Parsimonious classification via generalised linear mixed models

By G. Kauermann,

Faculty of Economics, University Bielefeld,
Postfach 300131, 33501 Bielefeld, Germany

gkauermann@wiwi.uni-bielefeld.de

J.T. Ormerod

School of Mathematics and Statistics,
University of New South Wales, Sydney 2052, Australia

johno@uow.edu.au

and M.P. Wand

mwand@uow.edu.au

School of Mathematics and Applied Statistics,
University of Wollongong, Northfields Avenue, Wollongong 2522, Australia

24th November, 2008

Summary

We devise a classification algorithm based on generalised linear mixed model (GLMM) technology. The algorithm incorporates spline smoothing, additive model-type structures and model selection. For reasons of speed we employ the Laplace approximation, rather than Monte Carlo methods. Tests on real and simulated data show the algorithm to have good classification performance. Moreover, the resulting classifiers are generally interpretable and parsimonious.

Keywords: Akaike Information Criterion; Feature selection; Generalised additive models; Penalised splines; Supervised learning; Model selection; Rao statistics; Variance components.

1 Introduction

Classification is a very old and common problem, where training data are used to guide the classification of future objects into two or more classes based on observed predictors. Examples include clinical diagnosis based on patient symptoms, handwriting recognition based on digitised images and financial credit approval based on applicant attributes. Classification has an enormous number of applications; arising in most areas of science, but also in business as evidenced by the ongoing growth of industries such as data mining and fraud detection. The literature on classification methodology and theory is massive and mature. Contemporary statistical perspectives include Breiman (2001), Hastie, Tibshirani and Friedman (2001) and Hand (2006). A substantial portion of the classification literature is within the field of Computing Science, where 'classification' is usually called 'supervised learning' and 'predictors' often called 'features' or 'variables'.

There is a multitude of criteria that could be considered when tuning and assessing the quality of a classification algorithm. Numerical criteria include test error, Brier score and area under the curve of the receiver operating characteristic. A non-numerical quality criterion which, depending on the application, can be of utmost importance is *interpretability*. Hastie *et al.* (2001, Section 10.7) state that 'data mining applications generally require interpretable models' and that 'black box' classifiers with good numerical performance are 'far less useful'. Nevertheless, a good deal of classification theory and methodology, within both Statistics and Computing Science, is oblivious to interpretability. Some exceptions include tree-based approaches (e.g. Breiman, Friedman, Olshen & Stone, 1984; Hastie *et al.*, 2001) and additive model-based approaches (e.g. Hastie *et al.*, 2001). Related to interpretability is *parsimony*, where superfluous predictors are sifted out. This corresponds to pruning of tree-type classifiers and variable selection in those based on additive models. In Computing Science the topics of *variable selection* and *feature selection* (e.g. Guyon & Elisseeff, 2003) have similar aims.

Another often neglected quality measure is *speed*. Again, depending on the application, speed can be crucial. Speed is invariably tied to the size of the training data but there are huge differences, some involving several orders of magnitude, between existing classification algorithms in this respect.

In this paper we develop a classification algorithm that strives for very good performance in terms of interpretation, parsimony and speed; while also achieving good classification performance. The algorithm, which we call KOW (after ourselves), performs classification via a semiparametric logistic regression model after undergoing variable selection on the predictors. In this respect, KOW is similar in spirit to variable selection algorithms for additive models such as BRUTO (Hastie & Tibshirani, 1990), those based on versions of the `R` function `step.gam()` (Chambers & Hastie, 1992; Hastie, 2006; Wood, 2006), and Markov Chain Monte Carlo approaches such as that developed by Yau, Kohn & Wood (2003). The additive structure aids interpretation, but can also lead to improved test errors; see e.g. Section 12.3.4 of Hastie *et al.* (2001).

The KOW algorithm performs fast fitting and variable selection by borrowing ideas from generalised linear mixed models (GLMM). This is a relatively young, but rapidly growing, area of research that has its roots in biostatistical topics such as longitudinal data analysis and disease mapping; see e.g. Breslow & Clayton (1993), Verbeke & Molenberghs (2000) and Wakefield, Best & Waller (2000). However GLMM can handle a much wider range of problems including generalised additive models (e.g. Zhao, Staudenmayer, Coull & Wand, 2006). The essence of KOW is to equate inclusion of a predictor with the significance of parameters in a GLMM. Linear terms correspond to fixed effect parameters, while non-linear terms correspond to variance components. KOW uses efficient score-based statistics, also known as *Rao statistics*, to choose among candidate predictors. A version of the Akaike Information Criterion is used to choose between fixed effect parameters and variance components, and also acts as a stopping rule. Unlike `step.gam()`, KOW has inbuilt automatic smoothing parameter selection for smooth function components.

When fitting a GLMM, whether for classification or not, the main obstacle is the presence of intractable integrals in the likelihood. Currently available methods for fitting a GLMM fall into three general categories: quadrature, Monte Carlo methods and analytic approximation (e.g. McCulloch & Searle, 2000). Quadrature is not viable for the size of integrals arising GLMM with additive model structure. Monte Carlo methods are generally ruled out by their slowness. KOW makes use of much faster Laplace approximation methods. Laplace approximation is sometimes criticised in GLMM analysis due to the substantial biases inherent in estimates of parameters of interest (e.g. McCulloch & Searle, 2000, p. 283). Recently, Kauermann, Krivobokova & Fahrmeir (2008) have shown that the Laplace approximation is asymptotically justifiable and hence unbiased for penalized spline smoothing in GLMMs. This also holds when the number of splines increases with the sample size. Hence, bias problems do not occur in our setting. Furthermore, such issues are less crucial in the classification context where minimizing classification error is paramount.

We have tested KOW on several real and simulated data sets and compared it with other additive model-based classifiers. Our implementation of KOW fits a classifier to data sets with 5–10 possible predictors in a few seconds on a typical 2008 computer. If the number of predictors is in the tens then computation is in the order of minutes. The penalised spline aspect of KOW means that training sample size only has a linear effect on computation times. KOW is generally much faster than `step.gam()`, although not as fast as BRUTO. However KOW can yield much better classification performance than BRUTO and is on par with `step.gam()`. Performances tend to be similar among algorithms in terms of interpretability and parsimony. On balance, we believe KOW has the potential for improved fast classification in contexts when interpretability and parsimony are important.

In Section 2 we develop a fast algorithm for fitting a GLMM. In keeping with the classification goals we concentrate on the logistic mixed model. Section 3 describes our model selection strategy based on Rao statistics and AIC. We report on some comparative numerical studies in Section 4. We conclude with some discussion in Section 5.

2 Fast Logistic Mixed Model Classifiers

Consider two-class classification with class labels denoted by $y \in \{0,1\}$ and let $\mathbf{x} = (x_1,\ldots,x_d)$ be the set of possible predictors. Logistic regression-type classification is based on models of general form

$$\text{logit}\{P(y=1|\mathbf{x})\} = \eta(\mathbf{x}). \tag{1}$$

Classification of a new observation with predictor vector $\mathbf{x}_{\text{new}}$ is performed according to

$$\text{sign}\{\widehat{\eta}(\mathbf{x}_{\text{new}})\}$$

where $\widehat{\eta}$ is an estimate of η based on training data $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$. Here $\mathbf{x}_i$ is a d-variate vector representing the ith observation on $\mathbf{x}$.

A key element is appropriate modelling of $\eta(\mathbf{x})$. Given our interpretability goals, we work with sums of smooth low-dimensional functions of the predictors such as:

$$\eta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \eta_2(x_2) + \eta_3(x_3)$$

and

$$\eta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \eta_{23}(x_2,x_3).$$

Smooth univariate functions are modelled using penalised splines with mixed model formulation as follows:

$$\eta_j(x_j) = \beta_j x_j + \sum_{k=1}^{K_j} u_{jk} z_{jk}(x_j)$$

where u_{jk} i.i.d. $N(0,\sigma_j^2)$ where the variance component σ_j^2 controls the amount of smoothing by acting as penalty parameter corresponding to a quadratic penalty on the u_{jk}s. The z_{jk} are spline basis functions appropriate for handling the non-linear component of η_j. There are several options for their choice; see e.g. Durbán & Currie (2003), Wand (2003), Welham, Cullis, Kenward & Thompson (2007) and Wand & Ormerod (2008). Bivariate functions will be of the form

$$\eta_{j\ell}(x_j,x_\ell) = \beta_j x_j + \beta_\ell x_\ell + \sum_{k=1}^{K_j} u_{j\ell k} z_{j\ell k}(x_j,x_\ell)$$

with $u_{j\ell k}$ i.i.d. $N(0,\sigma_{j\ell}^2)$. Appropriate bivariate spline functions $z_{j\ell k}(x_j,x_\ell)$ are described by Ruppert, Wand & Carroll (2003, Chapter 13) and Wood (2003). The extension to general multivariate functions is obvious from these references. However, due to interpretability and curse of dimensionality issues, it is rare to have more than 2 or 3 variables handled together.

One advantage of utilizing the above mixed model representation of penalized splines is that it easily allows for the incorporation of additional complexities such as longitudinal and spatial effects. A further advantage of the mixed model/penalized spline representation is that it allows use the well established framework of maximum likelihood and best prediction for estimation and inference (Ruppert *et al.*, 2003; Wand, 2003).

Models for $\boldsymbol{\eta} = [\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n)]^T$ can be written in the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ contains a column of ones, together with a subset of of the columns of $[\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]^T$, and $\mathbf{Z}$ are design matrices corresponding to spline bases. The covariance matrix of $\mathbf{u}$ takes the form

$$\mathbf{G}_{\boldsymbol{\sigma}^2} \equiv \underset{1 \le j \le r}{\text{blockdiag}}(\sigma_j^2 \mathbf{I}) \tag{3}$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_r^2)$ is the vector of variance components.

For the model defined by (1), (2) and (3) the log-likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) \ &= \ \log \int_{\mathbb{R}^q} \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T \log(1 + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}})\right\} \\ &\qquad \times (2\pi)^{-q/2} |\mathbf{G}_{\boldsymbol{\sigma}^2}|^{-1/2} \exp(-\tfrac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1} \mathbf{u})\, d\mathbf{u} \end{aligned} \tag{4}$$

where q is the dimension of $\mathbf{u}$.The integral (4) cannot be calculated in analytic form. This is usually dealt with via Monte Carlo methods or analytic approximations. In the interest of speed we work with the Laplace approximation of (4):

$$\ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\tfrac{1}{2} \log |\mathbf{I} + \mathbf{Z}^T \mathbf{W}_{\boldsymbol{\beta},\widehat{\mathbf{u}}} \mathbf{Z} \mathbf{G}_{\boldsymbol{\sigma}^2}| + \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\widehat{\mathbf{u}}) - \mathbf{1}^T \log(1 + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widehat{\mathbf{u}}}) - \tfrac{1}{2}\widehat{\mathbf{u}}^T \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1} \widehat{\mathbf{u}} \tag{5}$$

where $\mathbf{W}_{\boldsymbol{\beta},\mathbf{u}} \equiv \text{diag}\{\frac{e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}}}{(1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}})^2}\}$ and $\widehat{\mathbf{u}}$ is the maximiser of the integrand in (5) (e.g. Breslow & Clayton, 1993).

If exact calculation of the likelihood was possible then predictions for new data would be made by replacing $\boldsymbol{\beta}$ and $\mathbf{u}$ in η with the maximum likelihood estimate of $\boldsymbol{\beta}$ and the best predictor of $\mathbf{u}$, i.e. $\mathbb{E}(\mathbf{u}|\mathbf{y})$. Since we do not have a closed form for the likelihood we instead use the $\widehat{\boldsymbol{\beta}}$ obtained by maximizing ℓ_{Laplace} with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ and the mode $\widehat{\mathbf{u}}$ to approximate $\mathbb{E}(\mathbf{u}|\mathbf{y})$.

Maximising (5) is difficult due to non-linear expressions involving both $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ in the first and last terms of (5). We therefore pursue a backfitting idea by iteratively maximising (5) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, respectively. Note that $\widehat{\mathbf{u}}$ depends on $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, so that the Laplace approximation has to be updated in each estimation iteration as well. We do this by updating the estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ simultaneously. Let $\mathbf{B} \equiv \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})$, $\boldsymbol{\nu} \equiv [\boldsymbol{\beta}^T, \mathbf{u}^T]^T$, $\mathbf{C} \equiv [\mathbf{X}\ \mathbf{Z}]$ and

$$df_j(\sigma_j^2) \equiv \text{tr}\{\mathbf{E}_j(\mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z} + \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})^{-1} \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z}\}$$

where $\mathbf{E}_j$ is the diagonal matrix with ones in the diagonal positions corresponding to the spline basis functions for σ_j^2 and zeroes elsewhere. Note that $df_j(\sigma_j^2)$ has an 'effective degrees of freedom' (e.g. Buja, Hastie & Tibshirani, 1989) interpretation for the contribution from the spline terms attached to σ_j^2. We propose fitting logistic mixed model classifiers using Algorithm 1.

Algorithm 1 is similar to the algorithm developed by Breslow and Clayton (1993), commonly referred to as PQL (an acronym for Penalized Quasi-Likelihood) but differs in two respects. PQL uses Fisher scoring as the updating step for $\widehat{\boldsymbol{\nu}}$ while Algorithm

Algorithm 1 Fast Fitting of a Logistic Mixed Model Classifier

```
1. Initialise: $\widehat{\boldsymbol{\nu}}^{(0)}$ and $\widehat{\boldsymbol{\sigma}}^{2(0)}$. Set $L$ to be a small integer.

2. Cycle:
     for $\ell = 1, 2, \ldots$ do
       if $\ell \mod L = 1$ then
         $\mathbf{K} = \mathbf{C}^T \mathbf{W}_{\widehat{\boldsymbol{\nu}}^{(\ell)}} \mathbf{C}$
       end if
       $\widehat{\boldsymbol{\nu}}^{(\ell+1)} = \widehat{\boldsymbol{\nu}}^{(\ell)} + (\mathbf{K} + \mathbf{B})^{-1} \left\{ \mathbf{C}^T \left( \mathbf{y} - \frac{e^{\mathbf{C}\widehat{\boldsymbol{\nu}}^{(\ell)}}}{1+e^{\mathbf{C}\widehat{\boldsymbol{\nu}}^{(\ell)}}} \right) - \mathbf{B}\widehat{\boldsymbol{\nu}}^{(\ell)} \right\}$
       for $\ell' = 1, 2, \ldots$ do
         for $I \in \mathcal{I}$ do
           $\widehat{\sigma}_j^{2(\ell'+1)} = \|\widehat{\mathbf{u}}_j^{(\ell')^T}\|^2 / df_j(\widehat{\sigma}_j^{2(\ell')})$
         end for
       end for
     end for

   until: $\max \left\{ \frac{\|\widehat{\boldsymbol{\nu}}^{(\ell+1)} - \widehat{\boldsymbol{\nu}}^{(\ell)}\|}{\|\widehat{\boldsymbol{\nu}}^{(\ell)}\|}, \frac{\|\widehat{\boldsymbol{\sigma}}^{2(\ell'+1)} - \widehat{\boldsymbol{\sigma}}^{2(\ell')}\|}{\|\widehat{\boldsymbol{\sigma}}^{2(\ell')}\|} \right\}$ is below some small tolerance value.
```

1 for uses a *repeated Hessian* Newton's method (Ormerod, 2008, Appendix C). Here the Hessian is updated every second iteration and can be viewed as a slight modification of Fisher scoring. However, unlike PQL, the updating step for $\widehat{\boldsymbol{\sigma}}^2$ uses a fixed point iteration in order to avoid calculating to Hessian matrix of derivatives with respect to $\boldsymbol{\sigma}^2$. The fixed point updating formula arises from differentiation of $\ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ with respect to $\boldsymbol{\sigma}_j$. The PQL approach to updating $\widehat{\boldsymbol{\sigma}}^2$ is trickier to implement since more care is required to calculate the Hessian and ensuring positive definiteness in calculating Newton search directions for $\boldsymbol{\sigma}^2$.

Algorithm 1 is also quite fast compared to PQL. Solving for $\widehat{\boldsymbol{\nu}}^{(\ell+1)}$ for a fixed $\widehat{\boldsymbol{\sigma}}^2$ is a concave programming problem. Assuming that the function to be maximized is strictly concave and has a Lipschitz continuous Hessian and the current iterate is sufficiently close to the solution it is possible to show that the rate of convergence over two-steps of the algorithm is at up to cubic (Ormerod, 2008, Appendix C). Every odd iteration takes $O(nP^2 + P^3)$ while every even step only takes $O(nP + P^2)$ where P is the length of the $\widehat{\boldsymbol{\nu}}$ vector. Solving for $\widehat{\boldsymbol{\sigma}}^2$ is can be comprehended as a fixed-point iteration. Each $\boldsymbol{\sigma}^2$ update can be computed in $O(nP^2 + P^3)$ operations.

3 Model Selection

We now address the problem of choosing between the various models for the classifier $\boldsymbol{\eta}(\mathbf{x})$. Even for moderate d the number of such models can be very large. Our approach is driven by our previously stated goals of speed, parsimony and interpretability.

According to the spline models described in Section 2, the fullest model has fixed effects component

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

However, smooth function terms will not be appropriate for all predictors. For example, some of the x_i's may be binary. Let S be the subset of $\{1, \ldots, d\}$ such that x_i is to be modelled as smooth function for each $i \in S$. Then let $\mathcal{I}$ be a partition of S that specifies the type of non-linear modelling in the fullest model. For example if $d = 4$ then $\mathcal{I} = \{1, 3, 4\}$ corresponds to the fullest model being the additive model $\eta(x_1, x_2, x_3, x_4) = \beta_0 + \eta_1(x_1) + \beta_2 x_2 + \eta_3(x_3) + \eta_4(x_4)$, while $\mathcal{I} = \{\{1, 3\}, \{4\}\}$ corresponds to the model

$\eta(x_1, x_2, x_3, x_4) = \beta_0 + \eta_{13}(x_1, x_3) + \beta_2 x_2 + \eta_4(x_4)$. We will assume, for now, that S and $\mathcal{I}$ are specified in advance. A recommended default choice is

$$\mathcal{I} = \text{all singleton sets of elements of } S$$

corresponding to an additive model. Note that subscripting on the σ_j^2 corresponds to the elements of $\mathcal{I}$ rather than those of $\mathbf{x}$.

Description of our model selection strategy for the general set-up becomes notationally unwieldy. Therefore we will describe the algorithm via an example. Suppose that the set of possible predictors $\{x_1, x_2, x_3\}$ where x_1 is binary and x_2 and x_3 continuous, and that only additive models are to be considered. Then $S = \{2, 3\}$ and $\mathcal{I} = \{\{2\}, \{3\}\}$. The fullest model is

$$\eta(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sum_{k=1}^{K_2} u_{2k} z_{k2}(x_2) + \sum_{k=1}^{K_3} u_{3k} z_{k3}(x_3)$$

where u_{2k} i.i.d. $N(0, \sigma_1^2)$ and u_{k3} i.i.d. $N(0, \sigma_2^2)$. There are $2^5 = 32$ possible sub-models that include the intercept term. We propose the following forward selection approach to choosing among them:

1. Start with $\eta(x_1, x_2, x_3) = \beta_0$.
2. (a) Determine the 'best' linear component to add to the model from $\{\beta_1 x_1, \beta_2 x_2, \beta_3 x_3\}$. Let β_* denote the β_k corresponding to this choice.

   (b) Determine the 'best' non-linear (spline) component to add to the model from $\left\{\sum_{k=1}^{K_2} u_{k2} z_{k2}(x_2), \sum_{k=1}^{K_3} u_{k3} z_{k3}(x_3)\right\}$. Let σ_*^2 denote the σ_k^2 corresponding to this choice.
3. Add the component corresponding to β_* or σ_*^2 that leads to the bigger decrease in the marginal Akaike Information Criterion (mAIC). If there is no decrease then stop and use the current model for classification. Otherwise, add the new component to the model and return to Step 2; modified to have one less component. Continue while there are still unselected components.

We propose to choose the 'best' linear and non-linear components using approximate score-type test statistics that do not require fitting of the candidate models. This has an obvious speed advantage. The details are given in Sections 3.1 and 3.2. The mAIC criterion is described in Sections 3.3.

Before that we briefly give some required notation. For a general $d \times 1$ parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ with log-likelihood $\ell(\boldsymbol{\theta})$ the derivative vector of ℓ, $\mathsf{D}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$, is the $1 \times d$ with ith entry $\partial\ell(\boldsymbol{\theta})/\partial\theta_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \mathsf{D}_{\boldsymbol{\theta}}\{\mathsf{D}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})^T\}$. The information matrix of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is then $-E\{\mathsf{H}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\}$.

3.1 Choosing the 'best' linear component to add

Let $(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2)$ define the current model, with fitted values $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}}, \widehat{\boldsymbol{\sigma}}^2)$ as obtained via Algorithm 1, and let $\beta_k \mathbf{x}_k$ represent a generic linear component not already in the model. The log-likelihood corresponding to the new model with $\beta_k \mathbf{x}_k$ added is a modification of (4) with $\mathbf{X}\boldsymbol{\beta}$ replaced by $\mathbf{X}\boldsymbol{\beta} + \beta_k \mathbf{x}_k$ and is denoted by $\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \beta_k)$.

We propose to choose the 'best' $\beta_k \mathbf{x}_k$ among all candidates according to maximum absolute *Rao statistic* (also known as the *score statistic*) (e.g. Rao, 1973, Chapter 6). Exact

Rao statistics in GLMM are computationally expensive, so we make a number of convenient approximations. The first of these is to assume orthogonality between $(\boldsymbol{\beta}, \beta_k)$ and $\boldsymbol{\sigma}^2$ in the information matrix of the joint parameters. Strictly speaking, these parameters are not orthogonal (Wand, 2007), but such orthogonality arises in the approximate log-likelihoods with which we work. Under orthogonality, the Rao statistic for the hypotheses $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$ is

$$R_{\beta_k} = [\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)]_{p+1}\Bigg/\sqrt{1\Big/\left([E_{\mathbf{y}}\{-\mathsf{H}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)\}]^{-1}\right)_{p+1,p+1}}$$

where p is the length of $\boldsymbol{\beta}$. A practical approximation involves dropping the determinant term in (5) to obtain

$$\ell(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\beta_k) \simeq \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}) - \mathbf{1}^T\log(1+e^{\mathbf{X}\boldsymbol{\beta}+\beta_k\mathbf{x}_k+\mathbf{Z}\widehat{\mathbf{u}}}) - \tfrac{1}{2}\widehat{\mathbf{u}}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\widehat{\mathbf{u}}. \qquad (6)$$

As shown in the Appendix, this leads to

$$R_{\beta_k} \simeq \mathbf{x}_k^T\left(\mathbf{y}-\frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}\right)\Bigg/\sqrt{\mathbf{x}_k^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\{\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\}\mathbf{x}_k}. \qquad (7)$$

An advantage of this Rao statistic approach is that the candidate models corresponding to addition of the $\beta_k\mathbf{x}_k$ do not need to be fitted. This means that the R_{β_k} can be computed quickly even when there is a large number of candidate linear components. This strategy has been used successfully in fitting regression spline models; see e.g. Stone, Hanson, Kooperberg & Truong (1997).

3.2 Choosing the 'best' non-linear component to add

As in Section 3.1, let $(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2)$ define the current model and let $\mathbf{Z}_k\mathbf{u}_k$, $\mathbf{u}_k \sim N(\mathbf{0},\sigma_k^2\mathbf{I})$, represent a generic non-linear component not already in the model. The log-likelihood corresponding to the new model with σ_k^2 added is a modification of (4) with $\mathbf{Z}\mathbf{u}$ replaced by $\mathbf{Z}\mathbf{u}+\mathbf{Z}_k\mathbf{u}_k$ and is denoted by $\ell(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\sigma_k^2)$.

The Rao statistic for $H_0 : \sigma_k^2 = 0$ versus $H_1 : \sigma_k^2 > 0$ is

$$R_{\sigma_k^2} = [\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)]_{r+1}\Bigg/\sqrt{1\Big/[E_{\mathbf{y}}\{-\mathsf{H}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)\}]^{-1}_{r+1,r+1}} \equiv R^{\text{num}}_{\sigma_k^2}/R^{\text{den}}_{\sigma_k^2} \qquad (8)$$

where $R^{\text{num}}_{\sigma_k^2}$ and $R^{\text{den}}_{\sigma_k^2}$ respectively denote the numerator and denominator in $R_{\sigma_k^2}$ and r is the length of $\boldsymbol{\sigma}^2$. Test statistics of this type were studied by Cox & Koh (1989), Gray (1994), Lin (1997) and Zhang & Lin (2003), for example. We use the largest approximate $R_{\sigma_k^2}$ to choose the 'best' non-linear component not already in the model.

For practical reasons, we work with the Laplace approximation to $\ell(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\sigma_k^2)$:

$$\begin{aligned}\ell_{\text{Laplace}}(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\sigma_k^2) = &-\tfrac{1}{2}\log|\mathbf{I}+[\mathbf{Z}\ \mathbf{Z}_k]^T\mathbf{W}_{\boldsymbol{\beta},\widehat{\mathbf{u}},\widehat{\mathbf{u}}_k}[\mathbf{Z}\ \mathbf{Z}_k]\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2},\sigma_k^2\mathbf{I})|\\ &+\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k) - \mathbf{1}^T\log(1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k}) - \tfrac{1}{2}\widehat{\mathbf{u}}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\widehat{\mathbf{u}} - \tfrac{1}{2}\sigma_k^{-2}\widehat{\mathbf{u}}_k^T\widehat{\mathbf{u}}_k\end{aligned}$$

where $(\widehat{\mathbf{u}},\widehat{\mathbf{u}}_k)$ maximises

$$\mathbf{y}^T(\mathbf{Z}\mathbf{u}+\mathbf{Z}_k\mathbf{u}_k) - \mathbf{1}^T\log(1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\mathbf{Z}\mathbf{u}_k}) - \tfrac{1}{2}\mathbf{u}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{u} - \tfrac{1}{2}\mathbf{u}_k^T\mathbf{u}_k/\sigma_k^2. \qquad (9)$$

The dependence of $\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}$ on $(\boldsymbol{\sigma}^2,\sigma_k^2)$ is ignored in the differentiation. We show in the Appendix that

$$\begin{aligned}R^{\text{num}}_{\sigma_k^2} \simeq\ & -\tfrac{1}{2}\text{tr}[\mathbf{Z}_k^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\{\mathbf{I}-\mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2}(\mathbf{I}+\mathbf{Z}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2})^{-1}\mathbf{Z}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\}\mathbf{Z}_k]\\ &+\tfrac{1}{2}\left\|\mathbf{Z}_k^T\left(\mathbf{y}-\tfrac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}\right)\right\|^2.\end{aligned} \qquad (10)$$

Expression (10) has the computational advantage that the matrix inversion pertains to the current model and only needs to be performed once for selecting the 'best' non-linear component.

The denominator of $R_{\sigma_k^2}$ can be approximated using the arguments in Section 2.4 of Breslow & Clayton (1993). These lead to

$$R^{\text{den}}_{\sigma_k^2} \simeq \sqrt{1/[\mathcal{K}(\boldsymbol{\sigma}^2, 0)^{-1}]_{r+1,r+1}}$$

where $\mathcal{K}(\boldsymbol{\sigma}^2, \sigma_k^2)$ is the $(r+1) \times (r+1)$ matrix with (i, j) entry given by

$$\begin{aligned}\mathcal{K}_{ij}(\boldsymbol{\sigma}^2, \sigma_k^2) \equiv \tfrac{1}{2}\text{tr}\{&\mathbf{E}_i(\mathbf{I} + \widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2})^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\\ &\times\mathbf{E}_j(\mathbf{I} + \widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2})^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\},\end{aligned}$$

$\widetilde{\mathbf{Z}} \equiv [\mathbf{Z}\ \mathbf{Z}_k]$, $\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2} \equiv \text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}, \sigma_k^2\mathbf{I})$ and $\mathbf{E}_1, \ldots, \mathbf{E}_{r+1}$ are the diagonal matrices, with zeroes and ones on the diagonal, defined by $\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2} = \sum_{i=1}^r (\boldsymbol{\sigma}^2)_j\mathbf{E}_j + \sigma_k^2\mathbf{E}_{r+1}$. A more explicit formula for $R^{\text{den}}_{\sigma_k^2}$, that aids efficient computation of the $R_{\sigma_k^2}$, is given in the Appendix.

3.3 The mAIC criterion

For the model defined by $(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2)$ the marginal Akaike Information Criterion (mAIC) is

$$\text{mAIC}(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2) = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2) + 2\{\dim(\boldsymbol{\beta}) + \dim(\boldsymbol{\sigma}^2)\}$$

where $\dim(\mathbf{v})$ denotes the dimension, or length, of the vector $\mathbf{v}$. In practice we replace ℓ by ℓ_{Laplace}. The word 'marginal' is used to distinguish the criterion from conditional AIC (cAIC) introduced to mixed model analysis by Vaida & Blanchard (2005). In smooth function contexts, cAIC differs from mAIC in that the former used an 'effective degrees of freedom' measure (e.g. Buja *et al.*, 1989) in the second term rather than the number of fixed effects and variance components. Recently, Wager, Vaida & Kauermann (2007) compared mAIC and cAIC for model selection in Gaussian response models and concluded comparable performance in that context. While similar comparisons are yet to be made in the logistic context it is unlikely that one will significantly dominate the other. Our decision to use mAIC in the default KOW algorithm is driven by the high premium we are placing on computational speed.

3.4 Variants and extensions

The algorithm described near the start of this section, with details as laid out in Sections 3.1–3.3, is the 'default' version of the KOW algorithm for building a parsimonious classifier; optimised for speed and implementation simplicity. There are a number of variants and extensions that could be considered — albeit at the expense of speed and simplicity. Some of these are:

- Replace the mAIC-based model selection strategy with one that uses hypothesis testing and p-values. This involves approximate distribution theory for the Rao statistics. We have done some experimentation with this p-value approach and the results look promising. Appendix B provides details.

- Replace the simple forward selection algorithm with a more elaborate scheme. One option is to have forward selection up to the fullest model, followed by a backward selection phase, using Wald statistics, back to the smallest model. Such a strategy is used by Stone *et al.* (1997), for example.

- Automate the choice between univariate and multivariate functions of the continuous predictors corresponding to the set $\mathcal{I}$. The default version requires the user to either specify $\mathcal{I}$ or use only univariate functions.

- Decide whether a component should be added to the model based on criteria other than largest decrease in mAIC. Options include cAIC and versions of generalised cross-validation (e.g. Kooperberg, Bose & Stone, 1997).

- Insist that all non-linear components have a corresponding linear term. So if the non-linear component for x_k is selected for addition to the model then also add $\beta_k x_k$ if it is not already present.

4 Comparative Performance

We now compare KOW with algorithms similar in their aims including: BRUTO (Hastie and Tibshirani, 1990) and the functions `step.gam()` from the `R` libraries `gam` version 0.98 and `mgcv` version 1.3-27 (Chambers & Hastie, 1992; Hastie, 2006; Wood, 2006). The comparisons are made with respect to test error, parsimony and speed.

The `mgcv` package performs smoothing and model selection via optimization of the generalized cross-validation (GCV) criteria. However `mgcv` does not perform variable selection as such but uses the related concept of shrinkage (see Hastie *et al.*, 2001, Chapter 3 for instance). For the purposes of testing we treat variables with an estimated effective degrees of freedom smaller than 0.01 as not included in the model. The `step.gam()` function in the `gam` package requires the use to specify the number of degrees of freedom for each component. In our comparison studies we have set the number of degrees of freedom for each component to 3. Model selection is then performed by greedily selecting the component that gives the smallest AIC value. The BRUTO procedure uses least squares loss with smoothing splines where back-fitting model selection is based on an approximate GCV criteria.

Our comparison study involved both real and simulated datasets. All datasets were obtained from the following Internet locations in 2008:

Name	Location
banana	`users.rsise.anu.edu.au/~raetsch/data/index.html`
PID/spam	`cran.au.r-project.org/src/contrib/mlbench_1.1-0.tar.gz`
orange	`www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/orange`

Two simulated datasets were also used for comparison, *Orange* and *Banana*. In *Orange* ten predictors $X_1, \ldots, X_{10}$ are simulated from a univariate standard normal distribution with one class having the first four predictors conditioned on $9 \leq \sum_{i=1}^{4} X_i^2 \leq 16$. Thus *Orange* has 4 real predictors and 6 noise predictors. *Banana* is a 2 class 2-dimensional dataset simulated such that the points from four overlapping clusters two of which are banana shaped. A sub-sample of these points are displayed in Figure 1. For the *Banana* dataset we added 6 standard normal noise predictors to make a total of 8 predictors for the dataset used for testing. Note that the data from the *Banana* dataset is not simulated from an additive model structure.

For the *Orange* dataset each algorithm was run using 50 observations for each class (making a total of 100 observations), and the test error was attained by taking the average error from 50 simulations containing 500 observations for each class. For the *Banana* dataset each algorithm was run using 400 observations and the test error was attained by taking the average error from 100 simulations containing 4900 observations altogether.

The two real datasets used were the *Spam* dataset, containing 4601 observations and 57 predictors, and the *Pima Indians Diabetes (PID)* dataset, containing 768 observations and 8 predictors.

Testing on the real datasets was conducted using 10-fold cross-validation. This involves splitting the dataset into 10 different parts. For the ith part we fit the model using the other 9 parts of the data, and calculate the prediction error of the model when predicting the ith part of the data. We did this for all 10 parts and averaged the 10 estimates to obtain the test error.

For each variable we used a univariate O-Sullivan spline basis as described in Wand and Ormerod (2008). Twenty interior knots, equally spaced with respect to the quantiles, were used for each variable.

Dataset	Method	Without Noise Test Error (%)	With Noise Test Error (%)	Real	Noise	Mean Time (seconds)
Banana	mgcv	28.12 (0.15)	29.06 (0.16)	2.00	3.41	22.74 (1.25)
	gam	32.63 (0.28)	33.10 (0.24)	2.00	0.85	2.24 (0.07)
	BRUTO	28.13 (0.12)	28.29 (0.13)	1.85	0.35	0.81 (0.001)
	KOW	28.11 (0.15)	28.76 (0.15)	1.87	1.07	1.08 (0.05)
Orange	mgcv	13.18 (0.86)	12.00 (0.85)	4.00	1.10	57.46 (2.69)
	gam	9.16 (0.74)	9.64 (0.78)	4.00	0.32	17.62 (0.26)
	BRUTO	8.58 (0.65)	9.10 (0.71)	4.00	0.30	0.14 (0.001)
	KOW	9.45 (0.39)	11.92 (0.87)	3.92	0.78	1.82 (0.06)

Table 1: *Averages (standard deviation) results for the* Banana *and* Orange *study described in Section 4.*

Examining Table 1 we see that all methods are fairly robust classifiers when noise variables are added. Furthermore all methods appear to be fairly good at discerning the real predictors from the noise predictors. KOW appears to select more noise predictors than all of the other methods accept `mgcv`. BRUTO appears to give slightly better classification rates on the *Orange* dataset.

Dataset	Method	Test Error (%)	Mean No. Predictors Included	Mean Time (seconds)
Pima	mgcv	23.43 (1.90)	6.9	14.27 (1.08)
Indians	gam	23.69 (2.16)	4.6	4.13 (0.16)
Diabetes	BRUTO	50.64 (1.80)	5.3	0.12 (0.004)
	KOW	22.92 (1.62)	6.0	2.51 (0.11)
Spam	mgcv	5.89 (0.34)	50.7	21278.00 (4466.75)
	gam	failed	N/A	N/A
	BRUTO	failed	N/A	N/A
	KOW	5.38 (0.20)	37.6	1033.05 (98.93)
Reduced Spam	mgcv	6.15 (0.37)	28.4	4076.51 (694.35)
	gam	6.42 (0.22)	28.2	7521.10 (1467.74)
	BRUTO	16.86 (0.73)	25.7	1.01 (0.01)
	KOW	5.57 (0.25)	27.3	590.06 (62.13)

Table 2: *Averages (standard deviation) results for the* Pima Indians Diabetes *and* Spam *study described in Section 4.*

The `gam` and BRUTO procedures failed on the full *Spam* dataset. The `gam` procedure failed because it creates an object indicating whether each of the possible 2^d candidate models had been fitted. For high d the size of this object becomes too large. We could not ascertain why the BRUTO procedure failed. To allow comparison of all 4 methods we also worked with a reduced version of the *Spam* dataset based on the 29 variables most often selected by KOW.

Examining Table 2 we see that KOW seems to gives similar (possibly slightly better)

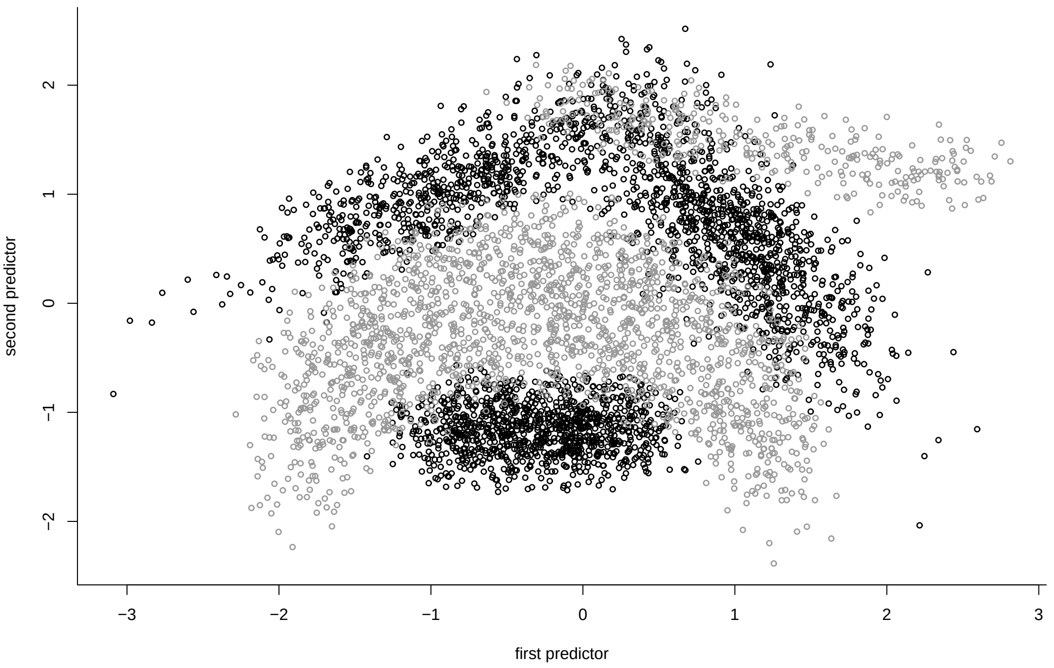

Figure 1: *Test sample 1 of 4900 data points from the Banana dataset.*

classification error compared to other methods on the real datasets. For the aforementioned reasons, the `gam` procedure becomes infeasible when a large number of predictors are used. Also when many predictors are used the computational time for `mgcv` may rule out its use on large data mining problems. BRUTO was faster than KOW, however the classification performance enjoyed by BRUTO on the simulated datasets did not seem to carry on to the *PID* and *Spam* datasets for which it fails miserably.

Figure 2 illustrates cross-sections from the fitted additive function $\widehat{\eta}(\mathbf{x})$ for the *Spam* dataset. The cross-section for each predictor corresponds to all other predictors set to their medians. When the curve moves above the zero line e-mails are more likely to be spam and when the curve moves below the zero line e-mails are less likely to be spam e-mails. For example when the proportion of number of times *business* is used to the total number of words is less than 2 there is nearly no effect but after the proportion is above 2 the probability that the e-mail is spam appears to increase (roughly) linearly. Curves that hover around the zero curve, for example the variable *our*, do not have a large effect on the predicted value.

5 Discussion

The KOW classification algorithm represents an appealing application of statistical inferential techniques to data mining and related problems. Parsimony and interpretability are delivered using likelihood-based inference ideas. Speed is obtained via Laplace approximation. Generalised linear mixed models, which have mainly been the providence of regression-type analyses of data from biostatistical studies, can be seen to have wider applicability.

While, in this article, we have concentrated on classification and logistic mixed models the methods presented are directly extendible to more general mixed models; e.g. those appropriate for count data, and non-classification problems such as variable selec-

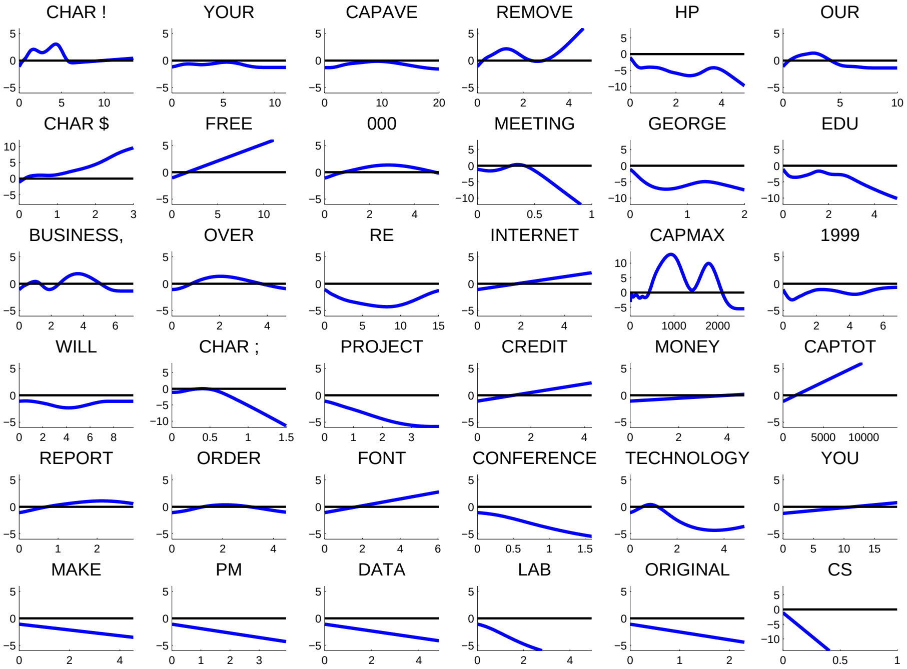

Figure 2: *A plot of fitted model for the Spam dataset using the predictors as chosen by the KOW algorithm.*

tion in generalised additive model analyses. We envisage several useful by-products of the KOW algorithm for semiparametric analysis of multi-predictor data.

Acknowledgements

The first author acknowledges support from the Deutsche Forschungsgemeinschaft (Project D 310 122 40). The second and third authors acknowledge support from the Australian Research Council (Project DP0556518).

Appendix A: Rao Statistic Derivations

Derivation of the R_{β_k} *expression*

Vector calculus methods (e.g. Wand, 2002) applied to the right hand side of (6) lead to

$$\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\beta_k) \simeq \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}}\right)^T [\mathbf{X}\ \mathbf{x}_k].$$

Therefore the approximate numerator of R_{β_k} is the last entry of this vector with β_k set to zero:

$$[\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)]_{p+1} \simeq \mathbf{x}_k^T\left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}\right).$$

The negative Hessian is approximately

$$-\mathsf{H}_{(\boldsymbol{\beta},\beta_k)}\ell(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\beta_k) \simeq [\mathbf{X}\ \mathbf{x}_k]^T \mathrm{diag}\left(\frac{e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}}{(1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}})^2}\right)[\mathbf{X}\ \mathbf{x}_k].$$

The approximate denominator of R_{β_k} is the square root of the bottom right entry of this matrix with β_k set to zero and $\boldsymbol{\beta}$ set to its estimate at the current model. Standard results on the inverse of partitioned matrices lead to (7).

Derivation of the $R_{\sigma_k^2}$ expression

Let $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2}$ and $\mathbf{E}_1,\ldots,\mathbf{E}_{r+1}$ be as defined in Section 3.2. Then vector calculus methods (e.g. Wand, 2002) applied to the right hand side of (6) lead to

$$[\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell_{\text{Laplace}}(\boldsymbol{\beta},\boldsymbol{\sigma}^2,\sigma_k^2)]_j=-\tfrac{1}{2}\text{tr}\{\mathbf{E}_j(\mathbf{I}+\widetilde{\mathbf{Z}}^T\mathbf{W}_{\boldsymbol{\beta},\widehat{\mathbf{u}},\widehat{\mathbf{u}}_k}\widetilde{\mathbf{Z}}\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2})^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\boldsymbol{\beta},\widehat{\mathbf{u}},\widehat{\mathbf{u}}_k}\widetilde{\mathbf{Z}}\}+\tfrac{1}{2}\|\widehat{\mathbf{u}}_j/\sigma_j^2\|^2. \tag{11}$$

Noting that $(\widehat{\mathbf{u}},\widehat{\mathbf{u}}_k)$ maximise (9) we get the relationships

$$\mathbf{G}_{\boldsymbol{\sigma}^2}\mathbf{Z}\left(\mathbf{y}-\frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k}}\right)=\widehat{\mathbf{u}}\qquad\text{and}\qquad\sigma_k^2\mathbf{Z}_k\left(\mathbf{y}-\frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}+\mathbf{Z}_k\widehat{\mathbf{u}}_k}}\right)=\widehat{\mathbf{u}}_k.$$

The second of these gives $\|\widehat{\mathbf{u}}_k/\sigma_k^2\|^2=\|\mathbf{Z}_k^T\left(\mathbf{y}-\frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}\right)\|^2$. Substitution of this equation into (11) and setting $(\boldsymbol{\beta},\boldsymbol{\sigma}^2)=(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2)$, $\sigma_k^2=0$, $\mathbf{u}_k=\mathbf{0}$ and $j=r+1$ then leads to

$$\begin{aligned}[\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell_{\text{Laplace}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\sigma}}^2,0)]_{r+1}=&-\tfrac{1}{2}\text{tr}\{\mathbf{E}_{r+1}\{\mathbf{I}+\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\,\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2},\mathbf{0})\}^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\}\\&+\tfrac{1}{2}\left\|\mathbf{Z}_k^T\left(\mathbf{y}-\tfrac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{1+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}\right)\right\|^2.\end{aligned}$$

Note that $\{\mathbf{I}+\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}\,\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2},\mathbf{0})\}^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\widetilde{\mathbf{Z}}$ has the explicit expression

$$\left[\begin{array}{cc}\mathbf{I}+\mathbf{Z}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2} & \mathbf{0}\\ \mathbf{Z}_k^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2} & \mathbf{I}\end{array}\right]^{-1}\left[\begin{array}{cc}\mathbf{Z}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z} & \mathbf{Z}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z}_k\\ \mathbf{Z}_k^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z} & \mathbf{Z}_k^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{Z}_k\end{array}\right].$$

The expression for $R^{\text{num}}_{\sigma_k^2}$ then follows from standard results on the inverse of a partitioned matrix and some straightforward matrix algebra.

We now provide a computationally efficient expression for $R^{\text{den}}_{\sigma_k^2}$. First, partition $\mathcal{K}(\boldsymbol{\sigma}^2,\sigma_k^2)$ as

$$\mathcal{K}(\boldsymbol{\sigma}^2,\sigma_k^2)=\left[\begin{array}{cc}\mathcal{K}_{11}(\boldsymbol{\sigma}^2,\sigma_k^2) & \mathcal{K}_{12}(\boldsymbol{\sigma}^2,\sigma_k^2)\\ \mathcal{K}_{12}(\boldsymbol{\sigma}^2,\sigma_k^2)^T & \mathcal{K}_{22}(\boldsymbol{\sigma}^2,\sigma_k^2)\end{array}\right]$$

where $\mathcal{K}_{11}(\boldsymbol{\sigma}^2,\sigma_k^2)$ is the $r\times r$ upper left-hand block corresponding to the current model. Then

$$R^{\text{den}}_{\sigma_k^2}\simeq\{\mathcal{K}_{22}(\widehat{\boldsymbol{\sigma}}^2,0)-\mathcal{K}_{12}(\widehat{\boldsymbol{\sigma}}^2,0)^T\mathcal{K}_{11}(\widehat{\boldsymbol{\sigma}}^2,0)^{-1}\mathcal{K}_{12}(\widehat{\boldsymbol{\sigma}}^2,0)\}^{1/2}.$$

Note that the matrix inversion $\mathcal{K}_{11}(\widehat{\boldsymbol{\sigma}}^2,0)^{-1}$ needs only be done once for the current model.

Appendix B: Variable Selection Via p-values

Let $\boldsymbol{\varepsilon}\equiv\mathbf{y}-\mu(\boldsymbol{\eta})$ where $\mu(\boldsymbol{\eta})\equiv\exp(\boldsymbol{\eta})/\{1+\exp(\boldsymbol{\eta})\}$ and set $\widehat{\boldsymbol{\varepsilon}}\equiv\mathbf{y}-\mu(\widehat{\boldsymbol{\eta}})$. The stochastic component of $R^2_{\beta_k}$ is $Q_k\equiv\widehat{\boldsymbol{\varepsilon}}^T\mathbf{x}_k^T\mathbf{x}_k\widehat{\boldsymbol{\varepsilon}}$. Note that $\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}\simeq(\mathbf{K}+\mathbf{B})^{-1}(\mathbf{C}^T\boldsymbol{\varepsilon}-\mathbf{B}\boldsymbol{\nu})$, where the notation of Section 2 is being used. First order expansion then yields

$$\begin{aligned}\widehat{\boldsymbol{\varepsilon}}&\simeq\boldsymbol{\varepsilon}-\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{C}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})\\&\simeq\left\{\mathbf{I}-\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}}}\mathbf{C}(\mathbf{K}+\mathbf{B})^{-1}\mathbf{C}^T\right\}\boldsymbol{\varepsilon}+\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}}}\mathbf{C}(\mathbf{K}+\mathbf{B})^{-1}\mathbf{B}\boldsymbol{\nu}=\mathbf{M}\boldsymbol{\varepsilon}+\mathbf{R}\mathbf{u}\end{aligned}$$

where $\mathbf{M} \equiv \left\{\mathbf{I} - \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}}}\mathbf{C}(\mathbf{K}+\mathbf{B})^{-1}\mathbf{C}^T\right\}\boldsymbol{\varepsilon},\ \mathbf{R} \equiv \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}}}\mathbf{C}\left[\begin{array}{c}\left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{12}\\ \left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{22}\end{array}\right]\mathbf{G}^{\mathbf{-1}}_{\boldsymbol{\sigma}^{\mathbf{2}}}$ and

$$(\mathbf{K}+\mathbf{B})^{-1} = \left[\begin{array}{cc}\left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{11} & \left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{12}\\ \left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{21} & \left\{(\mathbf{K}+\mathbf{B})^{-1}\right\}_{22}\end{array}\right],$$

with partitions corresponding to the components $\boldsymbol{\beta}$ and $\mathbf{u}$ respectively. We then obtain the approximation

$$Q_k \simeq \left[\boldsymbol{\varepsilon}^T, \mathbf{u}^T\right]\left[\begin{array}{c}\mathbf{M}^T\\ \mathbf{R}^T\end{array}\right]\mathbf{x}_k\mathbf{x}_k^T\left[\mathbf{M}, \mathbf{R}\right]\left[\begin{array}{c}\boldsymbol{\varepsilon}\\ \mathbf{u}\end{array}\right].$$

Making use of the simplifying assumption $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}_{\boldsymbol{\beta},\boldsymbol{u}})$ and noting that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2})$ Q_k is approximately distributed as $\rho\chi^2_1$ where $\rho \equiv \boldsymbol{x}_k^T\left(\mathbf{M}\mathbf{W}_{\boldsymbol{\beta},\mathbf{u}}\mathbf{M}^T + \mathbf{R}\mathbf{G}_{\boldsymbol{\sigma}^2}\mathbf{R}^T\right)\mathbf{x}_k$ and χ^2_1 denotes the chi-squared distribution with one degree of freedom (e.g. Imhof, 1961). This approximative distribution can be employed to calculate a p-value for $R^2_{\beta_k}$ based on the assumption that if $\mathbf{x}_k$ is not in the model then $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Similiarly, $R_{\sigma^2_k}$ contains the quadratic form $\widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{Z}_k^T\boldsymbol{Z}_k\widehat{\boldsymbol{\varepsilon}}$ which can be decomposed in the same way leading to a mixture of chi-squared distributions as derived in Zhang & Lin (2003).

References

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. New York: Cambridge University Press.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statistical Science*, **16**, 199–231.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth Publishing.

Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Buja, A., Hastie, T. & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**, 453–510.

Chambers, J. M. & Hastie, T. J. (1992). *Statistical Models in S.* New York: Chapman & Hall.

Cox, D. & Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression. *Annals of the Institute of Statistical Mathematics*, **41**, 383–400.

Durbán, M. & Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, **18**, 263–292.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics*, **50**, 640–652.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Hand, D.J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science*, **21**, 1–34.

Hastie, T. (2006). `gam 0.97`. R package. `http://cran.r-project.org`.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer-Verlag.

Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

Kauermann, G., Krivobokova, T. & Fahrmeir, L. (2008). Some Asymptotic Results on Generalized Penalized Spline Smoothing. *Journal of the Royal Statistical Society, Series B*, (to appear).

Kooperberg, C., Bose, S. & Stone, C.J. (1997). Polychotomous regression. *Journal of the American Statistical Association*, **92**, 117–127.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326.

McCulloch, C.E., & Searle, S.R. (2000). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

Ormerod, J.T. (2008). On Semiparametric Regression and Data Mining. *PhD Thesis.* School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons.

Ruppert, D., Wand, M. P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, **25**, 1371–1425.

Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effect models. *Biometrika*, **92**, 351–370.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Wager, C., Vaida, F. & Kauermann, G. (2007). Model selection for P-spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics*, **49**, 173–190.

Wakefield, J.C., Best, N.G. & Waller, L. (2000). Bayesian approaches to disease mapping. In Spatial Epidemiology, eds. Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. Oxford: Oxford University Press. 104–127.

Wand, M.P. (2002). Vector differential calculus in statistics. *The American Statistician*, **56**, 55–62.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223–249.

Wand, M.P. (2007). Fisher information for generalised linear mixed models. *Journal of Multivariate Analysis*, **98**, 1412–1416.

Wand, M.P. & Ormerod, J.T. (2008). On Semiparametric Regression with O'Sullivan penalised splines. *Australian and New Zealand Journal of Statistics*, **50**, 179-198.

Welham, S.J., Cullis, B.R., Kenward, M.G. & Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian and New Zealand Journal of Statistics*, **49**, 1–23.

Wood, S.N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society, Series B*, **65**, 95–114.

Wood, S.N. (2006). `mgcv 1.3`. R package. `http://cran.r-project.org`.

Yau, P., Kohn, R. & Wood, S. (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, **12**, 1–32.

Zhang, D. & Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**, 57–74.

Zhao, Y., Staudenmayer, J., Coull, B.A. & Wand, M.P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, **21**, 35–51.