Vector Generalized Additive Models
Author(s): T. W. Yee and C. J. Wild
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 3 (1996), pp. 481-493
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2345888
Accessed: 22-02-2019 15:27 UTC

J. R. Statist. Soc. B (1996)
58, *No.* 3, *pp.* 481–493

Vector Generalized Additive Models

By T. W. YEE† and C. J. WILD

University of Auckland, New Zealand

[Received April 1994. Final revision May 1995]

SUMMARY

Vector smoothing is used to extend the class of generalized additive models in a very natural way to include a class of multivariate regression models. The resulting models are called 'vector generalized additive models'. The class of models for which the methodology gives generalized additive extensions includes the multiple logistic regression model for nominal responses, the continuation ratio model and the proportional and non-proportional odds models for ordinal responses, and the bivariate probit and bivariate logistic models for correlated binary responses. They may also be applied to generalized estimating equations.

Keywords: BACKFITTING ALGORITHM; GENERALIZED ADDITIVE MODELS; NONPARAMETRIC MULTIVARIATE REGRESSION; SMOOTHING; VECTOR SPLINES

1. INTRODUCTION

Generalized linear models (GLMs) (Nelder and Wedderburn, 1972), where y has a distribution in the exponential family and the mean μ of y is related to p covariates $\mathbf{x} = (x_1, \ldots, x_p)^T$ by

$$g(\mu) = \eta(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

have been extended to form the class of generalized additive models (GAMs) in which

$$g(\mu) = \eta(\mathbf{x}) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p),$$

the f_j being arbitrary smooth functions. These were proposed by Hastie and Tibshirani in the mid-1980s, and a complete treatment can be found in Hastie and Tibshirani (1990). With GAMs, instead of constraining the relationship between each x_j and η to be linear as in GLMs, the relationship is merely constrained to be smooth. This allows non-linear features of the data to be revealed to the analyst automatically.

The usefulness of generalized additive modelling techniques has become well established, and practical use of these methods has spread widely. The arguments that have been used to establish the usefulness of additive extensions to GLMs (i.e. GAMs) apply with equal force to any class of models containing linear regression components. This paper proposes additive extensions to other classes of models with this feature. The methods proposed are very much within the GAM tradition and thus inherit most of the strengths and also the weaknesses of generalized additive modelling.

†*Address for correspondence*: Clinical Trials Research Unit, Department of Medicine, University of Auckland, Private Bag 92019, Auckland, New Zealand.
E-mail: yee@ctru.auckland.ac.nz

 0035–9246/96/58481

Suppose that for each individual under study a q-dimensional response vector $\mathbf{y}$ ($q \geqslant 1$) and a p-dimensional covariate vector $\mathbf{x}$ are observed. Vector GAMs (VGAMs) are defined as the additive model extension of a vector GLM, where a vector GLM is any model for which the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is of the form

$$f(\mathbf{y}|\mathbf{x};\, \boldsymbol{\beta}) = h(\mathbf{y},\, \eta_1,\, \ldots,\, \eta_M) \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}},\, \ldots,\, \boldsymbol{\beta}_M^{\mathrm{T}})^{\mathrm{T}}$ and $\eta_j = \eta_j(\mathbf{x}) = \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}$ for some function $h(\,)$. An example of a vector GLM is a bivariate probit model for correlated binary response variables y_1 and y_2 where

$$\Pr(y_j = 1|\mathbf{x}) = \Phi\{\eta_j(\mathbf{x})\}, \qquad j = 1,\, 2,$$

$$\Pr(y_1 = 1,\, y_2 = 1|\mathbf{x}) = \Phi_2\left\{\eta_1(\mathbf{x}),\, \eta_2(\mathbf{x});\, \rho = \frac{\exp\,\eta_3(\mathbf{x}) - 1}{\exp\,\eta_3(\mathbf{x}) + 1}\right\}.$$

Here, the correlation parameter ρ is modelled as a function of the covariates, $\Phi(\,)$ is the distribution function of a standard normal distribution and $\Phi_2(\ ,\ ;\, \rho)$ is the distribution function of a bivariate normal with zero means, unit variances and correlation ρ. Frequently, as in the above example, M does not coincide with the dimension of $\mathbf{y}$. By 'the additive model extension' it is meant that, for all j, $\eta_j(\mathbf{x}) = \beta_{0j} + \beta_{1j}x_1 + \ldots + \beta_{pj}x_p$ is replaced by

$$\eta_j(\mathbf{x}) = \beta_{(j)0} + f_{(j)1}(x_1) + \ldots + f_{(j)p}(x_p), \tag{2}$$

a sum of smooth functions of the individual covariates, just as with ordinary GAMs. Thus, the methods of this paper extend the class of models for which additive model extensions can be handled from models in which the likelihood depends on $\boldsymbol{\beta}$ only through a single linear predictor $\eta(\mathbf{x}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ to the class of models in which there are several.

Examples of models with this structure include multivariate linear regression with covariance matrices that can be treated as known, the bivariate logistic (or logit) model (see McCullagh and Nelder (1989), section 6.5.6, and Palmgren (1989)) and the bivariate probit model (Ashford and Sowden, 1970) for regression with correlated binary responses, the multiple logistic regression (or multinomial logit) model introduced by Nerlove and Press (1973) for regression with unordered categorical responses, the continuation ratio model (see Armstrong and Sloan (1989)), the proportional odds model (McCullagh, 1980) and non-proportional odds models (see Peterson (1990)). The methods of Section 4 allow us to extend the range of application from model (1) to models which also include parameters which are not in the form of linear predictors. Two examples of such parameters are the scale parameter σ in multiple regression for scalar y when it is wished to model both the mean and the variance as $E(y) = g\{\eta_1(\mathbf{x})\}$, $\mathrm{var}(y) = \sigma^2\, V\{\eta_2(\mathbf{x})\}$, and the elements of the unknown common covariance matrix $\boldsymbol{\Sigma}$ in multivariate linear regression.

2. VECTOR SPLINES

Smoothing splines and other scatterplot smoothers are discussed in Hastie and Tibshirani (1990), chapters 2 and 3. The methods of this paper depend critically on a

generalization of the cubic spline called a 'vector spline' (Fessler, 1991). Consider a vector response $\mathbf{y}_i$ of dimension M at each value of a scalar x_i, assumed to be a realization from the vector measurement model

$$\mathbf{y}_i = \mathbf{f}(x_i) + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,$$
$$\mathbf{f}(x_i), \boldsymbol{\epsilon}_i, \mathbf{y}_i \in \mathbf{R}^M, \quad E\{\boldsymbol{\epsilon}_i\} = \mathbf{0}, \quad E\{\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j^{\mathrm{T}}\} = \delta_{ij} \boldsymbol{\Sigma}_i$$

where $\boldsymbol{\Sigma}_i$ are known symmetric and positive definite error covariances. The smooth vector function $\mathbf{f}(x)$, written $(f_1(x), \ldots, f_M(x))^{\mathrm{T}}$, can be estimated in this vector smoothing problem by minimizing the generalized least squares criterion penalized for lack of smoothness in the component functions

$$\sum_{i=1}^{n} \{\mathbf{y}_i - \mathbf{f}(x_i)\}^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \{\mathbf{y}_i - \mathbf{f}(x_i)\} + \sum_{j=1}^{M} \lambda_j \int f_j''(t)^2 \, \mathrm{d}t. \tag{3}$$

Each component function f_j has a non-negative smoothing parameter λ_j which operates as with an ordinary cubic spline. The inverse of the variance–covariance matrix of $\mathbf{y}_i$ will be referred to as a weight matrix and written $\mathbf{W}_i$. Fessler (1991) has implemented a method of solution in an $O(nM^3)$ C program called VSPLINE. To fit VGAMs more effectively, we have enhanced a subset of VSPLINE called YEE-SPLINE.

As with ordinary splines, expression (3) can be translated into terms of the values of the vector function $\mathbf{f}$ at the observed x-values. If the xs are ordered so that $x_1 < x_2 < \ldots < x_n$ and $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{f} = (f_1(x_1), \ldots, f_M(x_1), \ldots, f_1(x_n), \ldots, f_M(x_n))^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n)$, then the penalized least squares criterion (3) is equivalent to

$$(\mathbf{y} - \mathbf{f})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{f}) + \mathbf{f}^{\mathrm{T}} \mathbf{K} \mathbf{f}$$

for some matrix $\mathbf{K}$ not depending on $\mathbf{y}$. This quadratic in $\mathbf{f}$ is minimized when

$$\hat{\mathbf{f}} = \mathbf{A}(\boldsymbol{\lambda}) \mathbf{y},$$

with

$$\mathbf{A}(\boldsymbol{\lambda}) = (\mathbf{I}_{nM} + \boldsymbol{\Sigma}\mathbf{K})^{-1}$$

being the so-called *influence* or *smoother matrix*. A vector spline is thus a linear smoother, a fact with important theoretical consequences. The theory of vector splines is further discussed by Fessler (1991) and Yee (1993). The latter has found that the results of Buja *et al.* (1989), which give a theoretical underpinning to GAMs, extend to cover vector splines and hence apply to VGAMs as well. Degrees of freedom and standard errors of vector smoothers are discussed in Sections 3.1 and 3.2.

3. VECTOR GENERALIZED ADDITIVE MODELS

Suppose that $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ are observations on n 'individuals' where the $\mathbf{y}_i$s are conditionally independent given the $\mathbf{x}_i$s. In what follows $\mathbf{x}_i$ is also used, without notational distinction, to represent $(1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$ where the vector has been augmented by

a unit term to allow for the intercept term in all the linear models. Recall also the partition of the $M(p+1)$-vector $\boldsymbol{\beta}$ into M $(p+1)$-vectors, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_M^{\mathrm{T}})^{\mathrm{T}}$.

For models of the form (1), the log-likelihood can be expressed in the form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i\{\eta_1(\mathbf{x}_i), \ldots, \eta_M(\mathbf{x}_i)\},$$

where $\eta_j = \eta_j(\mathbf{x}_i) = \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i$. Let $\mathbf{U}(\boldsymbol{\beta}) = \partial l/\partial\boldsymbol{\beta}$ denote the score vector for the model and $\mathcal{J}(\boldsymbol{\beta}) = -\partial^2 l/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^{\mathrm{T}}$ the (observed) information matrix. The Newton–Raphson algorithm for maximizing the likelihood is

$$\boldsymbol{\beta}^{(a+1)} = \boldsymbol{\beta}^{(a)} + \mathcal{J}(\boldsymbol{\beta}^{(a)})^{-1}\mathbf{U}(\boldsymbol{\beta}^{(a)})$$

which, in this case, can be written in iteratively reweighted least squares form as

$$\boldsymbol{\beta}^{(a+1)} = \left(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i\,\mathbf{z}_i\right). \tag{4}$$

Here $\mathbf{X}_i$ is an $M \times M(p+1)$ block diagonal matrix for which the diagonal blocks are copies of $\mathbf{x}_i^{\mathrm{T}}$, i.e. $\mathbf{X}_i = \mathrm{diag}(\mathbf{x}_i^{\mathrm{T}}, \ldots, \mathbf{x}_i^{\mathrm{T}})$, $\mathbf{W}_i$ is an $M \times M$ matrix with (j, k)th element

$$(\mathbf{W}_i)_{jk} = -\frac{\partial^2 l_i}{\partial\eta_j\,\partial\eta_k},$$

and $\mathbf{z}_i$ is an M-vector given by $\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta}^{(a)} + \mathbf{W}_i^{-1}\mathbf{d}_i$, where $\mathbf{d}_i$ has jth element $(\mathbf{d}_i)_j = \partial l_i/\partial\eta_j$. Strictly, $\mathbf{z}_i$, $\mathbf{d}_i$ and $\mathbf{W}$ should be written $\mathbf{z}_i^{(a)}$, $\mathbf{d}_i^{(a)}$ and $\mathbf{W}_i^{(a)}$ as they depend on the current iterate, but the superscript is suppressed for simplicity.

Above, $\boldsymbol{\beta}^{(a+1)}$ is the solution to the generalized least squares problem: minimize with respect to $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} (\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}_i(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}) = \sum_{i=1}^{n} \{\mathbf{z}_i - \boldsymbol{\eta}(\mathbf{x}_i)\}^{\mathrm{T}}\mathbf{W}_i\{\mathbf{z}_i - \boldsymbol{\eta}(\mathbf{x}_i)\}, \tag{5}$$

where $\boldsymbol{\eta}(\mathbf{x}_i)$ is an M-vector with jth element $\eta_j(\mathbf{x}_i) = \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i$. Focusing on $\boldsymbol{\eta}$ instead of $\boldsymbol{\beta}$, it can be seen that, at each step, this procedure updates $\boldsymbol{\eta}$ by minimizing equation (5) subject to the constraint that each $\eta_j(\mathbf{x})$ is a linear function of $\mathbf{x}$.

The linear constraints can be relaxed and $\boldsymbol{\eta}$ updated by minimizing expression (5) penalized for lack of smoothness in the component additive functions by using penalties that are appropriate for cubic splines, i.e. minimize

$$\sum_{i=1}^{n} \{\mathbf{z}_i - \boldsymbol{\eta}(\mathbf{x}_i)\}^{\mathrm{T}}\mathbf{W}_i\{\mathbf{z}_i - \boldsymbol{\eta}(\mathbf{x}_i)\} + \sum_{j=1}^{M}\sum_{k=1}^{p} \lambda_{(j)k}\int f''_{(j)k}(t)^2\,\mathrm{d}t, \tag{6}$$

where $\eta_j(\mathbf{x}) = \beta_{(j)0} + f_{(j)1}(x_1) + \ldots + f_{(j)p}(x_p)$ (see equation (2)). As with ordinary GAMs, the problem can be solved for a single covariate and then that solution extended for several covariates by using the so-called 'backfitting algorithm'.

Suppose that there is only a single covariate (let it be covariate k). Temporarily ignoring the intercept terms in the regression, the minimization criterion is

$$\sum_{i=1}^{n} \{\mathbf{z}_i - \mathbf{f}_k(x_{ik})\}^{\mathrm{T}} \mathbf{W}_i \{\mathbf{z}_i - \mathbf{f}_k(x_{ik})\} + \sum_{j=1}^{M} \lambda_{(j)k} \int f''_{(j)k}(t)^2 \, \mathrm{d}t, \tag{7}$$

where $\mathbf{f}_k(x_{ik}) = (f_{(1)k}(x_{ik}), \ldots, f_{(M)k}(x_{ik}))^{\mathrm{T}}$. Its solution is a vector spline with $\boldsymbol{\Sigma}_i^{-1} = \mathbf{W}_i$.

When there is more than a single covariate, backfitting (Hastie and Tibshirani (1990), section 4.4) can be applied to the *adjusted dependent vector* $\mathbf{z}_i$ but with vector smoothing instead of the usual y-scalar smoothing. We term this the *vector backfitting algorithm*. It fits the *vector additive model*

$$E(\mathbf{y}_i) = \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \mathbf{f}_j(x_{ij})$$

to a vector response $\mathbf{y}$. The vector backfitting algorithm, applied to the $\mathbf{z}_i$, can be written as follows.

Initialize: $\boldsymbol{\beta}_0 = \text{average}\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ and $\mathbf{f}_k^{(0)} \equiv \mathbf{0}$, $k = 1, \ldots, p$.
Iterate: for $b = 1, 2 \ldots$

(a) iterate—for $k = 1, \ldots, p$
  (i) compute the vector function $\mathbf{f}^*_{[k]}$, as the weighted vector smooth function of observations $(x_{ik}, \mathbf{z}_i^{[k]})$, (weights $\mathbf{W}_i$), $i = 1, \ldots, n$,
  (ii) adjust the intercept

$$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 + \text{average}\{\mathbf{f}^*_{[k]}(x_{1k}), \ldots, \mathbf{f}^*_{[k]}(x_{nk})\},$$

  (iii) adjust

$$\mathbf{f}_k^{(b)}(x_{ik}) = \mathbf{f}^*_{[k]}(x_{ik}) - \text{average}\{\mathbf{f}^*_{[k]}(x_{1k}), \ldots, \mathbf{f}^*_{[k]}(x_{nk})\}, \qquad i = 1, \ldots, n,$$

(b) stop if all changes from $\mathbf{f}^{(b-1)}$ to $\mathbf{f}^{(b)}$ are sufficiently small.

In the above, $\mathbf{z}_i^{[k]}$ is $\mathbf{z}_i$ adjusted so that the effects of all covariates except for the kth are removed. The adjustment of $\boldsymbol{\beta}_0$ is necessary for identifiability rather than estimation (see later). The algorithm is of the Jacobi type if

$$\mathbf{z}_i^{[k]} = \mathbf{z}_i - \boldsymbol{\beta}_0 - \sum_{l \neq k} \mathbf{f}_l^{(b-1)}(x_{il}),$$

whereas it is of the Gauss–Seidel type if use is made of updated functions from the same (bth) backfitting iteration, where available, to adjust $\mathbf{z}_i$. In principle, smoothers other than vector splines could be used in the above, just as the choice of smoother is largely unimportant with ordinary GAMs. Where the smoother used is a vector spline, Yee (1993) modifies the argument of Hastie and Tibshirani (1990), section 6.5.2, to show that the vector backfitting algorithm provides maximum penalized likelihood estimates for the penalty terms given in expression (6). For a general treatise of the ideas used in this section, see Green and Silverman (1994).

Computationally, $\mathbf{W}_i^{-1}\mathbf{d}_i$ can be calculated via the Cholesky decomposition of $\mathbf{W}_i$. For some types of VGAMs, the Cholesky decomposition is also passed into YEE-SPLINE for the efficient computation of $\mathbf{W}_i^{-1}$ in the smoothing algorithm. For other types of VGAMs, $\mathbf{W}_i^{-1}\mathbf{d}_i$ has a simple form and can therefore be computed directly.

Like ordinary GAMs, inference for VGAMs is based heavily on the degrees of freedom and standard errors of smoothers which are reviewed in the next two subsections.

3.1. *Degrees of Freedom*

The degrees of freedom of a smooth function is a measure of the amount of smoothing done. Hastie and Tibshirani (1990), section 3.5, described three definitions for the degrees of freedom of a y-scalar smoother that are motivated by relationships which hold in ordinary linear regression theory. Let $\mathbf{A}$ denote the influence matrix of a linear vector smoother. Then, following the notation of Section 2, we define df, df^{var} and df^{err} as $\text{tr}(\mathbf{A})$, $\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\text{T}})$ and $\text{tr}(\mathbf{I}_{nM} - 2\mathbf{A} + \mathbf{A}^{\text{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A}\boldsymbol{\Sigma})$ respectively as the degrees of freedom for the overall smooth $\mathbf{f}$, analogously with the y-scalar case. However, there is also a need to define the degrees of freedom for each of the M component functions of $\mathbf{f}$ as well. Intuitively, the latter should be the sum of the former. Correspondingly, we define the degrees of freedom $\text{df}_{(j)}$, $\text{df}_{(j)}^{\text{var}}$ and $\text{df}_{(j)}^{\text{err}}$ of the jth component function $f_j(\ )$ as the sum of those diagonal elements corresponding to the jth component function. Motivation for these definitions and further degrees of freedom issues are explored in Yee (1993).

3.2. *Standard Errors*

Since $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$ for a linear vector smoother, the variance–covariance matrix of $\hat{\mathbf{f}}$ in Section 2 is $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\text{T}}$. In theory, this can be used to form pointwise standard error bands for each of the M component functions of $\mathbf{f}$, something which is particularly useful in preventing the overinterpretation of plots of estimated component functions. Unfortunately, it is impractical if nM is large as the complete influence matrix must be computed. However, for vector splines, Yee and Wild (1994a) discussed the use of the alternative $\mathbf{A}(\boldsymbol{\lambda})\boldsymbol{\Sigma}$, the vector spline equivalent of a Bayesian derivation for cubic splines (see Wahba (1983) and Silverman (1985)). This is computationally quite cheap (only the central $2M-1$ bands of $\mathbf{A}(\boldsymbol{\lambda})$ are required, and these may be efficiently computed by using the Hutchinson and de Hoog (1985) algorithm in $O(nM^3)$ operations), and appears to give similar results to the above. This type of standard error is implemented in YEE-SPLINE.

4. CONSTRAINTS ON FUNCTIONS

In the linear models adapted above, $\eta_j(\mathbf{x}_i)$ is customarily denoted $\eta_j(\mathbf{x}_i) = \mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}$ whereas we have written $\eta_j(\mathbf{x}_i) = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_j$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\text{T}}, \ldots, \boldsymbol{\beta}_M^{\text{T}})^{\text{T}}$. This notational change has important consequences. In the earlier notation, a different set of covariates can be used for each linear predictor. Also, it is possible to constrain how covariates act. Probably the types of constraints that are most useful either constrain the effects of a single covariate to be the same for different j (e.g. constraining the effect of an environmental exposure to affect the right eye and left eye in the same way) or constrain a covariate to have no effect at all on linear predictor $\eta_j(\mathbf{x}_i)$ for some particular j (we simply would not include that covariate in $\mathbf{x}_{ij}$).

These two capabilities, which can be very important when modelling, are not available in a straightforward way in the $\eta_j(\mathbf{x}_i) = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_j$ notation. Under this notation, every element in the covariate set for individual i, namely $\mathbf{x}_i$, can affect every linear

predictor. Moreover, the effects of the covariates for the various linear predictors indexed by j are independent — they are not constrained to be related to one another in any way. These considerations carry over to the smoothed predictors $\eta_j(\mathbf{x}_i) = \Sigma_{k=1}^{p} f_{(j)k}(x_{ik})$. Fortunately, there is a simple unified way of catering for the above which is now presented.

Both constraining the effects of a (kth) covariate to be the same for two or more η_js and to have no effect for some η_js can be written in the form $\mathbf{f}_k(x_{ik}) = \mathbf{B}_k \mathbf{f}_k^*(x_{ik})$ where $\mathbf{f}_k^*(x)$ is of lower dimension than $\mathbf{f}_k(x_{ik})$ and $\mathbf{B}_k$ is of full column rank. By extending this to restrictions of the form

$$\mathbf{f}_k(x_{ik}) = \mathbf{B}_k \mathbf{f}_k^*(x_{ik}) + \mathbf{c}_k,$$

for some vector $\mathbf{c}_k$, the following theory also allows for the incorporation of offsets.

The criterion that is minimized to fit the kth variable within the backfitting algorithm (compare expression (7)) is

$$\sum_{i=1}^{n} \{\mathbf{z}_i^{[k]} - \mathbf{B}_k \mathbf{f}_k^*(x_{ik}) - \mathbf{c}_k\}^{\mathrm{T}} \mathbf{W}_i \{\mathbf{z}_i^{[k]} - \mathbf{B}_k \mathbf{f}_k^*(x_{ik}) - \mathbf{c}_k\} + \sum_{j=1}^{M} \lambda_{(j)k} \int f''_{(j)k}(t)^2 \, \mathrm{d}t$$

$$= \text{constant} + \sum_{i=1}^{n} \{\mathbf{z}_i^{*[k]} - \mathbf{f}_k^*(x_{ik})\}^{\mathrm{T}} \mathbf{W}_{i,k}^* \{\mathbf{z}_i^{*[k]} - \mathbf{f}_k^*(x_{ik})\} + \sum_{j} \lambda_{(j)k} \int f^{*\prime\prime}_{(j)k}(t)^2 \, \mathrm{d}t \tag{8}$$

where

$$\mathbf{z}_i^{*[k]} = (\mathbf{B}_k^{\mathrm{T}} \mathbf{W}_i \mathbf{B}_k)^{-1} \mathbf{B}_k^{\mathrm{T}} \mathbf{W}_i (\mathbf{z}_i^{[k]} - \mathbf{c}_k)$$

and

$$\mathbf{W}_{i,k}^* = \mathbf{B}_k^{\mathrm{T}} \mathbf{W}_i \mathbf{B}_k.$$

The weight matrix $\mathbf{W}_i$ notation has been used here rather than Σ_i^{-1} to emphasize that in almost all applications weight matrices that have been calculated directly are used rather than variance–covariance matrices that must be inverted. Note that the relevant minimization problem is again a vector spline problem but with a transformed adjusted dependent vector and weight matrix. Also note that $\mathbf{f}_k(x_{ik}) = \mathbf{B}_k \mathbf{f}_k^*(x_{ik}) + \mathbf{c}_k$ must be reconstructed so that it can be used in computing future $\mathbf{z}_i^{[k]}$-values, and that $\mathbf{z}_i^{*[k]}$ can be obtained by a weighted least squares regression of $\mathbf{z}_i^{[k]} - \mathbf{c}_k$ on the columns of $\mathbf{B}_k$.

The above solution is direct if $\mathbf{c}_k$ is wholly known. If $\mathbf{c}_k$ contains any unknown parameters (for example if η_j is constrained to be independent of all covariates: $\eta_j = \beta_{(j)0}$), then these may be estimated by switching between updating $\mathbf{f}_k^*$ in terms of $\mathbf{c}_k$ (as above) and updating $\mathbf{c}_k$ in terms of $\mathbf{f}_k^*$. To obtain an expression for the latter, suppose that the component functions are ordered so that $\mathbf{c}_k = (\mathbf{c}_k^{*\mathrm{T}}, \boldsymbol{\alpha}_k^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{c}_k^*$ is known and $\boldsymbol{\alpha}_k$ is a vector of unknown parameters. Correspondingly partition

$$\mathbf{W}_i = \begin{pmatrix} \mathbf{W}_{i11} & \mathbf{W}_{i12} \\ \mathbf{W}_{i21} & \mathbf{W}_{i22} \end{pmatrix} \text{ and } \mathbf{z}_i^{[k]} - \mathbf{B}_k \mathbf{f}_k^*(x_{ik}) \text{ as } \begin{pmatrix} \boldsymbol{\gamma}_{ik} \\ \boldsymbol{\delta}_{ik} \end{pmatrix}.$$

Then setting the partial derivative of equation (8) with respect to the unknown parameters $\boldsymbol{\alpha}_k$ to $\mathbf{0}$ gives

$$\boldsymbol{\alpha}_k = \left(\sum_{i=1}^{n} \mathbf{W}_{i22} \right)^{-1} \left\{ \sum_{i=1}^{n} \mathbf{W}_{i22} \boldsymbol{\delta}_{ik} + \sum_{i=1}^{n} \mathbf{W}_{i21} (\boldsymbol{\gamma}_{ik} - \mathbf{c}_k^*) \right\}. \tag{9}$$

If some η_j are to be constrained to be identical, there is also need to constrain intercepts to be equal. This can be achieved very simply by averaging over the appropriate components of $\mathbf{z}_i$ in the initialization step of the vector backfitting algorithm. For example, if $\eta_j = \eta_k$, then assign both $\beta_{(j)0}$ and $\beta_{(k)0}$ the average over i of $\{z_{ij}, z_{ik}\}$.

A fuller exposition of the methods described in this section is given in Yee and Wild (1994b).

4.1. *Application to Proportional Odds Model*

The non-proportional odds model for an ordinal response falling into one of K categories is given by

$$\Pr(y \geq j|\mathbf{x}) = \frac{\exp \eta_j(\mathbf{x})}{1 + \exp \eta_j(\mathbf{x})}, \qquad j = 2, \ldots, K. \tag{10}$$

The proportional odds model of McCullagh (1980) is a special case where $\eta_j(\mathbf{x})$ is constrained to be of the form

$$\eta_j(\mathbf{x}) = \alpha_j + \eta(\mathbf{x}).$$

In a nonparametric setting, the constraints that are necessary to obtain the proportional odds model from the non-proportional odds model can be imposed by letting $\mathbf{B}_k = \mathbf{1}$, a $(K-1)$-vector of 1s, and $\mathbf{c}_k$ be a $(K-1)$-vector consisting of $K-2$ unknown parameters (one of the intercepts is absorbed into $f_k^*(x_{ik})$). Alternatively, $\mathbf{c}_k$ could consist wholly of unknown parameters and constrain $E(f_k^*) = 0$. For the latter, the update for f_k^* is an ordinary weighted smoother fitted to the scalar observations

$$(x_i^* = x_{ik},\ z_i^* = \{\mathbf{1}^T \mathbf{W}_i (\mathbf{z}_i^{[k]} - \mathbf{c}_k)\}/(\mathbf{1}^T \mathbf{W}_i \mathbf{1})),$$

with $\sigma_i^{*2} = (\mathbf{1}^T \mathbf{W}_i \mathbf{1})^{-1}$, $i = 1, \ldots, n$, whereas the update for $\mathbf{c}_k$, equation (9), simplifies to

$$\mathbf{c}_k = \left(\sum_{i=1}^{n} \mathbf{W}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{W}_i \{\mathbf{z}_i^{[k]} - f_k^*(x_{ik})\mathbf{1}\}.$$

The resulting method is identical with that given by Hastie and Tibshirani (1987). The difference in development is that they smoothed the constrained proportional odds model, whereas we apply constrained smoothing to the unconstrained non-proportional odds model. Although Hastie and Tibshirani's derivation is special and, as they stated, 'intricate' (Hastie and Tibshirani (1990), p. 220), the proportional odds additive model is a simple and natural specialization within the vector smoothing framework. Another advantage of our development is that it makes obvious the need to examine the proportionality or parallelism assumption; see Peterson and Harrell (1992) (this assumption is often also made in the continuation ratio model). Informal testing could be done approximately here by using a likelihood ratio test, as discussed in the following example.

A practical problem with the fitting of linear non-proportional odds models is that the $\beta_j^T \mathbf{x}$ intersect, making negative probabilities unavoidable for some $\mathbf{x}$-values. This problem can also occur with the nonparametric version. In the linear case, the problem is not serious if it occurs outside the range of the data. However, it is more likely to occur in the nonparametric case, especially if the amount of smoothing is small and if $\eta_j(\mathbf{x}) - \eta_k(\mathbf{x}) \approx 0$ for some j, k and $\mathbf{x}$.

5. EXAMPLE

A single covariate nonparametric non-proportional odds model is illustrated with a data set collected from a large New Zealand cross-sectional study of a working population conducted during 1992–93. A confidential questionnaire was administered to employees of Fletcher Challenge, a widely dispersed company, for providing employees with information about their risk of heart disease. As answering was voluntary, the response rate was 76%, resulting in a sample of 7988 workers who were aged between 17 and 65 years. Of these, 21% were females, and approximately 80% were 'New Zealand European'. The questionnaire fielded questions relating to health, life style, emotions and diet. Overall, the data can be considered as a reasonable representation of the New Zealand working population.

In this example, we shall model the way that the frequency of alcohol consumption varies with age and discuss the analysis of deviance with VGAMs. The response was measured on an ordinal scale with seven levels but for demonstration the categories are combined into $K = 3$ levels: $y = 1$ corresponds to less than once a month, $y = 2$ to between once a month and once a week, and $y = 3$ to more than once a week, on average. The model is

$$\text{logit}\{\Pr(y \geqslant j|\text{age})\} = f_j(\text{age}), \qquad j = 2, 3.$$

Each fitted function f_j is then simply the probability of $y \geqslant j$ as a function of age viewed on a logit scale. Fig. 1 displays the fitted functions and probabilities after fitting all functions with $\text{df}_{(j)} = 5$ degrees of freedom. Convergence was achieved

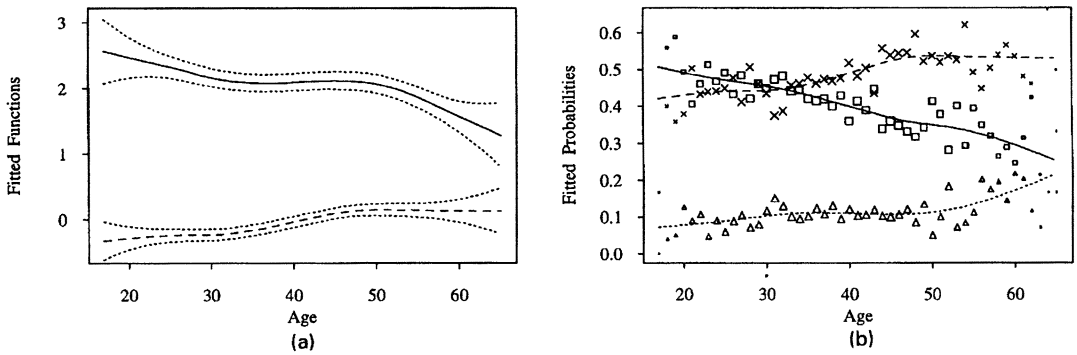

Fig. 1. Fitted nonparametric non-proportional odds model for alcohol frequency in a New Zealand working population against age (years): (a) fitting functions $\hat{f}_2$ (———) and $\hat{f}_3$ (– – – –) with ±2 standard errors curves (·····); (b) fitted probabilities $\Pr(y = 1|\text{age})$ (·····), $\Pr(y = 2|\text{age})$ (———) and $\Pr(y = 3|\text{age})$ (– – – –) (△, □, X, sample proportions with sizes proportional to sample size)

after five local scoring iterations, and none of the numerical problems described above were encountered.

The proportional odds assumption looks untenable as the two functions are not parallel. Function $\hat{f}_2$ is generally decreasing with a plateau at 35–50 years, showing a drop in drinking frequency for workers who are over 50 years of age. Possibly, it could be replaced by a cubic function in age. Function $\hat{f}_3$, which measures the highest frequency of drinking, is generally increasing and, interestingly, plateaus at 50 years of age. The fitted probabilities show that the proportion of those drinking less than once a month increases almost linearly with age.

As with ordinary GAMs, informal hypothesis tests can be performed by using the likelihood ratio test and its F-test modification. This has several important applications such as testing for linearity in the component functions, testing H_0: $f_j(x) \equiv 0$ and testing the proportional odds assumption described above. Even with ordinary GAMs, formal theory remains largely undeveloped. VGAMs have added complications such as the degrees of freedom being affected by the amount of correlation between the errors of the component functions. Intuitively, if the errors are independent so that the weight matrices are diagonal, then a likelihood ratio test for a VGAM is likely to behave very similarly in quality to that of an ordinary GAM. At the present stage, we have made use of the likelihood ratio test in a similar manner to how they have been used with ordinary GAMs as a very approximate test, and our limited experience (for instance in this example) has shown this to work reasonably well. Yee and Wild (1994a) addressed these issues with more generality and in greater detail.

The fitted function $\hat{f}_3$ appears roughly linear, and a test for this resulted in an increase in deviance of 9.81 for an increase in $\mathrm{df}_{(j)}^{\mathrm{err}}$ of 4.04. A likelihood ratio test would therefore yield a p-value of $\Pr(\chi^2_{4.04} > 9.81) = 0.045$ providing evidence of non-linearity. Following Hastie and Tibshirani (1990), section 10.2, we simulated the change in deviance in 500 data sets generated from the constrained model. Fig. 2

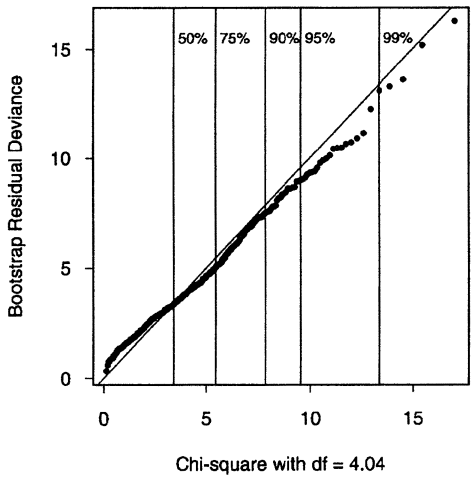

Fig. 2. Quantile–quantile plot of the distribution of 500 bootstrapped residual deviances against a $\chi^2_{4.04}$-distribution (———, $y = x$; |, quantiles of a $\chi^2_{4.04}$-distribution)

displays the quantile–quantile plot of the bootstrapped deviances. Like ordinary GAMs, the distribution appears to be approximately χ^2. 16 residual deviances were larger than 9.81, suggesting an empirical p-value of about 3.2%. Unfortunately, computing $\mathrm{df}_{(j)}^{\mathrm{err}}$ directly as done here is usually impractical owing to its expense. Our limited experience has shown that the approximation

$$\mathrm{tr}(2\mathbf{A} - \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A}\boldsymbol{\Sigma}) \approx 1.25\,\mathrm{tr}(\mathbf{A}) - 0.5M,$$

generalizing Hastie and Tibshirani (1990), appendix B, can be used componentwise to estimate $\mathrm{df}_{(j)}^{\mathrm{err}}$. The alternative method of testing for linearity in f_3 by fitting an ordinary logistic regression GAM resulted in a p-value of 0.007.

As a second example of a likelihood ratio test, the corresponding proportional odds model was fitted. This gave an increase of 76.14 in deviance for about 4 degrees of freedom, thus emphatically confirming the visual impression. The untenability of the proportional odds assumption was also confirmed by bootstrapping. Of 1000 bootstrap deviances obtained, all were much smaller than 76.14, suggesting a p-value considerably less than 0.001.

6. DISCUSSION

Many data analysts have found that the additive extensions to GLMs provided by GAMs (Hastie and Tibshirani, 1990) are invaluable for exploring data. The main intent of this paper was to demonstrate how vector smoothing can be used to extend this methodology to a multivariate setting in a very natural way. There is further breadth of applicability not discussed here; for example, in Wild and Yee (1996), we have extended the results of this paper to the Liang and Zeger type of estimating equation methods (Liang and Zeger, 1986; Prentice, 1988; Liang *et al.*, 1992) which are proving to be very useful for clustered and longitudinal data.

With the capability to compel individual covariates to have linear effects (either explicitly within the backfitting process or by manipulating the smoothing parameters), GAMs provide a seamless transition between parametric and nonparametric modelling. With very little effort, we can begin by allowing the data to reveal the nature of an effects curve in a fairly unconstrained way, and very often end by constraining it to have a suitable parametric form. Where this can be done with all the covariates, we can use standard parametric inference, but with the added assurance that comes from having explored wider possibilities. These capabilities should prove to be as useful in the broader class of models considered here as they have already proved to be for generalized additive extensions of GLMs.

In the estimation of the $\mathbf{f}_k$ in the vector backfitting algorithm, there is the interesting issue of whether vector smoothing may be replaced by ordinary y-scalar smoothing, and if so when, i.e. using $\mathrm{diag}(\boldsymbol{\Sigma}_i)$ instead of $\boldsymbol{\Sigma}_i$, so that the correlations between the η_j are ignored and the vector backfitting algorithm using vector splines simplifies to M separate applications of the ordinary backfitting algorithm with cubic splines. In such a situation, there is a need to know what is lost and how much. Clearly, if all the $\boldsymbol{\Sigma}_i$ are 'nearly' diagonal, there would be little loss of efficiency in the estimated smooth functions. Computationally, there is a gain. Assuming an equal number of iterations required for convergence, the cost of ordinary backfitting will be $M\,O(n) = O(Mn)$ using cubic splines, so that the cost of true vector backfitting

with vector splines is $O(M^3n)/O(Mn) = O(M^2)$ times greater. Although it may be argued that ignoring the correlation is not so important in nonparametric models, this is another area for future work. This idea of approximating $\mathbf{\Sigma}_i$ by $\text{diag}(\mathbf{\Sigma}_i)$ has been used before (Hastie and Tibshirani (1990), chapter 8).

ACKNOWLEDGEMENTS

We wish to thank Dr M. F. Hutchinson for referring us to vector splines and Dr J. A. Fessler for VSPLINE. The data were kindly provided by the management committee of the Fletcher Challenge–University of Auckland Heart and Health Study. Helpful comments from the Editor and referee are also acknowledged. The first author gratefully acknowledges the support of a Universities Grants Committee post-graduate scholarship, William Georgetti scholarship and C. Alma Baker post-graduate scholarship during his doctoral studies.

REFERENCES

Armstrong, B. G. and Sloan, M. (1989) Ordinal regression models for epidemiologic data. *Am. J. Epidem.*, **129**, 191–204.

Ashford, J. R. and Sowden, R. R. (1970) Multi-variate probit analysis. *Biometrics*, **26**, 535–546.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models (with discussion). *Ann. Statist.*, **17**, 453–555.

Fessler, J. A. (1991) Nonparametric fixed-interval smoothing with vector splines. *IEEE Trans. Signal Process.*, **39**, 852–859.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Hastie, T. and Tibshirani, R. (1987) Non-parametric logistic and proportional odds regression. *Appl. Statist.*, **36**, 260–276.

——— (1990) *Generalized Additive Models*. London: Chapman and Hall.

Hutchinson, M. F. and de Hoog, F. R. (1985) Smoothing noisy data with spline functions. *Numer. Math.*, **47**, 99–106.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc.* B, **54**, 3–40.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

Nerlove, M. and Press, S. J. (1973) Univariate and multivariate log-linear and logistic models. *Report R-1306-EDA/NIH*. Rand Corporation, Santa Monica.

Palmgren, J. (1989) Regression models for bivariate binary responses. *Technical Report 101*. Department of Biostatistics, University of Washington, Seattle.

Peterson, B. (1990) Ordinal regression models for epidemiologic data. *Am. J. Epidem.*, **131**, 745–746.

Peterson, B. and Harrell, Jr, F. E. (1992) Proportional odds model. *Biometrics*, **48**, 325–326.

Prentice, R. L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Statist. Soc.* B, **47**, 1–52.

Wahba, G. (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.

Wild, C. J. and Yee, T. W. (1996) Additive extensions to generalized estimating equation methods. *J. R. Statist. Soc.* B, **58**, in the press.

Yee, T. W. (1993) The analysis of binary data in quantitative plant ecology. *Doctoral Thesis*. Department of Mathematics and Statistics, University of Auckland, Auckland.

Yee, T. W. and Wild, C. J. (1994a) Vector splines and the vector additive model.

——— (1994b) Vector generalized additive models. *Technical Report STAT04*. Department of Statistics, University of Auckland, Auckland.