An Introduction to Generalized Linear Array Models

Currie, Iain
Heriot-Watt University, Department of Actuarial Mathematics & Statistics
Edinburgh EH14 4AS, UK
E-mail: I.D.Currie@hw.ac.uk

1 Motivating example

We use Swedish data downloaded from the Human Mortality Database (HMD) to motivate the need for a high speed, low storage method of smoothing. We suppose that the matrix $\boldsymbol{Y} = (y_{i,j})$, $i = 1, \ldots, n_a$, $j = 1, \ldots, n_y$, contains the number of deaths in the j th year at the i th age. In our example, we consider ages $\boldsymbol{x}_a = (x_{a,1}, \ldots, x_{a,n_a})' = (10, \ldots, 90)'$ and years $\boldsymbol{x}_y = (x_{y,1}, \ldots, x_{y,n_y})' = (1900, \ldots, 2000)'$. The matrix $\boldsymbol{E} = (e_{i,j})$, $n_a \times n_y$, contains the corresponding exposed to risk, ie, the total time lived at the age $x_{a,i}$ in year $x_{y,j}$. Let $\boldsymbol{M} = (m_{i,j}) = (\log(y_{i,j}/e_{i,j}))$, $n_a \times n_y$, be the matrix of the observed forces of mortality (measured on the log scale). The rows and columns of $\boldsymbol{Y}$, $\boldsymbol{E}$ and $\boldsymbol{M}$ are indexed by age $\boldsymbol{x}_a$ and year $\boldsymbol{x}_y$ respectively. The left panel of Figure 1, a plot of the observed mortality surface, suggests the following: the force of mortality (1) has fallen steadily throughout the 20th century, (2) it increases steadily with age, except that (3) it increases then decreases rapidly around the age of twenty (4) it exhibited a very large systematic increase around the end of the first world war and (5) it is subject to random departure from these underlying patterns.

It is natural to suppose that underlying the observed mortality surface in Figure 1 is a smooth mortality surface. We estimate this smooth surface with the P-spline system of Eilers & Marx (1996). Let $\{B_{a,1}, \ldots, B_{a,c_a}\}$ be a B-spline basis of dimension c_a defined along age and let $\boldsymbol{B}_a = \boldsymbol{B}_a(\boldsymbol{x}_a) = (B_{a,j}(x_{a,i}))$, $n_a \times c_a$, be the resulting regression matrix. Similarly, let $\{B_{y,1}, \ldots, B_{y,c_y}\}$ be a B-spline basis of dimension c_y defined along year and let $\boldsymbol{B}_y = \boldsymbol{B}_y(\boldsymbol{x}_y) = (B_{y,j}(x_{y,i}))$, $n_y \times c_y$, be the resulting regression matrix. Then a suitable model matrix for 2-d smoothing is given by the Kronecker product

$$(1) \quad \boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a,\ n_a n_y \times c_a c_y.$$

The right panel of Figure 1 is a simplified plot of the underlying 2-d basis. Each of the $c_a c_y$ regression coefficients is associated with the summit of a 2-d B-spline in the right panel of Figure 1. Thus, it is natural to think of these coefficients as arranged in a $c_a \times c_y$ matrix, say $\boldsymbol{\Theta}$. The P-spline system now consists of choosing a rich basis in both age and year. Smoothness is ensured by penalizing adjacent coefficients in the rows and columns of $\boldsymbol{\Theta}$; the appropriate penalty matrix (Currie *et al.*, 2004) is

$$(2) \quad \boldsymbol{P} = \lambda_a \boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a + \lambda_y \boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a},$$

where $\boldsymbol{I}_n$ is the identity matrix of size n, $\boldsymbol{D}_a$, $(c_a - d_a) \times c_a$, and $\boldsymbol{D}_y$, $(c_y - d_y) \times c_y$, are difference matrices of order d_a and d_y respectively (often we take $d_a = d_y = 2$). The difference matrices $\boldsymbol{D}_a$ and $\boldsymbol{D}_y$ penalize the coefficients in the columns and rows of $\boldsymbol{\Theta}$ and λ_a and λ_y are the smoothing parameters in the age and year directions; notice that the P-spline system allows non-isotropic smoothing. Finally we suppose that the numbers of deaths $y_{i,j}$ are realizations of independent Poisson distributions with means $\mu_{i,j} = e_{i,j}\phi_{i,j}$ where $\phi_{i,j}$ is the force of mortality at age $x_{a,i}$ in year $x_{y,j}$.

The standard approach to fitting this model is to *vectorize the data* and interpret the model as a penalized generalized linear model (PGLM). We have the following *model structure* and *estimation algorithm*

- **Data:** *vectors* $\boldsymbol{y} = \operatorname{vec} \boldsymbol{Y}$, deaths, and $\boldsymbol{e} = \operatorname{vec} \boldsymbol{E}$, exposures

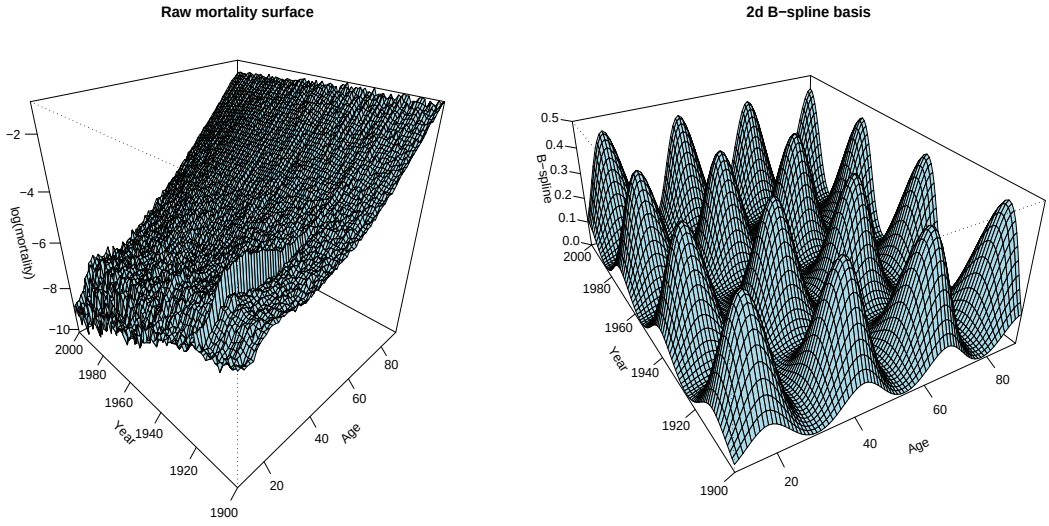

Figure 1: Left: Observed mortality surface for Swedish males. Right: Simplified 2-d B-spline basis.

- **Model:** a *model matrix* $\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a$ of B-splines, a parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\theta} = \text{vec}\,(\boldsymbol{\Theta})$ and a link function

  (3) $\quad \boldsymbol{\mu} = E(\boldsymbol{y}), \quad \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{B}\boldsymbol{\theta}.$

- **Error distribution:** Poisson.

- **Algorithm:** Penalized scoring algorithm

  (4) $\quad (\boldsymbol{B}'\tilde{\boldsymbol{W}}_\delta \boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}_\delta \tilde{\boldsymbol{z}}$

  where $\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{W}}_\delta{}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})$ is the working vector, $\tilde{\boldsymbol{W}}_\delta$ is a diagonal matrix of weights and $\boldsymbol{P}$ is the penalty matrix, (2).

This is a medium size problem: the length of $\boldsymbol{y}$ is $81 \times 101 = 8181$ and a typical model matrix with $c_a = 15$ and $c_y = 20$ is 8181×300. The smoothing parameters, λ_a and λ_y, must now be chosen in this PGLM framework. This is quite possible but we can see that, with a larger problem in higher dimensions where three or more smoothing parameters must be chosen, we may start to run into serious difficulties both in computational time and even in storage.

From an aesthetic point of view there is something unnatural about the above structure. Our data and coefficients are matrices, and our model matrix is a product, yet the above analysis makes no use of these structures. We show in the next section that by adopting an array approach we can solve the storage problem and reduce computational time by up to orders of magnitude.

2 Generalized linear array models or GLAM

The key formula follows from a well known property of Kronecker products

(5) $\quad [\boldsymbol{B}_y \otimes \boldsymbol{B}_a]\boldsymbol{\theta},\ n_a n_y \times 1 \equiv \boldsymbol{B}_a \boldsymbol{\Theta} \boldsymbol{B}_y',\ n_a \times n_y,$

where '$\equiv$' indicates both sides have the same elements, although their dimensions are different. On the left we use the full model matrix, $\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a$ to evaluate $\boldsymbol{B}\boldsymbol{\theta}$; on the right, we operate on the matrix of coefficients first by $\boldsymbol{B}_a$ and then by $\boldsymbol{B}_y$. This solves the storage problem. The number

of multiplications required to evaluate the right hand side is also very substantially smaller than is required on the left hand side. In a large problem this computational saving can be a factor of several orders of magnitude. We now give the *model structure* and *estimation algorithm* for the generalized linear array model or GLAM approach. In 2-dimensions an array is simply a matrix.

- **Data:** *matrices* $\boldsymbol{Y}$, deaths, and $\boldsymbol{E}$, exposures.
- **Model:** a *model matrix* $\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a$ of B-splines, a parameter matrix $\boldsymbol{\Theta}$ and a link function

  $$\log E(\boldsymbol{Y}) = \log \boldsymbol{E} + \boldsymbol{B}_a \boldsymbol{\Theta} \boldsymbol{B}_y'. \tag{6}$$

- **Error distribution:** Poisson.
- **Algorithm:** Penalized scoring algorithm (4) with

  $$\boldsymbol{B}\boldsymbol{\theta},\ n_a n_y \times 1 \ \equiv\ \boldsymbol{B}_a \boldsymbol{\Theta} \boldsymbol{B}_y',\ n_a \times n_y, \tag{7}$$
  $$\boldsymbol{B}'\boldsymbol{W}_\delta \boldsymbol{B},\ c_a c_y \times c_a c_y \ \equiv\ G(\boldsymbol{B}_a)' \boldsymbol{W} G(\boldsymbol{B}_y),\ c_a^2 \times c_y^2, \tag{8}$$

  where $\boldsymbol{W}$, $n_a \times n_y$, is the matrix of weights, ie, $\operatorname{vec} \boldsymbol{W} = \operatorname{diag} \boldsymbol{W}_\delta$. In (8), the function $G(\cdot)$ is the *row-tensor* function defined for any matrix $\boldsymbol{X}$, $n \times c$, as

  $$G(\boldsymbol{X}) = \left[\boldsymbol{X} \otimes \mathbf{1}_c'\right] * \left[\mathbf{1}_c' \otimes \boldsymbol{X}\right],\ n \times c^2, \tag{9}$$

where $\mathbf{1}_c$ is a vector of 1's of length c. We make two comments on the matrix form of this algorithm. First, (8) achieves the same sequential computation for the weighted inner product $\boldsymbol{B}'\boldsymbol{W}_\delta\boldsymbol{B}$ in the scoring algorithm (4) as (6) achieves for the linear predictor. Second, the matrix forms on the right-hand sides of (7) and (8) need to be rearranged into the corresponding vector forms on their respective left-hand sides. Details of how this is achieved are given in Currie *et al,* (2006); here, we simply remark that such rearrangements are very efficient. In summary, GLAM

- is conceptually attractive (it takes advantage of both the data and the model structure),
- has a low footprint (a result of the sequential nature of the algorithm),
- is very fast, and
- generalizes to d-dimensions.

The sequential nature of the GLAM algorithm was first described in Eilers *et al.* (2006). The array nature of these algorithms was described in Currie *et al.* (2006); the acronym GLAM was coined in this second paper.

3 Examples of GLAMs

In the remainder of this paper we describe a number of applications of GLAM. We will describe the data, give the linear predictor in GLAM form and report some results.

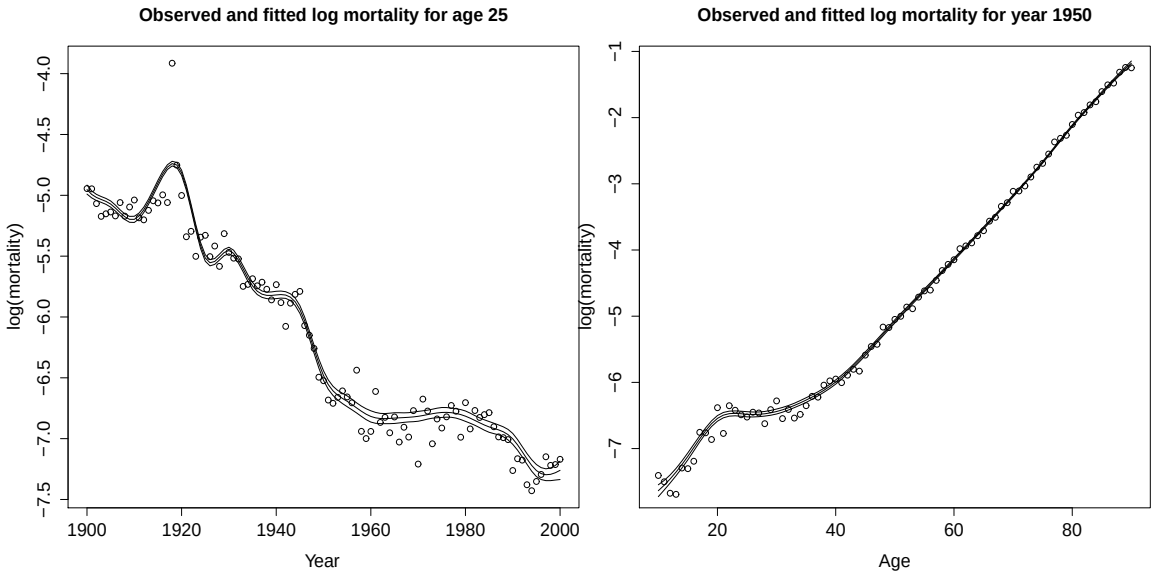

Figure 2: Smoothed log mortality for Swedish males

3.1 Modelling mortality

We continue with our introductory example on smoothing Swedish mortality data. We ignore for the moment the one-off behaviour at the end of the first world war and fit a simple smooth surface $\boldsymbol{B}_a\boldsymbol{\Theta}\boldsymbol{B}_y'$. Smoothing parameters are chosen by minimizing the Bayesian Information Criterion (BIC). Figure 2 gives two cross-sections of the fitted surface. The fitted surface has successfully picked up the feature around age 20 (often described as the 'accident hump') but has failed to catch the feature in 1918. We will return to this point in section 3.3.

There is an R-package, MortalitySmooth, (Camarda, 2009) for fitting a 2-d mortality surface; overdispersion of the Poisson counts, a common feature of mortality data, is allowed for. The coding uses the GLAM algorithms.

3.2 Joint modelling of mortality by lives and amounts

Mortality data in life insurance has some features of its own. Deaths are the number of claims on policies and exposure is the total time at risk; this is usually referred to as *data by lives*. Alternatively, deaths are the total amount claimed and exposure is the total amount at risk; this is usually referred to as *data by amounts*. Thus we have deaths and exposure matrices $\boldsymbol{D}_l$ and $\boldsymbol{E}_l$ on lives, and $\boldsymbol{D}_a$ and $\boldsymbol{E}_a$ on amounts. The force of mortality can then be computed by lives (as in the previous example) or by amounts. It is generally found that mortality by lives is heavier than mortality by amounts since those with better mortality have larger insured amounts; this observation is borne out in Figure 3.

Actuaries are interested in forecasting mortality for the purpose of pricing and reserving of annuities and pensions. The penalty function allows forecasting (see Currie *et al.* (2004) for details of how this is done). Here, with two measures of mortality we must take care that any such forecasts are consistent, ie, do not cross over in the future. We achieve this joint forecasting by taking $\boldsymbol{B}_a\boldsymbol{\Theta}\boldsymbol{B}_y'$ as the mortality surface by lives and $\boldsymbol{B}_a\boldsymbol{\Theta}\boldsymbol{B}_y' + \boldsymbol{B}_a\breve{\boldsymbol{\Theta}}\mathbf{1}_{n_y}'$ as the mortality surface by amounts. This is an additive GLAM with the property that the surface by amounts differs from the surface by lives by a constant amount in time; this constant is a smooth function of age. There are three smoothing parameters to be chosen (two for the component $\boldsymbol{B}_a\boldsymbol{\Theta}\boldsymbol{B}_y'$ and one for the component $\boldsymbol{B}_a\breve{\boldsymbol{\Theta}}\mathbf{1}_{n_y}'$. Figure 3 shows the results of fitting this model. See Currie *et al.* (2004) and Djeundje and Currie (2011) for further details.

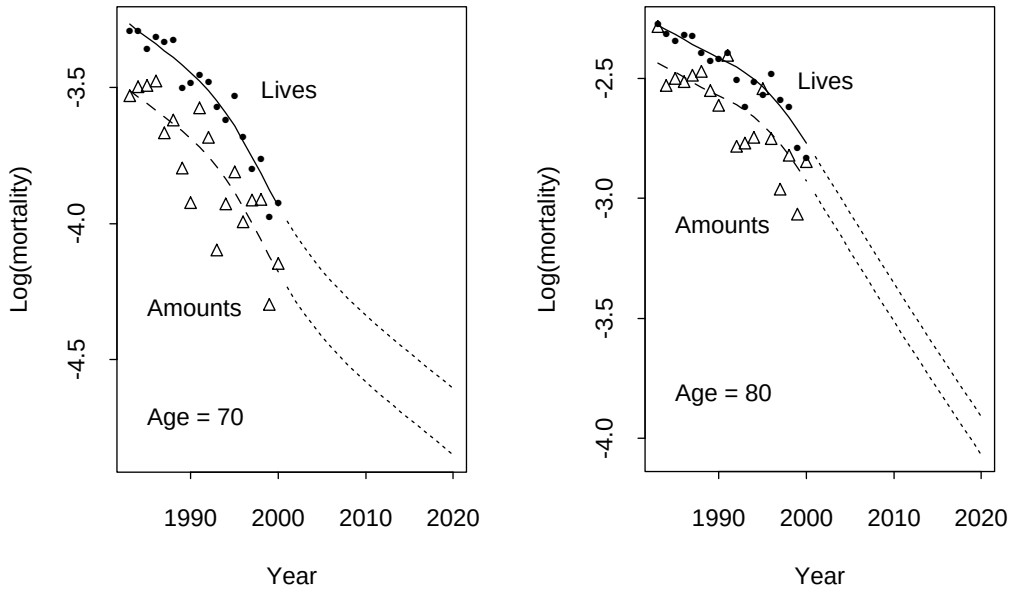

Figure 3: Smoothed and forecast log mortality by lives and amounts

3.3 Modelling mortality with period shocks

We return to the Swedish mortality data and the problem with the one-off feature in 1918. Rather than use a model with the observed feature in 1918 built-in we propose a more general model in which each year can potentially have such a feature. We refer to such features as *period shocks* or simply as shocks. We assume that any period shocks are smooth functions of age; we leave the modelling process to identify them. Kirkby and Currie (2010) proposed an additive GLAM

$$(10) \quad \boldsymbol{B}_a \boldsymbol{\Theta} \boldsymbol{B}_y' + \breve{\boldsymbol{B}}_a \breve{\boldsymbol{\Theta}}$$

which splits the mortality surface into a smooth 2-d surface and (possibly) a number of smooth shocks. This is a large computationally demanding model since each of n_y years has $\breve{c}_a$ coefficients. Thus, the three smoothing parameters must be chosen in the context of a very large model matrix; in Kirkby and Currie (2010) the model matrix was 8181×1346.

Figure 4 shows the two components of the additive model (10). The smooth surface in the left panel of Figure 4 is much smoother than that fitted in section 3.1 since the one-off behaviour in 1918 is now modelled by a shock. The shock to the mortality surface in 1918 was caused by the Spanish Influenza pandemic which predominately affected the young. The 1918 shock and other smaller shocks can be seen in the right panel of Figure 4.

3.4 2-d density estimation

GLAM can also be applied in some situations when the data do not obviously lie on a grid, and 2-d density estimation is a good example. We consider the classic 2-d data set on the waiting times between and durations of eruptions of the Old Faithful geyser. The left panel of Figure 5 shows the data from which it is evident that data falls into two main regions. We have 272 data points scattered over a rectangular region. We form a fine 2-d grid of counts. In the example we used waiting time by

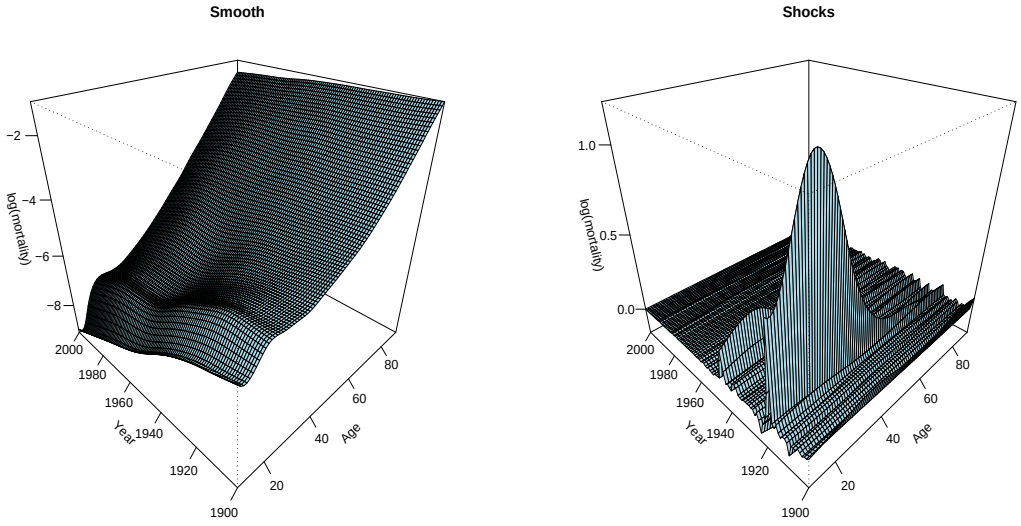

Figure 4: Components of the shock model. Left: the smooth 2-d surface; right: the annual shocks

duration bins of size 1 minute by 1 second which gives $60 \times 217 = 13020$ bins. Our data then consists of the frequencies of observation in cells: we have 238 counts of 1, 17 of 2, and 12765 (98%) of 0. We apply 2-d P-spline smoothing with Poisson errors, log link, model matrix $\boldsymbol{B}_w(\boldsymbol{x}_w) \otimes \boldsymbol{B}_d(\boldsymbol{x}_d)$ where $\boldsymbol{x}_w$ and $\boldsymbol{x}_d$ are the mid-points of the waiting time and duration bins respectively. Figure 5 shows the fitted contour surface and fitted density. Further details on density estimation with GLAM can be found in Eilers and Marx (2006). Durban *et al.* (2006) combine GLAM and mixed models to estimate multi-dimensional densities while Lambert and Eilers (2006) show how GLAM and Bayesian methods can also be used here.

3.5 Other applications

One important area where GLAM can be used with advantage is in spatio-temporal smoothing. Lee and Durban (2011) give an example of smoothing ozone measurements. Data are located at scattered locations but are measured at monthly intervals. This gives this 3-dimensional problem (two space dimensions and one time dimension) a 2-dimensional GLAM structure where one GLAM dimension is space and the other GLAM dimension is time. For further details see Lee and Durban (2011).

We have emphasized the computational advantage of GLAM in this introductory paper. There is another important aspect: if data have an array structure and models of interest have a row and column structure then GLAM is the correct way to think about modelling. GLAM is more than a computational device, it is a structure for modelling.

REFERENCES

Camarda, C. G. (2009). MortalitySmooth: Smoothing Poisson counts with P-splines. R-package: 1.0.

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

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

Djeundje, V. A. B. and Currie, I. D. (2011). Smoothing dispersed counts with applications to mortality data. *Annals of Actuarial Science*, **5**, 33–52.

Durban, M., Currie, I. D. and Eilers, P. H. C. (2006). Mixed models, array methods and multidimensional density estimation. *Proceedings of 21st International Workshop on Statistical Modelling*, Galway, 143-150.

Eilers, P. H. C., Currie, I. D. and Durban, M. (2006). Fast and compact smoothing on large multidimen-

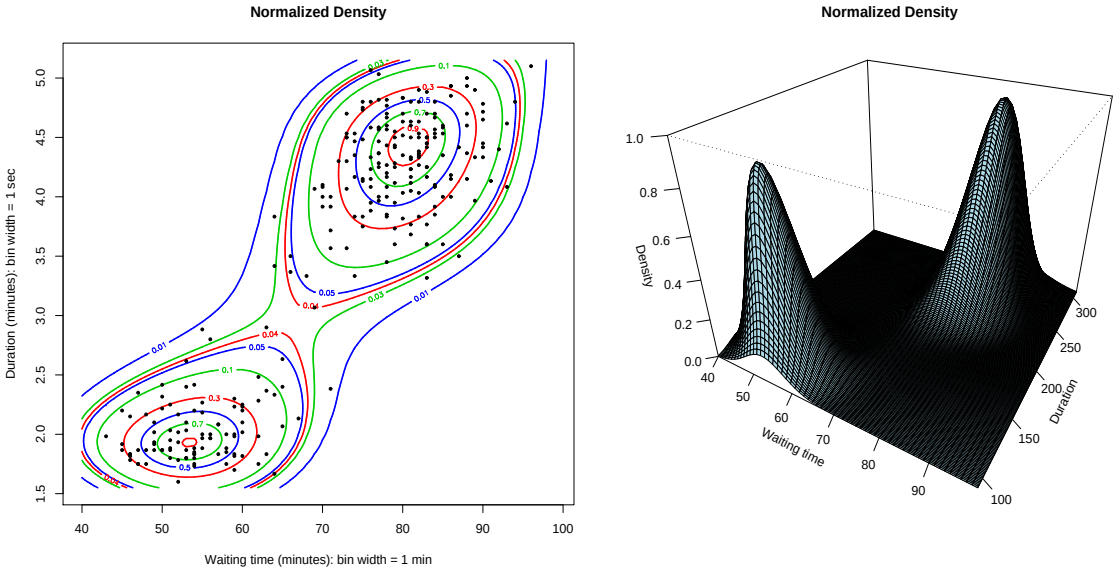

Figure 5: Data and estimate of 2-d density of waiting and durations times of eruptions of the Old Faithful geyser

sional grids. *Computational Statistics and Data Analysis*, **50**, 61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Eilers, P. H. C. and Marx, B. D. (2006). Multidimensional density smoothing with P-splines. *Proceedings of 21st International Workshop on Statistical Modelling*, Galway, 151-158.

Kirkby, J. G. and Currie, I. D. (2010). Smooth models of mortality with period shocks. *Statistical Modelling*, **10**, 177–196.

Lambert, P. and Eilers, P. H. C. (2006). Bayesian multi-dimensional density estimation with P-splines. *Proceedings of 21st International Workshop on Statistical Modelling*, Galway, 313-320.

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

ABSTRACT

Data with an array structure are common in statistics (mortality tables and spatio-temporal data are two important examples). Such data often require smoothing to remove noise and estimate trends. One natural and attractive approach is to use penalized regression where (a) the basis for the regression is a Kronecker product of B-splines and (b) the penalty is a roughness penalty on regression coefficients; this is the P-spline approach of Eilers & Marx. However, such an approach is particularly susceptible to runaway problems with (a) storage and (b) computational time. Generalized linear array models (GLAM) were developed precisely to address both these issues. In a conventional GLM you store the model matrix and then fit the model. Unfortunately, with large amounts of data this model matrix can get rather large: computation and even storage can be a problem. In GLAM the model matrix is not stored; the GLAM algorithm works sequentially with the factors of the Kronecker product. Further, the GLAM algorithm is very fast and can be orders of magnitude quicker than the usual GLM approach in a large problem. In this paper we first describe the GLAM algorithms and then give an introduction to a range of applications. These applications include various models for smoothing and forecasting of mortality tables, density estimation and spatio-temporal smoothing.