Using P-splines to smooth two-dimensional Poisson data

Maria Durbán¹, Iain Currie² and Paul Eilers³

¹ Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Madrid, Spain
² Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
Email: I.D.Currie@ma.hw.ac.uk
³ Department of Medical Statistics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

Abstract: Eilers & Marx (1996) used P-splines to smooth one-dimensional count data with Poisson errors. We extend their methods to two dimensions. We consider a variety of models and highlight some of the computational problems that arise with large data sets. We illustrate our remarks with the analysis of a large set of mortality data indexed by age of death and year of death.

Keywords: Mortality; P-splines; Poisson data; smoothing; two-dimensions.

1 Introduction

Eilers and Marx (1996) introduced P-splines as a method of smoothing in generalized linear models (GLMs). The method has two main ideas: (a) use B-splines as the basis for the regression, and (b) modify the log-likelihood by a difference penalty on the regression coefficients. The method has many attractive properties amongst which we mention only its flexibility, its ease of computation and its connection to smoothing splines and polynomial regression.

Eilers and Marx (1996) gave a number of examples including that of smoothing count data with a Poisson error. In this paper we show how their method can be extended to two-dimensional Poisson data. We discuss a number of models and pay particular attention to the computational problems that may be encountered with large data sets. We illustrate our remarks with the analysis of a large set of mortality data indexed by year of death (1947 to 1999) and age of death (11 to 100).
2 Description of the data

The Continuous Mortality Investigation Bureau (CMIB) collates information on claims from the UK life insurance business. In this paper we consider a data set for male policyholders. For each calendar year (1947 to 1999) and each age (11 to 100) we have the number of years lived (the exposure) and the number of policy claims (deaths). The mortality of male policyholders has improved rapidly over the last thirty years (as indeed has that of the general population) and this has important financial implications for the insurance industry. Thus we will be particularly interested in mortality trends over time and the extent to which these trends depend on age.

3 Some models for two-dimensional Poisson data

We set up our notation by describing the one dimensional case. Let \( E_x \) be the number of lives aged \( x \), \( y_x \) be the corresponding number of deaths and suppose that \( Y_x \) has a Poisson distribution with mean \( \mu_x = E_x \theta_x \); we write this as \( Y_x \sim \text{Poisson}(E_x \theta_x) \). Let \( x' = (x_1, \ldots, x_m) \) with similar definitions for \( y, \mu \), etc. We assume that the distribution of the \( Y_x \) follows a GLM with \( \log \mu = B \alpha \) (canonical link) where \( B = (B_1(x), \ldots, B_k(x)) \) is an \( n \times k \) matrix of \( B \)-splines (\( k \) depends on the number of knots and the degree of the \( B \)-spline). We choose \( \alpha \) by maximising the penalized log-likelihood \( \ell(\alpha; y) - \frac{1}{2} \lambda \alpha'D'D\alpha \) where \( D \) is a difference matrix. With Poisson errors the usual GLM iterative scheme for solving for the estimate of \( \alpha \) is

\[
\hat{\alpha}_{t+1} = \left( B'W_tB + \lambda D'D \right)^{-1}B'W_tz_t
\]

(1)

where \( z = \eta + W^{-1}(y - \mu) \) is the working variable and \( W = \text{diag}(\mu) \) is the diagonal matrix of weights. We now look at various models for two-dimensional data.

3.1 A generalized additive model

Let \( Y = (y_{ij}) \) be the matrix of deaths at age \( i, i = 1, \ldots, m \) and year \( j, j = 1, \ldots, n \). We arrange the matrix of deaths by column order into a vector \( y \) (this corresponds to how Splus stores a matrix) of length \( N = m \times n \). Similarly we arrange the matrices of exposures \( E = (E_{ij}) \), mortalities \( \Theta = (\theta_{ij}) \), ages, years, etc into vectors \( e, \theta \), etc. Let \( B_\alpha, N \times n_\alpha, \) be the set of \( B \)-splines for age and \( B_y, N \times n_y, \) be the set of \( B \)-splines for years. We write \( \log \theta = B \alpha \) where \( \alpha' = (\alpha_1, \alpha_2) \), \( B = (1 : B_\alpha : B_y) \) and \( 1 \) is a vector of ones. The model is a generalized additive model (GAM) and fits a single smooth effect for age and a single smooth effect for year. The fitted log mortalities against year for different ages are parallel (the same is true for the log mortalities against age for different years).
In smoothing the usual method of fitting an additive model is by back-fitting. However, Marx & Eilers (1998) showed that the above model can be fitted directly (this is a consequence of the essentially regression approach of $P$-splines). Equation (1) becomes

$$a_{t+1} = (B'W_tB + P)^{-1}B'W_tz_t$$

where $P = \text{blockdiag}(0, P_a, P_y)$; $P_a = \lambda_a D_a' D_a$ and $P_y = \lambda_y D_y' D_y$ are the penalty matrices for age and year respectively.

We make two comments on the computational aspects of fitting this model. First, since each of the columns of both $B_a$ and $B_y$ sum to one the system (2) is singular. We can get round this difficulty by the use of a ridge penalty; in our example a penalty of $10^{-4}$ worked well. Second, and more importantly, this is a large data set with $90 \times 53 = 4770 = N$ observations. Usually it is not possible to consider smoothing models, e.g., smoothing spline or kernel smoothing models, for such a large value of $N$. With $P$-splines this problem often disappears; in the present case we found models with the number of columns in $B$ around 40 satisfactory.

We considered two methods of smoothing parameter selection: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Both these criteria have the form

$$\text{dev}(y; a, \lambda_a, \lambda_y) + \delta \text{tr}(H)$$

where $\text{tr}(H)$ is the trace of the hat-matrix $H$: for AIC we have $\delta = 2$ while for BIC we have $\delta = \log(N)$. There is evidence that use of AIC can result in undersmoothing (Hurvich, et al. (1998), Shibata (1976)); in contrast, use of BIC introduces a much heavier penalty in our example since $N = 4770$. With cubic $B$-splines ($\text{bdeg} = 3$), second order penalties ($\text{pord} = 2$) and number of internal knots 20 for age and 10 for years ($\text{ndx}_a = 20$ and $\text{ndx}_y = 10$) use of BIC gave a model with 30.6 d.f. while AIC gave one with 33.6 d.f. The fit of the model selected by BIC is illustrated in Fig. 1 where the parallel nature of the model is evident; the fit of the model selected by AIC is omitted from Fig. 1 since it is scarcely distinguishable from the plot for BIC. The fit at age sixty does not seem satisfactory and plots at ages around sixty show the same pattern: underestimating the mortality in the early years and overestimating it in later years.

### 3.2 Two-dimensional smoothing with penalties

In this section we suppose that the matrix $A = \log \Theta$ of log mortalities is an $m \times n$ matrix of parameters. We smooth the entries in $A$ by imposing penalties on its rows and columns. This conceptually simple method works well in our example but has the disadvantage that we must work with a very
large parameter set. There is no explicit mention of $P$-splines. However we will see in 3.3 that the method is a special case of $P$-spline smoothing. We consider general $P$-spline smoothing in 3.3 when we will take full advantage of the dimension reducing properties of $P$-splines.

We write the columns and rows of the $m \times n$ matrix of mortalities as

$$\log \Theta = A = (a_1, \ldots, a_n), \quad A' = (a_1', \ldots, a_m')$$

and impose a smoothness condition on each row and column of $A$. We now choose $A$ by maximizing the penalized log-likelihood

$$\ell(A; Y) - \frac{1}{2} \lambda_a \sum_{j=1}^{n} a_j' D_a a_j - \frac{1}{2} \lambda_y \sum_{i=1}^{m} a_i' D_y a_i$$

where $D_a$ and $D_y$ are the difference matrices on columns (age) and rows (years) respectively. In column form (5) becomes

$$\ell(a; y) - \frac{1}{2} \lambda_a a'(I_n \otimes D_a) a - \frac{1}{2} \lambda_y a'(D_y' D_y \otimes I_m) a$$

where $a = (a_1', \ldots, a_n')$, $P_a = I_n \otimes D_a'$, $P_y = D_y' D_y$ and $\otimes$ denotes the Kronecker or tensor product of two matrices. We will show in 3.3 that this model is a special case of a $P$-spline model.

In theory, we can now use (1) to choose $a$. This will be possible in moderate sized problems but in the present case the parameter vector $a$ has length $mn = 4770$ and this requires that we handle $4770 \times 4770$ matrices; this will be too large for most machines. The following iterative scheme is a practical alternative. Differentiating (5) we obtain the likelihood equations:

$$Y - M = \lambda_a D_a' D_a A + \lambda_y A D_y' D_y$$

FIGURE 1. Log hazard against year for ages 34 and 60: model in 3.1
where $M = (\mu_1, \ldots, \mu_n) = E(Y)$. We solve this set of equations by iterating between rows and columns using equation (16) in Eilers & Marx (1996) adjusted for the column and row penalties in turn. For example, to update the column estimates we define $U = (u_1, \ldots, u_n)$ by

$$U = Y - M + M \star A - \lambda_y AD y' D y.$$  \hspace{1cm} (9)

where $A$ and $M$ are current estimates, and $\star$ indicates element by element multiplication. The updated estimate of $a_j$, $j = 1, \ldots, n$, is

$$a_j = (\text{diag}(\mu_j) + \lambda_a D a' D a)^{-1} u_j.$$  \hspace{1cm} (10)

The algorithm copes with the potential computational problems associated with two-dimensional smoothing, at least with this large data set. We found that convergence was reached after about 50 iterations in about ten minutes on a PC running at 500 MHz. The computation of the dimension of the fitted model via the trace of the hat-matrix remains a problem and we were unable to select good values of the penalties via AIC or BIC. Instead we used plots like Fig. 2 as a guide. Figure 2 is very different from Fig. 1 and shows that the pattern in mortality at younger ages is different from those at older ages. The fit at age sixty is good with the present model, in contrast to the unsatisfactory fit at age sixty in Fig. 1; in particular, the mortality at age sixty improves more rapidly in the later years with the present model than with the model in 3.1.
3.3 Dimension reduction using $P$-splines

Let $B_a, m \times n_a$, be the one-dimensional $B$-spline basis for smoothing by age for a single year; similarly, let $B_y, n \times n_y$, be the one-dimensional $B$-spline basis for smoothing by year for a single age. We take advantage of the fact that the data are defined over a regular grid and define a regression model with regressor matrix $B$ given by the tensor product

$$B = B_y \otimes B_a.$$  \hspace{1cm} (11)

Thus, we assume that $\log \theta = Ba$ (where, as before, we assume the data matrix is arranged in column order). We note that the dimension of $B$ is $mn \times n_a n_y$ and so the length of $a$ is $n_a n_y$. Corresponding to (4) we write $a$ in matrix form as

$$A = (a_1, \ldots, a_{n_y}), \quad A' = (a_1', \ldots, a_n'').$$  \hspace{1cm} (12)

Expressions (5), (6) and (7) now hold but with $n \rightarrow n_y$ and $m \rightarrow n_a$. In particular, the penalty matrix $P = \lambda A P_a + \lambda_y P_y$ where $P_a = I_{n_y} \otimes D_a D_a$ and $P_y = D_y' D_y \otimes I_{n_a}$.

We make two comments on this approach. First, if we take the degree of the $B$-splines to be zero, i.e., $bdeg = 0$, and the number of knots to be $m$ for age and $n$ for years, i.e., $ndz_a = m$ and $ndz_y = n$ then $B = I_{mn}$. Thus the model in 3.2 is a special case of the present model. Second, and more importantly, suppose that we use cubic $B$-splines, i.e., $bdeg_a = bdeg_y = 3$, and that we follow the recommendation in Eilers and Marx (1996) that we use one knot for every four or five ages and one for every four or five years. In the present case we might take $ndz_a = 20$ and $ndz_y = 10$; this results in $n_a = 23$ and $n_y = 13$ and the parameter vector has length 299. This is an example of the dimension reduction that can be achieved with $P$-splines; the model can now be fitted directly with (1).

There are some computational issues in fitting models with large numbers of parameters. In the previous section the matrix $B'WB$ had dimension $4770 \times 4770$ and there was little prospect of storing this matrix. In this section the largest models we considered had 299 parameters so $B'WB$ can easily be stored. However, $B$ can be as large as $4770 \times 299$ and this alone gives storage problems. We wrote $B = [B_1, B_2, B_3]$ and assembled $B'WB$ from this partition. This had the further advantage that we could take advantage of the banded nature of $B$ to speed up the computations.

With 299 parameters the model was fitted in about 5 minutes.

Table 1 gives minimum values of BIC and AIC for various values of $ndz_a$ and $ndz_y$. With BIC the selected model has $ndz_a = ndz_y = 10$ and 169 free parameters. We note the large difference in the effective dimensions of the fitted models selected by BIC and AIC: BIC: 64 with BIC but 124 with AIC. Figure 3 is a plot of the fitted log hazards for ages 34 and 60. There is very little to choose between the two fits at age 60 but we much prefer the stiffer fit given by BIC at age 34.
TABLE 1. Values of AIC and BIC with optimal $\lambda_a$ and $\lambda_y$ for $bdeg_a = bdeg_y = 3$, $pord_a = pord_y = 2$ and various values of $ndx_a$ and $ndx_y$; $npar$ is the total number of fitted parameters. Minimum values of AIC and BIC are in bold.

<table>
<thead>
<tr>
<th>$ndx_a$</th>
<th>$ndx_y$</th>
<th>$npar$</th>
<th>$\lambda_a$</th>
<th>$\lambda_y$</th>
<th>tr</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
<td>104</td>
<td>0.0018</td>
<td>0.023</td>
<td>89</td>
<td>8339</td>
<td>8913</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0075</td>
<td>18</td>
<td>58</td>
<td>8414</td>
<td>8790</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>184</td>
<td>0.00002</td>
<td>0.006</td>
<td>167</td>
<td>8174</td>
<td>9252</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>25</td>
<td>76</td>
<td>8323</td>
<td>8814</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>169</td>
<td>0.0005</td>
<td>0.7</td>
<td>123</td>
<td>8282</td>
<td>9078</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>300</td>
<td>64</td>
<td>8359</td>
<td>8774</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>299</td>
<td>0.00003</td>
<td>0.1</td>
<td>245</td>
<td>8126</td>
<td>9711</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>600</td>
<td>85</td>
<td>8260</td>
<td>8809</td>
</tr>
</tbody>
</table>

FIGURE 3. Log hazard against year for ages 34 and 60: model in 3.3. Smoothing parameter selection: BIC, solid line; AIC, dotted line.

4 Concluding remarks

We have shown how $P$-splines can be used to model two-dimensional Poisson data. The dimension reduction property of $P$-splines is fundamental in keeping the computational side of the fitting process under control. Judged by Figure 3 and similar plots the fitted model is successful in describing the underlying pattern in mortality.

We have presented this work in the context of the smoothing of two-
dimensional Poisson data. However, it is clear that the algorithms presented here apply in a full GLM setting. Future work includes the following:

(a) It is common practice in demographic models to fit cohort effects explicitly. Our attitude is to assume that our approach will fit any cohort effects implicitly. Indeed cohort plots support this view. Comparison of our two-dimensional approach with, for example, age-period-cohort models would be of interest (Ogata, et al. (2000)).

(b) The regression matrix $B$ in 3.3 is banded and is about 90\% sparse. We have made only rudimentary use of this. Computations could be both extended and speeded up if full advantage was taken of the banded nature of $B$. Matlab is one system that could be used.

(c) We were unable to compute $\text{tr}(H)$ in 3.2. It would be useful to develop a method of at least approximating this value. This would allow model selection and comparison via, for example, BIC.

(d) One complication is that some individuals have more than one policy. However it is not known which deaths give rise to multiple claims so it would be useful to develop methods of dealing with the resulting over-dispersion.

(e) The models could be fitted in a mixed model framework (GLMM). This is an active area of current research (see, for example, Parise et al. (2001)) and would provide an alternative approach to model selection.

References


