value can be chosen as the best model for describing the data under consideration.

When there are several maximum likelihood models, one can compare the performance of alternative models based on several likelihood measures which have been proposed in the statistical literature. The AIC and BIC are two of the most regularly used measures. The AIC is defined as

$$\mathrm{AIC} = -2l + 2p \tag{16}$$

where $l$ denotes the log likelihood evaluated for estimated parameters $\mu$ and $p$ the number of parameters. For this measure, the smaller the AIC, the better the model is.

## 5. An application

The state wildlife biologists want to model how many fish[1] are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there are no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish. We have data on 250 groups that went to a park. Each group was questioned about how many fish they caught (*count*), how many children were in the group (*child*), how many people were in the group (*persons*), and whether or not they brought a camper to the park (*camper*).

We will use the variables $child, persons$ and $camper$ in our model. Table 1 shows the descriptive statistics of using variables and also the *camper* variable has two values, zero and one as Table 2. In addition, Figure 1 shows the histogram of the *count* variable before censoring.

We have considered the model as follow

$$\begin{aligned} \log(\mu) &= b_0 + b_1 * camper + b_2 * persons + b_3 * child, \\ \mathrm{logit}(w_0) &= a_0 + a_1 * child \end{aligned}$$

Furthermore, we put two censoring points, $c_1 = 3, c_2 = 5$. Table 3 shows the estimation of the parameters according to different censoring constants. Also, the $-2\log L$ and AIC are presented as the goodness-of-fit measures.

***Table 1:*** *Descriptive statistics for the fish data.*

| Variable | Mean | Std Dev | Min | Max | Variance |
|---|---|---|---|---|---|
| *count* | 3.296 | 11.635028 | 0 | 149 | 135.373880 |
| *child* | 0.684 | 0.850315 | 0 | 3 | 0.723036 |
| *persons* | 2.528 | 1.112730 | 1 | 4 | 1.238169 |

[1] The fish dataset is available at the UCLA Academic Technology Services website, http://www.ats.ucla.edu.

***Table 2:*** *camper variable description.*

| *camper* | Frequency | Percent |
|---|---|---|
| 0 | 103 | 41.2 |
| 1 | 147 | 58.8 |

***Table 3:*** *Parameter estimation.*

| Parameter | $c_1 = 3$ | $c_2 = 5$ |
|---|---|---|
| $b_0$ | $-1.0922$ | $-0.9616$ |
| | (0.5998) | (0.4764) |
| $b_1$ | 0.7043 | 0.6079 |
| | (0.3235) | (0.2702) |
| $b_2$ | 0.7397 | 0.7227 |
| | (0.2086) | (0.1533) |
| $b_3$ | $-0.9130$ | $-0.9266$ |
| | (0.3449) | (0.2807) |
| $a_0$ | $-0.3843$ | $-0.3843$ |
| | (0.1703) | (0.1703) |
| $a_1$ | 1.1110 | 1.1110 |
| | (0.2049) | (0.2049) |
| $\alpha$ | 0.5673 | 0.6225 |
| | (0.4388) | (0.3412) |
| $-2\log L$ | 540.9 | 618.1 |
| AIC | 554.9 | 632.1 |

According to the censoring points, there is 22.8% censored data when $c_1 = 3$. It means that 22.8% of the values of the response variable (*count*) are 0, 1, 2, 3 and the remaining 77.2% of values of the response variable are greater than 3, that is censored in the model. Also the percentage of the censoring for $c_2 = 5$ is 12%. For example, the 25th value of the response variable is $count_{25} = 30$, and the values of the independent variables are as follow

$$camper_{25} = 1,\ persons_{25} = 3,\ child_{25} = 0$$

So we want to censor only the value of the response variable ($count_{25} > censoredpoint$).

The estimated parameter for camper variable of the model is a positive value for both censoring points, it means that while being a camper ($camper = 1$), the expected $\log(count)$ will be increased by 0.7043 and 0.6079, respectively when $c_1 = 3$ and $c_2 = 5$. Also, the effect of *persons* is positively associated with the number of fish caught for both censoring points, and the expected $\log(count)$ will be increased for a unit increase in *persons* for the first and the second censoring point, respectively, by 0.7397 and 0.7227. But, the expected $\log(count)$ will be decreased for a unit increase in *child* for by 0.9130 and 0.9266, respectively for the first and the second censoring point. Further-

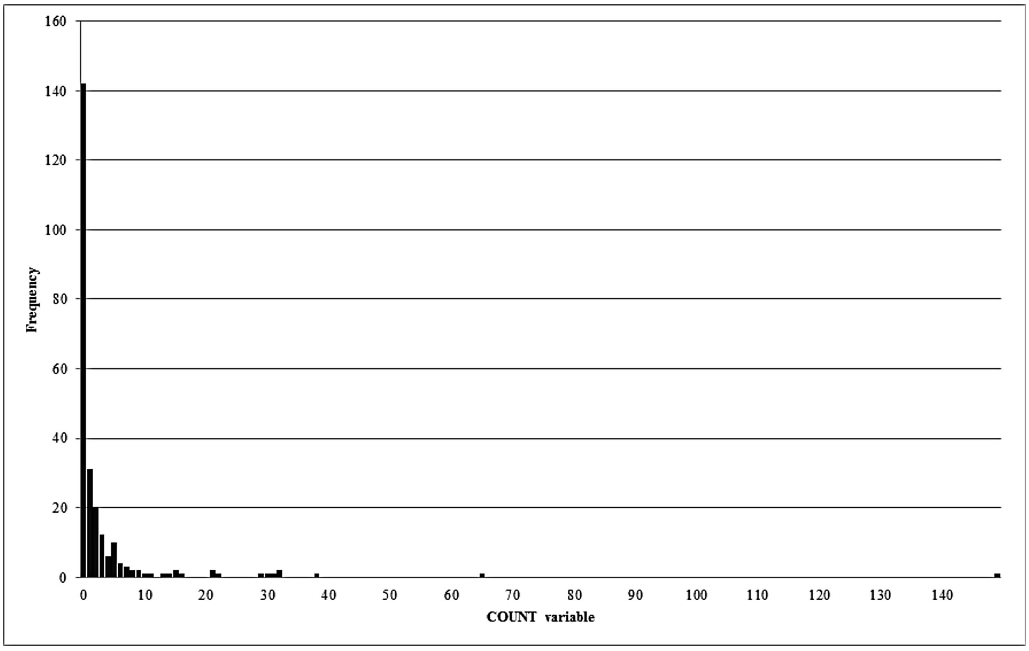

***Figure 1:*** *Histogram of the response variable.*

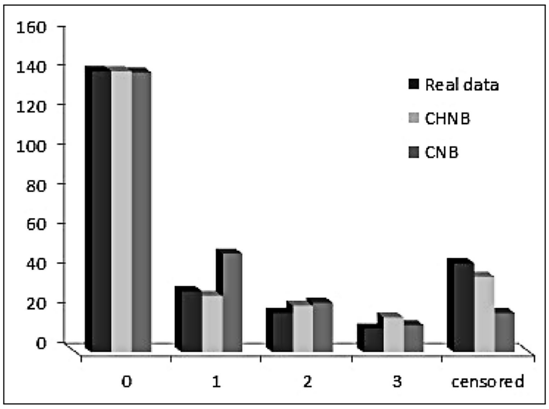

***Figure 2:*** *Estimate vs Real when the censoring point is 3.*

more, the estimated parameter for *child* variable of the logit part of the model is positive for both censoring points, i.e, the group with more children was less successful in fishing. The estimated dispersion parameter suggested overdispersion in the model for both censoring points.

We have compared the censored hurdle negative binomial (CHNB) regression model with the censored negative binomial (CNB) regression model and the results are shown in Figure 2 and Figure 3 when the censoring points are $c_1 = 3$, $c_2 = 5$, respectively. The CHNB regression model shows a better estimation than the CNB regression model except when the value of the count variable is 3 in Figure 2 and 4 in Figure 3. The CHNB

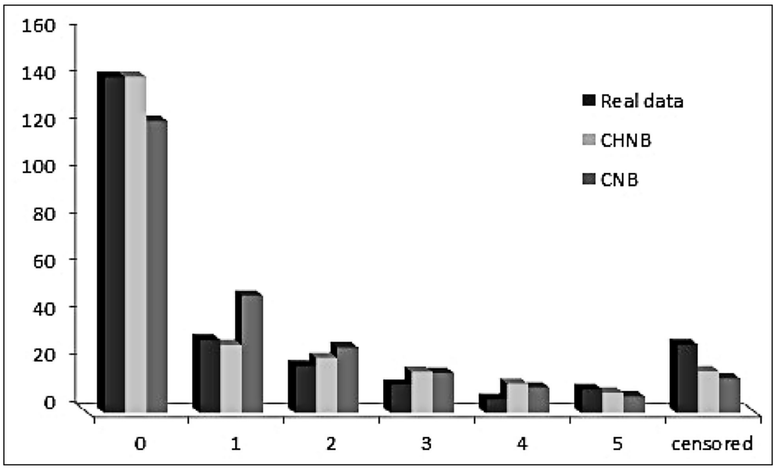

***Figure 3:** Estimate vs Real when the censoring point is 5.*

regression model estimated the number of zeros as 142 (the same as number of zeros in fish data) in both cases ($c_1 = 3, c_2 = 5$), but the CNB regression model estimated 141 and 123, respectively, when the censoring points are $c_1 = 3, c_2 = 5$. Also, the CHNB regression model shows a closer estimate for the censored values than the CNB regression model. For example, when there are 45 censored values in the fish data, the CHNB regression model presented 38 censored values and the CNB regression model presented 20 censored values.

## 6. Conclusion

In this article we want to show that the hurdle negative binomial regression model can be used to fit right censored data. In fact, the proposed model is suitable to solve the excess zeros problem in the response variable when the data are censored from the right side. The results from the fish data are summarized in Table 1-3. The goodness-of-fit measures are presented in the Table 3 according to different censoring points and it is obvious that we have a smaller value for $-2\log L$ or AIC when the percentage of censoring increase and that is because of the number of the data which are used in the model. Also, the censored hurdle negative binomial model shows a better fit with respect to the censored negative binomial model for different censoring points as shown in Figure 2 and Figure 3.

## Acknowledgment

We would like to acknowledge the financial support from Universiti Teknologi Malaysia for the Research University Grant (Q.J1300000.7126.02J67) and Ministry Of Higher Education (MOHE) of Malaysia.

## References

Arulampalam, W. and Booth, A. (1997). Who gets over the training hurdle? a study of the training experiences of young men and women in britain. *Journal of Population Econometrics*, 10, 197–217.

Boucher, J. P. and Guillen, M. (2009). A survey on models for panel count data with applications to insurance. *RACSAM, Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas*, 103(2), 277–294.

Chou, N. T. and Steenhard, D. (2009). A Flexible Count Data Regression Model Using SAS. *PROC NLMIXED. Proceedings SAS Global Forum*, paper 250-2009.

Dalrymple, M., Hudson, I. and Barnett, A. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41, 491–504.

Dionne, G. and Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with regression component. *Astin Bulletin*, 19, 199–212.

Greene, W. (2005). Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*, 1(2), 113–218.

Grootendorst, P. V. (1995). A comparison of alternative models of prescription drug utilization. *Health Economics*, 4, 183–198.

Gurmu, S. and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics*, 14, 469–477.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58, 263–268.

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

Pohlmeier, W. and Ulrich, V. (1995). An econometric model of the two-part decision-making process in the demand for health care. *The Journal of Human Resources*, 30, 339–361.

Saffari, S. E. and Robiah, A. (2011a). Zero-inflated poisson regression models with right censored count data. *Matematika*, 27(1), 21–29.

Saffari, S. E. and Robiah, A. (2011b). Zero-inflated negative binomial regression model with right censoring count data. *Journal of Materials Science and Engineering*, B 1, 551–554.

Saffari, S. E., Robiah, A. and Greene, W. (2011). Handling of over-dispersion of count data via truncation using poisson regression model. *Journal of Computer Science and Computational Mathematics*, 1(1), 1–4.

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

Winkelmann, R. (2003). Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics*, 19, 455–472.

## Appendix:

### *SAS code to estimate the parameters for Fish data*

```
data fish;
set fish;
bound=3;
if count > bound then count=bound+1; * This is probably
 how you would see the data if it was actually censored;

proc nlmixed TECH=NRRIDG;
parms a_0=-0.4 a_1=1 b_0=-1 b_1=1 b_2=1 b_3=-1 alpha=0.5;
bounds alpha>0;
lin = a_0 + a_1* child;
w = exp(lin)/(1+exp(lin));
eta = b_0 + b_1 *camper + b_2* persons + b_3* child;
   mu = exp(eta);
phi=1/alpha;
pdf=(gamma(count+phi)/(gamma(count+1)*gamma(phi)))
*((1/(1+alpha*mu))**phi*(alpha*mu/(1+alpha*mu))** count);
l_1 =  w;
l_2 = (1-w) * pdf / (1-(1+alpha*mu)**(-phi));
cdf=0;
do t=1 to bound;
cdf=cdf+(gamma(t+phi)/(gamma(t+1)*gamma(phi)))
*((1/(1+alpha*mu))**phi*(alpha*mu/(1+alpha*mu))**t);
end;
l_3= (1-w)*(1-cdf/(1-(1+alpha*mu)**(-phi)));
if count = 0 then ll = log(l_1);
if  0 < count <= bound then ll = log(l_2);
if count <= bound then d=0; else d=1;
      ll=(1-d)*ll+d*log(l_3);
model count ~general(ll);
predict mu out=hnbmu;
predict w out=hnbw;
run;
```

SORT 36 (2) July-December 2012, 195-218

# Statistical modelling and forecasting of outstanding liabilities in non-life insurance

María Dolores Martínez-Miranda[1,2], Jens Perch Nielsen[2]
and Mario V. Wüthrich[3]

## Abstract

Non-life insurance companies need to build reserves to meet their claims liability cash flows. They often work with aggregated data. Recently it has been suggested that better statistical properties can be obtained when more aggregated data are available for statistical analysis than just the classical aggregated payments. When also the aggregated number of claims is available one can define a full statistical model of the nature of the number of claims, their delay until payment and the nature of these payments. In this paper we provide a new development in this direction by entering yet another set of aggregated data, namely the number of payments and when they occurred. A new element of our statistical analysis is that we are able to incorporate inflationary trends of payments in a direct and explicit way. Our new method is illustrated on a real life data set.

*MSC:* 62P05 Applications to actuarial sciences and financial mathematics, 91B30 Risk theory, insurance.

*Keywords:* Outstanding loss liabilities, claims settlement process, claims reserving, chain ladder method, individual claims data, prediction uncertainty, bootstrap, early warning systems.

## 1. Introduction

Non-life insurance companies need to forecast future payments arising from claims where the companies already received the insurance premium. The discounted aggregate of these future payments is called the reserve (outstanding liabilities) and is one of the most important components in the accounts of a non-life company. The reserve is most often set by actuaries and the reserving problem is omnipresent in the literature of actuarial science. However, the history of the reserving problem is not a mathematical

[1] University of Granada, Department of Statistics and O.R., Spain.

[2] Cass Business School, City University, London, UK.

[3] ETH Zurich, RiskLab, Department of Mathematics, Zurich, Switzerland.

Received: June 2012
Accepted: August 2012

statistical history even though it clearly is a mathematical statistical forecasting problem. The history is a practical one, where actuaries have had to develop methodologies to set reserves at a time when mathematical statistics was not well developed. The most popular reserving method used by almost all insurance companies is called the chain ladder method by actuaries. Most practical actuaries would talk about chain ladder as a method rather than as a mathematical statistical model even though the actuarial literature has shown a close connection between the chain ladder method and the multiplicative Poisson model. It was only just recently that this multiplicative Poisson model was identified as belonging to the class of exponential families implying well defined solutions to the maximum likelihood estimators and it was also only recently that the explicit expressions of the entering parameters were derived, see Kuang, Nielsen and Nielsen (2009). While practical actuaries work with chain ladder forecasts identical to the forecasts provided by a multiplicative Poisson model, they do not work with the distributional properties of the multiplicative Poisson model. Other distributional properties are preferred, often based on *ad hoc* bootstrap type of procedures. In this paper we build on theory recently derived in three interconnected papers. The main underlying idea of these three papers is that more data (aggregated reported number of claims) should be added to classical actuarial data to allow for a better and more precise formulation of the underlying mathematical statistical model driving the claims development process defining the reserve. The first of these papers (Verrall, Nielsen and Jessen, 2010) defines the simplest possible version of such a model, the second (Martínez-Miranda, Nielsen, Nielsen and Verrall, 2011) develops a bootstrap methodology to assess the distribution of such a model, but the most important of these three papers is perhaps the third one (Martínez-Miranda, Nielsen and Verrall, 2012). This paper shows that a slight modification of Verrall *et al.* (2010)'s model, with one particular moment type estimation method, provides us with a well-defined mathematical statistical model exactly replicating the reserving estimates one would obtain using the classical chain ladder method. This model has trustable distributional properties that can be used in practice by actuaries. In this paper we take the ideas of the above three papers one step further and add yet another piece of data (aggregated number of payments) to our data set and we show that important insights of the claim development process result when incorporating this extra piece of information in our mathematical statistical model. We follow in this paper Martínez-Miranda *et al.* (2012) and work with moment type of estimators. Our hope is that this paper provides information to the mathematical statistician wishing to use their excellent tools on this important real life problem and can perhaps be helpful in bringing mathematical statisticians into this important field. The notation and vocabulary of this paper are deliberately closely related to classical actuarial terminology while describing a well-defined mathematical statistical model. This is a deliberate attempt to bridge the gap between classical actuarial terminology, often obscure to mathematical statisticians, and standard mathematical statistical model formulations that might seem unrelated to classical reserving for many actuaries.

The general post credit crunch atmosphere in the financial sector emphasizes a better understanding of outstanding loss liabilities of non-life insurance companies, with reserving models as one of the essential technical building blocks. However, the insurance industry is also gaining new territory in new markets, where better early warning reserving systems are required than that provided by the old chain ladder methodology. In this paper we introduce a new reserving methodology with an automatic early warning system to detect important irregularities in the claims development process. Our methodology requires more detailed data than classical reserving methods. The point of view taken is that the aggregated payments do not provide us with sufficient mathematical statistical information, we argue that also the number of payments and the number of reported claims are needed. This enables us to embed a variety of new claims inflation type of information in our overall model. We consider severity inflation, underwriting year inflation and claims delay inflation and show how to incorporate those in the reserving process. The calendar inflation is not treated in detail in this paper, but it could have been extracted up front using the Kuang, Nielsen and Nielsen (2011) methodology of calendar inflation (see also Kuang, Nielsen and Nielsen 2008a,b) .

In the next section we define the model on the micro-level. The basis of our model is the compound Poisson processes studied in Norberg (1993, 1999) and Jessen, Mikosch and Samorodnitsky (2011). We show how we need to structure these compound Poisson processes on the micro-level so that we obtain a chain ladder claims reserving method on the aggregate level. Such connection is proved from first moments calculations which are provided in Appendix A. In Section 3 we provide estimates of the parameters of the model. From the estimated model, point forecasts for the reserve are given in Section 4. Using bootstrap methods we provide in Section 6 (together with Appendix B) an approximation of the full predictive distribution of the outstanding loss liabilities. The methods proposed in this paper are illustrated using a dataset from the insurance industry, given in Appendix C. The focus of this application is to provide an estimate of the claims reserves and to detect irregularities in the data.

## 2. Model setup

### 2.1. Data and micro-level structure

In classical reserving methods the data upon which projections of future claims are usually represented by so called run-off triangles. This format tabulates the claim data (payments, numbers of reported or paid claims, etc.) according to the period in which the claim arose (called underwriting or accident period) and the period in which the payment (or other action) was made. The difference between the payment period and the accident period is referred to as the development period. The data are usually aggregated in years or quarters of years, but other time periods can also be used depending on the business line. Hereafter we write years as the periods considered in the aggregation. We denote accident years by $i = 1, \ldots, m$, and development years by $j = 0, \ldots, m-1$, where $m \in \mathbb{N}$ denotes the last observed accident year. Then the available data lie in the triangle $\mathfrak{I}_m = \{(i, j);\ i = 1, \ldots, m; j = 0, \ldots, m-1; i + j \leq m\}$. In Appendix C we show an example of this type of data.

The methods proposed in this paper consider such run-off triangles as input data. In fact we will need more triangles to provide a more precise formulation of the mathematical statistical micromodel that underlies the claims development process defining the reserve. It is a parametric model that is deliberately formulated in such a way that the entering parameters are identifiable and estimable from three aggregated data sets: number of reported claims, number of payments and aggregated paid amounts. These stochastic variables are described in the following:

- Number of reported claims of accident year $i$ with a reporting delay of $j$ years, denoted by $N_{i,j}$.
- Number of payments. Each of these $N_{i,j}$ reported claims generates a claims payment cash flow. We denote by $R_{i,j,l}$ the number of payments generated by these $N_{i,j}$ reported claims that have a payment delay of $l \geq 0$ years. That is, $R_{i,j,l}$ is the number of payments in accounting year $i + j + l$ for claims that have occurred in accounting year $i$ and were reported in accounting year $i + j$.
- Individual claims payments. Each of these $R_{i,j,l}$ claims payments has size $Y_{i,j,l}^{(k)}$, for $k = 1, \ldots, R_{i,j,l}$.

Often, claims payment data is not available on the micro-level structure described by $\{N_{i,j};\ (i, j) \in \mathfrak{I}_m\} \cup \{R_{i,j,l}, Y_{i,j,l}^{(k)};\ (i, j) \in \mathfrak{I}_m,\ i + j + l \leq m,\ k \geq 1\}$. Therefore, we define the following aggregate claims payment information. The total number of payments in accounting year $i + j$ from claims with accident year $i$ is given by

$$R_{i,j} = \sum_{l=0}^{j} R_{i,j-l,l}. \tag{1}$$

These $R_{i,j}$ are the number of payments in accounting year $i + j$ generated by all claims with accident year $i$ which where reported prior to (and including) accounting year $i + j$, i.e. these are payments from the $N_{i,j-l}$ reported claims, with $l = 0, \ldots, j$. The payments (total quantity paid) in accounting year $i + j$ from claims with accident year $i$ are then given by

$$X_{i,j} = \sum_{l=0}^{j} \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)}. \tag{2}$$

From these definitions we assume that the available information at time $m$ consists of the following three $\sigma$-fields (upper claims development triangles):

$$\mathcal{N}_m = \sigma \{N_{i,j};\ (i,j) \in \mathcal{I}_m\},$$
$$\mathcal{R}_m = \sigma \{R_{i,j};\ (i,j) \in \mathcal{I}_m\},$$
$$\mathcal{X}_m = \sigma \{X_{i,j};\ (i,j) \in \mathcal{I}_m\},$$

and the aim is to predict the total payments in the future:

$$\mathcal{X}_m^c = \{X_{i,j};\ (i,j) \in \mathcal{J}_m\},$$

where $\mathcal{J}_m = \{(i,j); i = 2,\dots,m,\ j = 0,\dots,m-1,\ i+j > m\}$ is the lower (inexperienced) triangle.

Classical reserving methods as the chain ladder method provide predictions for $\mathcal{X}_m^c$. However, a better description of the reserving problem would be provided if we are able to separate these future payments in the lower triangle into payments for claims that have been already reported (prior to and including accounting year *m*) and claims that will be reported after accounting year *m*. The first class of claims is contained in the number of reported claims $\mathcal{N}_m$, and constitutes what is called the reported but not settled (RBNS) claims reserves. The latter class contains the so-called incurred but not reported (IBNR) claims and constitutes the IBNR claims reserves. Such a distinction is often important, for example, in the calculation of unallocated loss adjustment expenses (ULAE), see Wüthrich, Bülmann and Furrer (2010, Section 5.6). If we apply the classical chain ladder method then we predict $\mathcal{X}_m^c$ based solely on the information $\mathcal{X}_m$, thus, we predict the outstanding loss liabilities on a rather aggregate level, which does not allow a distinction between RBNS and IBNR claims reserves.

### 2.2. Model assumptions

With the above definitions we assume the following hypotheses about the micro-level structure.

(A1) All random variables in different accident years $i \in \{1,\dots,m\}$ are independent.

(A2) The numbers of reported claims $N_{i,0},\dots,N_{i,m-1}$ are independent and Poisson distributed with cross-classified means $\mathbb{E}[N_{i,j}] = \vartheta_i \beta_j$, for given parameters $\vartheta_i > 0$, $\beta_j > 0$ with normalization $\vartheta_1 = 1$.

(A3) The claims payments

$$X_{i,j,l} = \sum_{k=1}^{R_{i,j,l}} Y_{i,j,l}^{(k)}$$

are, conditionally given $N_{i,0},\dots,N_{i,m-1}$, independent (in $l \geq 0$) and compound Poisson distributed with

- $R_{i,j,l}|_{\{N_{i,0},\ldots,N_{i,m-1}\}} \sim \mathrm{Poi}(N_{i,j}\pi_l)$ with given parameter $\pi_l > 0$;
- $Y_{i,j,l}^{(k)}|_{\{N_{i,0},\ldots,N_{i,m-1}\}} \overset{(d)}{=} Y_{i,j,l}^{(k)}$ are i.i.d. for $k \geq 1$ with the first two moments given by

$$\mathbb{E}\left[Y_{i,j,l}^{(1)}\right] = \nu_i\, \mu_{j,l} \qquad \text{and} \qquad \mathbb{E}\left[\left(Y_{i,j,l}^{(1)}\right)^2\right] = \nu_i^2\, s_{j,l}^2,$$

for parameters $\nu_i, \mu_{j,l}, s_{j,l} \in \mathbb{R}_+$ with normalization $\nu_1 = 1$.

One crucial point in assumption (A3) is that the claim size (or severity) distribution of $Y_{i,j,l}^{(k)}$ can be split into an accident year dependent part $\nu_i$ which models claims inflation in the accident year direction, and a development year dependent part $\mu_{j,l}$ which takes care of reporting delay $j \geq 0$ and payment delay $l \geq 0$. Note that assumption (A3) implies that the payments $Y_{i,j,l}^{(k)}$ are independent from the number of reported claims $N_{i,j}$ as well as from the number of payments $R_{i,j,l}$ (conditional compound Poisson model assumption).

The choices $\vartheta_1 = \nu_1 = 1$ will make the parameters identifiable in the estimation procedure. One can also use other normalizations, such as e.g. $\sum_j \beta_j = 1$ (normalized claims reporting pattern). However, our choice is rather simple to implement and other normalizations are obtained by rescaling.

## 3. Parameter estimation

The estimation of the model parameters, $\{\vartheta_i, \beta_j, \pi_l, \nu_i, \mu_{j,l}; i = 1, \ldots, m, j, l = 0, \ldots, m-1\}$, can be solved just using the simple chain ladder method on the three input triangles. The only requirement is to demonstrate that the random variables $N_{i,j}$, $R_{i,j}$ and $X_{i,j}$ all have the same cross-classified mean structure, which is the chain ladder mean structure. As was discussed in Martínez-Miranda *et al.* (2012) this can be done from model specifications about just the first moment of the underlying stochastic components. Further purposes about deriving the distribution of the future payments requires conditions on higher order moments and also a more detailed specification including distributional assumptions (see Martínez-Miranda *et al.* 2012 for further explanation). Under the distributional model proposed here, we suggest in Section 6 an estimator for the second moment parameters $s_{j,l}$ $(j, l = 0, \ldots, m-1)$ to derive then the predictive distribution.

Therefore we next provide estimates of the parameters based in the first moment of the random variables, $N_{i,j}$, $R_{i,j}$ and $X_{i,j}$. We have deferred such calculations to Appendix A in order to facilitate the reading of the paper. Specifically in Propositions 2 and 3 we have obtained that the first moments of the three sets of random variables $N_{i,j}$, $R_{i,j}$ and $X_{i,j}$ all have the same cross-classified mean structure. Also we have established connections among the parameters in the model through the following equations:

$$\alpha_i = \vartheta_i \; \nu_i, \tag{3}$$

$$\lambda_j = \sum_{l=0}^{j} \beta_{j-l} \; \pi_l, \tag{4}$$

$$\gamma_j = \sum_{l=0}^{j} \beta_{j-l} \; \pi_l \; \mu_{j-l,l}, \tag{5}$$

From these initial steps our aim is to estimate the corresponding parameters based on the information in $\mathcal{N}_m$, $\mathcal{R}_m$ and $\mathcal{X}_m$, and by applying the simple chain ladder method to each triangle. As an example, we demonstrate the estimation for the observed number of reported claims $\mathcal{N}_m$ and the parameters $\vartheta_i$ and $\beta_j$. The remaining parameters are estimated in the same way, but based on $\mathcal{R}_m$ and $\mathcal{X}_m$, respectively. In a distribution-free approach we rely on moment estimators. If we aggregate rows and columns, respectively, over the set of information $\mathcal{I}_m$ we obtain the first moment equalities

$$\sum_{k=0}^{m-i} \mathbb{E}[N_{i,k}] \;=\; \vartheta_i \sum_{k=0}^{m-i} \beta_k \qquad \text{for } i = 1,\dots,m, \tag{6}$$

$$\sum_{k=1}^{m-j} \mathbb{E}[N_{k,j}] \;=\; \beta_j \sum_{k=1}^{m-j} \vartheta_k \qquad \text{for } j = 0,\dots,m-1. \tag{7}$$

Unbiased estimators for the right-hand side of these equalities are obtained by replacing the moments $\mathbb{E}[N_{i,j}]$, $(i,j) \in \mathcal{I}_m$, by their observations $N_{i,j} \in \mathcal{N}_m$. Then the resulting system of linear equations is solved for $\vartheta_i$ and $\beta_j$, which provides the corresponding estimators for these parameters. This is in the spirit of the "total marginals" method of Bailey (1963) and Jung (1968). Kremer (1985) and Mack (1991) have shown that in the case of triangular data $\mathcal{N}_m$ this leads to the chain ladder estimators that can be calculated in closed form. Thus,

$\mathcal{N}_m$ provides the chain ladder estimators $\widehat{\vartheta}_i^{(1)}$ and $\widehat{\beta}_j$ for $\vartheta_i$ and $\beta_j$,
$\mathcal{R}_m$ provides the chain ladder estimators $\widehat{\vartheta}_i^{(2)}$ and $\widehat{\lambda}_j$ for $\vartheta_i$ and $\lambda_j$,
$\mathcal{X}_m$ provides the chain ladder estimators $\widehat{\alpha}_i$ and $\widehat{\gamma}_j$ for $\alpha_i$ and $\gamma_j$,

with $\widehat{\vartheta}_1^{(1)} = \widehat{\vartheta}_1^{(2)} = \widehat{\alpha}_1 = 1$ (initialization in cross-classified means). Note that we obtain two different estimators $\widehat{\vartheta}_i^{(1)}$ and $\widehat{\vartheta}_i^{(2)}$ for the same parameter $\vartheta_i$. However, their values should not be too different, otherwise this indicates that the model may not fit to the claims reserving problem. In order to estimate $\vartheta_i$ we could now take a credibility weighted average between $\widehat{\vartheta}_i^{(1)}$ and $\widehat{\vartheta}_i^{(2)}$. For simplicity we set $\widehat{\vartheta}_i$ as the arithmetic mean between $\widehat{\vartheta}_i^{(1)}$ and $\widehat{\vartheta}_i^{(2)}$. Anyway, the appropriateness of this choice should always be checked on the data. Using equality (3) we can estimate the accident year inflation parameter $\nu_i$ by

$$\widehat{\nu}_i = \widehat{\alpha}_i / \widehat{\vartheta}_i \qquad \text{for } i = 1, \dots, m. \tag{8}$$

Thus, it remains to estimate the parameters $\pi_l$ and $\mu_{j,l}$ $(j,l = 0,\dots,m-1)$. There are different ways to estimate these parameters. We start with $\pi_l$ using the equality (4). If we rewrite this equation in vector notation we have

$$(\lambda_0, \dots, \lambda_{m-1})^{\mathsf{T}} \;=\; B_\beta \, (\pi_0, \dots, \pi_{m-1})^{\mathsf{T}},$$

for an appropriate matrix $\mathbf{B}_\beta = \mathbf{B}_{\beta_0,\dots,\beta_{m-1}} \in \mathbb{R}^{m\times m}$. This matrix is estimated by $\widehat{\mathbf{B}}_\beta = \mathbf{B}_{\widehat{\beta}_0,\dots,\widehat{\beta}_{m-1}} \in \mathbb{R}^{m\times m}$ and then we can provide estimates, $\widehat{\pi}_0, \dots, \widehat{\pi}_{m-1}$, by solving the following system:

$$(\widehat{\pi}_0, \dots, \widehat{\pi}_{m-1})^{\mathsf{T}} \;=\; \widehat{\mathbf{B}}_\beta^{-1} \left(\widehat{\lambda}_0, \dots, \widehat{\lambda}_{m-1}\right)^{\mathsf{T}}. \tag{9}$$

The estimation of $\mu_{j,l}$ needs more care because the model is over-parametrized. In order to reduce the number of parameters we make one of the following two assumptions

$$\mu_{j,l} \equiv \mu_l \tag{10}$$

or

$$\mu_{j,l} \equiv \mu_j. \tag{11}$$

Using the condition (10) and the equality (5) we have that

$$(\gamma_0, \dots, \gamma_{m-1})^{\mathsf{T}} \;=\; \mathbf{B}_\beta \, (\pi_0\mu_0, \dots, \pi_{m-1}\mu_{m-1})^{\mathsf{T}},$$

for matrix $\mathbf{B}_\beta = \mathbf{B}_{\beta_0,\dots,\beta_{m-1}} \in \mathbb{R}^{m\times m}$. If this matrix is again estimated by $\widehat{\mathbf{B}}_\beta = \mathbf{B}_{\widehat{\beta}_0,\dots,\widehat{\beta}_{m-1}}$ we obtain estimates $\widehat{\pi\mu}_0, \dots, \widehat{\pi\mu}_{m-1}$ as the solution of the following system:

$$\left(\widehat{\pi\mu}_0, \dots, \widehat{\pi\mu}_{m-1}\right)^{\mathsf{T}} \;=\; \widehat{\mathbf{B}}_\beta^{-1} \, (\widehat{\gamma}_0, \dots, \widehat{\gamma}_{m-1})^{\mathsf{T}}, \tag{12}$$

and, finally, the estimator for $\mu_{j,l}$ assumption (10) is given by $\widehat{\mu}_{j,l} = \widehat{\mu}_l = \widehat{\pi\mu}_l / \widehat{\pi}_l$.

On the other hand, using assumption (11) and rewriting (5) we have the following system

$$(\gamma_0, \dots, \gamma_{m-1})^{\mathsf{T}} \;=\; \mathbf{B}_\pi \, (\beta_0\mu_0, \dots, \beta_{m-1}\mu_{m-1})^{\mathsf{T}},$$

for matrix $\mathbf{B}_\pi = \mathbf{B}_{\pi_0,\dots,\pi_{m-1}} \in \mathbb{R}^{m\times m}$. And again plugging in the estimated matrix $\widehat{\mathbf{B}}_\pi = \mathbf{B}_{\widehat{\pi}_0,\dots,\widehat{\pi}_{m-1}} \in \mathbb{R}^{m\times m}$, we obtain the estimates, $\widehat{\beta\mu}_0, \dots, \widehat{\beta\mu}_{m-1}$, by solving the system

$$\left(\widehat{\beta\mu}_0, \dots, \widehat{\beta\mu}_{m-1}\right)^{\mathsf{T}} \;=\; \widehat{\mathbf{B}}_\pi^{-1} \, (\widehat{\gamma}_0, \dots, \widehat{\gamma}_{m-1})^{\mathsf{T}}. \tag{13}$$

This yields the estimator $\widehat{\mu}_{j,l} = \widehat{\mu}_j = \widehat{\beta\mu}_j / \widehat{\beta}_j$.

The above procedure provides estimates for all the parameters required for point prediction purposes, under the additional assumption (10) or (11). In the next section we are going to describe how they are used to predict the outstanding loss liabilities $\mathcal{X}_m^c$ at time $m$. Moreover, we will also discuss further adjustments to these estimators in practise.

## 4. Point forecasts

Point predictions for the outstanding loss liabilities can be derived as estimated unconditional (or conditional) means of the aggregated payments, $X_{i,j}$, in the lower triangle, $\mathfrak{J}_m$. In the previous section we have estimated all the parameters in the model from the observations $\mathcal{N}_m$, $\mathcal{R}_m$ and $\mathcal{X}_m$. It only remains to estimate the second moment parameters $s_{j,l}$ $(j,l = 0,\dots,m-1)$ of the size of the individual payments. But, as we pointed in the previous section, such higher order moments are not involved in the point forecasts. Therefore, we have all that is necessary to predict the outstanding liabilities, $\mathcal{X}_m^c$. At time $m$ the conditionally expected outstanding loss liability cash flows in $\mathcal{X}_m^c$ are given by

$$Z_m = \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \mathbb{E}\left[X_{i,j} | \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m\right].$$

If we only rely on the observations $\mathcal{X}_m$, then we can only estimate the parameters $\alpha_i$ and $\gamma_j$. Thus, in this case we set

$$\widehat{Z}_m^{CL} = \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \widehat{\alpha}_i\, \widehat{\gamma}_j,$$

which provides an estimator for $Z_m$. The crucial property of this estimator $\widehat{Z}_m^{CL}$ is that it provides the chain ladder reserves exactly (see Kremer 1985, Mack 1991 and Section 2.4 in Wüthrich and Merz 2008). Having additional information $\mathcal{N}_m$ and $\mathcal{R}_m$ we can refine this estimate. We have

$$\begin{aligned} Z_m &= \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \sum_{l=0}^{j} \mathbb{E}\left[ \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)} \middle| \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m \right] \\ &= \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \sum_{l=i+j-m}^{j} \mathbb{E}\left[ \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)} \middle| \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m \right] \\ &\quad + \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \sum_{l=0}^{i+j-m-1} \mathbb{E}\left[ \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)} \middle| \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m \right]. \end{aligned}$$

Note that the decoupling separates RBNS and IBNR claims: if $i+j-l \le m$ then the payment $Y_{i,j-l,l}^{(k)}$ belongs to a claim that has been reported prior to (and including) accounting year $m$, and henceforth is an RBNS claim at time $m$. Therefore, we define

$$Z_m^{\text{RBNS}} = \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \sum_{l=i+j-m}^{j} \mathbb{E}\left[ \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)} \middle| \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m \right],$$

$$Z_m^{\text{IBNR}} = \sum_{i=2}^{m} \sum_{j=m-i+1}^{m-1} \sum_{l=0}^{i+j-m-1} \mathbb{E}\left[ \sum_{k=1}^{R_{i,j-l,l}} Y_{i,j-l,l}^{(k)} \middle| \mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m \right].$$

Using assumptions (A1)–(A3) we obtain the following result.

**Proposition 1**

$$Z_m^{\text{RBNS}} = \sum_{i=2}^{m} \nu_i \sum_{j=m-i+1}^{m-1} \sum_{l=i+j-m}^{j} N_{i,j-l}\ \pi_l\ \mu_{j-l,l}, \tag{14}$$

$$Z_m^{\text{IBNR}} = \sum_{i=2}^{m} \vartheta_i\ \nu_i \sum_{j=m-i+1}^{m-1} \sum_{l=0}^{i+j-m-1} \beta_{j-l}\ \pi_l\ \nu_i\ \mu_{j-l,l}. \tag{15}$$

Using the previous expressions we can estimate the RBNS claims reserve by plugging estimates of the parameters in (14) and similarly the IBNR reserve using (15). Denote the resulting predictions by $\widehat{Z}_m^{\text{RBNS}}$ and $\widehat{Z}_m^{\text{IBNR}}$, respectively. Then the total reserve can be estimated by $\widehat{Z}_m = \widehat{Z}_m^{\text{RBNS}} + \widehat{Z}_m^{\text{IBNR}}$. A straightforward calculation demonstrates that the model defined in (A1)–(A3) can provide the same reserve as the classical chain ladder just by making a particular choice. This result is stated in the following corollary.

**Corollary 1** *Under the additional assumptions that* $\widehat{\vartheta}_i^{(1)} = \widehat{\vartheta}_i^{(2)}$*, for all* $i = 2, \ldots, m$*, and* $N_{i,j} = \widehat{\vartheta}_i \widehat{\beta}_j$*, for all* $(i,j) \in \mathcal{I}_m$*, we have* $\widehat{Z}_m = \widehat{Z}_m^{CL}$*.*

Often claims development goes beyond the latest development period $m-1$, which has been observed at time $m$. Therefore, in practice, one needs to add a tail estimate to the claims reserves in order to also cover these additionally expected outstanding loss liability cash flows. The entire tail can be estimated under assumptions (A1)–(A3) if we additionally assume that $\beta_j = \pi_j = 0$ for $j = 1, \ldots, m-1$. In this particular case, we know that all claims are reported after development period $j = m-1$. Thus, we define the claims reserves including the tail by (re-arranging the summations)

$$\widehat{Z}_m^{\text{RBNS+}} = \sum_{i=1}^{m} \widehat{\nu}_i \sum_{j=0}^{m-i} N_{i,j} \sum_{l=m-(i+j)+1}^{m-1} \widehat{\pi}_l\ \widehat{\mu}_{j,l},$$

$$\widehat{Z}_m^{\text{IBNR+}} = \sum_{i=2}^{m} \widehat{\vartheta}_i\ \widehat{\nu}_i \sum_{j=m-i+1}^{m-1} \widehat{\beta}_j \sum_{l=0}^{m-1} \widehat{\pi}_l\ \widehat{\mu}_{j,l},$$

and the total reserves including the tail are defined by $\widehat{Z}_m^{+} = \widehat{Z}_m^{\text{RBNS+}} + \widehat{Z}_m^{\text{IBNR+}}$.

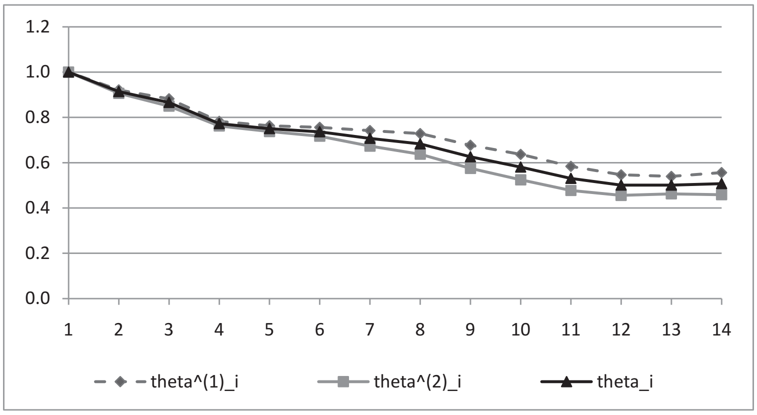

***Figure 1:*** *Real data example: estimates for $\vartheta_i$. Estimates $\widehat{\vartheta}_i^{(1)}$ are based on $\mathcal{N}_m$, estimates $\widehat{\vartheta}_i^{(2)}$ are based on $\mathcal{R}_m$ and $\widehat{\vartheta}_i$ is the arithmetic mean between the latter two estimates.*

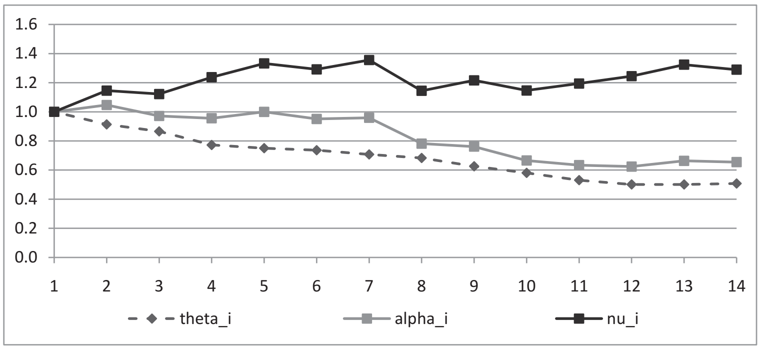

***Figure 2:*** *Real data example: estimates for $\alpha_i$, $\vartheta_i$ and $\nu_i$.*

## 5. An example with real data

We illustrate the methods proposed in this paper using a real data example provided in Tables 6, 7 and 8 in Appendix C. The first step is to estimate the parameters according to Section 3.

In Figure 1 we give the estimates for $\vartheta_i$ for $i = 1,\ldots,m = 14$. We see that both data sets $\mathcal{N}_m$ and $\mathcal{R}_m$ provide similar estimates $\widehat{\vartheta}_i^{(1)}$ and $\widehat{\vartheta}_i^{(2)}$ for $\vartheta_i$ which confirms the model assumptions (A1)–(A3). Moreover, we see a strong decrease in the volume in this portfolio, since the exposure parameters $\widehat{\vartheta}_i$ decrease from 1 to roughly 0.5.

We could now proceed as described above and use the estimates $\widehat{\beta}_j$ and $\widehat{\lambda}_j$. However, we slightly deviate from this approach. Namely, if we plug in the resulting (adjusted) exposure estimates $\widehat{\vartheta}_i$ from (8) into (6) and (7) we get adjusted estimates $\widetilde{\beta}_j$ for $\beta_j$ and similarly $\widetilde{\lambda}_j$ for $\lambda_j$. We prefer to work with these adjusted estimates because they assure that the overall level is correct if calculate the cross-classified means of $N_{i,j}$ and $R_{i,j}$, see Proposition 2.

In Figure 2 we show the estimates for the exposures $\alpha_i$ and $\vartheta_i$, and the resulting inflation estimate $\widehat{\nu}_i$ is provided by the ratio of the latter two estimates. In general, we see an increase in the time-series $\widehat{\nu}_1, \ldots, \widehat{\nu}_{14}$, however accident year $i = 8$ seems conspicuous and needs further analysis on single claims data. It may indicate that there is a change in the underlying product (if it only acts on horizontal axis in the claims development triangle). Indeed we observe a substantial decrease in average payments per reported claim in accident year $i = 8$ which supports the argument of changes in the product (or portfolio).

Finally, we estimate $\pi_l$ and $\mu_{j,l}$ from $\widetilde{\beta}_j$, $\widetilde{\lambda}_j$ and $\widehat{\gamma}_j$. We solve the estimation problem under assumption (10), i.e. $\mu_{j,l} \equiv \mu_l$. In that case we set $\widetilde{\mathbf{B}}_\beta = \mathbf{B}_{\widetilde{\beta}_0,\ldots,\widetilde{\beta}_{m-1}}$ and then we estimate $\pi_l$ and $\pi_l \mu_l$ from equations (9) and (12). Figure 3 provides the estimates $\widehat{\pi}_l$. First of all we observe that all $\widehat{\pi}_l > 0$ except $\widehat{\pi}_2 < 0$ and $\widehat{\pi}_{12} < 0$ which contradicts the model assumptions (A1)–(A3). Thus, at this point we might ask for a more sophisticated model. However, this would also ask for more micro-level observations. We refrain from doing so but correct this value. In our particular case, we choose correction

$$\widetilde{\pi}_l = \begin{cases} \widehat{\pi}_l - 2|\widehat{\pi}_{l+1}| & l = 1, \\ |\widehat{\pi}_l| & l = 2, \\ 0 & l = 12, 13, \\ \widehat{\pi}_l & \text{otherwise.} \end{cases}$$

The resulting adjusted estimates $\widetilde{\pi}_l$ are also given in Figure 3. Note that we have $\sum_l \widetilde{\pi}_l \approx \sum_l \widehat{\pi}_l = 0.7251$, which says that on average we expect 0.7251 payments per reported claim, and in the average almost half of the claims can be settled without a payment. An analysis of payments per reported claim shows that this figure is decreasing

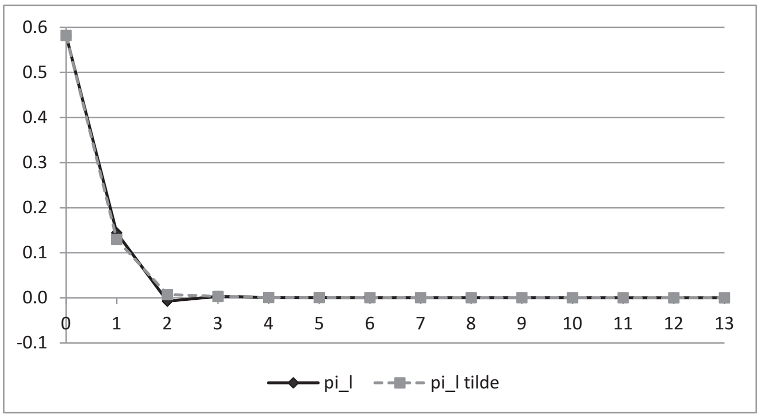

***Figure 3:*** *Real data example: estimates $\widehat{\pi}_l$ and $\widetilde{\pi}_l$ for $\pi_l$.*

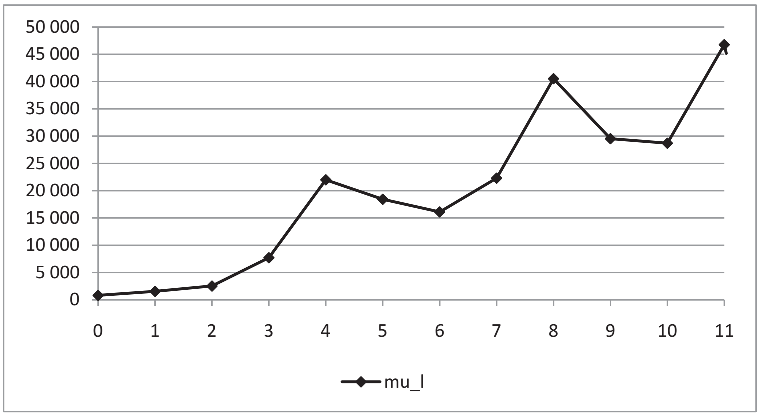

***Figure 4:*** *Real data example: estimates $\widehat{\mu}_{j,l}$ for $\mu_{j,l} \equiv \mu_l$ for $l = 0, \ldots, 11$.*

over time. This decrease can have various reasons such as changes in reporting philosophy, changes in the claims handling process, but it could also be related to changes in the portfolio (we have already mentioned that the volume is strongly decreasing).

We then estimate $\mu_l \pi_l$ from $\widehat{\mu\pi}_l$, which is the solution of the system (12). And, under (10), we estimate $\mu_{j,l} = \mu_l$ by $\widehat{\mu\pi}_l / \widetilde{\pi}_l$.

The results are presented in Figure 4. We see that the average payments $\mu_{j,l}$ are increasing in the payment delay $l$. We could now further smooth this curve for the expected payments $\mu_{j,l}$, but we refrain from doing so. There are also other issues, for example that the payments $Y_{i,j,l}^{(k)}$ may not only depend on the accident year $i$ and the payment delay $l$ but also on the reporting delay $j$. However, as described in Section A we cannot model all directions simultaneously because this would lead to an over-parametrization.

Finally, in Table 1 we present the resulting claims reserves. We observe that under assumptions (A1)–(A3) and (10) we obtain higher claims reserves than classical chain ladder (see the last two columns in Table 1). One reason for this more conservative estimate is that we judge the upper right corner of the triangle $\mathcal{X}_m$ differently. The estimate for later development periods, say $j = 11, 12, 13$, is based on a rather small set observations in the chain ladder method (and hence not very reliable). In our model we use the additional model structure for the estimation of payments in later development periods which, in this case, is more conservative. The influence of the tail estimate is only minor, specifically $\widehat{Z}_m^+ - \widehat{Z}_m = 7\,074$. This has to do with the fact that we have a rather short payout pattern $\widetilde{\pi}_l$ (see Figure 3).

Another possible approach in the previous calculations is to use condition (11), i.e. $\mu_{j,l} \equiv \mu_j$. However, the resulting claims reserves derived in this case seemed to be too low and we have decided not to include this in the paper. The reason is because the main driver of late payments is the payment delay $\pi_l$ and not the reporting delay $\beta_j$. This implies that under (11) we underestimate the amounts of late payments because they are attached too strongly to the reporting pattern $\beta_j$ compared to the payment pattern $\pi_l$.

***Table 1:*** *Real data example: resulting claims reserves under (10).*

| a.y. $i$ | $\widehat{Z}_m^{\text{RBNS+}}$ | $\widehat{Z}_m^{\text{IBNR+}}$ | $\widehat{Z}_m^{+}$ | $\widehat{Z}_m^{CL}$ | difference | in % |
|---|---|---|---|---|---|---|
| 1 | 536 | 0 | 536 |  | 536 |  |
| 2 | 1 540 | 0 | 1 540 | 0 | 1 540 |  |
| 3 | 23 799 | 0 | 23 799 | 2 220 | 21 579 | 971.8% |
| 4 | 162 275 | 0 | 162 275 | 147 434 | 14 841 | 10.1% |
| 5 | 291 122 | 790 | 291 912 | 280 056 | 11 855 | 4.2% |
| 6 | 415 955 | 1 590 | 417 545 | 408 154 | 9 391 | 2.3% |
| 7 | 584 991 | 3 300 | 588 291 | 569 060 | 19 231 | 3.4% |
| 8 | 605 767 | 3 676 | 609 443 | 583 785 | 25 658 | 4.4% |
| 9 | 704 687 | 5 039 | 709 726 | 675 363 | 34 363 | 5.1% |
| 10 | 803 884 | 6 343 | 810 228 | 764 373 | 45 855 | 6.0% |
| 11 | 1 054 124 | 10 037 | 1 064 161 | 1 004 331 | 59 829 | 6.0% |
| 12 | 1 397 607 | 22 068 | 1 419 675 | 1 352 819 | 66 856 | 4.9% |
| 13 | 1 999 243 | 84 680 | 2 083 922 | 2 076 674 | 7 248 | 0.3% |
| 14 | 4 221 084 | 1 474 793 | 5 695 877 | 5 487 650 | 208 227 | 3.8% |
| total | 12 266 615 | 1 612 315 | 13 878 930 | 13 351 921 | 527 009 | 3.9% |

## 6. Bootstrap predictive distribution

### 6.1. Conditional mean square error of prediction

In addition to the claims reserves estimates $\widehat{Z}_m^{+}$ we also need to assess the corresponding prediction uncertainty. We briefly describe this with the help of the conditional mean square error of prediction (MSEP) uncertainty measure which is defined by

$$\operatorname{msep}_{X_m|\{\mathcal{N}_m,\mathcal{R}_m,\mathcal{X}_m\}}\left(\widehat{Z}_m^{+}\right)=\mathbb{E}\left[\left(X_m-\widehat{Z}_m^{+}\right)^2\middle|\mathcal{N}_m,\mathcal{R}_m,\mathcal{X}_m\right], \tag{16}$$

where the aggregate cash flow in the lower triangle is defined by $X_m=\sum_{i=2}^{m}\sum_{j=m-i+1}^{m-1}X_{i,j}$. Thus, the conditional MSEP describes the possible fluctuations of the true outstanding loss liability cash flows $X_m$ around the predictor $\widehat{Z}_m^{+}$. Since the predictor is $\sigma\{\mathcal{N}_m,\mathcal{R}_m,\mathcal{X}_m\}$-measurable the conditional MSEP can be decoupled into process variance and parameter estimation error, see (3.1) in Wüthrich and Merz (2008),

$$\operatorname{msep}_{X_m|\{\mathcal{N}_m,\mathcal{R}_m,\mathcal{X}_m\}}\left(\widehat{Z}_m^{+}\right)=\operatorname{Var}\left(X_m|\mathcal{N}_m,\mathcal{R}_m,\mathcal{X}_m\right)+\left(Z_m-\widehat{Z}_m^{+}\right)^2. \tag{17}$$

The first term (process variance) can be calculated explicitly under our model assumptions, the second term (parameter estimation error) is more difficult to assess. Often, one derives approximations for this latter term. However, in our case this is too involved, therefore we rely on the bootstrap simulation method to quantify the prediction uncertainty. In order to apply the bootstrap method there is the parameter $s^2_{j,l}$ that still needs to be estimated.We do this under calibration (10), i.e. we set

$$s^2_{j,l} \equiv s^2_l \tag{18}$$

to avoid over-parameterization. In view of Proposition 5 we have

$$\mathbb{E}\left[\frac{X_{i,j}-\alpha_i\,\gamma_j}{\sqrt{\alpha_i\;\nu_i}}\right] \;=\; 0 \qquad \text{and} \qquad \operatorname{Var}\left(\frac{X_{i,j}-\alpha_i\,\gamma_j}{\sqrt{\alpha_i\;\nu_i}}\right) \;=\; \sigma^2_j.$$

The sample estimator then provides estimates

$$\widehat{\sigma}^2_j = \frac{1}{m-j-1}\sum_{i=1}^{m-j}\left(\frac{X_{i,j}-\widehat{\alpha}_i\,\widehat{\gamma}_j}{\sqrt{\widehat{\alpha}_i\;\widehat{\nu}_i}}\right),$$

for $j=0,\ldots,m-2$ and we set $\widehat{\sigma}^2_{m-1}=\widehat{\sigma}^2_{m-2}$. In view of (21) we have a second description for $\sigma^2_j$. If we solve this for $s^2_l$ and replace all parameters by their estimates we obtain estimates

$$(\widehat{(\pi s^2)}_0,\ldots,\widehat{(\pi s^2)}_{m-1})^{\mathsf{T}} = \widetilde{B}_{\beta}^{-1}(\widehat{\sigma}^2_0,\ldots,\widehat{\sigma}^2_{m-1})^{\mathsf{T}} - (\widetilde{\pi}^2_0\,\widehat{\mu}^2_0,\ldots,\widetilde{\pi}^2_{m-1}\,\widehat{\mu}^2_{m-1})^{\mathsf{T}},$$

and finally we set

$$\widehat{s}^2_l = \widehat{(\pi s^2)}_l/\widetilde{\pi}_l, \qquad \text{for all } \; l=0,\ldots,m-1. \tag{19}$$

If we apply this procedure to Example 1 we obtain the result in Table 2. In order to justify these estimates we calculate the estimates of the corresponding coefficients of variation given by $\widehat{\mathrm{vco}}=\widehat{s}_l/\widehat{\mu}_l$. Table 2 shows that these estimated coefficients of variation are in the interval $[1.5, 5.5]$, i.e. the coefficients of variation for single claims payouts $Y^{(k)}_{i,j,l}$ are of order 1.5 to 5.5. These are reasonable values, for instance in the Swiss Solvency Test (SST) the coefficients of variation for single claim sizes (not payouts) are estimated between 2.25 and 11 depending on the underlying line of business, see Section 8.4.4 in FINMA (2006). These estimators now allow for applying bootstrap methods which are

***Table 2:*** *Real data example: resulting standard deviation estimates $\widehat{s}_l$ together with the mean estimates $\widehat{\mu}_l$ and the corresponding coefficient of variation estimates $\widehat{\mathrm{vco}}$.*

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{s}_l$ | 2 862 | 8 511 | 11 651 | 26 688 | 93 291 | 28 083 | 52 846 | 43 333 | 104 714 | 59 276 | 75 632 | 104 701 |
| $\widehat{\mu}_l$ | 818 | 1 561 | 2 534 | 7 712 | 21 993 | 18 435 | 16 113 | 22 300 | 40 529 | 29 540 | 28 704 | 46 764 |
| $\widehat{\mathrm{vco}}$ | 3.50 | 5.45 | 4.60 | 3.46 | 4.24 | 1.52 | 3.28 | 1.94 | 2.58 | 2.01 | 2.63 | 2.24 |

***Table 3:** Real data example: process standard deviation (first row) and rooted conditional MSEP (second row) under model (A1)-(A3) for the predicted RBNS, IBNR and the total claim reserves. The last column gives the results of the Mack formula (Mack 2010).*

| | RBNS | IBNR | total | Mack (1993) |
|---|---|---|---|---|
| process standard deviation | 1 511 860 | 293 166 | 1 545 503 | 1 521 713 |
| conditional MSEP$^{1/2}$ | 2 273 294 | 326 382 | 2 324 966 | 2 182 722 |

close to those proposed by Martínez-Miranda *et al.* (2011, 2012). Specifically, we derive the predictive distribution using a parametric bootstrapping procedure which exploits the model assumptions in Section 2. In a first step we define a bootstrapping scheme based on Monte Carlo simulation from the model (A1)–(A3) where the unknown parameters are simply replaced by the estimated parameters (ignoring the parameter estimation uncertainty). This gives an estimation of the process variance defined as the first term in equation (17). The resulting process variances for RBNS, IBNR and total reserves (for all the years) are given in the first row of Table 3. To quantify the second term in equation (17), i.e. the parameter estimation error, we consider a more general bootstrap algorithm which also simulates the distribution of the involved parameters. From such general bootstrap method – formally described below – we derive the desired conditional MSEP. The resulting errors are displayed in the second row of Table 3. The last column displays the same uncertainties obtained from the Mack's distribution-free chain ladder model Mack (1993). We observe that our bootstrap results are slightly more conservative compared to the classical Mack formula.

## 6.2. Bootstrapping the RBNS and IBNR reserve

The predictive distribution which describes the possible fluctuations of the true outstanding loss liability cash flows can be derived using parametric bootstrap methods. By exploiting the distributional assumptions (A1)–(A3) we describe in the Appendix B an explicit algorithm to derive separately the predictive distribution of the RBNS and IBNR cash flows, $X_m^{\mathrm{RBNS}}$ and $X_m^{\mathrm{IBNR}}$, respectively. With this resampling scheme the RBNS and IBNR cash flows can be simulated using Monte Carlo methods. We have derived these cash flows for the data in Example 1. Table 4 shows the median and the upper quantiles separately for the RBNS and IBNR cash flows. Here we consider $B = 10\,000$ replications in the resampling scheme. As we expect the means imitate the predicted reserves given in Table 1. The calculated medians however are slightly lower, which reveals that the derived distribution is negatively skewed.

For comparison purposes we also consider the double chain ladder method (DCL) proposed by Martínez-Miranda *et al.* (2012). This method is defined under a simpler distributional model which makes the following assumptions on the first two moments $\mathbb{E}[Y_{i,j,l}^{(1)}] = \nu_i\mu$ and $\mathbb{E}[(Y_{i,j,l}^{(1)})^2] = \nu_i^2\sigma^2$. Table 5 reports the summary of the distribution for

***Table 4:*** *Real data example. Simulation of predictive distribution of RBNS and IBNR reserves by accident year: mean, median and* 95% *and* 99% *quantiles over* $B = 10\,000$ *repetitions. Column 2–5 give the RBNS reserves, Column 6–9 give the IBNR reserves.*

| | RBNS | | | | IBNR | | | |
|---|---|---|---|---|---|---|---|---|
| a.y. $i$ | mean | median | 95% | 99% | mean | median | 95% | 99% |
| 1 | 522 | 0 | 0 | 882 | 0 | 0 | 0 | 0 |
| 2 | 1658 | 0 | 0 | 38 893 | 0 | 0 | 0 | 0 |
| 3 | 23 947 | 0 | 140 730 | 352 637 | 0 | 0 | 0 | 0 |
| 4 | 165 490 | 73 036 | 633 033 | 1 172 022 | 0 | 0 | 0 | 0 |
| 5 | 297 554 | 199 030 | 932 461 | 1 579 736 | 202 | 0 | 18 | 3 910 |
| 6 | 418 854 | 321 105 | 1 106 734 | 1 840 653 | 688 | 0 | 2 509 | 16 514 |
| 7 | 586 159 | 476 807 | 1 435 075 | 2 194 285 | 1 617 | 0 | 8 521 | 30 389 |
| 8 | 609 403 | 522 311 | 1 377 117 | 2 047 477 | 2 312 | 8 | 12 839 | 34 068 |
| 9 | 712 294 | 615 028 | 1 548 731 | 2 236 150 | 3 750 | 92 | 18 937 | 49 613 |
| 10 | 809 344 | 716 227 | 1 660 795 | 2 374 073 | 5 108 | 639 | 22 475 | 58 796 |
| 11 | 1 056 515 | 953 340 | 2 092 864 | 2 990 605 | 9 096 | 2 896 | 37 051 | 78 573 |
| 12 | 1 410 137 | 1 295 048 | 2 537 813 | 3 437 648 | 21 487 | 13 153 | 69 200 | 125 271 |
| 13 | 2 008 886 | 1 899 042 | 3 271 259 | 4 179 189 | 86 354 | 72 050 | 188 811 | 327 241 |
| 14 | 4 211 291 | 4 126 027 | 5 463 231 | 6 402 499 | 1 552 438 | 1 512 135 | 2 074 502 | 2 514 402 |
| total | 12 312 055 | 12 040 963 | 16 325 473 | 18 860 539 | 1 683 054 | 1 640 097 | 2 222 831 | 2 709 200 |

the RBNS, IBNR and total claims reserves. The resulting reserves are similar when we consider the sum over all accident years. However, we observe more variability in the method proposed in this paper, under assumptions (A1)–(A2), compared to the DCL method. This is due to the fact that in DCL method there is the assumption that a claim is settled by a single payment and hence there is less volatility in the cash flow process. Besides, the model in this paper involves more parameters than the DCL model and therefore it increases the uncertainty of the parameters, which we are taking into account in the resampling scheme (see algorithm in Appendix B).

***Table 5:*** *Real data example. Bootstrap predictive distribution: RBNS, IBNR and total claims reserves. The first three columns give the summary of the distribution under model (A1)–(A3). The last three columns provide the bootstrap distribution from the DCL method proposed in Martínez-Miranda* et al. *(2012).*

| | model (A1)-(A3) | | | DCL | | |
|---|---|---|---|---|---|---|
| | RBNS | IBNR | total | RBNS | IBNR | total |
| mean | 12 312 055 | 1 683 054 | 13 995 109 | 11 758 152 | 1 585 151 | 13 343 303 |
| MSEP$^{1/2}$ | 2 273 294 | 326 382 | 2 324 966 | 1 881 154 | 485 312 | 2 018 112 |
| 1% | 8 090 717 | 1 131 376 | 9 615 040 | 8 081 739 | 687 623 | 9 314 398 |
| 5% | 9 088 207 | 1 262 754 | 10 685 634 | 9 012 040 | 897 886 | 10 408 658 |
| 50% | 12 040 963 | 1 640 097 | 13 723 567 | 11 637 796 | 1 532 079 | 13 243 493 |
| 95% | 16 325 473 | 2 222 831 | 18 101 695 | 14 869 197 | 2 448 915 | 16 729 435 |
| 99% | 18 860 539 | 2 709 200 | 20 660 941 | 16 516 558 | 2 941 469 | 18 487 830 |

## 7. Conclusions

In this paper we have defined the claims reserving model on an individual claims processes basis (micro-level). The definition of the model on this micro-level has been done such that on the aggregate level we re-discover the classical chain ladder reserving method. Under this model we show how extended data collection can provide us with more and better information to act in time on unforeseen patterns of outstanding liabilities. In particular we have focused on how various claims delays impact severities and how to incorporate this information in the reserve. Our approach in this paper shares the simplicity and intuitive appeal which have popularized the chain ladder method in claims reserving. But, with a little more effort in calculations and data requirements, our approach reports several other advantages. Since chain ladder is only based in the aggregated payments triangles, it cannot provide the split of the claims reserves into RBNS and IBNR and the tail as we do. Such split is required for the calculation of unallocated loss adjustment expenses ULAE and it gives valuable information to the insurer. In addition, to work under a well-defined and firm statistical model provides a suitable framework to develop consistent bootstrap methods to quantify the uncertainty in the predictions. In future work we will also consider simulation of coefficients of variation following the insights of for example Gulhar, Kibria, Albatineh and Ahmed (2012).

## Acknowledgement

This research was financially supported by a pump-priming grant by the Actuarial Profession. The first author also acknowledges the support of the European Commission under the Marie Curie Intra-European Fellowship FP7-PEOPLE-2011-IEF Project number 302600, and the Spanish Ministerio de Ciencia e Innovación, Project MTM2008-03010.

## References

Bailey, R. A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society*, 4–11.

FINMA (2006). *Swiss Solvency Test*. Technical Document, Version October 2, 2006.

Gulhar, B. M. M., Kibria, G., Albatineh, A. N. and Ahmed, N. U. (2012) A comparison of some confidence intervals for estimating the population coefficient of variation: a simulation study. *SORT*, 36/1, 45–68.

Jessen, H. A., Mikosch, T. and Samorodnitsky, G. (2011). Prediction of outstanding payments in a Poisson cluster model. *Scandinavian Actuarial Journal*, 2011/3, 214–237.

Jung, J. (1968). On automobile insurance ratemaking. *ASTIN Bulletin*, 5/1, 41–48.

Kremer, E. (1985). *Einführung in die Versicherungsmathematik*. Vandenhoek & Ruprecht, Göttingen.

Kuang, D., Nielsen, B. and Nielsen, J. P. (2008a). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95, 979–986.

Kuang, D., Nielsen, B. and Nielsen, J. P. (2008b). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95, 987–991.

Kuang, D., Nielsen, B. and Nielsen, J. P. (2009). Chain-Ladder as maximum likelihood revisited. *Annals of Actuarial Science*, 4/1, 105–121

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

Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating IBNR claims reserves. *ASTIN Bulletin*, 21/1, 93–109.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23/2, 213–225.

Mack, T. (2008). The prediction error of Bornhuetter/Ferguson. *ASTIN Bulletin*, 38/1, 93–109.

Martínez-Miranda, M. D., Nielsen, B., Nielsen, J. P. and Verrall, R. J. (2011). Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers. *ASTIN Bulletin*, 41/1, 107–129.

Martínez-Miranda, M. D., Nielsen, J. P. and Verrall, R. J. (2012). Double chain ladder. *ASTIN Bulletin*, 42/1, 59–76.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, 23/1, 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model variations and extensions. *ASTIN Bulletin*, 29/1, 5–25.

Saluz, A., Gisler, A. and Wüthrich, M. V. (2011). Development pattern and prediction error for the stochastic Bornhuetter-Ferguson claims reserving model. *ASTIN Bulletin*, 41/2, 279–313.

Verrall, R., Nielsen, J. P. and Jessen, A. (2010). Including count data in claims reserving. *ASTIN Bulletin*, 40/2, 871–887.

Wüthrich, M. V., Bühlmann, H. and Furrer, H. (2010). *Market-Consistent Actuarial Valuation*. 2nd edition. Springer.

Wüthrich, M. V. and Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley.

# A. Moments calculations

Here we provide calculations about the two first moments of the stochastic variables in the triangles $\mathcal{N}_m$, $\mathcal{R}_m$ and $\mathcal{X}_m$. Hereafter we work under the model assumptions (A1)–(A3) formulated in Section 2.2.

## A.1. Calculation of means

We start with the claims payments $X_{i,j,l}$ given in (A3). The conditional and unconditional means are given by

$$\mathbb{E}[X_{i,j,l}|N_{i,0},\ldots,N_{i,m-1}] = N_{i,j}\ \pi_l\ \nu_i\ \mu_{j,l},$$
$$\mathbb{E}[X_{i,j,l}] = \mathbb{E}[\mathbb{E}[X_{i,j,l}|N_{i,0},\ldots,N_{i,m-1}]] = \vartheta_i\ \beta_j\ \pi_l\ \nu_i\ \mu_{j,l}.$$

The total number of payments $R_{i,j}$ of accident year $i$ in accounting year $i+j$ has, conditionally given $\{N_{i,0},\ldots,N_{i,m-1}\}$, a Poisson distribution with conditional mean

$$\mathbb{E}[R_{i,j}|N_{i,0},\ldots,N_{i,m-1}] = \sum_{l=0}^{j}\mathbb{E}[R_{i,j-l,l}|N_{i,j-l}] = \sum_{l=0}^{j}N_{i,j-l}\ \pi_l.$$

This implies for the unconditional mean

$$\mathbb{E}[R_{i,j}] = \mathbb{E}[\mathbb{E}[R_{i,j}|N_{i,0},\ldots,N_{i,m-1}]] = \vartheta_i\ \sum_{l=0}^{j}\beta_{j-l}\ \pi_l.$$

Define $\lambda_j = \sum_{l=0}^{j}\beta_{j-l}\ \pi_l$, for $j=0,\ldots,m-1$, then we have just proved the following proposition.

**Proposition 2** $\mathbb{E}[N_{i,j}] = \vartheta_i\,\beta_j$ *and* $\mathbb{E}[R_{i,j}] = \vartheta_i\,\lambda_j.$

Thus, the pair $(N_{i,j},R_{i,j})$ satisfies the double chain ladder model of Martínez-Miranda *et al.* (2012) with inflation parameter set equal to 1. $\vartheta_i$ describes an exposure measure for accident year $i$, $(\beta_j)_j$ gives the claims reporting pattern and $(\lambda_j)_j$ provides the number of payment count pattern.

The accounting year payments $X_{i,j}$ for accident year $i$ in accounting year $i+j$ have, conditionally given $\{N_{i,0},\ldots,N_{i,m-1}\}$, a compound Poisson distribution with conditional mean

$$\mathbb{E}[X_{i,j}|N_{i,0},\ldots,N_{i,m-1}] = \sum_{l=0}^{j}N_{i,j-l}\ \pi_l\ \nu_i\ \mu_{j-l,l}.$$

This provides the unconditional mean for $X_{i,j}$ given by

$$\mathbb{E}[X_{i,j}] = \vartheta_i \; \nu_i \sum_{l=0}^{j} \beta_{j-l} \; \pi_l \; \mu_{j-l,l}.$$

We define the parameter $\gamma_j$ which only depends on the development period $j$ given by $\gamma_j = \sum_{l=0}^{j} \beta_{j-l} \; \pi_l \; \mu_{j-l,l}$. Thus, we obtain a cross-classified unconditional first moment for $X_{i,j}$ which is stated in the following proposition.

**Proposition 3** *We have for $\alpha_i = \vartheta_i \; \nu_i$ that $\mathbb{E}[X_{i,j}] \; = \; \alpha_i \, \gamma_j$.*

This moment property is similar to the Bornhuetter-Ferguson models used by Mack (2008) and Saluz, Gisler and Wüthrich (2011), Models 4.11 and 4.16. Moreover, Proposition 3 explains how the claims development reporting pattern $(\beta_j)_j$ for $N_{i,j}$ is related to the claims development pattern $(\gamma_j)_j$ for claims payments $X_{i,j}$.

### A.2. Calculation of variances

In a similar fashion to the first moments we calculate the variances. First we have under the conditional compound Poisson assumptions (A3)

$$\mathrm{Var}(X_{i,j,l} | N_{i,0}, \ldots, N_{i,m-1}) \; = \; N_{i,j} \; \pi_l \; \nu_i^2 \; s_{j,l}^2,$$

and for the unconditional variance we have

$$\begin{aligned}\mathrm{Var}(X_{i,j,l}) &= \mathrm{Var}(\mathbb{E}[X_{i,j,l} | N_{i,0}, \ldots, N_{i,m-1}]) + \mathbb{E}[\mathrm{Var}(X_{i,j,l} | N_{i,0}, \ldots, N_{i,m-1})] \\ &= \vartheta_i \; \beta_j \; \nu_i^2 \left(\pi_l^2 \; \mu_{j,l}^2 + \pi_l \; s_{j,l}^2\right).\end{aligned}$$

The total number of payments $R_{i,j}$ of accident year $i$ in accounting year $i+j$ has, conditionally given $\{N_{i,0}, \ldots, N_{i,m-1}\}$, a Poisson distribution with conditional variance

$$\mathrm{Var}(R_{i,j} | N_{i,0}, \ldots, N_{i,m-1}) = \sum_{l=0}^{j} \mathrm{Var}(R_{i,j-l,l} | N_{i,j-l}) = \sum_{l=0}^{j} N_{i,j-l} \; \pi_l.$$

This implies for the unconditional variance

$$\begin{aligned}\mathrm{Var}(R_{i,j}) &= \mathrm{Var}(\mathbb{E}[R_{i,j} | N_{i,0}, \ldots, N_{i,m-1}]) + \mathbb{E}[\mathrm{Var}(R_{i,j} | N_{i,0}, \ldots, N_{i,m-1})] \\ &= \; \vartheta_i \sum_{l=0}^{j} \beta_{j-l} \; \pi_l^2 + \vartheta_i \sum_{l=0}^{j} \beta_{j-l} \; \pi_l.\end{aligned}$$

Define for $j = 0, \dots, m-1$

$$t_j^2 = \sum_{l=0}^{j} \beta_{j-l}\ \pi_l\,(1+\pi_l) \geq \lambda_j, \tag{20}$$

then we have just proved the following proposition.

**Proposition 4** $\operatorname{Var}(N_{i,j}) \ = \ \vartheta_i\ \beta_j$ *and* $\operatorname{Var}(R_{i,j}) \ = \ \vartheta_i\ t_j^2.$

In view of Proposition 2 we see that for the number of payments $R_{i,j}$ we obtain over-dispersion parameter

$$\phi_j = \frac{t_j^2}{\lambda_j} = 1 + \frac{\sum_{l=0}^{j} \beta_{j-l}\ \pi_l^2}{\sum_{l=0}^{j} \beta_{j-l}\ \pi_l} \geq 1.$$

Note that $R_{i,j}$ has a mixed Poisson distribution which is exactly reflected in this over-dispersion parameter $\phi_j \geq 1$.

The accounting year payments $X_{i,j}$ for accident year $i$ in accounting year $i+j$ have, conditionally given $\{N_{i,0}, \dots, N_{i,m-1}\}$, a compound Poisson distribution with conditional variance

$$\operatorname{Var}(X_{i,j} | N_{i,0}, \dots, N_{i,m-1}) = \sum_{l=0}^{j} N_{i,j-l}\ \pi_l\ \nu_i^2\ s_{j-l,l}^2.$$

This provides the unconditional variances for $X_{i,j}$ given by

$$\begin{aligned}\operatorname{Var}(X_{i,j}) &= \operatorname{Var}(\mathbb{E}[X_{i,j} | N_{i,0}, \dots, N_{i,m-1}]) + \mathbb{E}[\operatorname{Var}(X_{i,j} | N_{i,0}, \dots, N_{i,m-1})] \\ &= \vartheta_i\ \nu_i^2\ \sum_{l=0}^{j} \beta_{j-l}\ \pi_l^2\ \mu_{j-l,l}^2 + \vartheta_i\ \nu_i^2\ \sum_{l=0}^{j} \beta_{j-l}\ \pi_l\ s_{j-l,l}^2.\end{aligned}$$

We define the parameter $\sigma_j^2$ which only depends on the development period $j$ given by

$$\sigma_j^2 = \sum_{l=0}^{j} \beta_{j-l}\ \pi_l\ \mu_{j-l,l} \left( \pi_l\ \mu_{j-l,l} + \frac{s_{j-l,l}^2}{\mu_{j-l,l}} \right). \tag{21}$$

Thus, we obtain a cross-classified model for $X_{i,j}$ with first moment given by $\mathbb{E}[X_{i,j}] = \alpha_i\ \gamma_j$ and variance given in the following proposition:

**Proposition 5** $\operatorname{Var}(X_{i,j}) \ = \ \alpha_i\ \nu_i\ \sigma_j^2.$

Again it is similar to the claims reserving models used in Mack (2008) and Saluz *et al.* (2011), Models 4.11 and 4.16, but now the parameters have an explicit meaning.

## B. Resampling schemes

Here we provide the algorithm to derive the predictive distribution of the RBNS and IBNR cash flow: $X_m^{\text{RBNS}}$ and $X_m^{\text{IBNR}}$. We denote by $\theta = \{\pi_l, \mu_l, s_l, \nu_i;\ l = 0, \ldots, m-1, i = 1, \ldots, m\}$ the set of parameters involved in the model, under calibration (10). Moreover, let $\widehat{\theta}$ denote the parameters estimated from the data $(\mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m)$ which can be calculated using the methods described in Section 3 and expression (19).

*Algorithm RBNS*

Step 1. *Estimation of the parameters and distributions.* From the observed data $(\mathcal{N}_m, \mathcal{R}_m, \mathcal{X}_m)$ estimate the model parameters $\theta$ by the estimator $\widehat{\theta} = \{\widehat{\pi}_l, \widehat{\mu}_l, \widehat{s}_l, \widehat{\nu}_i;\ l = 0, \ldots, m-1, i = 1, \ldots, m\}$, as described above. The payment delay distribution is estimated by a Poisson distribution with estimated parameter, i.e. $R_{i,j,l}|_{\{N_{i,0},\ldots,N_{i,m-1}\}} \sim \text{Poi}(N_{i,j}\widehat{\pi}_l)$. The distribution of the individual payments, $Y_{i,j,l}^{(1)}$ is estimated by a gamma distribution with shape parameter $\widehat{\lambda} = \widehat{\mu}_l^2/(\widehat{s}_l^2 - \widehat{\mu}_l^2)$ and scale parameter $\widehat{\kappa} = (\widehat{s}_l^2 - \widehat{\mu}_l^2)\widehat{\nu}_i/\widehat{\mu}_l$.

Step 2. *Bootstrapping the data.* Conditional on the observed number of reported claims $\mathcal{N}_m$ generate new bootstrapped triangles $\mathcal{R}_m^* = \{R_{i,j}^*;\ i+j \leq m\}$ and $\mathcal{X}_m^* = \{X_{i,j}^*;\ i+j \leq m\}$ as follows:

(i) Simulate the payment delay: from each $N_{i,j}$, $i+j \leq m$, generate the number of payments, $R_{i,j,l}^*$ from a Poisson distribution with parameter $N_{i,j}\widehat{\pi}_l$ estimated in Step 1. Calculate the bootstrapped total number of payments, $\mathcal{R}_m^* = \{R_{i,j}^*;\ i+j \leq m\}$ from expression (1).

(ii) Get the bootstrapped aggregated payments, $\mathcal{X}_m^* = \{X_{i,j}^*;\ i+j \leq m\}$, from the gamma distribution estimated in Step 1 and using expression (2) but replace $R_{i,j-l,l}$ by $R_{i,j-l,l}^*$.

Step 3. *Bootstrapping the parameters.* From the bootstrap data, $(\mathcal{R}_m^*, \mathcal{X}_m^*)$, and the original $\mathcal{N}_m$, estimate again the parameters and get bootstrapped parameters $\theta^*$.

Step 4. *Bootstrapping the RBNS predictions.* Simulate the RBNS cash flow, $X_m^{\text{RBNS}*}$, in the lower triangle using similar specifications to (i) and (ii) in Step 2 but with bootstrapped parameters $\theta^*$.

Step 5. *Monte Carlo approximation.* Repeat Steps 2-4 $B$ times and get the empirical bootstrap distribution of the RBNS cash flows $\{X_m^{\text{RBNS},b}; b = 1, \ldots, B\}$.

The IBNR algorithm to simulate the IBNR cash flows $X_m^{\text{IBNR}*}$ follows the same steps as the algorithm RBNS but, in addition, involves the estimation and the simulation of the number of reported claims $N_{i,j}$ in the lower triangle $\mathcal{J}_m$. In this case and under assumption (A1), we simulate $\mathcal{N}_m^* = \{N_{i,j}^*;\ (i,j) \in \mathcal{I}_m\}$ from a Poisson distribution with parameters estimated by the chain ladder estimates $\{\widehat{\vartheta}_i, \widehat{\beta}_j; i, j+1 = 1, \ldots, m\}$ (for more details we also refer to Martínez-Miranda *et al.* 2011).

## C. Run-off triangles

***Table 6:*** *Example 1, number of reported claims $N_{i,j}$, $(i,j) \in \mathfrak{I}_m$.*

| ay / dy | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 247 | 3 083 | 124 | 22 | 5 | 5 | 3 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 2 | 17 098 | 2 567 | 98 | 25 | 6 | 1 | 1 | 3 | 0 | 1 | 0 | 0 | 0 | |
| 3 | 16 110 | 2 700 | 107 | 18 | 7 | 5 | 4 | 1 | 4 | 0 | 0 | 0 | | |
| 4 | 14 426 | 2 253 | 103 | 17 | 10 | 3 | 2 | 1 | 1 | 1 | 0 | | | |
| 5 | 14 142 | 2 173 | 62 | 11 | 7 | 4 | 0 | 1 | 1 | 0 | | | | |
| 6 | 14 275 | 1 850 | 86 | 25 | 6 | 2 | 0 | 0 | 1 | | | | | |
| 7 | 14 019 | 1 797 | 97 | 19 | 5 | 1 | 1 | 1 | | | | | | |
| 8 | 13 933 | 1 602 | 84 | 24 | 6 | 3 | 1 | | | | | | | |
| 9 | 12 962 | 1 503 | 65 | 11 | 2 | 2 | | | | | | | | |
| 10 | 12 226 | 1 352 | 74 | 18 | 7 | | | | | | | | | |
| 11 | 11 124 | 1 347 | 57 | 12 | | | | | | | | | | |
| 12 | 10 360 | 1 307 | 56 | | | | | | | | | | | |
| 13 | 10 371 | 1 141 | | | | | | | | | | | | |
| 14 | 10 435 | | | | | | | | | | | | | |

***Table 7:*** *Example 1, number of payments $R_{i,j}$, $(i,j) \in \mathfrak{I}_m$.*

| ay / dy | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 761 | 4 800 | 324 | 71 | 39 | 14 | 10 | 6 | 3 | 5 | 5 | 2 | 2 | 0 |
| 2 | 10 927 | 4 077 | 303 | 60 | 28 | 12 | 13 | 5 | 8 | 4 | 5 | 5 | 0 | |
| 3 | 9 856 | 4 168 | 294 | 71 | 23 | 23 | 16 | 10 | 9 | 4 | 4 | 3 | | |
| 4 | 8 915 | 3 682 | 246 | 70 | 27 | 16 | 7 | 7 | 4 | 7 | 4 | | | |
| 5 | 8 854 | 3 340 | 265 | 46 | 33 | 9 | 4 | 6 | 2 | 5 | | | | |
| 6 | 8 881 | 3 000 | 199 | 70 | 22 | 15 | 8 | 8 | 4 | | | | | |
| 7 | 8 170 | 2 983 | 221 | 46 | 18 | 8 | 5 | 6 | | | | | | |
| 8 | 7 827 | 2 741 | 184 | 55 | 22 | 15 | 3 | | | | | | | |
| 9 | 6 999 | 2 540 | 166 | 44 | 18 | 7 | | | | | | | | |
| 10 | 6 240 | 2 420 | 184 | 45 | 18 | | | | | | | | | |
| 11 | 5 652 | 2 210 | 184 | 45 | | | | | | | | | | |
| 12 | 5 223 | 2 317 | 148 | | | | | | | | | | | |
| 13 | 5 627 | 2 024 | | | | | | | | | | | | |
| 14 | 5 483 | | | | | | | | | | | | | |

***Table 8:*** *Example 1, claims payments $X_{i,j}$, $(i,j) \in \mathfrak{I}_m$.*

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 829 717 | 5 690 608 | 874 882 | 420 112 | 154 884 | 55 497 | 46 239 | 313 960 | 290 204 | 12 936 | 6 218 | 18 755 | 4 678 | 0 |
| 2 | 9 263 718 | 5 004 173 | 971 523 | 660 324 | 208 000 | 531 391 | 495 368 | 48 367 | 566 099 | 49 905 | 362 747 | 388 190 | 0 | |
| 3 | 9 402 126 | 5 625 116 | 805 027 | 322 263 | 325 505 | 101 469 | 160 747 | 310 837 | 30 754 | 69 395 | 8 123 | 51 756 | | |
| 4 | 8 650 875 | 5 150 702 | 752 354 | 802 485 | 209 590 | 466 859 | 197 654 | 41 763 | 25 349 | 367 750 | 123 091 | | | |
| 5 | 8 848 118 | 4 748 516 | 1 390 699 | 1 140 610 | 412 090 | 359 991 | 20 169 | 220 227 | 54 395 | 240 967 | | | | |
| 6 | 9 070 691 | 5 890 678 | 519 808 | 539 202 | 127 701 | 86 472 | 122 060 | 83 853 | 6 660 | | | | | |
| 7 | 8 763 254 | 4 293 444 | 1 339 396 | 292 330 | 1 515 615 | 155 402 | 28 210 | 36 709 | | | | | | |
| 8 | 7 777 082 | 4 145 234 | 642 816 | 504 127 | 92 030 | 101 250 | 6 620 | | | | | | | |
| 9 | 7 212 984 | 3 498 230 | 778 132 | 354 855 | 626 442 | 342 182 | | | | | | | | |
| 10 | 6 265 457 | 3 737 631 | 546 644 | 182 490 | 297 995 | | | | | | | | | |
| 11 | 5 737 447 | 3 281 469 | 748 102 | 456 983 | | | | | | | | | | |
| 12 | 5 612 232 | 3 495 586 | 593 774 | | | | | | | | | | | |
| 13 | 6 386 024 | 3 289 703 | | | | | | | | | | | | |
| 14 | 6 110 750 | | | | | | | | | | | | | |

SORT 36 (2) July-December 2012, 221-230

# Flexible geoadditive survival analysis of non-Hodgkin lymphoma in Peru

Claudio Flores[1], Mar Rodríguez-Girondo[2,3], Carmen Cadarso-Suárez[3], Thomas Kneib[4], Guadalupe Gómez[1] and Luis Casanova[5]

**Abstract**

Knowledge of prognostic factors is an important task for the clinical management of Non Hodgkin Lymphoma (NHL). In this work, we study the variables affecting survival of NHL in Peru by means of geoadditive Cox-type structured hazard regression models while accounting for potential spatial correlations in the survival times. We identified eight covariates with significant effect for overall survival. Some of them are widely known such as age, performance status, clinical stage and lactic dehydrogenase, but we also identified hemoglobin, leukocytes and lymphocytes as covariates with a significant effect on the overall survival of patients with NHL. Besides, the effect of continuous covariates is clearly nonlinear and hence impossible to detect with the classical Cox method. Although the spatial component does not show a significant effect, the results show a trend of low risk in certain areas.

*MSC:* 62P10 (Applications to biology and medical science) 62N01 (Censored data models) 6207 (Data analysis)

*Keywords:* Non-Hodgkin lymphoma, structured regression models, survival analysis.

## 1. Introduction

Non-Hodgkin lymphomas (NHLs) are a group of lymphoproliferative malignancies of the lymphatic system defined by different morphological, immunophenotypic and genetic features. This heterogeneity determines different patterns of prognosis in the NHL patients that should be considered to optimize their treatment benefit (Friedberg *et al.*, 2008).

[1] Department of Statistics, Universitat Politècnica de Catalunya, Spain.

[2] SiDOR Research Group, University of Vigo, Spain. margirondo@uvigo.es

[3] Department of Statistics, University of Santiago de Compostela, Spain.

[4] Department of Economics, Georg August University, Göttingen, Germany.

[5] Instituto Nacional de Enfermedades Neoplásicas, Lima, Peru.

Received: November 2011
Accepted: May 2012

Traditionally, the International Prognostic Index (IPI) has been used to classify the NHL patients into four risk groups (low, intermediate low, intermediate high, high) considering five variables of prognostic significance (age, performance status, clinical stage, lactic dehydrogenase and extranodal sites) derived from a Cox regression analysis based on categorical covariates (Shipp *et al.*, 1993). However, a relatively important group of patients presents poor survival, despite being classified as good prognosis according to the IPI.

Several aspects can be related to the observed inaccuracy of the IPI. It is possible that important prognostic factors are not being included in the analysis such as new genetic and biological markers currently under investigation. Another important issue that could lead to implausible results refers to the categorization of the continuous covariates included in the IPI (age and lactic dehydrogenase).

Beyond the IPI, many studies of prognostic factors for NHL have been performed using the classical Cox's proportional hazard model. Within this framework, the effect of the continuous covariates is assumed to have a linear functional form, however it is important to note that when this assumption is not satisfied, the Cox model may lead to biased inferences, loss of statistical power and incorrect conclusions (Therneau and Grambsch, 2000).

In addition, in databases based on hospital records, referral centres, population studies or multicenter clinical trials, the results may be affected by spatial correlations. These complexities in the covariates affecting survival are not covered by the Cox model and hence a more general and flexible regression framework is required.

A variety of flexible methods have been developed in recent years. An up-to-date review of Cox-type models extensions can be found in Buchholz and Sauerbrei (2011). In this article we use geoadditive Cox-type structured hazard models to inspect the functional form of several covariates effects, including a spatial component, on the overall survival of the patients with NHL.

The rest of the paper is organized as follows. In Section 2 structured geoadditive Cox-type hazard regression models for modelling survival data are revisited. Section 3 presents the results of the analysis of the data set of NHL in Peru and finally, a discussion concludes the paper.

## 2. Methodology

### 2.1. Geoadditive survival models

In many clinical studies, the common target of analysis is to model the effect of several covariates (prognostic factors) on the survival time. A classical tool for studying the effect of a vector of covariates $\boldsymbol{v}$ on continuous survival times is the Cox proportional hazards model (Cox, 1972):

$$\lambda_i(t, \boldsymbol{v}) = \lambda_0(t)\exp(\boldsymbol{v}_i^\top \boldsymbol{\gamma}) \tag{1}$$

However, this specification is often not flexible enough for the correct modelling of variables affecting survival in many applications.

In our analysis, we used structured geoadditive survival models (Hennerfeind *et al.*, 2006; Kneib and Fahrmeir, 2007), a flexible spatial generalization of the Cox model. Specifically, the linear predictor of equation (1) was extended to a structured geoadditive predictor, including a spatial component for geographical effects and nonparametric terms for modelling unknown functional forms of the log-baseline hazard rate and nonlinear effects of continuous covariates. Specifically, individual hazard rates are given by:

$$\lambda_i(t) = \exp(\eta_i(t)), i = 1, \ldots, n \tag{2}$$

with geoadditive predictor

$$\eta_i(t) = g_0(t) + \boldsymbol{v}_i^\top \boldsymbol{\gamma} + \sum_{k=1}^{q} s_k(x_{ik}) + f_{\text{spat}}(s) \tag{3}$$

where $g_0(t) = \log(\lambda_0(t))$ represents the log-baseline hazard rate, the vector $\boldsymbol{\gamma}$ contains the usual linear effects, $s_k(x_k)$ refers to the nonlinear effect of a continuous covariate $x_k$, and $f_{\text{spat}}(s)$ is the spatial effect in region $s \in \{1, \ldots, S\}$.

In this representation, all the nonparametric effects, including the log-baseline hazard are modeled using penalized splines (P-splines, Eilers and Marx, 1996). Thus, the nonparametric problem is replaced by a parametric equivalent, in which a vector of regression coefficients is estimated under a smoothness penalty (details are given in Section 2.2.). The general idea is to approximate the functions $g_0$ and $s_k$ by linear combinations of B-splines basis functions,

$$s_k(x) = \sum_{j=1}^{d_k} \beta_j B_j(x) \tag{4}$$

where vector $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_{d_k})$ is the vector of unknown coefficients corresponding to the B-splines basis of degree $a$ and defined over a grid of $k$ knots lying on the domain of $x$, with $d_k = a + k - 1$. Specifically, we considered B-splines basis of degree 3 and a grid of 20 equidistant knots in our analyses.

At the same time, the spatial effect of each region $s$ is split up into a structured part and an unstructured part:

$$f_{\text{spat}}(s) = f_{\text{str}}(s) + f_{\text{unstr}}(s) \tag{5}$$

With this division of the spatial effects, we aim to distinguish between two types of geographical influential factors. On the one hand, the structured effect refers to a general smooth spatial effect along the whole studied area. On the other hand, the unstructured effect accounts for possible effects that may be present only locally. The structured spatial effects are modeled by means of Markov random fields, assuming that the effect of an area $s$ is conditionally Gaussian, where the expectation is the mean of the effects of neighbouring areas and the variance is inversely proportional to the number of neighbours, specifically

$$f_{\mathrm{str}}(s) = \beta_s^{\mathrm{str}} = \frac{1}{N_s} \sum_{s' \epsilon \delta_s} \beta_{s'}^{\mathrm{str}} + u_s, \;\; u_s \sim N\left(0, \frac{1}{\lambda_{\mathrm{str}} N_s}\right) \tag{6}$$

where $\delta_s$ denotes the set of neighbouring areas of $s$ and $N_s$ the corresponding number of areas falling in $\delta_s$. As for the unstructured spatial effects, a Gaussian region specific i.i.d. random effect is assumed.

As a result, we can express each of the predictor components as the product of an appropriate design matrix $\boldsymbol{Z}_j$ and a vector $\boldsymbol{\beta}_j$ of regression coefficients, and consequently we can represent the predictor vector $\boldsymbol{\eta}$ in a generic matrix notation as $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_1 \boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_q \boldsymbol{\beta}_q + \boldsymbol{Z}_{\mathrm{str}} \boldsymbol{\beta}_{\mathrm{str}} + \boldsymbol{Z}_{\mathrm{unstr}} \boldsymbol{\beta}_{\mathrm{unstr}}$, where $\boldsymbol{V}$ is the design matrix of parametric effects.

Interestingly, from equations (2) and (3) we can extend the concept of hazard ratio with respect to a reference value $x_{\mathrm{ref}}$. In contrast to the linear hazard ratios derived from the Cox model, the structured geoadditive survival specification provides flexible hazard ratio curves. Hence, for a given smooth effect $s$ associated to a continuous covariate $X$, the adjusted hazard ratio for a subject with covariate $x$ compared to a subject with covariate $x_{\mathrm{ref}}$ is given by the smooth curve:

$$\mathrm{HR}(x, x_{\mathrm{ref}}) = \exp(s(x) - s(x_{\mathrm{ref}})) \tag{7}$$

### 2.2. Estimation of the parameters

Under the usual assumptions about non-informative censoring, the log likelihood, given the vectors of all parametric effects $\boldsymbol{\gamma}$ and all nonparametric and spatial effects $\boldsymbol{\beta}$, is $\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \boldsymbol{\delta}^{\mathsf{T}} \boldsymbol{\eta} - \mathbf{1}^{\mathsf{T}} \boldsymbol{\Lambda}$, where $\boldsymbol{\eta}$ denotes the linear predictor defined in (3) and $\boldsymbol{\delta}$ and $\boldsymbol{\Lambda}$ are, respectively, the vector of censoring indicators and cumulative hazard rates.

However, instead of obtaining the estimates of $\boldsymbol{\beta}$ by means of the unpenalized likelihood, a penalty term is added to control the level of smoothness by penalizing wiggly functions. The most commonly used penalization term is based on the integral of the second derivative of the smooth functions, $s_k$:

$$\text{pen}(s_k) = \frac{1}{2}\lambda_i \int_0^{\infty} [s_k''(z_i)]^2 dz_i \tag{8}$$

Since equation (8) is a quadratic form of the corresponding vector of regression coefficients $\boldsymbol{\beta}_j$, it can be written as $\frac{1}{2}\lambda_j \boldsymbol{\beta}_j \boldsymbol{K}_j \boldsymbol{\beta}_j$, where the penalty matrix $\boldsymbol{K}_j$ is a positive semidefinite matrix and $\lambda_j$ a smoothing parameter. Furthermore, the smooth functions for the nonlinear effects are represented in terms of B-splines and it allows to approximate the penalty term in terms of the squared differences of coefficients associated with adjacent basis functions (Eilers and Marx, 1996). As a result, the difference penalty matrix can be written as $\boldsymbol{K}_j = \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}$, with $\boldsymbol{D}$ the second order difference matrix of neighbouring coefficients.

A special remark about the spatial smoothing is required. In this case, the smoothing referees to the intuitive idea that risk in neighbouring areas should be close to each other. We define as neighbour areas those sharing a common boundary and analogously to the nonlinear effects, we penalize large deviations between neighbouring coefficients $\boldsymbol{\beta}_{\text{str}}$, where $\lambda_{\text{str}}$ from equation (6) is considered as the corresponding smoothing parameter. Hence, the corresponding penalty matrix $\boldsymbol{K}_{\text{str}}$ is defined as an adjacency matrix. For the unstructured spatial effect, the penalty matrix is simply the identity matrix corresponding to independent and identically distributed random effects for the regions.

As a result, the estimation of the regression effects is based on the penalized log-likelihood to ensure a compromise between fidelity to data (in terms of the likelihood) and smoothness (in terms of the penalty terms):

$$l_{\text{pen}}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = l(\boldsymbol{\gamma}, \boldsymbol{\beta}) - \sum_{j=1}^{q} \lambda_j \boldsymbol{\beta}_j^{\mathsf{T}} \boldsymbol{K}_j \boldsymbol{\beta}_j - \lambda_{\text{str}} \boldsymbol{\beta}_{\text{str}}^{\mathsf{T}} \boldsymbol{K}_{\text{str}} \boldsymbol{\beta}_{\text{str}} - \lambda_{\text{unstr}} \boldsymbol{\beta}_{\text{unstr}}^{\mathsf{T}} \boldsymbol{K}_{\text{unstr}} \boldsymbol{\beta}_{\text{unstr}} \tag{9}$$

Empirical Bayes inference was used to fit the model. This inferential procedure is based on a mixed model representation of equation (9) where the smoothing parameters ($\lambda_j$) are considered as variance components corresponding to the vector of regression coefficients ($\boldsymbol{\beta}_j$). It allows for the simultaneous estimation of the regression coefficients and the smoothing parameters corresponding to each unknown function $g_0$, $s_k$ or $f_{\text{spat}}$ using restricted maximum likelihood (REML) estimation. See Kneib and Fahrmeir (2007) for details.

The analysis was conducted using BayesX statistical software (Brezger *et al.*, 2005) freely available online from `www.bayesx.org`. Empirical Bayes inference was performed due to its equivalence to the penalized splines likelihood in the frequentist framework but BayesX also allows for a full Bayesian inference by means of MCMC simulation techniques (Hennerfeind *et al.*, 2006). To check the consistency of our results with regard to the inference procedure, the corresponding full Bayesian analysis was also conducted.

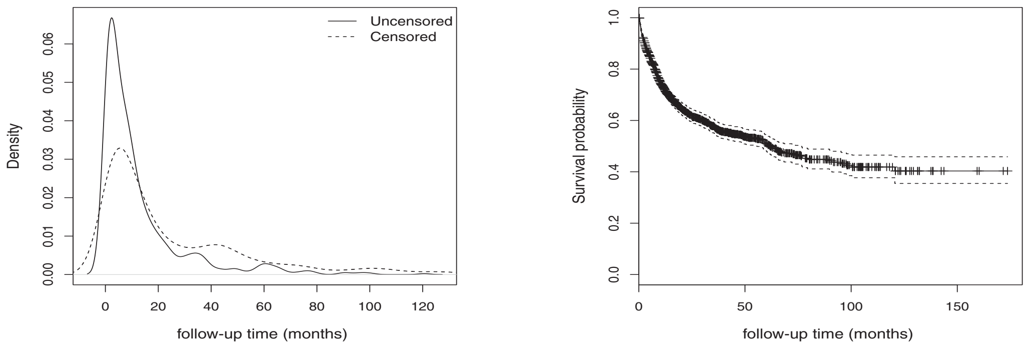

***Figure 1:*** *Density function of survival time (left) and Kaplan-Meier estimate of the overall survival curves with corresponding 95% confidence bands(right).*

## 3. Application to NHL

We analyzed survival data for 2160 patients diagnosed NHL, older than 14 years and treated at the Instituto Nacional de Enfermedades Neoplásicas (INEN), Lima, Peru, between 1990 and 2002. The clinical features evaluated were age, sex, performance status (zubrod), primary disease, clinical stage (CS), B symptoms, hemoglobin (Hbg), log leukocytes (ln(WBC)), lymphocytes and log lactic dehydrogenase (ln(LDH)).

***Table 1:*** *Fixed and random estimates of the fitted model.*

| Variables | df | HR (95% CI) |
|---|---|---|
| Fixed effects: | | |
| Sex: male | 1 | 1.25 (1.07,1.46) |
| Zubrod: 2-4 | 1 | 1.88 (1.59,2.22) |
| Primary: nodal | 1 | 0.90 (0.76,1.06) |
| CS: III-IV | 1 | 1.44 (1.22,1.71) |
| B-symptoms | 1 | 1.16 (0.99,1.37) |
| Non-parametric effecs: | | |
| g(t) | 3.52 | see figure 2 |
| Age | 2.67 | see figure 2 |
| Hbg | 4.16 | see figure 2 |
| ln(WBC) | 2.63 | see figure 2 |
| Lymphocytes | 5.79 | see figure 2 |
| ln(LDH) | 4.16 | see figure 2 |
| Spatial effects: | | |
| Random component | 5.41 | see figure 3 |
| Spatial component | 1.06 | see figure 3 |
| AIC | | 6549.95 |
| BIC | | 6750.99 |

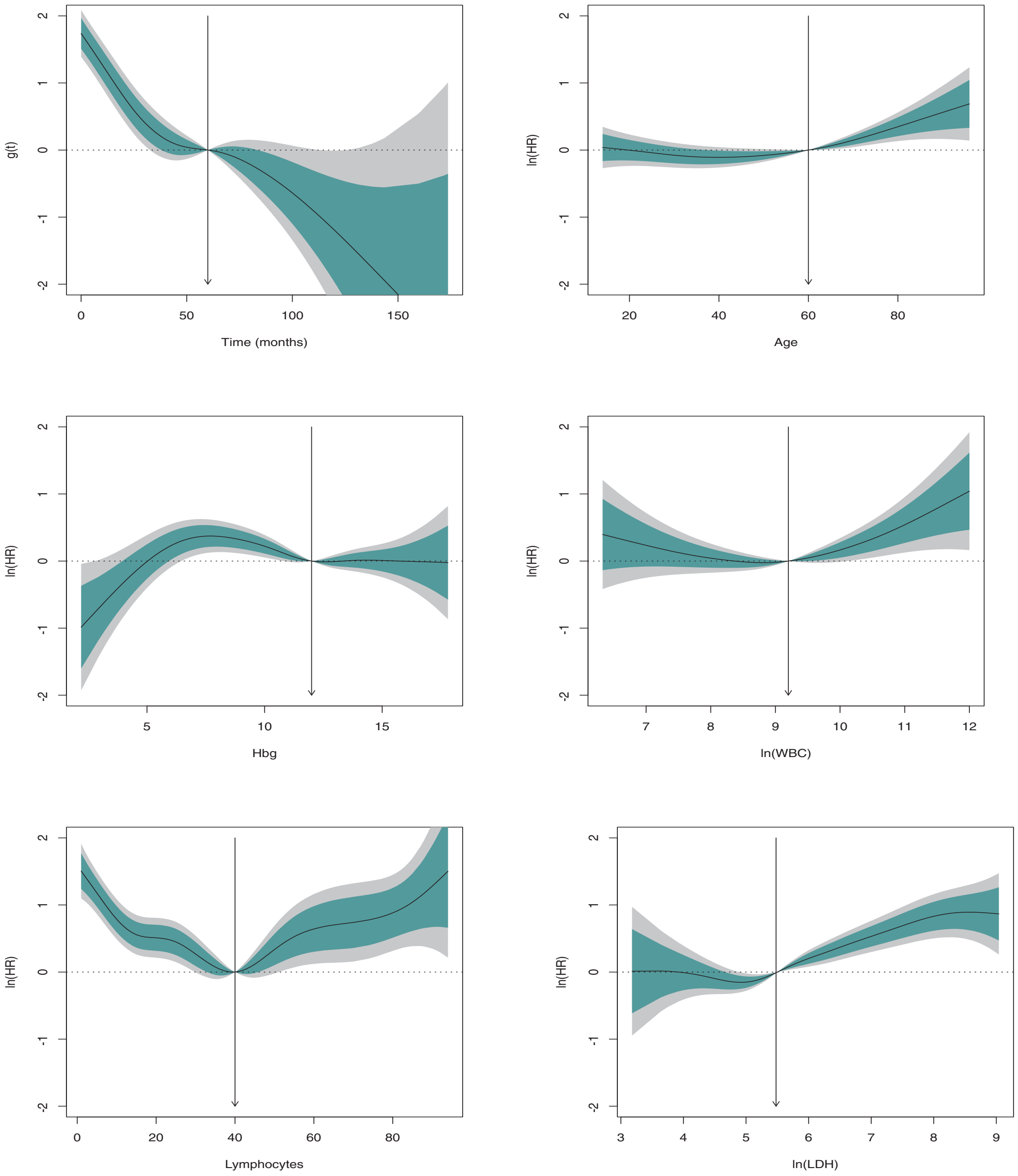

***Figure 2:*** *Logarithm of the smooth hazard ratio curves derived from the geoadditive model fit. Clinical cut-off values were used as reference points in the analysis. 80% (green) and 95% (grey) credible intervals are shown.*

The median age was 54.0 years (range: 14-96 years). Most patients presented advanced-stage disease at diagnosis: 50.8% presented Stage I-II and 49.2% presented Stage III-IV. Thirty-eight percent of the patients had B symptoms at diagnosis. The median length of follow-up for the patients was 12.6 months. Among all the patients, 32.8% had died before the end of the follow-up period (uncensored cases) and 67.2%

remained alive (censored cases). Figure 1 shows the distribution of survival time of patients with and without censoring (left) and the Kaplan-Meier estimate of the overall survival curve (right). According to the structured geoadditive Cox-type hazard analysis, eight prognostic factors were identified associated with worse survival (Table 1). Three categorical covariates: male patients, zubrod 2-4 and clinical stage III-IV at diagnosis were associated with worse prognosis for overall survival. The location of the disease described as primary nodal or extranodal, and the symptoms B had no significant effect on the overall survival.

A significant nonlinear relationship was identified for the effects of all continuous covariates: age, Hbg, ln(WBC), lymphocytes and ln(DHL). Figure 2 shows the functional form of the covariate effects in the log hazard ratio. Usual clinical cut-off values were used as reference points: 60 years (age), 12 $g/dL$ (Hbg), $10^3$ counts/dL (WBC), 40% (lymphocytes) and 240 $UI/L$ (DHL). Note that a strong nonlinear effect ($df = 5.79$) was found for Lymphocytes with increased hazard ratios for lowest and highest values. Risk geographical pattern is presented in Figure 3. Although Lima and Apurimac areas were identified as increased risk areas, the spatial effect was not significant according to the included variables.

As for the inference procedure, the results obtained from the full Bayesian inference (not shown) are very similar to the ones derived from REML estimation, hence we can assess that both inferential methods perform equivalently to our data.

## 4. Conclusions

The study of new covariates (with possible non-linear functional forms) in a flexible way and the existence of spatial correlation are examples of new challenges that the traditional tools of survival analysis do not allow to manage in an efficient way. Recent development of flexible methods for survival analysis allow for a deeper investigation of the variables affecting survival.

We used structured geoadditive survival models, a nonparametric approach that allows for the joint estimation of the baseline and covariates effects by means of a modelling through P-splines. Specifically, we considered nonlinear effects for the continuous covariates and we also account for possible geographical correlation.

In this work we identified eight covariates with significant effect for overall survival by means of the fitted geoadditive Cox-type structured hazard model. Age, zubrod, CS and DHL are prognostic factors reported in many published series, but we also identified hemoglobin, leukocytes and lymphocytes as covariates with a significant effect on the overall survival of patients with NHL.

Besides, the effect of continuous covariates is clearly non linear and hence impossible to detect with the classical Cox method. Nicely, the concept of hazard ratio is extended to obtain smooth hazard ratio curves for each of the continuous covariates.

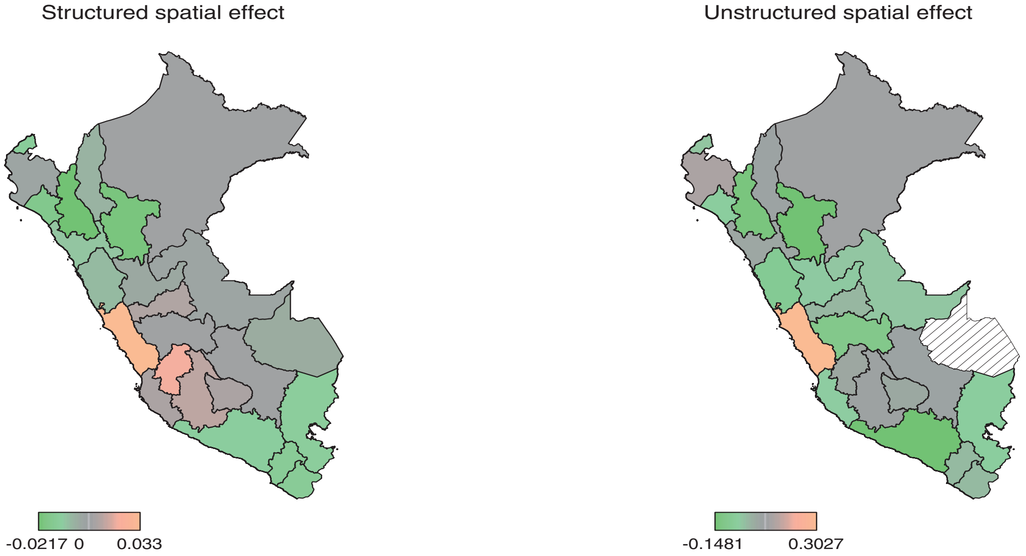

***Figure 3:*** *Spatial effect estimates.*

Although the spatial component does not show a significant effect, the results show a trend of low risk in certain areas. This phenomenon could be associated with certain subtype of NHL more frequent in these areas. So, the spatial analysis points out that further inspection of the NHL subtypes is required.

Still, it is noteworthy that more general specifications of the predictor are possible in the structured geoadditive Cox-type hazard regression framework, such as the inclusion of time-varying effects which allows to relax the proportional hazards assumption or the inclusion of interactions between covariates. In fact, possible extensions of the present work considering time-varying prognostic factors and interactions between them are currently under investigation.

To sum-up, geoadditive Cox-type structured hazard regression is a useful tool for assessing prognostic factors for the survival in a flexible way. This methodology allows to detect variables that may affect the risk of mortality while taking the possible spatial correlation of data into account.

## 5. Acknowledgments

This work was partially supported by grants MTM2008-01603, MTM2008-03129, MTM2008-06747-C02-00, MTM 2011-2304 and MTM2011-28285-C02-01 from the Spanish Ministry of Science and Innovation and the INBIOMED project (DXPCTSUG, Ref. 2009/063).

## References

Brezger, A., Kneib, T. and Lang, S. (2005). BayesX: Analyzing Bayesian structural additive regression models. *Journal of Statistical Software*, 14, i11.

Buchholz, A. and Sauerbrei, W. (2011). Comparison of procedures to assee non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal*, 53(2), 308–331.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34, 187–220.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties. *Statistical Science*, 11, 89–121.

Friedberg, J. W., Mauch, P. M., Rimsza, L. M. and Fisher, R. I. (2008). Non-Hodgkin's lymphomas. In: DeVita, V. T., Lawrence, T. S., Rosenberg, S. A., eds. DeVita, Hellman, and Rosenberg's Cancer: *Principles and Practice of Oncology*. 8th ed. Philadelphia, Pa: Lippincott Williams & Wilkins; 2278–2292.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive survival models. *Journal Of the American Statistical Association*, 101, 1065–1075.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34 207–228.

Shipp, M. A., Harrington, D. P. and Aderson, J. R. *et al.* (1993). A predictive model for aggressive non-Hodgkin's lymphoma. The International non-Hodgkin's lymphoma prognostic factors project. *The New England Journal of Medicine*, 329, 987–994.

Therneau, T. M. and Grambsch, P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. New York: Springer.