Multidimensional density smoothing with P-splines: a mixed model approach

María Durbán¹, Iain Currie² and Paul Eilers³

¹ Dept. Statistics, Universidad Carlos III de Madrid
² Dept. Actuarial Mathematics and Statistics, Heriot-Watt University
³ Dept. Medical Statistics, Leiden University

Abstract: We show how array methods and mixed models give rise to an efficient method of estimating multidimensional densities.

Keywords: GLMM; P-splines; Multidimensional density.

1 P-splines as mixed models

In the companion papers Eilers & Marx (2006) and Lambert & Eilers (2006) described efficient methods for multidimensional density smoothing using P-splines; in these papers, the smoothing parameters are chosen by maximizing some information criterion, such as AIC, or by Bayesian methods respectively. This paper tackles the same problem but uses a mixed model representation of P-splines to choose the smoothing parameters.

There are several approaches to the formulation of splines, and in particular of P-splines, as mixed models. These approaches differ mainly in the bases used for regression and in the penalty, and so yield mixed models with different fixed and random parts (although the fitted model is equivalent).

We begin with a model in 1-dimension with normal errors:

\[ y = B\theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \]  \hspace{1cm} (1) \hspace{1cm} \text{eq:smooth1}

where \( B \) is the \( B \)-spline regression matrix; the smoothness of the fitted curve is obtained by means of a penalty matrix \( P = \lambda D'D \) where \( D \) is a difference matrix acting on the coefficients \( \theta \). The aim is to look for a new basis which allows the representation of (1) as a mixed model:

\[ y = X\beta + Z\alpha + \epsilon, \quad \alpha \sim N(0, G), \quad \epsilon \sim N(0, \sigma^2 I) \]  \hspace{1cm} (2) \hspace{1cm} \text{eq:mixed1}

where \( G \) is a diagonal matrix which depends on \( \lambda \).

One way to find the new basis is to use the singular value decomposition of the penalty \( D'D = V\Sigma V' \) to partition the penalty into a null penalty (corresponding to the fixed part) and a diagonal penalty (for the random part). We partition the matrix \( V = [V_s : V_n] \), where \( V_s \) corresponds to
the non-zero eigenvalues and $V_n$ to the zero eigenvalues. Then, we reparameterize the model with

$$X = BV_n, \quad Z = BV_s, \quad G = \lambda^{-1} \sigma^2 \Sigma$$

where $\Sigma$ is the diagonal matrix which consists of the non-zero eigenvalues in $\Sigma$ (multiplying what(???) by $\Sigma^{-1/2}$ we get $G = \lambda^{-1} \sigma^2 I$).

2 Multidimensional smoothing

Previous approaches to multidimensional smoothing (models based on thin plate splines, or, in the context of P-splines as mixed models, models based on radial basis) presented computational problems or where restricted to isotropic models. With data on a multidimensional array, Durban et al. (2002) proposed the Kronecker product of B-splines as a basis for smoothing in two or more dimensions with a separate penalty for each covariate. In the two-dimensional case, the basis for regression is:

$$B = B_2 \otimes B_1$$

where $B_1$ and $B_2$ are the individual basis for each dimension and the penalty

$$P = \lambda_1 I_{c_2} \otimes D_1' D_1 + \lambda_2 D_2' D_2 \otimes I_{c_1}$$

where $c_1$ and $c_2$ are the ranks of $B_1$ and $B_2$ respectively. With this additional structure (data on an array and the regression matrix as a Kronecker product) Currie et al. (2006) generalized the mixed model representation in 1-dimension given above. The regression matrix $X$ ($Z$) for the fixed (random) part is essentially the Kronecker product of the fixed (random) matrices of the individual basis. The variance matrix $G$ is diagonal with entries that are functions of the eigenvalues of $P_1$ and $P_2$.

3 Penalized generalized smoothing as GLMMs

We have shown above that penalized smooth models can be represented as linear mixed models. The natural extension to cases with non-Gaussian responses is the generalized linear mixed model (GLMM). We focus on the case of Poisson data, since this model will be applied for density estimation. Other responses such as negative binomial, binary data, etc, require small changes from the model show here. For simplicity we will work with a single variable; the extension to the multidimensional case is immediate. Is this true - what about $V$ in (9)?
A GLMM with Poisson error and linear predictor \( \eta = B\theta = X\beta + Z\alpha \) has pdf

\[
f(y|\alpha) = \exp\{y'(X\beta + Z\alpha) - 1' \exp(X\beta + Z\alpha) - 1' \log(\Gamma(y + 1))\}
\]

conditional on \( \alpha \) where \( \alpha \sim \mathcal{N}(0, \lambda^{-1}I_{c-2}) \) where \( c \) is the rank of \( B \). We use the penalized quasi-likelihood (PQL) of Breslow (1993), which leads to

\[
\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y 
\]

\[
\hat{\alpha} = \lambda^{-1}Z'V^{-1}(y - X\hat{\beta}) 
\]

\[
V = W^{-1} + ZZ'\lambda^{-1}.
\]

with \( W = \text{diag}(\exp(X\beta + Z\alpha)) \). Then, conditional on the estimates obtained above, we use REML to estimate the variance parameters,

\[
\text{ql}(\hat{\beta}(\lambda), \lambda) \approx -\frac{1}{2}V^{-1} - \frac{1}{2}X'V^{-1}X - \frac{1}{2}(y - X\hat{\beta} - Z\hat{\alpha})V^{-1}(y - X\hat{\beta} - Z\hat{\alpha})',
\]

and iterate between (7), (8) and (10) until convergence.

The estimates given by PQL correspond to the estimates in a weighted linear Gaussian model where the error variance is \( \sigma^2 = 1 \). This point makes possible the use of the \texttt{lme()} in \texttt{Splus} and \texttt{R} in the one-dimensional case. The function \texttt{glmm.PQL()} in \texttt{R} also may be adapted to fit these models. In cases where we need to fit a surface, the functions mentioned above run into computational models. The authors have written functions in \texttt{Splus/R}, following Pinheiro (2000), but based on the array algorithms presented in Currie (2006), which are easily adapted to the case of GLMMs. These algorithms are extremely fast and have low storage requirements (since they avoid the computation of the full regression matrix \( B \)).

4 Density estimation: The Old Faithful data

The “Old Faithful” geyser data set consists of 299 pairs of eruption duration times and waiting times till the next eruption of the “Old Faithful” geyser in Yellowstone National Park. To apply the present method we first construct a two-dimensional histogram over a fine grid; we used 100 \( \times \) 100 in this example. Note that the great majority of the histogram bins have a zero count. We take \( B \)-spline bases \( B_1 \) and \( B_2 \) each of rank 13 and fit anisotropic and isotropic models. Selection of the smoothing parameters and estimation of the smoothed density took less than 40 seconds in both cases. Figure 1 shows the density contour plots.

References


