



Working Paper 08-39  
Statistics and Econometrics Series 09  
September 2008

Departamento de Estadística  
Universidad Carlos III de Madrid  
Calle Madrid, 126  
28903 Getafe (Spain)  
Fax (34) 91 624-98-49

## GOODNESS OF FIT IN MODELS FOR MORTALITY DATA

CARLO GIOVANNI CAMARDA<sup>1</sup> AND MARÍA DURBÁN<sup>2</sup>

### Abstract

---

Mortality data on an aggregate level are characterized by very large sample sizes. For this reason, uninformative outcomes are evident in common Goodness-of-Fit measures. In this paper we propose a new measure that allows comparison of different mortality models even for large sample sizes. Particularly, we develop a measure which uses a null model specifically designed for mortality data. Several simulation studies and actual applications will demonstrate the performances of this new measure with special emphasis on demographic models and P-spline approach.

---

**Keywords:** Goodness of fit, P-splines,  $R^2$ , mortality.

<sup>1</sup> Carlo Giovanni Camarda, Universidad Carlos III de Madrid, Departamento de Estadística, Avda. Universidad s/n, 28911, Leganés, Madrid, España.  
Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, 18057, Rostock, Alemania, email: [camarda@demogr.mpg.de](mailto:camarda@demogr.mpg.de)

<sup>2</sup> María Durbán, Universidad Carlos III de Madrid, Departamento de Estadística, Avda. Universidad s/n, 28911, Leganés, Madrid, España, email: [mdurban@est-econ.uc3m.es](mailto:mdurban@est-econ.uc3m.es)

# Goodness of fit in models for mortality data

Carlo G. Camarda\*

Max Planck Institute for Demographic Research and  
Departamento de Estadística, Universidad Carlos III de Madrid

María L. Durbán Reguera†

Departamento de Estadística, Universidad Carlos III de Madrid

## Abstract

Mortality data on an aggregate level are characterized by very large sample sizes. For this reason, uninformative outcomes are evident in common Goodness-of-Fit measures. In this paper we review the common measures of Goodness-of-Fit, and propose a new measure that allows comparison of different mortality models even for large sample sizes. Particularly, we develop a measure which uses a null model specifically designed for mortality data. Several simulation studies and actual applications will demonstrate the performances of this new measure with special emphasis on demographic models and P-spline approach.

KEYWORDS: Goodness of fit, P-splines,  $R^2$ , mortality.

---

\*Max Planck Institute for Demographic Research, Konrad-Zuse-Strasse 1, 18057 Rostock, Germany; email: [camarda@demogr.mpg.de](mailto:camarda@demogr.mpg.de)

†Departamento de Estadística, Universidad Carlos III de Madrid; C/ Madrid, 126 - 28903 Getafe, Spain; email: [mdurban@est-econ.uc3m.es](mailto:mdurban@est-econ.uc3m.es)

# 1 Introduction

Models such as,  $P$ -splines (Penalized splines), improve model fit with respect of traditional demographic models, uses less degrees of freedom and it is particularly flexible for modelling mortality surface. Nevertheless, demography can build on large samples, and this has implications for the statistical analysis of demographic data, including mortality studies. As Keyfitz (1966, p. 312) argues, in demographic studies, the “mere fact that over a period of a year the population is likely to be fifty or a hundredfold the deaths will result in a higher order of precision”.

The consequences of this peculiarity are relevant in the residual analysis, and in construction of confidence intervals in a  $P$ -spline approach for mortality surface. Specific and uninformative outcomes are also evident in common goodness-of-fit (gof) measures. However, such measures are a necessary statistical tool to compare mortality developments and, especially, to assess different models.

In this paper we will first present the common measure of GoF in the framework of GLMs (generalized linear models). Extensions and adjustments of the classic  $R^2$  are needed in models for non-Normal distributed data. Section 3 will introduce further extensions of gof measures in non-parametric settings such as the  $P$ -spline approach and effective dimension of the smoother will be considered when adjusting classic measures for GLMs. The presence of a large counts in the mortality surface makes simple corrections essentially uninformative, in Section 4 we will, thus, propose a new effective gof measure in models for mortality data:  $R_{mort}^2$ . The basic idea is to consider a null model which is specifically appropriate for mortality data. Particular emphasis will be given to the behaviour of this measure in the Lee-Carter model and the  $P$ -spline approach. Simulation studies in one- and two-dimensional settings, and applications for actual mortality surfaces will be presented in Section 5 and 6.

## 2 Goodness-of-fit measure for Generalized Linear Models

The goodness-of-fit (gof) measures examine how well a statistical model fits a set of observations. Measures of gof typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, or when investigating whether outcomes follow from a specified distribution.

In classic linear models the most frequently used measure to express how well the model

summaries feature of the data is the well-known  $R^2$ . It is also called the “coefficient of determination” or the “percentage of variance explained”. Its range is  $0 \leq R^2 \leq 1$ , with values closer to 1 indicating a better fit. It was developed to measure gof for linear regression models with homoscedastic errors. The concept of explained variation was generalized to heteroscedastic errors (Buse, 1973) and for logit, probit and tobit models (Veall and Zimmermann, 1996; Windmeijer, 1995).

One of the first studies on measuring explained variation in a GLM setting was undertaken by Cameron and Windmeijer (1997). They proposed an  $R^2$ -measure of goodness of fit for the exponential family. As a starting point they defined a measure that took into account the proportional reduction in uncertainty due to the inclusion of regressors. Since in GLMs we have generalized the Normal distribution, the coefficient of determination should be interpreted as the fraction of uncertainty explained and no longer as a percentage of variance explained.

More specifically Cameron and Windmeijer (1997) defined the  $R^2$  for an exponential family regression model based on the Kullback-Leibler (KL) divergence (Kullback, 1959). A standard measure of the information from observations in a density  $f(Y)$  is the expected information  $E[\log(f(Y))]$  with the KL divergence measuring the discrepancy between two densities. Let  $f_{\mu_1}(y)$  and  $f_{\mu_2}(y)$  be two densities differing in mean  $\mu$  only. The KL divergence is defined as

$$K(\mu_1, \mu_2) \equiv 2E_{\mu_1} \log \left[ \frac{f_{\mu_1}(y)}{f_{\mu_2}(y)} \right],$$

where the factor 2 is multiplied for convenience and  $E_{\mu_1}$  denotes that the expectation is taken with respect to the density  $f_{\mu_1}(y)$ . The KL measures how close  $\mu_1$  is to  $\mu_2$  and  $K(\mu_1, \mu_2) \geq 0$  with equality iff  $f_{\mu_1}(y) \equiv f_{\mu_2}(y)$ .

If we define  $f_y(y)$  the density for which the mean is set to the realized  $y$ , the deviation of  $y$  from the mean  $\mu$  is given by

$$K(y, \mu) \equiv 2E_y \log \left[ \frac{f_y(y)}{f_\mu(y)} \right] = 2 \int f_y(y) \log \left[ \frac{f_y(y)}{f_\mu(y)} \right] dy. \quad (2.1)$$

Hastie (1987) and Vos (1991) proved that if  $f_y(y)$  is within the exponential family equation (2.1) is reduced to

$$K(y, \mu) = 2 \log \left[ \frac{f_y(y)}{f_\mu(y)} \right].$$

In an estimated regression model, with  $n$  individual estimated means  $\hat{\mu}_i = \mu(x'_i \hat{\beta})$ , the KL

divergence between vectors  $\mathbf{y}$  and  $\boldsymbol{\mu}$  is equal to twice the difference between the maximum log-likelihood achievable,  $l(\mathbf{y}, \mathbf{y})$ , and the log-likelihood achieved by the fitted model  $l(\hat{\boldsymbol{\mu}}, \mathbf{y})$

$$K(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^n [\log f_{y_i}(y_i) - \log f_{\hat{\mu}_i}(y_i)] = 2 [l(\mathbf{y}, \mathbf{y}) - l(\hat{\boldsymbol{\mu}}, \mathbf{y})] \quad (2.2)$$

A particular case would be the constant only model where the fitted mean would be a  $n$ -vector  $\hat{\boldsymbol{\mu}}_0$  and the KL divergence,  $K(\mathbf{y}, \hat{\boldsymbol{\mu}}_0)$ , can be interpreted as the information in the sample data on  $\mathbf{y}$  potentially recoverable by inclusion of expectation with respect to the observed values  $\mathbf{y}$ .

Using the mentioned attributes of the KL divergence, Cameron and Windmeijer (1997) proposed an  $R^2$  for the class of exponential family regression models

$$R_{KL}^2 = 1 - \frac{K(\mathbf{y}, \hat{\boldsymbol{\mu}})}{K(\mathbf{y}, \hat{\boldsymbol{\mu}}_0)} \quad (2.3)$$

given that  $K(\mathbf{y}, \hat{\boldsymbol{\mu}}_0)$  is minimized when  $\hat{\boldsymbol{\mu}}_0$  is the maximum likelihood estimate. Furthermore,  $R_{KL}^2$  is proved to satisfy all the above mentioned criteria.

Since the expression for  $K(\mathbf{y}, \hat{\boldsymbol{\mu}})$  in (2.2) is equivalent to the definition of the deviance (McCullagh and Nelder, 1989, p. 33),  $R_{KL}^2$  can be interpreted as being based on deviance residuals. Therefore,  $R_{KL}^2$  is related to the analysis of deviance in the same way as the classic  $R^2$  is related to the analysis of variance. We can re-write the  $R_{KL}^2$  as

$$R_{KL}^2 = \frac{K(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0)}{K(\mathbf{y}, \hat{\boldsymbol{\mu}}_0)}, \quad (2.4)$$

which reveals another interesting aspect of this measure: using the canonical link in exponential family models,  $R_{KL}^2$  measures the fraction of uncertainty explained by the fitted model, if uncertainty is quantified by the deviance.

As we have seen, standard tools to quantify the discrepancy between observed and fitted values for Poisson models are deviance and the Pearson statistics. These concepts can be used to define different  $R^2$  measures. One idea is to compare the sum of squared Pearson residuals for two different models: the fitted model and the most restricted model in which only an intercept is included, which is estimated by  $\bar{y}$ . Cameron and Windmeijer (1996) proposed

$$R_{PEA}^2 = 1 - \frac{\sum_i^n (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{\sum_i^n (y_i - \bar{y})^2 / \bar{y}} \quad (2.5)$$

The choice of  $\bar{y}$  as weight in the denominator is a generalization for the Poisson case of the weighted  $R^2$  propose by Buse (1973).

Instead of using Pearson residuals, we alternatively can construct an  $R^2$ -measure based on Deviance Residuals. Let  $\bar{y}$  be the predicted mean for a Poisson model with just an intercept, then, the deviance is  $\text{Dev}(\mathbf{y}, \bar{\mathbf{y}}) = 2 \sum_i^n y_i \log(y_i/\bar{y})$ . From this formulation, the deviance- $R^2$  for the Poisson model is

$$R_{DEV}^2 = 1 - \frac{\sum_i^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{\sum_i^n y_i \log(y_i/\bar{y})}. \quad (2.6)$$

For the canonical link, the term  $\sum_i (y_i - \hat{\mu}_i)$  reduces to 0. Though equation (2.6) is equivalent to the  $R^2$  based on the KL divergence in (2.3), we opt for a different subscript to emphasize that deviance residuals are the basic quantities in (2.6).

## 2.1 Adjustments according to the number of parameters

The  $R^2$ -measures we presented so far do not consider the number of covariates used in the regression models. Some studies underline this aspect in situations with small sample sizes relative to the number of covariates in the model (Mittlböck and Waldhör, 2000; Waldhör et al., 1998). In these cases  $R^2$ -measures may be seriously inflated and may need to be adjusted. As we have seen, in a non-parametric setting effective dimensions are an equivalent concept to the number of covariates. Hence gof measures for non-parametric models should take into account the effective dimensions of the fitted model.

In a GLM context Waldhör et al. (1998) proposed to correct both the deviance and Pearson  $R^2$  in the following way:

$$R_{PEA,adj}^2 = 1 - \frac{(n-p-1)^{-1} \sum_i^n (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{(n-1)^{-1} \sum_i^n (y_i - \bar{y})^2 / \bar{y}}$$

$$R_{DEV,adj}^2 = 1 - \frac{(n-p-1)^{-1} \sum_i^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{(n-1)^{-1} \sum_i^n y_i \log(y_i/\bar{y})}$$

where  $p$  is the number of estimated covariates additional to the intercept. This type of adjustment stems from the normal linear model and is appropriate when using sum-of-squares approach to quantify deviation. In Poisson regression models, this adjustment would be just an approximation.

Two adjusted  $R^2$ -measures for Poisson regression models based on deviance residuals are

presented in Mittlböck and Waldhör (2000):

$$\begin{aligned}
R_{DEV,adj1}^2 &= 1 - \frac{l(\mathbf{y}) - [l(\hat{\boldsymbol{\mu}}) - p/2]}{l(\mathbf{y}) - l(\bar{\mathbf{y}})} \\
&= 1 - \frac{l(\mathbf{y}) - l(\hat{\boldsymbol{\mu}}) + p/2}{l(\mathbf{y}) - l(\bar{\mathbf{y}})} \\
R_{DEV,adj2}^2 &= 1 - \frac{l(\mathbf{y}) - [l(\hat{\boldsymbol{\mu}}) - (p+1)/2]}{l(\mathbf{y}) - [l(\bar{\mathbf{y}}) - 1/2]} \\
&= 1 - \frac{l(\mathbf{y}) - l(\hat{\boldsymbol{\mu}}) + (p+1)/2}{l(\mathbf{y}) - l(\bar{\mathbf{y}}) + 1/2}
\end{aligned} \tag{2.7}$$

It is easy to see how  $R_{DEV,adj2}^2$  is always closer to zero than  $R_{DEV,adj1}^2$ .

Mittlböck and Waldhör (2000) compared these two measures by simulation with different population values. They showed that  $R_{DEV,adj1}^2$  performs remarkably well where the Poisson regression is based on a small sample and/or many covariates. Moreover, while the equations in (2.7) work well in a GLM setting, further extensions are needed in the case of smoothing models, these are presented in the next section.

### 3 Extending $R^2$ -measures for smoothers

In the previous section we presented  $R^2$ -measures for GLM. Although the usage of likelihood ratio statistics in a smoothing context needs particular care (Ruppert et al., 2003), we follow the same arguments and construct  $R^2$ -measures for Poisson distributed mortality data fitted with  $P$ -splines.

Deaths counts are Poisson distributed data and, therefore, we use a measure based on deviance residuals since Pierce and Schafer (1986) illustrated that deviance residuals are more suitable for this type of data. We replace in equations (2.7) the number of covariates  $p$  by the effective dimension ED:

$$R_{DEV,SMO,1}^2 = 1 - \frac{\sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} + \frac{ED-1}{2}}{\sum_{i=1}^n \{y_i \log(y_i/\bar{y})\}} \tag{3.1}$$

and

$$R_{DEV,SMO,2}^2 = 1 - \frac{\sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} + \frac{ED}{2}}{\sum_{i=1}^n \{y_i \log(y_i/\bar{y})\} + \frac{1}{2}} \quad (3.2)$$

where ED is estimated. Equations (3.1) and (3.2) can be naturally computed in both unidimensional and two-dimensional contexts. Distinct smoothing techniques can be compared using these measures as long as they allow an easy computation of the used effective dimension of the model.

As a first example, the two  $R^2$ -measure are computed for Danish mortality data to which a surface is fitted with two-dimensional  $P$ -splines.  $R_{DEV,SMO,1}^2$  and  $R_{DEV,SMO,2}^2$  are equal to 0.995722 and 0.995721, respectively. The difference between these values is minimal and it seems that the fraction of uncertainty explained by the model is close to 100%. Table 1 presents values of the  $R^2$ -measure given in (3.2) for different period and age ranges. For comparison results of the  $R^2$ -measure from the Lee-Carter (LC) model are also given. All the values in Table 1 are in the range [0.989038, 0.995721] with marginally smaller values for the LC model. Of course the similarity between these outcomes do not reveal the important differences in explaining variation between  $P$ -spline and LC model. For a specific Danish mortality surface, such differences between the two approaches are evident from the fitted values and from the residual pattern.

Danish Data			$P$ -splines	Lee-Carter
females	1930–2006	10-100	0.995721	0.992671
males	1930–2006	10-100	0.995166	0.992583
females	1930–2006	50-100	0.993654	0.989038
males	1930–2006	50-100	0.993951	0.990885
females	1950–2006	50-100	0.994261	0.991518
males	1950–2006	50-100	0.994089	0.991426

Table 1:  $R_{DEV,SMO,2}^2$  values for the Danish population by different period and age ranges as well as models.

The presence of a large number of death counts in the mortality surface leads to rather small deviance residuals which are the basic elements of these  $R^2$ -measures. Consequently, equations (3.1) and (3.2) will always generate figures significantly close to 1, which are essentially uninformative. An explanation for this drawback of equations (3.1) and (3.2) refers directly to the null model in the denominators of these measures. The model in the denominators incorporates only the intercept, and is a reasonable null model in a GLM framework. A different



and peculiar null model is needed in models for mortality data. Specifically, it is appropriate to compare different mortality models to a “limit” model, which is nested in all the selected models, and which is more complex than a simple constant plane. For instance, the mortality surface for the Danish population is a  $91 \times 77$  matrix. It is pleonastic and uninformative to check whether a  $P$ -spline model explains the variation in the data more than the overall mean of the matrix.

## 4 Alternative $R^2$ -measures for mortality models

An alternative strategy for constructing a gof measure for mortality data is to choose, as null model, the linear and bilinear models for unidimensional and two-dimensional models, respectively. That is we consider a model where age and time and possibly their interaction. This approach is appealing since both  $P$ -splines and the Lee-Carter can be seen as extensions of this proposed null model. This will allow comparison of different models relative to the linear or bilinear null model.

In the next sections we will show that a linear model is nested within a  $P$ -spline model or a Lee-Carter model, and therefore it is natural to use it as a null model. In Section 4.3 we will demonstrate that this decomposition can be used for the alternative  $R^2$ -measure.

### 4.1 $P$ -splines with a transformed basis

In a  $P$ -splines setting, one can extract a linear component from the fitted trend, and fit the remaining variation by fitting a smooth curve with a penalty that penalized departures from zero. An early reference about this topic can be found in Green (1985). Verbyla et al. (1999) and Currie et al. (2006) discussed this idea, too.

We can write a  $P$ -splines model in the following way:

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (4.1)$$

where  $\mathbf{B} = \mathbf{B}(x)$ ,  $n \times k$  is the regression matrix of the  $B$ -splines and an additional difference penalty  $\mathbf{P}$  on the coefficients  $\mathbf{a}$  is used to enforce smoothness. Given these components, the

smoothed function is found by minimizing

$$S^* = \|\mathbf{y} - \mathbf{B}\mathbf{a}\|^2 + \mathbf{P}.$$

where  $\mathbf{P} = \lambda \mathbf{D}'_d \mathbf{D}_d$  and  $\mathbf{D}_d$  is the difference matrix acting on the coefficients  $\mathbf{a}$ .

A linear or bilinear model can be seen as a nested model in the more general  $P$ -spline framework, and in the following we will demonstrate explicitly this association. Specifically, we will present how a  $P$ -spline model can be decomposed in two unique and distinct components, one of which is the linear model.

In particular, we will show how to represent equation (4.1) in the following alternative way:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \epsilon, \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (4.2)$$

where  $\mathbf{G}$  is a covariance matrix which depends on  $\lambda$  and in the following we will assume the simple structure  $\mathbf{G} = \sigma_\alpha^2 \mathbf{I}$ , with unknown variance  $\sigma_\alpha^2$ . In this way we separated the fixed part, which does not depend on the smoothing parameter, and the remaining variation which will be smoothed via  $P$ -splines. We will show how the fixed part can be a simple linear (bilinear) model in a unidimensional (two-dimensional) setting.

It is worth pointing out that this is the common representation of  $P$ -splines as mixed models (Currie and Durban, 2002; Currie et al., 2006; Durban et al., 2006) and that the fixed-effects model-matrix  $\mathbf{X}$  is nested in the full model-matrix  $\mathbf{B}$ .

We assume that a second order penalty is used, i.e.  $d = 2$ . The aim is to find the unique matrix  $\mathbf{T}$  such that

$$\mathbf{B}\mathbf{T} \equiv [\mathbf{X} : \mathbf{Z}] \Rightarrow \mathbf{B}\mathbf{a} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} \quad (4.3)$$

The idea is to use the Singular Value Decomposition (SVD, Good, 1969) of the penalty  $\mathbf{P}$  to partition the difference penalty into a null penalty (for the fixed part) and a diagonal penalty (for the random part).

The SVD of the square matrix  $\mathbf{D}'\mathbf{D}$  can be written as

$$\mathbf{D}'\mathbf{D} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}',$$

where  $\mathbf{V}$  can be decomposed into two matrices:  $\mathbf{V}_n$  and  $\mathbf{V}_s$ . The former is the part of the matrix

$\mathbf{V}$  corresponding to the zero eigenvalues of  $\mathbf{D}'\mathbf{D}$  (fixed part). Since we are using a second order penalty we will have only two zero eigenvalues. The matrix  $\mathbf{V}_s$  is corresponding to the nonzero eigenvalues (random part). Therefore the fixed part would be:

$$\mathbf{X} = \mathbf{B}\mathbf{V}_n \quad (4.4)$$

For the random part we consider the diagonal matrix  $\mathbf{\Lambda}$  where we remove the elements corresponding to the fixed part  $\mathbf{X}$ :

$$\mathbf{\Lambda} = \left[ \begin{array}{c|c} \tilde{\Sigma} & \\ \hline & \mathbf{0}_{2 \times 2} \end{array} \right]$$

The new diagonal matrix  $\tilde{\Sigma}$  contains the non-zero eigenvalue, therefore the random-effects part can be written as:

$$\mathbf{Z} = \mathbf{B}\mathbf{V}_s \tilde{\Sigma}^{-\frac{1}{2}} \quad (4.5)$$

The mentioned matrix  $\mathbf{T}$  will then be:

$$\mathbf{T} = \left[ \mathbf{V}_n : \mathbf{V}_s \tilde{\Sigma}^{-\frac{1}{2}} \right]$$

and consequently  $\mathbf{TB} = [\mathbf{X} : \mathbf{Z}]$  giving the parameterization in equation (4.3) where

$$\boldsymbol{\beta} = \mathbf{V}_n' \mathbf{a} \quad \text{and} \quad \boldsymbol{\alpha} = [\mathbf{V}_s \tilde{\Sigma}^{\frac{1}{2}}] \mathbf{a}.$$

and the penalty term is given by

$$\mathbf{a}' \mathbf{D}' \mathbf{D} \mathbf{a} \rightarrow \boldsymbol{\alpha}' \boldsymbol{\alpha}$$

In this way quadratic and cubic fixed-effects can be chosen with  $d = 3$  and  $d = 4$ , respectively (Verbyla et al., 1999, p. 308). This representation can be generalized in a Poisson case in a straightforward manner with the additional weight matrix and link function. In a one-dimension setting, the fixed part for mortality data which can be used as null model for constructing gof

measure will be a simple linear term such that

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad (4.6)$$

where the second column of  $\mathbf{X}$  will be either the age or year values.

In a two dimensional setting the previous considerations can be easily generalized as in the  $P$ -spline approach. We now have  $\mathbf{B} = \mathbf{B}_y \otimes \mathbf{B}_a$  with penalty  $\mathbf{P}$  given by

$$\mathbf{P} = \lambda_a \mathbf{I}_{c_y} \otimes \mathbf{D}'_a \mathbf{D}_a + \lambda_y \mathbf{D}'_y \mathbf{D}_y \otimes \mathbf{I}_{c_a}. \quad (4.7)$$

Taking the SVD of  $\mathbf{D}'_a \mathbf{D}_a$  we obtain  $\mathbf{V}_a \boldsymbol{\Sigma}_a \mathbf{V}'_a$  and partitioning the matrix

$$\mathbf{V}_a = [\mathbf{V}_{as} : \mathbf{V}_{an}]$$

where  $\mathbf{V}_{as}$  corresponds to the non-zero eigenvalues and  $\mathbf{V}_{an}$  to the zero eigenvalues.

Assuming a second order penalty in both dimensions,  $\boldsymbol{\Sigma}_a$  has two zero eigenvalues and  $\mathbf{V}_{an}$  has two columns. Let  $\boldsymbol{\Sigma}_{as}$  contains the positive eigenvalues of  $\boldsymbol{\Sigma}_a$ . In the same way we decompose  $\mathbf{D}'_y \mathbf{D}_y$  obtaining  $\mathbf{V}_y = [\mathbf{V}_{ys} : \mathbf{V}_{yn}]$  and  $\boldsymbol{\Sigma}_{ys}$ .

Then we have the fixed part:

$$\begin{aligned} \mathbf{X} &= \mathbf{B}(\mathbf{V}_{yn} \otimes \mathbf{V}_{an}) \\ &= \mathbf{B}_y \mathbf{V}_{yn} \otimes \mathbf{B}_a \mathbf{V}_{an} \\ &= \mathbf{X}_y \otimes \mathbf{X}_a. \end{aligned} \quad (4.8)$$

And the random part is given by

$$\begin{aligned} \mathbf{Z} &= \mathbf{B}(\mathbf{V}_{ys} \otimes \mathbf{V}_{as}) \tilde{\boldsymbol{\Sigma}}^{-1/2} \\ &= (\mathbf{B}_y \mathbf{V}_{ys} \otimes \mathbf{B}_a \mathbf{V}_{as}) \tilde{\boldsymbol{\Sigma}}^{-1/2} \end{aligned} \quad (4.9)$$

where the diagonal matrix  $\tilde{\Sigma}$  is a block-diagonal containing the non-zero eigenvalues in both dimensions.

The new basis  $\mathbf{T}$  is given by the combination of equations (4.8) and (4.9):

$$\mathbf{T} = [\mathbf{V}_{yn} \otimes \mathbf{V}_{an} : \mathbf{V}_{ys} \otimes \mathbf{V}_{an} : \mathbf{V}_{yn} \otimes \mathbf{V}_{as} : \mathbf{V}_{ys} \otimes \mathbf{V}_{as}]$$

We can prove that  $\mathbf{T}$  is orthogonal so we can represent for the two-dimensional case in (4.3) where

$$\begin{aligned}\boldsymbol{\beta} &= (\mathbf{V}_{yn} \otimes \mathbf{V}_{an})' \mathbf{a} \\ \boldsymbol{\alpha} &= (\mathbf{V}_{ys} \otimes \mathbf{V}_{an} : \mathbf{V}_{yn} \otimes \mathbf{V}_{as} : \mathbf{V}_{ys} \otimes \mathbf{V}_{as})' \mathbf{a}.\end{aligned}$$

The penalty is then given by

$$\mathbf{a}' \mathbf{P} \mathbf{a} = \boldsymbol{\omega}' \mathbf{T}' \mathbf{P} \mathbf{T} \boldsymbol{\omega}$$

where  $\boldsymbol{\omega}' = [\boldsymbol{\beta}' : \boldsymbol{\alpha}']$ .

The fixed part for our null model in a two-dimensional case for mortality surfaces is then

given by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & a_1 & y_1 & a_1 \cdot y_1 \\ 1 & a_2 & y_1 & a_2 \cdot y_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_1 & a_m \cdot y_1 \\ 1 & a_1 & y_2 & a_1 \cdot y_2 \\ 1 & a_2 & y_2 & a_2 \cdot y_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_2 & a_m \cdot y_2 \\ 1 & a_1 & y_3 & a_1 \cdot y_3 \\ 1 & a_2 & y_3 & a_2 \cdot y_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_3 & a_m \cdot y_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_1 & y_n & a_1 \cdot y_n \\ 1 & a_2 & y_n & a_2 \cdot y_n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_n & a_m \cdot y_n \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \quad (4.10)$$

where  $a_i, i = 1, \dots, m$  and  $y_j, j = 1, \dots, n$  are age and year values, respectively. In this way we have as fixed part, and, consequently as null model, a bilinear surface in which age and time interact. The dimension of the model is equal to four, i.e. the number of columns of  $\mathbf{X}$ . The linear model can be easily fitted using a two-dimensional  $P$ -spline framework by choosing large smoothing parameters for both age and year. In the example we considered  $\lambda_a = \lambda_y = 10^8$  worked well and lead to effective dimension of about four.

## 4.2 The Lee-Carter as a simple bilinear model

In a two-dimensional setting the Lee-Carter (LC) model is widely used in modelling mortality surface and it is a commonly used model for mortality developments. Therefore, it is useful to also apply alternative gof measures for mortality data to this model. In this section we will show how the basic bilinear structure is nested in the LC model, too.

The LC model is given by:

$$Y_{ij} \sim \text{Poisson}(E_{ij} \cdot \exp(\alpha_i + \beta_i \cdot \gamma_j))$$

where  $\alpha_i$ ,  $\beta_i$  and  $\gamma_j$  are vectors of parameters that have to be estimated. Using the canonical link function for Poisson distributed data, the linear predictor of the LC model is given by

$$\eta_{ij} = \alpha_i + \beta_i \cdot \gamma_j. \quad (4.11)$$

It can be proved that equation (4.11) is a general case of the fixed part of the model in (4.3) in the two dimensional case, where the linear predictor is given in equation (4.10). We let the Lee-Carter vectors of parameters,  $\alpha_i$ ,  $\beta_i$  and  $\gamma_j$ , vary linearly over ages and years, that is

$$\begin{aligned} \alpha_i &= \theta_1 + \theta_2 \cdot a_i \\ \beta_i &= \theta_3 + \theta_4 \cdot a_i \\ \gamma_j &= \theta_5 + \theta_6 \cdot y_j. \end{aligned} \quad (4.12)$$

The linear predictor in (4.11) then becomes:

$$\begin{aligned} \eta_{ij} &= \theta_1 + \theta_2 a_i + (\theta_3 + \theta_4 a_i)(\theta_5 + \theta_6 y_j) \\ &= (\theta_1 + \theta_3 \theta_5) + (\theta_2 + \theta_4 \theta_5) a_i + \theta_3 \theta_6 y_j + \theta_4 \theta_6 a_i y_j, \end{aligned}$$

which is equivalent to linear part of the mixed model representation of  $P$ -splines models in equation (4.10) if

$$\begin{aligned} \beta_1 &= \theta_1 + \theta_3 \theta_5 \\ \beta_2 &= \theta_2 + \theta_4 \theta_5 \\ \beta_3 &= \theta_3 \theta_6 \\ \beta_4 &= \theta_4 \theta_6. \end{aligned}$$

### 4.3 $R_{mort}^2$ : a goodness-of-fit measure for mortality data

In Sections 4.1 and 4.1 we showed that both the two-dimensional  $P$ -spline model and the Lee-Carter model can be seen as extensions of a bilinear surface, where age and year interact. Such parsimonious surface can be used as a null model to compare the explained variability from more sophisticated models in an appropriate way. That is, we replace the constant surface as a null model in the denominator in (3.2), by either an estimated linear or bilinear model as given in (4.6) and (4.10).

We define by  $\hat{\mu}_i^1$  and  $\hat{\mu}_i^0$  the estimated values by the fitted model and the null model, respectively. In a similar fashion,  $ED^1$  and  $ED^0$  denote the effective dimension of the two models. Recalling explicitly equation (3.2), we propose as appropriated gof measure for mortality models the following equation:

$$R_{mort}^2 = 1 - \frac{\sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i^1) - (y_i - \hat{\mu}_i^1)\} + \frac{ED^1}{2}}{\sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i^0) - (y_i - \hat{\mu}_i^0)\} + \frac{ED^0}{2}} \quad (4.13)$$

where  $n$  denotes the total number of measurement values in the data. As mentioned, the null model is defined by the linear predictor in (4.10) or (4.6). The variation explained by the fitted model is now compared to the bilinear model.

Equation (4.13) can be alternatively written as

$$R_{mort}^2 = 1 - \frac{\text{Dev}^1(\mathbf{y}; \mathbf{a}^1, \lambda) + \frac{ED^1(\mathbf{a}^1, \lambda)}{2}}{\text{Dev}^0(\mathbf{y}; \mathbf{a}^0) + \frac{ED^0(\mathbf{a}^0)}{2}} \quad (4.14)$$

where  $\mathbf{a}$  are the coefficients of the model and  $\lambda$  is the smoothing parameter, for smoothing models. Again the superscripts 0 and 1 denote quantities computed from the null and fitted model, respectively.

As in the other  $R^2$ -measures (see Section 2), values of (4.14) closer to 1 indicate a better fits compared to the bilinear null model. Moreover,  $\frac{ED^0(\mathbf{a}^0)}{2}$  is equal to 1 and 2 in unidimensional and two-dimensional setting, respectively, and it does not depend to the smoothing parameter  $\lambda$  in a smoothing setting.

#### 4.3.1 Relations between $R_{mort}^2$ and information criteria

Actual associations between our  $R_{mort}^2$  and information criteria for smoother can shed additional light the meaning and implications of the proposed  $R^2$ -measure. We already presented several



information criteria for selection of smoothing parameters for different data. Here we focus on the commonly used criteria in case of Poisson data: Akaike Information Criterion and Bayesian Information Criterion. Let recall their formulas:

$$\begin{aligned} \text{AIC}(\lambda) &= \text{Dev}(\mathbf{y}; \mathbf{a}, \lambda) + 2 \cdot \text{ED}(\mathbf{a}, \lambda) \\ \text{BIC}(\lambda) &= \text{Dev}(\mathbf{y}; \mathbf{a}, \lambda) + \ln(n) \cdot \text{ED}(\mathbf{a}, \lambda), \end{aligned}$$

and, consequently, the deviance can be written equivalently as:

$$\begin{aligned} \text{Dev}(\mathbf{y}; \mathbf{a}, \lambda) &= \text{AIC}(\lambda) - 2 \cdot \text{ED}(\mathbf{a}, \lambda) \\ \text{Dev}(\mathbf{y}; \mathbf{a}, \lambda) &= \text{BIC}(\lambda) - \ln(n) \cdot \text{ED}(\mathbf{a}, \lambda). \end{aligned} \tag{4.15}$$

One can show that AIC and BIC are linked with  $R_{mort}^2$ . Substituting the first equation of (4.15) in the measure presented in (4.14), we obtain:

$$\begin{aligned} R_{mort}^2 &= 1 - \frac{\text{AIC}^1 - 2 \cdot \text{ED}^1 + \frac{\text{ED}^1}{2}}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\ &= 1 + \frac{\frac{3}{2} \cdot \text{ED}^1 + \text{AIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\ &= \left[ 1 + \frac{\frac{3}{2} \cdot \text{ED}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \right] - \left[ \frac{1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \right] \cdot \text{AIC}^1 \end{aligned} \tag{4.16}$$

which shows the relation between  $R_{mort}^2$  and AIC. In a similar fashion, substituting the second equation of (4.15) in (4.14), we obtain the relationship between  $R_{mort}^2$  and BIC:

$$\begin{aligned} R_{mort}^2 &= 1 - \frac{\text{BIC}^1 - \ln(n) \cdot \text{ED}^1 + \frac{\text{ED}^1}{2}}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\ &= 1 + \frac{[\ln(n) - \frac{1}{2}] \cdot \text{ED}^1 + \text{BIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\ &= 1 + \frac{[\ln(n) - \ln(e^{\frac{1}{2}})] \cdot \text{ED}^1 + \text{BIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\ &= \left[ 1 + \frac{\ln\left(\frac{n}{\sqrt{e}}\right) \cdot \text{ED}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \right] - \left[ \frac{1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \right] \cdot \text{BIC}^1 \end{aligned} \tag{4.17}$$

For simplicity of notation, on the right side of (4.16) and (4.17) we have dropped the arguments in the brackets.

From equations (4.16) and (4.17), the  $R_{mort}^2$  is a linear transformation of both AIC and BIC. It is noteworthy that the slope of this transformation depends merely on the null model. As consequence, if we would use  $R_{mort}^2$  as a criterion for smoothing parameter selection, i.e. maximizing  $R_{mort}^2$ , we would obtain an optimal value that is different from the one obtained by minimizing the AIC and BIC.

The important point to note here is the presence of the deviance of the null model,  $Dev^0$ , in the second terms of the intercepts in equations (4.16) and (4.17). Especially in a two-dimensional setting and with mortality data,  $Dev^0$  can be substantially higher than any quantity in the numerator, leading to intercept close to 1. Therefore, in presence of larger  $Dev^0$ , the profiles  $R_{mort}^2$  and AIC (or BIC) will be more and more similar over a grid of smoothing parameters, that only differ in sign.

Furthermore, fitted values picked by minimizing AIC will always result in a larger  $R_{mort}^2$  with respect to those picked by minimizing BIC, especially for large mortality surface.

## 5 Simulation studies

In this section we will present different simulation studies which demonstrate the performance of the proposed  $R_{mort}^2$ . Its features will be considered in both unidimensional and two-dimensional context. Simulation settings are chosen such that they resemble mortality data, based on Poisson data and different sample sizes.

### 5.1 The unidimensional case

Though the proposed measure  $R_{mort}^2$  reveals its capability to identify how well a mortality model fits in a two-dimensional setting, in this section we will illustrate the performances of  $R_{mort}^2$  over a single variable only. A univariate  $P$ -spline model is fit based on a smoothing parameters selected by BIC.

The  $R_{mort}^2$  measure is constructed from a given fitted model and a null model. In a unidimensional setting, the latter is represented in equation (4.6). Equation (4.6) is a simple linear predictor where the time axis is the only covariate. As mentioned above, we will fit this null

model just applying a  $P$ -spline approach with sufficiently large smoothing parameter. Consequently  $ED^0$  will be always equal to 2.

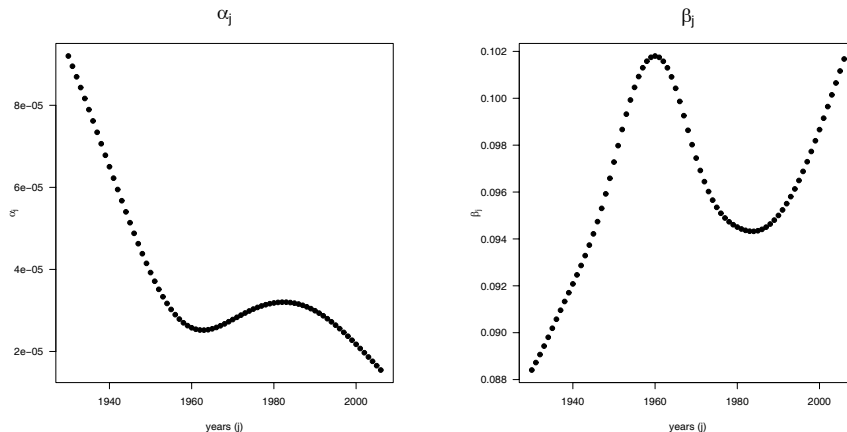


Figure 1: Gompertz parameters,  $\alpha_j$  and  $\beta_j$ , over time  $j$  used in the simulation setting, cf. equations (5.1) and (5.2)

Death counts were simulated from a Poisson distribution with rates following a Gompertz distribution. In particular, we simulated a mortality surfaces from the following setting

$$\begin{aligned}
 Y_{ij} &\sim \text{Poisson}(E_{ij} \cdot \exp(\eta_{ij})) & i &= 30, \dots, 100 \\
 & & j &= 1930, \dots, 2006,
 \end{aligned} \tag{5.1}$$

where the linear predictor varies over time  $j$ :

$$\eta_{ij} = \ln(\alpha_j) + \beta_j \cdot i \tag{5.2}$$

and  $i$  are the ages,  $30, \dots, 100$ . The values of the parameters  $\alpha_j$  and  $\beta_j$  over year  $j$  are shown in Figure 1 and they are chosen to mimic a realistic scenario.

In order to understand the role of the sample size in the outcomes of  $R_{mort}^2$ , we simply modify the values into the matrix of exposures  $E_{ij}$ . Specifically, we designed two different matrices in which  $E_{ij} = 5,000$  and  $E_{ij} = 25,000$  for all  $i = 30, \dots, 100$  and  $j = 1930, \dots, 2006$ . In this way we can generate two different mortality surfaces in which the true model is the same, whereas the variability is different.

We will pick two particular ages only ( $i = 40$  and  $i = 80$ ) and analyse the performance of  $R_{mort}^2$  over time,  $j = 1930, \dots, 2006$ . We have 4 series of death rates (2 ages and 2 exposures) which will be smoothed using  $P$ -spline methodology with 18 cubic  $B$ -spline bases. The proposed  $R_{mort}^2$  is then computed for each fitted model.

We repeated the procedure 1,000 times. Figure 2 shows the outcomes of a single simulation at the given ages  $i = 40$  and  $i = 80$  for the different exposure matrices. Both true, fitted and null models are plotted. Values for  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$  are also presented.

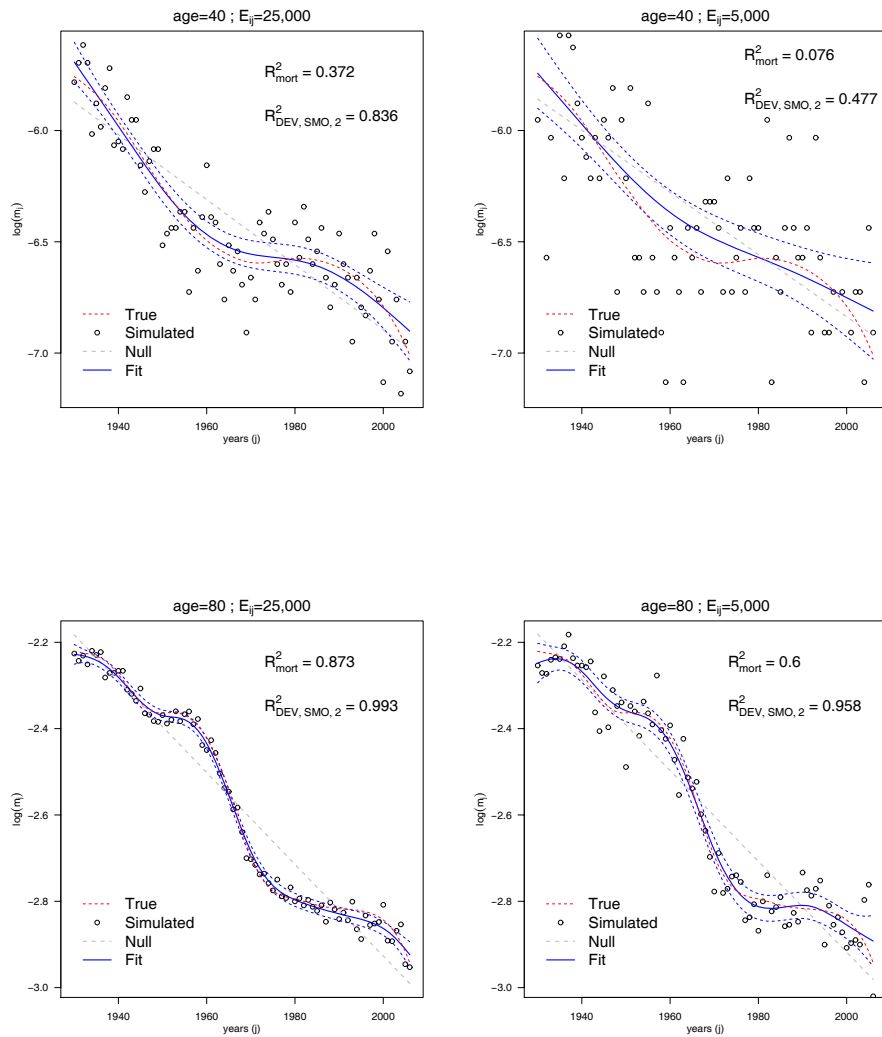


Figure 2: True, simulated and fitted deaths rates (with 95% conf. int.) along with the null model at age 40 and 80 over years  $j = 1930, \dots, 2006$ , logarithmic scale.  $P$ -spline approach is used to fit the data, and BIC for selecting the smoothing parameters.

Table 2 presents the median values for  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$  as well as median values for

deviance, effective dimensions and selected smoothing parameter from the 1,000 simulations by different exposures and ages. An overview of the distributions of these parameters is given in Figure 3.

Simulated data	Median value of:				
	$R_{mort}^2$	$R_{DEV,SMO,2}^2$	$\text{Dev}(\mathbf{y}; \mathbf{a}, \lambda)$	$\text{ED}(\mathbf{a}, \lambda)$	$\lambda$
$i = 40, E_{ij} = 25,000$	0.325	0.822	74.338	4.441	1000.000
$i = 40, E_{ij} = 5,000$	0.061	0.477	77.424	2.778	3162.278
$i = 80, E_{ij} = 25,000$	0.860	0.991	68.655	9.633	630.957
$i = 80, E_{ij} = 5,000$	0.545	0.957	72.280	7.280	630.957

Table 2: Median values of  $R_{mort}^2$ ,  $R_{DEV,SMO,2}^2$ ,  $\text{Dev}(\mathbf{y}; \mathbf{a}, \lambda)$ ,  $\text{ED}(\mathbf{a}, \lambda)$  and  $\lambda$  from the 1,000 simulations at age 40 and 80 over years  $j = 1930, \dots, 2006$ . Different exposure matrices are used, cf. equations (5.1) and (5.2).

As expected,  $R_{mort}^2$  always is smaller than  $R_{DEV,SMO,2}^2$ .  $R_{mort}^2$  measures how much more variation is captured by the model relative to the null linear model. The outcomes of  $R_{DEV,SMO,2}^2$  are close to 1, especially at age 80 due to the large number of deaths at this age (0.991 and 0.956).

Moreover, Figure 3 shows that  $R_{mort}^2$  differs more strongly between the setting than  $R_{DEV,SMO,2}^2$ . In particular results from  $R_{DEV,SMO,2}^2$  at age  $i = 80$  are all very close to 1, i.e. all models capture almost 100% of the variation.

The smoothing parameter at age 40 with  $E_{ij} = 5,000$  is on average considerably larger, leading to fitted values often similar to the linear null model. This might be also due to the fact that the variability in the data is larger for smaller exposure. In these cases the  $R_{mort}^2$  will generally be close to 0, i.e. the fitted model does not capture more variability than the linear null model.

It is a easy to see that the values of  $R_{mort}^2$  are mainly influenced by the different of effective dimensions in the fitted models. On the other hand, the deviance does not show substantial differences between the four settings (see Figure 3, bottom-left panel).

In conclusion, we consider it is often more meaningful and appealing to perceive how much our fitted model improves with respect to a known model, instead that to the overall mean.

## 5.2 The two-dimensional case

In this section we will present results from both  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$  in a simulated two-dimensional setting. As explained in Sections 4.1 and 4.2 both the Lee-Carter model and the

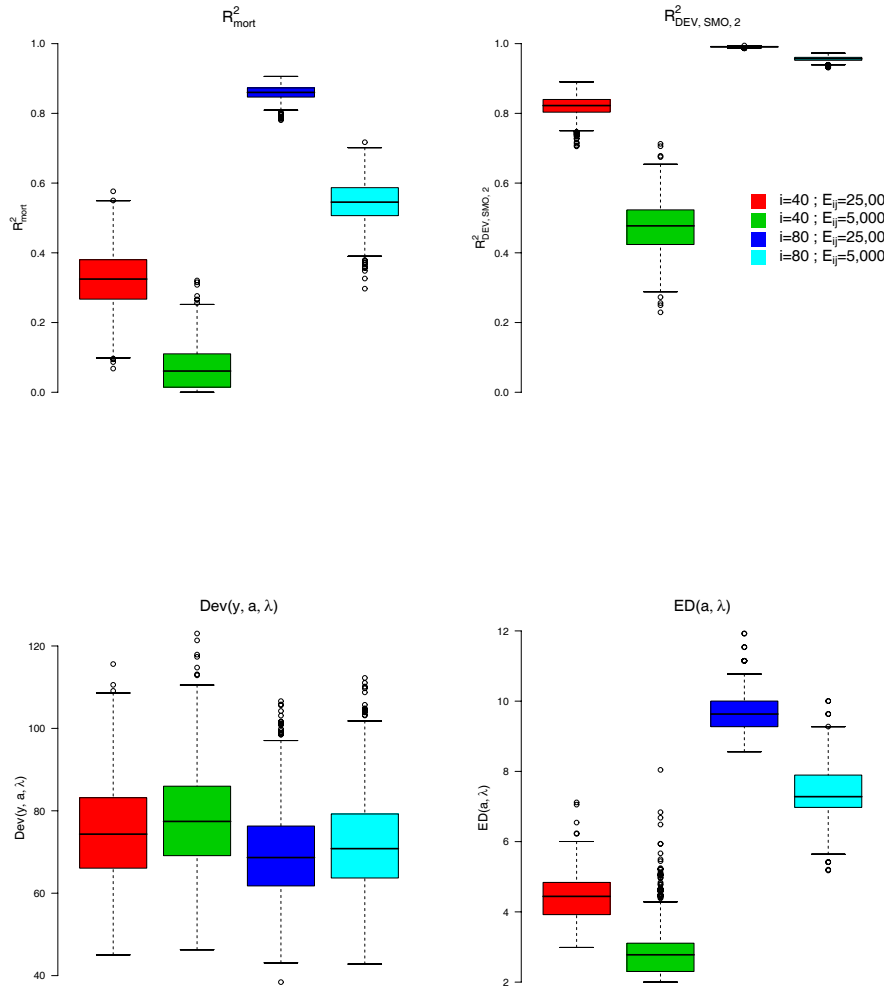


Figure 3: Summary of 1,000 simulations. Box-plots of  $R^2_{mort}$ ,  $R^2_{DEV,SMO,2}$ ,  $Dev(\mathbf{y}; \mathbf{a}, \lambda)$  and  $ED(\mathbf{a}, \lambda)$  for ages  $i = 40$  and  $i = 80$  and different exposure matrices, cf. equations (5.1) and (5.2).

two-dimensional regression  $P$ -splines can be considered as extension of the simple bilinear model over age and time as specified in equation (4.10). Such a bilinear model will be used in this simulation setting as null model for the  $R_{mort}^2$  measure.

Our study in a two-dimensional case employs equations (5.1) and (5.2) for simulating 1,000 mortality surfaces which follow Gompertz distribution with parameters  $\alpha_j$  and  $\beta_j$  varying over time as display in Figure 1. Figure 4 presents an example of such simulation in which the true mortality surface is along with possible simulated surfaces with different exposure matrices.

These mortality surfaces are then fitted by two-dimensional regression  $P$ -splines and the LC model. In particular, we selected the smoothing parameters by BIC in the  $P$ -spline approach and we followed the Poisson likelihood approach given by Brouhns et al. (2002) for fitting the LC model. Finally both  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$  are computed for each of the 1,000 simulations.

Table 3 shows the median values from the 1,000 simulations of both  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$  as well as the median deviances from both the  $P$ -spline and LC approach. Two-dimensional regression  $P$ -splines allows distinct smoothing parameters for each surface and consequently different effective dimensions (Table 3 presents also the median values of the effective dimensions). Note that the LC model employ always  $2 \cdot m + n - 2 = 215$  parameters, this will have a important impact on  $R_{mort}^2$  and  $R_{DEV,SMO,2}^2$ .

$R_{mort}^2$  is substantially higher for the two-dimensional  $P$ -spline approach than for the LC model (0.931 vs. 0.750 and 0.734 vs. 0.591 for the two simulation settings). We might conclude that the LC model performs much worse than two-dimension  $P$ -spline regression on the given simulation setting. This difference is more evident than looking directly at  $R_{DEV,SMO,2}^2$  in which all the values are close to 1.

median values of		Simulation setting	
		$E_{ij} = 25,000$	$E_{ij} = 5,000$
$P$ -splines	$R_{mort}^2$	0.931	0.734
	$R_{DEV,SMO,2}^2$	0.999	0.999
	$\text{Dev}(\mathbf{y}; \mathbf{a}, \lambda)$	5406.800	5443.667
	$\text{ED}(\mathbf{a}, \lambda)$	23.224	21.366
Lee-Carter	$R_{mort}^2$	0.750	0.591
	$R_{DEV,SMO,2}^2$	0.999	0.998
	$\text{Dev}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$	19656.059	8133.536
	$\text{ED}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$	215.000	215.000

Table 3: Median values of  $R_{mort}^2$ ,  $R_{DEV,SMO,2}^2$ , deviance and effective dimensions/parameters from the 1,000 simulations fitted with  $P$ -splines approach and LC. Different exposure matrices are used, cf. equations (5.1) and (5.2).

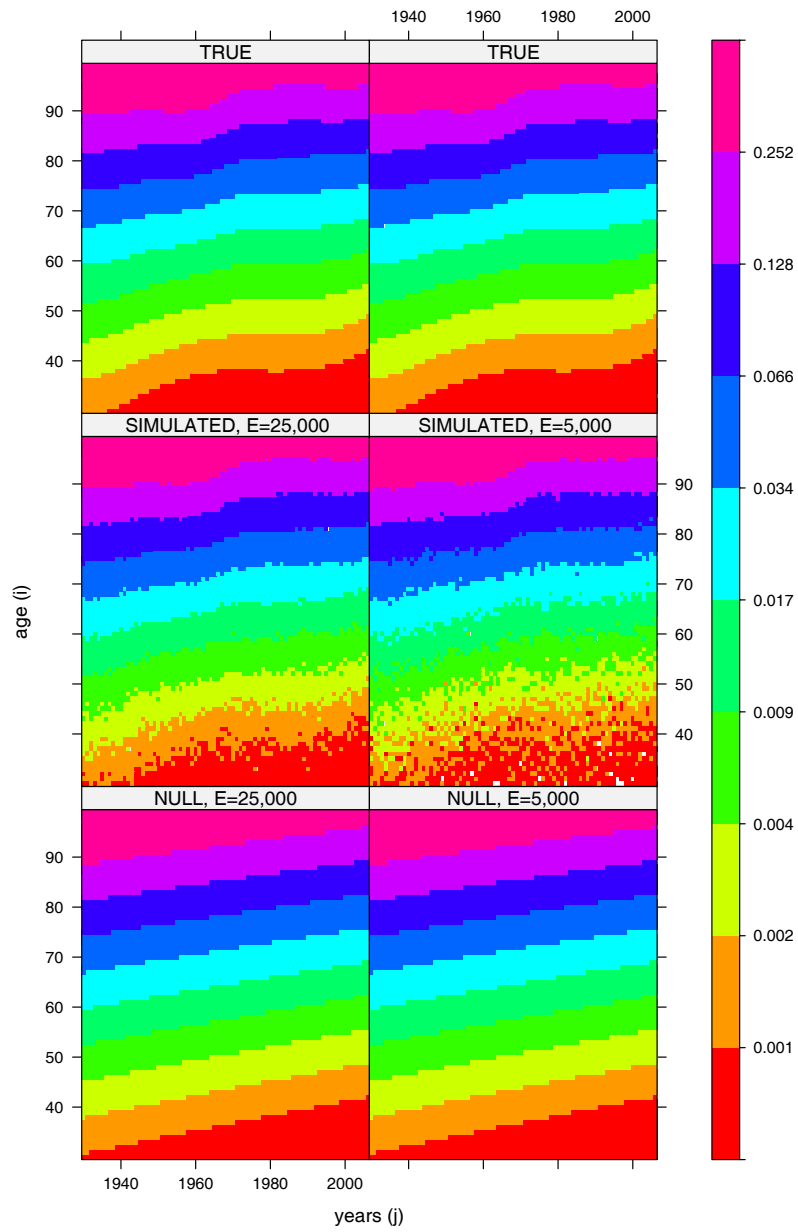


Figure 4: True and simulated death rates over age and years with different exposure matrices. Bilinear model from the simulation setting is also plotted.



Regarding models' comparison, one could check the differences in the deviances between the two approaches (5407 vs. 19656 and 5444 vs. 8134). Again in this case the discrepancy between simulated and fitted values is much larger in the LC model. This is due to the comparatively rigid structure of the LC model. Though the effective dimensions in the LC model is substantially larger as compare to the  $P$ -spline model, the LC model is not capable of capturing variability in the data better.

It is worth pointing out that  $R_{mort}^2$  combines both deviance and effective dimensions of a fitted model in a single value, which reveals straightforwardly the gof of the model. Also in this measure, the sample size plays an important role in the results.

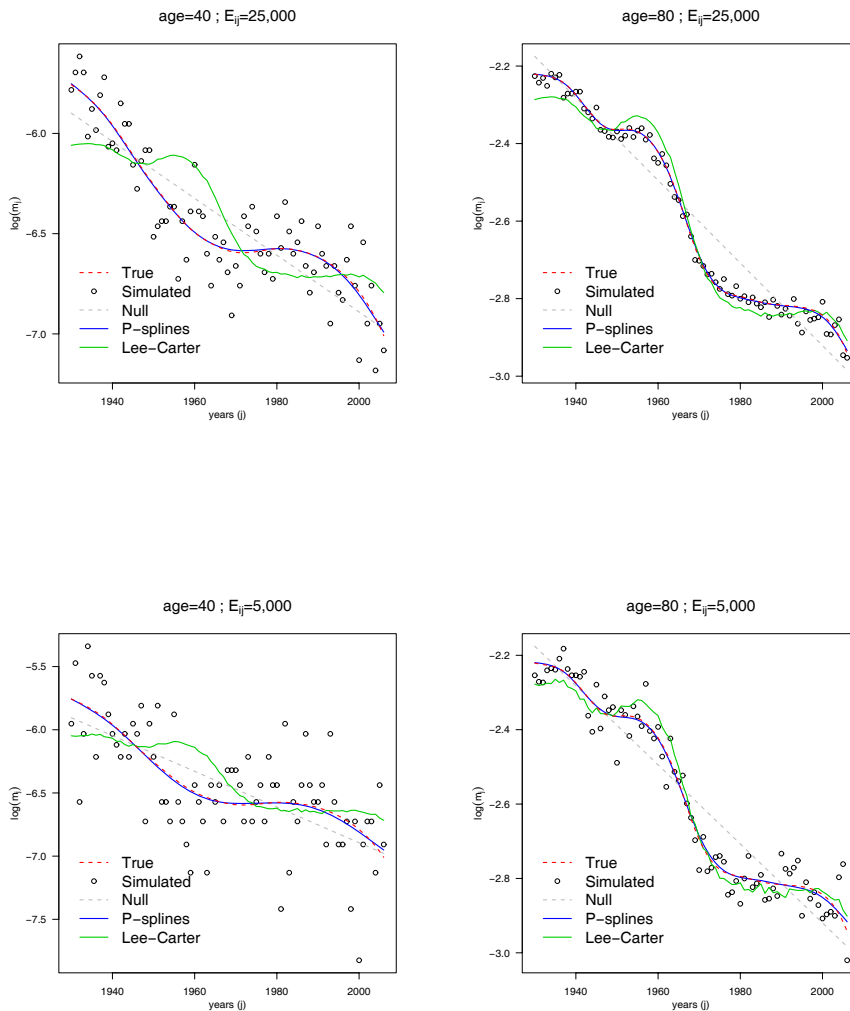


Figure 5: True, simulated and fitted deaths rates with  $P$ -splines and LC model along with the null model, logarithmic scale. Age 40 and 80 over years  $j = 1930, \dots, 2006$ .

The differences between the LC model and the  $P$ -spline approach are even clearer in Figure 5 where two particular ages from the two different simulated mortality surfaces are portrayed. The LC model clearly misfit the data and, additionally, it produces under-smoothed fitted values.

## 6 Applications to the Danish data

In this section we study the performance of  $R_{mort}^2$  for the Danish female population.

It was already clear from the residual analysis, that the  $P$ -spline approach outperformed the LC model, even though the former employs fewer effective dimension. Nevertheless in Table 1 modified  $R^2$ -measures for smoothers did not reveal remarkable differences in gof between these two approaches.

The proposed  $R_{mort}^2$  aims to overcome this issue. Figure 6 shows shaded contour maps of the actual Danish female death rates along with the fitted values and the null model. Figure 7 illustrates, for selected ages, the actual death rates and fitted values from the LC model and  $P$ -spline approach together with the fitted values from the null model.  $P$ -splines follow the mortality development over years more closely than the LC model and the LC model clearly under-smoothed the actual death rates.

$R_{mort}^2$  values are 0.8279138 and 0.7052449 for the  $P$ -spline approach and LC model, respectively. The difference is here more perceptible and informative than with  $R_{DEV,SMO,2}^2$ .

Finally, Table 4 presents different outcomes from (4.14) for the Danish population taking into consideration different period and age ranges. The different values between  $P$ -splines and LC model is clear in all the fitted mortality surfaces from which we conclude that  $P$ -splines give a better fit to these data in all scenarios. In contrast to Table 1, the range of the outcomes in Table 4 is [0.454684,0.827914].

	Danish Data		$P$ -splines	Lee-Carter
females	1930–2006	10-100	0.827914	0.705245
males	1930–2006	10-100	0.822210	0.727210
females	1930–2006	50-100	0.702440	0.486016
males	1930–2006	50-100	0.638110	0.454684
females	1950–2006	50-100	0.720131	0.586349
males	1950–2006	50-100	0.684898	0.542944

Table 4:  $R_{mort}^2$  values for the Danish population by different period and age ranges.

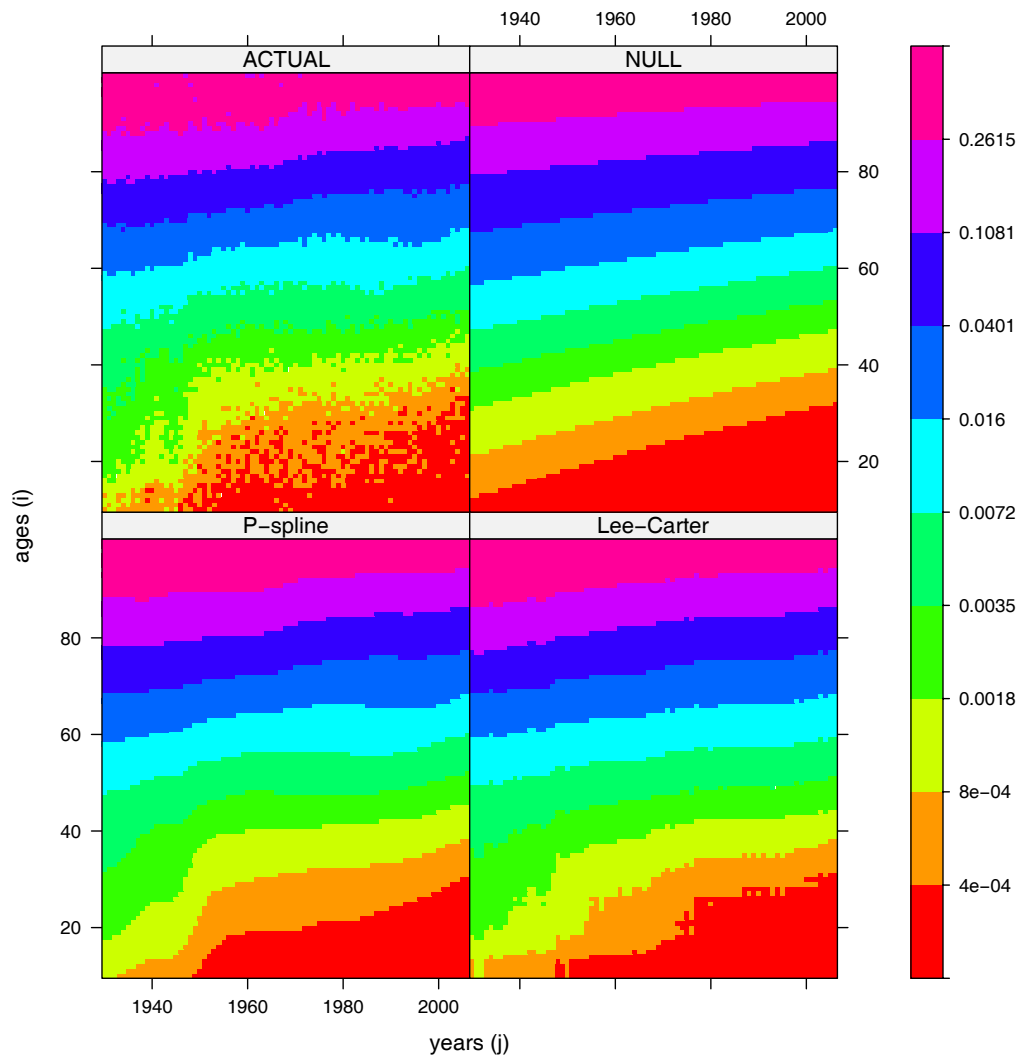


Figure 6: True, simulated and fitted deaths rates with  $P$ -splines and LC model along with the null model. Age 40 and 80 over years  $j = 1930, \dots, 2006$ .

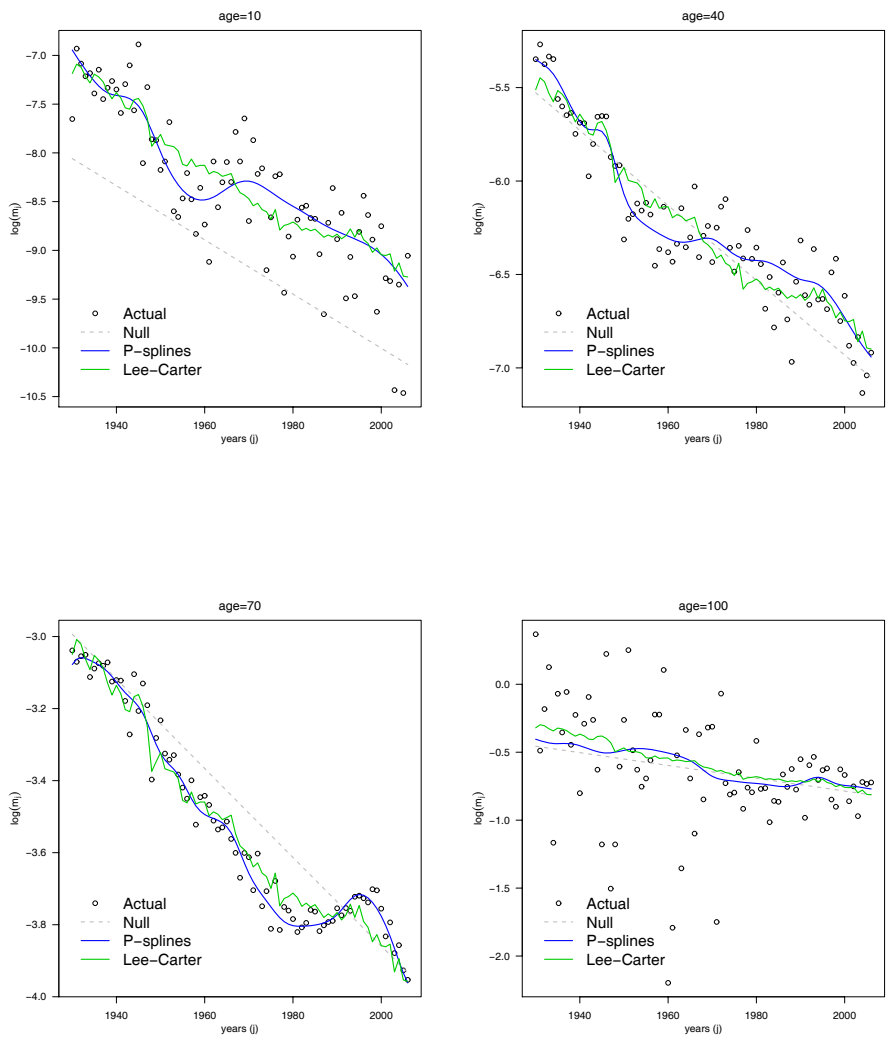


Figure 7: Actual and fitted death rates at selected ages over years, logarithmic scale. 2D smoothing with  $P$ -splines and LC model used for the estimation. Null model given in equation (4.10). Denmark, females.

## 6.1 $R_{mort}^2$ and information criteria

In Section 4.3.1 we presented the relations between  $R_{mort}^2$  and the information criteria for selecting smoothing parameters in a smoothing context, that is, AIC and BIC. For the Danish female population Figure 8 shows the contour plots of AIC, BIC and  $R_{mort}^2$  over the same grid of  $\lambda_a$  and  $\lambda_y$ .

BIC is more convenient to select smoothing parameters in a mortality setting, i.e. death counts are so large that effective dimension of the model needs to be penalized more. Smoothing parameters selected by AIC and  $R_{mort}^2$  (Fig. 8) are smaller than the smoothing parameters picked by BIC.

As explained in equations (4.16) and (4.17), outcomes from  $R_{mort}^2$  are closer to the AIC than the BIC. Consequently,  $R_{mort}^2$  will always be slightly higher for a fitted mortality surface in which  $\lambda_a$  and  $\lambda_y$  are selected by AIC. In this case  $R_{mort}^2$  is equal to 0.835897 when AIC is used, in contrast with 0.827914 of the fitted mortality surface selected by BIC.

## 7 Conclusions

In this paper we proposed a new gof measure for mortality data,  $R_{mort}^2$ , an important tool for comparing models and data in this specific setting. First, we noticed that classic measures are essentially uninformative in the mortality context due to two specific reasons. On one hand, mortality data present often large number of death counts. On the other hand, classic gof measure aim to compare fitted model with a null model which is a simple overall mean of the data. Consequently, commonly used  $R^2$  measures give outcomes always close to zero, regardless of the used model and of the actual data.

We also presented various generalization of the  $R^2$  for normal data which are suitable for non-normal data. Further corrections are also needed to account the number of parameters in the fitted model. Moreover, working with smoothing techniques, effective dimension of the model have to be included in a gof measures. Nevertheless, none of these adjustments is enough to allow informative comparison of explained variability of different models in mortality data.

The proposed measure is based on an alternative null model, specifically designed for mortality data, that is a linear or bilinear model for unidimensional or two-dimensional models, respectively. The equation (4.14) is a particular variant of commonly used gof measures in

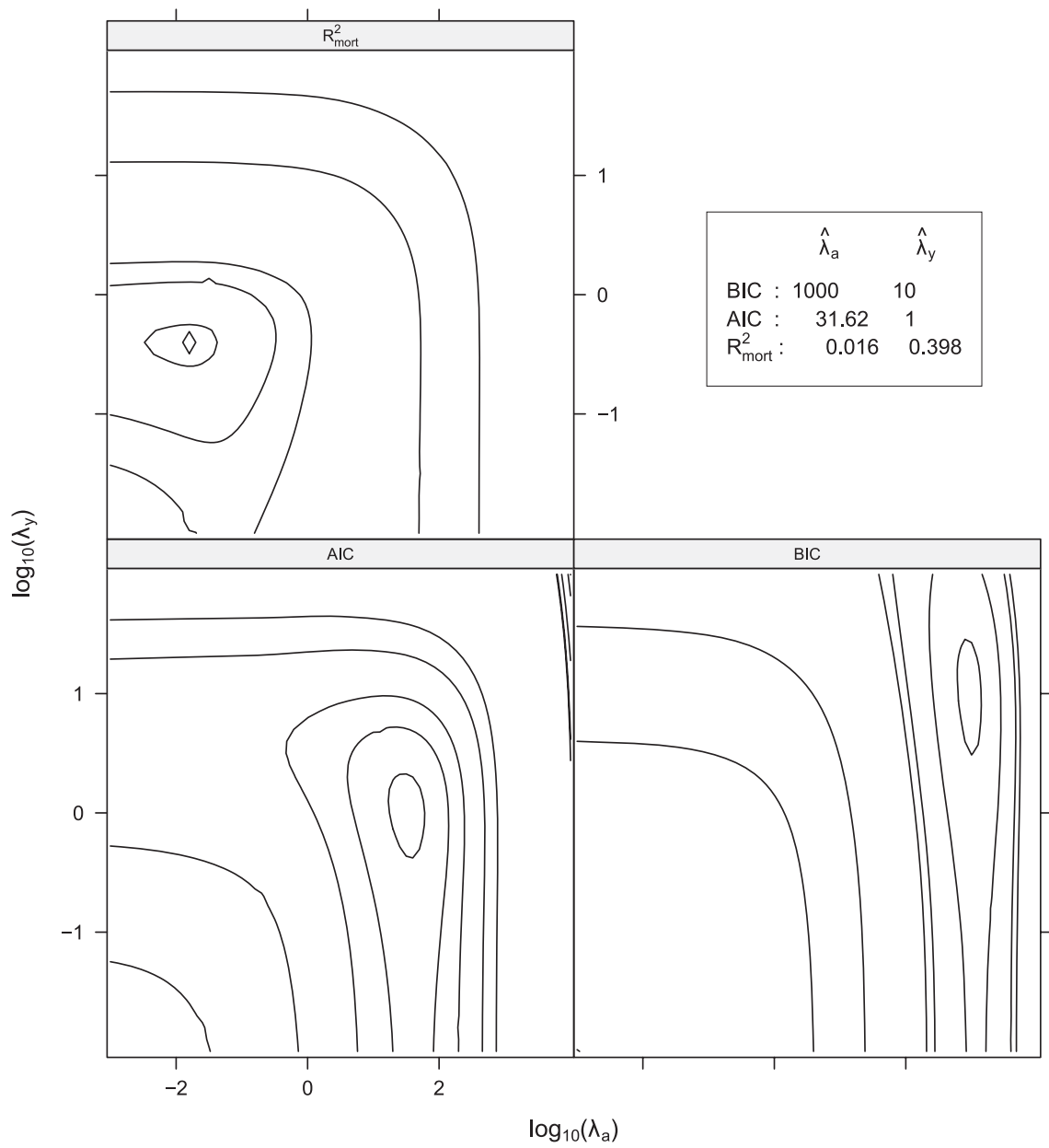


Figure 8: BIC, AIC and  $R^2_{mort}$  over a two-dimensional grid of  $\lambda_a$  and  $\lambda_y$ .

which we incorporate the effective dimension (number of parameters) used by the fitted model and an original denominator which includes deviance and effective dimension of a linear model.

The attractive feature of this new measure lies in the fact that the selected null model is nested in widely used demographic and smoothing models for analysis of mortality development. Specifically, both the Lee-Carter model and the two-dimensional regression  $P$ -splines can be viewed as extensions of the bilinear model over age and time.

Whereas differences in the classic gof measures, even after several adjustments, are hardly perceptible,  $R_{mort}^2$  can be easily used to select and assess models and mortality data. In particular, we showed that, though the Lee-Carter model employs a considerable large number of parameters,  $P$ -spline methodology can capture much better changes in mortality. For instance, for Danish females for the years 1930–2006 and for ages 10–100, the LC model explains 70% more variability present in the actual data than the bilinear model, while two-dimensional  $P$ -splines improves the bilinear null model by 83%. This difference in the outcomes summarizes remarkably well what can be seen in the residual analysis and was not evident in common gof measures.

## References

- Brouhns, N., M. Denuit, and J. K. Vermunt (2002). A Poisson log–bilinear regression approach to the construction of projectedlifetables. *Insurance: Mathematics & Economics* 31, 373–393.
- Buse, A. (1973). Goodness of Fit in Generalized Least Squares Estimation. *The American Statisticians* 27, 106–108.
- Cameron, A. C. and F. A. G. Windmeijer (1996). R-Squared Measures for Count Data Regression Models with Applicationsto Health–Care Utilization. *Journal of Buniness & Economic Statistics* 14, 209–220.
- Cameron, A. C. and F. A. G. Windmeijer (1997). An R-squared measure of goodness of fit for some common nonlinear regressionmodels. *Journal of Econometrics* 77, 329–342.
- Currie, I. D. and M. Durban (2002). Flexible Smoothing with P–splines: a unified approach. *Statistical Modelling* 2, 333–349.
- Currie, I. D., M. Durban, and P. H. C. Eilers (2006). Generalized linear array models with applications to multidimensionalsmoothing. *Journal of the Royal Statistical Society* 68, 259–280.
- Durban, M., I. Currie, and P. H. C. Eilers (2006). Mixed models, array methods and multidimensional density estimation. In J. Hinde, J. Einbeck, and J. Newell (Eds.), *Proceedings of the 21st International Workshop of Statistical Modelling*.
- Good, I. J. (1969). Some Applications of the Singular Decomposition of a Matrix. *Technometrics* 11(4), 823–831.
- Green, P. J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika* 72, 527–537.
- Hastie, T. (1987). A Closer Look at the Deviance. *The American Statisticians* 41, 16–20.
- Keyfitz, N. (1966). Sampling Variance of Standardized Mortality Rates. *Human Biology* 38, 309–317.
- Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.



- McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Model* (Second ed.). Monographs on Statistics Applied Probability. Chapman & Hall.
- Mittlböck, M. and T. Waldhör (2000). Adjustments for  $R^2$ -measures for Poisson regression models. *Computational Statistics & Data Analysis* 34, 461–472.
- Pierce, D. A. and D. W. Schafer (1986). Residuals in Generalized Linear Models. *Journal of the American Statistical Journal* 81(396), 977–986.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.
- Veall, M. R. and K. F. Zimmermann (1996). Pseudo- $R^2$  Measures for Some Common Limited Dependent Variable Models. *Journal of Economic Surveys* 10, 241–59.
- Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham (1999). The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines (with discussion). *Applied Statistics* 48, 269–311.
- Vos, P. W. (1991). A geometric approach to detecting influential cases. *The Annals of Statistics* 19, 1570–1581.
- Waldhör, T., G. Haidinger, and E. Schober (1998). Comparison of  $R^2$  measures for Poisson regression by simulation. *Journal of Epidemiology and Biostatistics* 3(2), 209–215.
- Windmeijer (1995). Goodness-of-Fit Measures in Binary Choice Models. *Econometric Reviews* 14, 101–116.