Smoothing Spline ANOVA for Multivariate Bernoulli Observations, With Application to Ophthalmology Data

Fangyu GAO, Grace WAHBA, Ronald KLEIN, and Barbara KLEIN

We combine a smoothing spline analysis of variance (SS-ANOVA) model and a log-linear model to build a partly flexible model for multivariate Bernoulli data. The joint distribution conditioning on the predictor variables is estimated. The log odds ratio is used to measure the association between outcome variables. A numerical scheme based on the block one-step successive over relaxation SOR–Newton-Ralphson algorithm is proposed to obtain an approximate solution for the variational problem. We extend the generalized approximate cross validation (GACV) and the randomized GACV for choosing smoothing parameters to the case of multivariate Bernoulli responses. The randomized version is fast and stable to compute and is used to adaptively select smoothing parameters in each block one-step SOR iteration. Approximate Bayesian confidence intervals are obtained for the flexible estimates of the conditional logit functions. Simulation studies are conducted to check the performance of the proposed method, using the comparative Kullback–Leibler distance as a yardstick. Finally, the model is applied to two-eye observational data from the Beaver Dam Eye Study, to examine the association of pigmentary abnormalities and various covariates.

KEY WORDS: Log-linear models; Multivariate responses; Odds ratio; Penalized likelihood; Repeated measurements; Representers; Reproducing kernel Hilbert space; Risk factor estimation; Semiparametric regression; Smoothing spline analysis of variance.

1. INTRODUCTION

Correlated Bernoulli outcomes may come from many applications. One motivation of this study is to develop a flexible model for analyzing typical data from ophthalmological studies. Usually, paired observations are available for both eyes of the same person. Both person-specific and eye-specific covariates may be available as predictor variables. The outcomes for the same person are expected to be correlated even after adjustment for the available predictor variables. This association reflects the consequence of unmeasured or unmeasurable genetic, behavioral or other risk factors. Other examples involving correlated outcomes include two-period cross-over designs (Jones and Kenward 1989), twin studies (Cessie and Houwelingen 1994) and typical longitudinal studies (Diggle, Liang, and Zeger 1994). Sometimes it is also of interest to model several closely related endpoints simultaneously. For example, Liang, Zeger, and Qaqish (1992) considered two endpoints from the Indonesian Children's Study, respiratory and diarrheal infections, in the same model.

We are interested in finding the relation between the outcome variables and the predictor variables, including conditional correlations between the outcome variables. Due to the complexity of biological processes, linear parametric assumptions on some scale, or even quadratic or cubic assumptions might not be adequate. When such an assumption is far away from the truth, the results obtained under it may even be misleading. Hence we are interested in building a flexible statistical model. A nonparametric model of the type considered in this article can also serve as an automated diagnostic tool for parametric fitting. The model should also have readily interpretable results for multivariate function estimate and a reasonable assessment of accuracy after the model has been fitted. This property is especially important for medical researchers, as the investigators are usually interested in understanding the cause of certain outcomes.

It is not enough to simply estimate the marginal distribution separately for the individual outcome variables. The dependence structure can be useful for the efficient estimation of the mean values, or it can be of direct scientific interest. This is a very active research topic, with numerous schemes proposed to study it. For example, Cox (1972) expressed the likelihood function in terms of the multivariate exponential family distribution, and Qu, Williams, Beck, and Goormastic (1987) considered conditional logistic models. Lipsitz, Laird, and Harrington (1991), Williamson, Kim, and Lipsitz (1995), and Heagerty and Zeger (1996) considered marginal models and used the (global) odds ratio as a measure of association. Liang et al. (1992) discussed the difference between log-linear and marginal models. Molenberghs and Ritter (1996) proposed a likelihood-based marginal model and established the connection with the second-order generalized estimating equations (GEEs). McCullagh and Nelder (1989) and Golenk and McCullagh (1995) proposed a multivariate marginal logistic regression model. Other related work has been done by Zhao, Prentice, and Self (1992), Fitzmaurice and Laird (1993), and Carey, Zeger, and Diggle (1993). Katz, Zeger, and Liang (1994) specifically discussed approaches to account for the association between fellow eyes.

Researchers have already realized the merit of a nonparametric approach to model correlated data. We note that additive smoothing splines with fixed smoothing parameters have been used by Wild and Yee (1996) and Yee and Wild (1996). These authors gave a nonparametric extension to both GEE and likelihood approaches. Heagerty and Zeger (1998) used the log odds ratio as a measurement of dependence and

Fangyu Gao is Senior Economist, Freddie Mac, McLean, VA 22102 (E-mail: *fangyu_gao@freddiemac.com*). Grace Wahba is Bascom Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *wahba@stat.wisc.edu*). Ronald Klein, MD, is Professor, and Barbara Klein, MD is Professor, Department of Ophthalmology, University of Wisconsin, Madison, WI 53705 (E-mail: *kleinr@epi.ophth.wisc.edu; kleinb@epi.ophth.wisc.edu*).

smoothing splines with fixed degrees of freedom to estimate it. Their model was fitted using GEEs. Lin and Zhang (1999) proposed a generalized additive mixed effects model and used smoothing splines to estimate the additive fixed effects term. Berhane and Tibshirani (1998) also provided a generalized additive model for the GEE method.

In this article we explore how to combine smoothing spline analysis of variance (SS-ANOVA) and log-linear models to build a partly flexible model for multivariate Bernoulli observations. We also propose a method for choosing the smoothing parameters adaptively in this multivariate outcome case. Classical log-linear models have been widely used to estimate joint conditional probabilities (see Bishop, Fienberg, and Holland 1975). SS-ANOVA provides a general framework for multivariate nonparametric function estimation that allows both main effects and interaction terms. The popular additive spline models discussed by Hastie and Tibshirani (1990) and implemented in S (Chambers and Hastie 1992) are special cases of SS-ANOVA models restricted to main effects. Hastie and Tibshirani (1990) also commented on interaction spaces.

SS-ANOVA models have been studied extensively for Gaussian data. Recently, Y. Lin (2000) obtained some general convergence results for the tensor product space ANOVA model and showed that the SS-ANOVA model achieves the optimal convergence rate. Wahba, Wang, Gu, Klein, and Klein (1995) gave a general setting for applying SS-ANOVA to data from exponential families. They successfully applied their method to analyze demographic medical data with univariate Bernoulli outcomes. X. Lin (1998) proposed using SS-ANOVA to analyze data with polychotomous responses. Wang (1998a) developed a mixed effects smoothing spline model for correlated Gaussian data. Brumback and Rice (1998) proposed smoothing spline models for correlated curves (see also Wang 1998b; Wang and Brown 1996). Interesting connections between SS-ANOVA models and graphical models, as discussed by Whittaker (1990), Jordan (1998), and others also may be observed.

It is of particular interest to us to explore the nonlinearity of the conditional logit functions. We use the log odds ratio to model the association among multivariate Bernoulli outcomes. We restrict the log odds ratios to simple parametric forms and estimate them using maximum likelihood estimation. However, the methods that we propose here can be easily generalized to estimate log odds ratios nonparametrically, and to other cases when the log-likelihood can be fully specified. An extension of the generalized approximate cross-validation (GACV) proposed by Xiang and Wahba (1996) to multivariate responses is derived for choosing the smoothing parameters. It is derived starting with an approximate leaving-out-one-subject argument, in contrast to the leaving-out-one-observation argument used by Xiang and Wahba (1996) in the original derivation of the GACV. Then the randomized trace technique for computing the GACV obtained by Wahba, Lin, Gao, Xiang, Klein, and Klein (1999) and Lin, Wahba, Xiang, Gao, Klein, and Klein (2000) is generalized to the extension of the GACV just derived. An efficient numerical approximation scheme and iterative algorithm involving a tailored block one-step successive over relaxation SOR–Newton-Raphson algorithm is proposed to compute the conditional logit functions and log odds ratios along with the smoothing parameter estimates. Simulation studies are presented to demonstrate the efficacy of the methods, and finally, the results are applied to two-eye observational data from the Beaver Dam Eye study to examine the association of pigmentary abnormalities and various covariates.

2. LOG-LINEAR MODEL FOR MULTIVARIATE BERNOULLI OBSERVATIONS

We first present the log-linear model for multivariate Bernoulli data. Assuming that there are J different endpoints, and K_j repeated measurements for the jth endpoint, let Y_{jk} denote the kth measurement of the jth endpoint. For example, in ophthalmological studies, we have two repeated measurements for each disease: left eye and right eye. In a typical longitudinal study, we have repeated measurements over time. $\mathbf{Y} = (Y_{jk}, j = 1, \ldots, J, k = 1, \ldots, K_j)$ is a multivariate Bernoulli outcome variable. Let $\mathbf{X}_{jk} = (X_{jk1}, X_{jk2}, \ldots, X_{jkD})$ be a vector of predictor variables ranging over the subset $\mathcal{X}$ of $\mathcal{R}^D$, where X_{jkd} denotes the dth predictor variable for the kth measurement of the jth endpoint. Some predictor variables may take different values for different measurements, whereas others may be the same for all Y_{jk}'s. For example, in ophthalmology studies both person-specific predictors and eye-specific predictors may be present. The person-specific predictors are the same for each person. For the eye-specific predictors, the set of predictor variables is the same, but the variables may take different values for the left and right eyes. We can treat observations from both eyes as correlated repeated measurements in our model. Let $\mathbf{X} = (\mathbf{X}_{jk}, j = 1, \ldots, J, k = 1, \ldots, K_j)$. Then $(\mathbf{X}, \mathbf{Y})$ is a pair of random vectors. For a response vector $\mathbf{y} = (y_{jk}, j = 1, \ldots, J, k = 1, \ldots, K_j)$, its joint probability distribution conditioning on the predictor variables $\mathbf{X}$ can be written as

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}|\mathbf{X}) &= \exp\Bigg\{\sum_{j=1}^{J}\sum_{k=1}^{K_j} f_{jk}y_{jk} + \sum_{j=1}^{J}\sum_{k_1<k_2} \alpha_{jk_1,\,jk_2}y_{jk_1}y_{jk_2} \\ &\quad + \sum_{j_1<j_2}\sum_{k_1,\,k_2} \alpha_{j_1k_1,\,j_2k_2}y_{j_1k_1}y_{j_2k_2} + \cdots \\ &\quad + \alpha_{11,12,\ldots,JK_J}y_{11}y_{12}\ldots y_{JK_J} - b(\mathbf{f}, \boldsymbol{\alpha})\Bigg\}, \end{aligned} \tag{1}$$

where

$$\begin{aligned} b(\mathbf{f}, \boldsymbol{\alpha}) = \log\Bigg(1 + \sum_{j,k} e^{f_{jk}} + \sum_{j_1,k_1}\sum_{j_2,k_2} e^{(f_{j_1k_1} + f_{j_2k_2} + \alpha_{j_1k_1,\,j_2k_2})} \\ + \cdots + e^{(\sum_{\text{all } f} f + \sum_{\text{all } \alpha} \alpha)}\Bigg). \end{aligned} \tag{2}$$

Let $M = \sum_{j=1}^{J} K_j$ be the length of the vector $\mathbf{Y}$. There are a total of $2^M - 1$ parameters: $(\mathbf{f}, \boldsymbol{\alpha}) = (f_{11}, f_{12}, \ldots, f_{JK_J}, \alpha_{11,12}, \ldots, \alpha_{11,12,\ldots,JK_J})$, which may depend on $\mathbf{X}$. The parameter space is unconstrained. They have straightforward interpretations in terms of conditional probabilities; for example,

$$f_{jk} = \text{logit}(P(Y_{jk} = 1|\mathbf{Y}^{(-jk)} = 0, \mathbf{X})) \tag{3}$$

is the conditional logit function;

$$\alpha_{j_1k_1,\,j_2k_2} = \log \mathrm{OR}(Y_{j_1k_1}, Y_{j_2k_2} | \mathbf{Y}^{(-j_1k_1,\,-j_2k_2)} = 0, \mathbf{X}) \quad (4)$$

is the conditional log odds ratio, which is a meaningful way to measure pairwise association; and

$$\begin{aligned}\alpha_{j_1k_1,\,j_2k_2,\,j_3k_3} = {} & \log \mathrm{OR}(Y_{j_1k_1}, Y_{j_2k_2} | Y_{j_3k_3} = 1, \\ & \mathbf{Y}^{(-j_1k_1,\,-j_2k_2,\,-j_3k_3)} = 0, \mathbf{X}) \\ & - \log \mathrm{OR}(Y_{j_1k_1}, Y_{j_2k_2} | Y_{j_3k_3} = 0, \\ & \mathbf{Y}^{(-j_1k_1,\,-j_2k_2,\,-j_3k_3)} = 0, \mathbf{X}) \end{aligned} \quad (5)$$

measures three-way association. Here $\mathbf{Y}^{(-*)}$ denotes the subset of vector $\mathbf{Y}$, except $\mathbf{Y}_*$, and

$$\begin{aligned}\mathrm{logit}(p) &= \log \frac{p}{1-p}, \\ \mathrm{OR}(v, w) &= \frac{P(v=1, w=1)P(v=0, w=0)}{P(v=1, w=0)P(v=0, w=1)}. \end{aligned} \quad (6)$$

Now assume that we have n independent observations $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n$, where $\mathbf{y}_i = (y_{i11}, y_{i12}, \ldots, y_{iJK_J})$ and $\mathbf{x}_i = (\mathbf{x}_{i11}, \mathbf{x}_{i12}, \ldots, \mathbf{x}_{iJK_J})$. Here y_{ijk} and $\mathbf{x}_{ijk} = (x_{ijk1}, x_{ijk2}, \ldots, x_{ijkD})$ are the outcome variable and predictor vector for the kth measurement of the jth endpoint of the ith subject. From now on we use $\mathbf{f}_i$ and $\boldsymbol{\alpha}_i$ to denote the parameters for the ith subject, while $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$. We can write the log-likelihood function based on the observed data

$$\begin{aligned}&\mathcal{L}(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha}) \\ &\quad = \sum_{i=1}^{n} \Bigg\{ \sum_{j=1}^{J} \sum_{k=1}^{K_j} f_{ijk} y_{ijk} + \sum_{j=1}^{J} \sum_{k_1<k_2} \alpha_{ijk_1,\,ijk_2} y_{ijk_1} y_{ijk_2} \\ &\qquad + \sum_{j_1<j_2} \sum_{k_1,\,k_2} \alpha_{ij_1k_1,\,ij_2k_2} y_{ij_1k_1} y_{ij_2k_2} + \cdots \\ &\qquad + \alpha_{i11,\,i12,\,\ldots,\,iJK_J} y_{i11} y_{i12} \cdots y_{iJK_J} - b(\mathbf{f}_i, \boldsymbol{\alpha}_i) \Bigg\}. \end{aligned} \quad (7)$$

We call (7) the log-linear model for multivariate logistic regression. Here f_{ijk} is the conditional logit function for the kth measurement of the jth endpoint of the ith subject. Scientifically, except for the possibility that the f_{ijk} may take different predictor values from measurement to measurement, there is little reason to believe they will take different functional forms for the same endpoint. Hence we can assume that $f_{ijk} = f_j(\mathbf{x}_{ijk})$. The same reasoning applies to the association terms; for example, we can assume that $\alpha_{ij_1k_1,\,ij_2k_2} = \alpha_{j_1j_2}(\mathbf{x}_{ij_1k_1}, \mathbf{x}_{ij_2k_2})$.

In practice, the number of parameters to be estimated can be reduced in many ways. For example, in many situations scientific interest is focused primarily on the conditional logit function f_{ijk} and log odds ratio $\alpha_{ij_1k_1,\,ij_2k_2}$, which measures pairwise association. The existence of three-way associations $\alpha_{ij_1k_1,\,ij_2k_2,\,ij_3k_3}$ and higher-order associations are usually difficult to verify in practical situations and may attract less scientific interest. Hence it is possible to set all higher-order associations to be 0 and to fit only a parsimonious model instead of the saturated one described in (7). The reduced model is a member of the quadratic exponential model of Zhao and Prentice (1990).

3. PENALIZED MULTIVARIATE LOGISTIC REGRESSION USING SMOOTHING SPLINE ANALYSIS OF VARIANCE

To simplify notation, we consider a parsimonious model here, by setting all higher-order associations than pairwise to be 0 in (7). Thus the negative log-likelihood function simplifies to

$$\begin{aligned}\mathcal{L}(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} l\Big(\mathbf{f}(\mathbf{x}_i), \boldsymbol{\alpha}(\mathbf{x}_i)\Big) = -\sum_{i=1}^{n} \Bigg\{ & \sum_{j=1}^{J} \sum_{k=1}^{K_j} f_j(\mathbf{x}_{ijk}) y_{ijk} \\ & + \sum_{j=1}^{J} \sum_{k_1<k_2} \alpha_{jj}(\mathbf{x}_{ijk_1}, \mathbf{x}_{ijk_2}) y_{ijk_1} y_{ijk_2} + \sum_{j_1<j_2} \sum_{k_1,\,k_2} \\ & \alpha_{j_1j_2}(\mathbf{x}_{ij_1k_1}, \mathbf{x}_{ij_2k_2}) y_{ij_1k_1} y_{ij_2k_2} - b(\mathbf{f}_i, \boldsymbol{\alpha}_i), \Bigg\}, \end{aligned} \quad (8)$$

where

$$\begin{aligned} b(\mathbf{f}_i, \boldsymbol{\alpha}_i) = \log\Bigg(& 1 + \sum_{j,\,k} \exp(f_j(\mathbf{x}_{ijk})) \\ & + \sum_{j_1,\,k_1} \sum_{j_2,\,k_2} \exp(f_{j_1}(\mathbf{x}_{ij_1k_1}) + f_{j_2}(\mathbf{x}_{ij_2k_2}) \\ & + \alpha_{j_1j_2}(\mathbf{x}_{ij_1k_1}, \mathbf{x}_{ij_2k_2})) + \cdots + \exp\Bigg(\sum_{j,\,k} f_j(\mathbf{x}_{ijk}) \\ & + \sum_{j_1,\,k_1} \sum_{j_2,\,k_2} \alpha_{j_1j_2}(\mathbf{x}_{ij_1k_1}, \mathbf{x}_{ij_2k_2}) \Bigg) \Bigg) \Bigg). \end{aligned} \quad (9)$$

We are interested in relaxing the parametric assumptions to build a flexible log-linear model. We are particularly interested in exploring the nonlinearity of the conditional logit functions f. We also could model the $\boldsymbol{\alpha}$ nonparametrically, but because a larger number of observations would be needed to estimate and tune many multivariate smooth functions nonparametrically, we assume in this article that the $\boldsymbol{\alpha}$ are of simple parametric form possibly depending on some set of parameters $\boldsymbol{\beta}$, and leave more general $\boldsymbol{\alpha}$ for future study. We use the penalized likelihood method to model the f_j nonparametrically. To relax the parametric assumptions, the penalized likelihood method (O'Sullivan 1983) assumes that the function to be estimated is smooth in some sense. This is done by assuming that the function to be estimated is in a given reproducing kernel Hilbert space of "smooth" functions, where "roughness" is measured by the size of some quadratic functional, typically a square norm or seminorm, and a roughness penalty is imposed in obtaining the estimate. The reproducing kernel Hilbert space theory (see Aronszajn 1950; Kimeldorf and Wahba 1971) can then used to characterize the solutions of very general penalized likelihood problems. We assume then that $f_j \in \mathcal{H}^j$, where $\mathcal{H}^j$ is a given reproducing kernel Hilbert space of functions on $\mathcal{X}$. Hence, $\mathbf{f} = (f_1, f_2, \ldots, f_J) \in \mathcal{H}^1 \times \cdots \times \mathcal{H}^J$. The penalized multivariate logistic regression estimate of $\mathbf{f}$ and $\boldsymbol{\alpha} =$

$(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{JJ})$ is the minimizer of the variational problem

$$\mathcal{L}_\Lambda(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} l(\mathbf{f}(\mathbf{x}_i), \boldsymbol{\alpha}(\mathbf{x}_i)) + \frac{n}{2}\mathcal{J}_\Lambda(f_1, \ldots, f_J), \quad (10)$$

where the first part is the negative log-likelihood, the second part is the roughness penalty, and $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_J)$ are the smoothing parameters that control the smoothness of the estimated f_j's. We assume an additive form of the penalty functional for simplicity and easy interpretation,

$$\mathcal{J}_\Lambda(f_1, \ldots, f_J) = \sum_{j=1}^{J} \mathcal{J}^j_{\Lambda_j}(f_j). \quad (11)$$

We consider the orthogonal decomposition $\mathcal{H}^j = \mathcal{H}^j_0 \oplus \mathcal{H}^j_1$. Here $\mathcal{H}^j_0$ is finite dimensional (the "parametric" part, usually polynomials), and $\mathcal{H}^j_1$ (the "smooth" part) is the orthocomplement of $\mathcal{H}^j_0$ in $\mathcal{H}^j$. The penalty functional will be related only to the smooth part of the functional, $\mathcal{J}^j_{\Lambda_j}(f_j) = \|P^j_1 f_j\|^2_{\Lambda_j}$, where P^j_1 is the orthogonal projection operator in $\mathcal{H}^j$ onto $\mathcal{H}^j_1$. The penalized likelihood has the expression

$$\mathcal{L}_\Lambda(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} l(\mathbf{f}(\mathbf{x}_i), \boldsymbol{\alpha}(\mathbf{x}_i)) + \frac{n}{2}\sum_{j=1}^{J} \|P^j_1 f_j\|^2_{\Lambda_j}. \quad (12)$$

The following theorem shows the existence and uniqueness of the solution to the variational problem (10). Letting $\mathcal{H}_0 = \mathcal{H}^1_0 \times \cdots \times \mathcal{H}^J_0$ denote the null space of $\mathcal{H}^1 \times \cdots \times \mathcal{H}^J$ with respect to the penalty function $\mathcal{J}_{\boldsymbol{\Lambda}}$, we have:

Theorem 1. If the minimizer of (12) exists in $\mathcal{H}_0$, then it uniquely exists in $\mathcal{H}^1 \times \cdots \times \mathcal{H}^J$.

Proof. See Appendix A.

If $\phi_\nu, \nu = 1, \ldots, M$ span $\mathcal{H}_0$, then this corresponds to the identifiability of the ordinary parametric model, which is a linear combination of the ϕ_ν. In the later examples, the ϕ_ν are low-degree polynomials.

3.1 Smoothing Spline Analysis of Variance Models for f_j

Here we briefly review some of the well-known properties of SS-ANOVA models. For any j $(1 \le j \le J)$, let $g = f_j$. Here g is assumed to be a function of D variables, $g = g(x_1, \ldots, x_D)$ with $x_d \in \mathcal{X}^{(d)}$, $d = 1, \ldots D$; thus $\mathbf{x} = (x_1, \ldots, x_D) \in \mathcal{X} = \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(D)}$. Given probability measures $d\mu_d$ on $\mathcal{X}^{(d)}$, define the averaging operators $\mathcal{E}_d$ on $\mathcal{X}$ as

$$(\mathcal{E}_d)g(\mathbf{x}) = \int_{\mathcal{X}^{(d)}} g(x_1, x_2, \ldots, x_D)\, d\mu_d(x_d). \quad (13)$$

These averaging operators define a unique (functional ANOVA) decomposition of g as

$$g(x_1, \ldots, x_D) = \mu + \sum_{d=1}^{D} g_d(x_d) + \sum_{d_1 < d_2} g_{d_1 d_2}(x_{d_1}, x_{d_2}) + \cdots + g_{1,\ldots,D}(x_1, \ldots, x_D), \quad (14)$$

where $\mu = (\prod_d \mathcal{E}_d)g$, $g_d = ((I - \mathcal{E}_d)\prod_{d_1 \ne d} \mathcal{E}_{d_1})g$, $g_{d_1 d_2} = ((I - \mathcal{E}_{d_1})(I - \mathcal{E}_{d_2})\prod_{d_3 \ne d_1, d_2} \mathcal{E}_{d_3})g$, and so forth. The averaging operators and norms on $\mathcal{H}^j$ can be chosen so that the components of this decomposition are projections of g onto orthogonal subspaces of $\mathcal{H}^j$ (see Gu and Wahba 1993; Wahba 1990; Wahba et al. 1995). In practice, the ANOVA decomposition in (14) will be truncated at some point. Assuming that we have already decided which subspaces (equivalently, components of g) will be included in our model $\mathcal{M}(\subset \mathcal{H}^j)$, and suppressing the superscript j on the subspaces for the rest of this section, we can regroup and write the model space as

$$\mathcal{M} = \mathcal{H}_0 \oplus (\mathcal{H}_{1,1} \oplus \cdots \oplus \mathcal{H}_{1,q}) = \mathcal{H}_0 \oplus \mathcal{H}_1. \quad (15)$$

Usually we will let $\mathcal{H}_0$ be a finite-dimensional space containing functions that will not be penalized. The norms on the composite $\mathcal{H}_{1,l}, 1 \le l \le q$ are the tensor product norms induced by the norms on the component subspaces, and $\|g\|^2_\Lambda = \|P_0 g\|^2 + \sum_{l=1}^{q} \lambda_l \|P_l g\|^2$, where P_l is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}_{1,l}$. Now we can use reproducing kernel Hilbert space methods to explicitly impose roughness penalties. For the penalty functional in (11), it can be further rewritten as

$$\mathcal{J}^j_{\Lambda_j}(f_j) = \sum_{l=1}^{q} \lambda_{jl} \|P^j_l f_j\|^2. \quad (16)$$

It is well known (see, e.g., Kimeldorf and Wahba 1971; Wahba 1990) that the minimizer of the variational problem (12) is in a particular finite-dimensional subspace spanned by a set of basis functions obtained from the reproducing kernel associated with the model space. It is known that the minimizers f_j of (12) have the form

$$f_j(\cdot) = \boldsymbol{\phi}^j(\cdot)^T \mathbf{d}^j + \boldsymbol{\xi}^j(\cdot)^T \mathbf{c}^j, \quad (17)$$

where $\mathbf{c}^j$ and $\mathbf{d}^j$ are vectors of coefficients to be found. Here $\{\phi^j_\nu\}_{\nu=1}^{p_j}$ is a set of p_j basis functions spanning the null space $\mathcal{H}^j_0$. $\boldsymbol{\phi}^j(\cdot)^T = (\phi^j_1(\cdot), \ldots, \boldsymbol{\phi}^j_{p_j}(\cdot))$. $\boldsymbol{\xi}^j(\cdot)^T = (\xi_{ijk}(\cdot),\ 1 \le i \le n, 1 \le k \le K_j)$, where $\xi_{ijk}(\cdot)$ is the representer of the evaluation functional at $\mathbf{x}_{ijk}$ in $\mathcal{H}^j_1$. It is not necessary to be familiar with the concept of the representer of an evaluation functional in a reproducing kernel Hilbert space to follow the rest of the article, if one is willing to accept the result that our estimates will be of the form (17) and that the penalty functionals are quadratic forms in the $\mathbf{c}^j$ as given following (19) below. (Interested readers can find definitions in Wahba 1990, pp. 1–2.) Descriptions of the particular ϕ^j_ν and ξ_{ijk} used in this article are given in Section 5.

The smoothing parameters $\boldsymbol{\Lambda}_j = (\lambda_{j1}, \ldots, \lambda_{jq})$ are incorporated into $\boldsymbol{\xi}^j(\cdot)^T$. Details have been given by Gu and Wahb (1993), and Wahba (1990, chap. 10). and Wahba et al. (1995). The penalty functionals are known quadratic forms in the $\mathbf{c}^j$. We give examples later. Given the $\{\lambda_{jl}\}$ and the result (17), the numerical problem is to obtain the minimizer of (10).

3.2 The Fitting Algorithm

For computational reasons, only an approximate minimizer of (12) will be obtained, using only a subset of the basis functions $\{\xi_{ijk}(\cdot), 1 \le i \le n, 1 \le k \le K_j\}$. Usually, when the design points are close, their representers are also very close. Hence when the dataset is large, it is very likely that many of the basis functions are nearly linearly dependent. On the other hand, if by some "prior" knowledge the structure of the true f_j is thought to be not very complicated, then it may be well approximated by a small number of basis functions. As a result, if we select a subset of the design points having maximum separation, then their corresponding representers are expected to have less correlation. We choose a fixed V and use the SAS procedure FASTCLUS to cluster the nK_j values of the design points $\{\mathbf{x}_{ijk}, 1 \le i \le n, 1 \le k \le K_j\}$ into V clusters. Within the vth cluster, $v = 1, \ldots, V$, we randomly select one design point and call it $\mathbf{x}_{j,v}$. The representer of evaluation in $\mathcal{H}^j$ at $\mathbf{x}_{j,v}$, called $\xi_{j,v}$, is used to form the approximating subspace. An approximate solution is of the form $f_j(\cdot) = \boldsymbol{\phi}^j(\cdot)^T\mathbf{d}^j + \boldsymbol{\xi}_V^j(\cdot)^T\mathbf{c}_V^j$, where $\boldsymbol{\phi}^j(\cdot)^T$ is as before and $\boldsymbol{\xi}_V^j(\cdot)^T = (\xi_{j,v}(\cdot), v = 1, \ldots, V)$, and is computed for (12) by a block one-step SOR–Newton-Ralphson algorithm. (see Ortega and Rheinboldt 1970). When the number of basis functions V increases, the approximate solution converges to the exact solution. We have found that for moderately large V, it is not important which point in each cluster is chosen.

Recall that the $\boldsymbol{\alpha}$ could depend on a set of parameters $\boldsymbol{\beta}$. The iterative algorithm for the approximate solution is given in Table 1.

Table 1. Iterative Algorithm for Approximate Spline

```
V ← initial value
do
  Cluster the data points into V groups
  Randomly select one data point from each group
  Generate the corresponding basis functions
  f_j ← initial values, j = 1,2,...,J
  β ← initial values
  do
    do j = 1 to J
      f_j ← updated values in the approximating subspace
    end
    β ← Newton-Ralphson update for β
  until (convergence)
  V ← 2 × V
until ( ||f_new − f_old|| / ||f_old|| < prec_1 and ||β_new − β_old|| / ||β_old|| < prec_2 )
```

Here prec_1 and prec_2 are prespecified thresholds, which we set to 10^{-7} and 10^{-8} in the examples to follow. We suggest that the initial value for V be at least 25. Combined with the iterative method for choosing smoothing parameters proposed in the next section, the foregoing algorithm usually converges very rapidly. From our experience, for medical data, 50–100 basis functions usually yield a very good approximation.

Next, we discuss in detail how to use the block one-step SOR algorithm to obtain an approximate solution for fixed V. Numerically, we need to solve a large nonlinear system. To speed up the computation, we break down the nonlinear system to several "blocks." When updating one "block" in the nonlinear system, we fix all other "blocks" at the most recently updated values. The natural breakdown in this application is to update the f_j's and $\boldsymbol{\beta}$ iteratively. To update the $\boldsymbol{\beta}$, we use the Newton-Raphson algorithm and iterate until convergence. However, the updating step for f_j is much more computationally intensive. The block nonlinear SOR algorithm requires that the Newton-Ralphson algorithm be run until convergence. However, we use only a one-step updating formula, sacrificing the overall convergence rate a bit to reduce the computational complexity in each iteration. Let $\mathbf{S}_j$ be the $nK_j \times p_j$ matrix with entries $\phi_\nu^j(\mathbf{x}_{ijk})$. Returning to our fixed V and the selected design points $\mathbf{x}_{j,v}$, $v = 1, \ldots, V$ and the associated representers $\xi_{j,v}(\cdot)$, we define

$$\mathbf{Q}_{j,V} = \begin{pmatrix} \xi_{j,1}(\mathbf{x}_{1j1}) & \xi_{j,2}(\mathbf{x}_{1j1}) & \cdots & \xi_{j,V}(\mathbf{x}_{1j1}) \\ \xi_{j,1}(\mathbf{x}_{1j2}) & \xi_{j,2}(\mathbf{x}_{1j2}) & \cdots & \xi_{j,V}(\mathbf{x}_{1j2}) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{j,1}(\mathbf{x}_{njK_j}) & \xi_{j,2}(\mathbf{x}_{njK_j}) & \cdots & \xi_{j,V}(\mathbf{x}_{njK_j}) \end{pmatrix}_{nK_j \times V}$$

and

$$\mathbf{Q}_{j,V}^* = \begin{pmatrix} \xi_{j,1}(\mathbf{x}_{j,1}) & \xi_{j,2}(\mathbf{x}_{j,1}) & \cdots & \xi_{j,V}(\mathbf{x}_{j,1}) \\ \xi_{j,1}(\mathbf{x}_{j,2}) & \xi_{j,2}(\mathbf{x}_{j,2}) & \cdots & \xi_{j,V}(\mathbf{x}_{j,2}) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{j,1}(\mathbf{x}_{j,V}) & \xi_{j,2}(\mathbf{x}_{j,V}) & \cdots & \xi_{j,V}(\mathbf{x}_{j,V}) \end{pmatrix}_{V \times V.} \tag{18}$$

Let $\mathbf{c}_V^j = (c_{j,1}, c_{j,2}, \ldots, c_{j,V})^T$. Our goal is to compute the approximate solution with the representation

$$f_j(\cdot) = \boldsymbol{\phi}^j(\cdot)^T\mathbf{d}^j + \boldsymbol{\xi}_V^j(\cdot)^T\mathbf{c}_V^j, \tag{19}$$

which involves the use of $\mathbf{S}_j$, $\mathbf{Q}_{j,V}$ and $\mathbf{Q}_{j,V}^*$. It is easy to verify that the penalty for f_j has the quadratic form $\|P_1 f_j\|_{\Lambda_j}^2 = \mathbf{c}_V^{j\,T}\mathbf{Q}_{j,V}^*\mathbf{c}_V^j$. Therefore, to update f_j, provided all other estimated values are fixed at the solutions from the previous iteration, and recalling that $f_{ijk} = f_j(\mathbf{x}_{ijk})$, the variational problem is to minimize

$$I_{j,V} = -\sum_{i=1}^{n}\left\{\sum_{k=1}^{K_j} f_{ijk}y_{ijk} - b(\mathbf{f}_i, \boldsymbol{\alpha}_i)\right\} + \frac{n}{2}\mathbf{c}_V^{j\,T}\mathbf{Q}_{j,V}^*\mathbf{c}_V^j. \tag{20}$$

Denote $b_i = b(\mathbf{f}_i, \boldsymbol{\alpha}_i)$. To update f_j, we need the following, where l_j is defined in Lemma C.1

$$\begin{aligned}
\mu_{ijk} &= \frac{\partial b_i}{\partial f_{ijk}} = EY_{ijk} = \left(e^{f_{ijk}} + \sum_{k_3 \neq k} e^{f_{ijk} + f_{ijk_3} + \alpha_{ijk, ijk_3}} + \cdots + e^{\sum_{j_3, k_3} f_{ij_3k_3} + \sum_{j_3, k_3} \sum_{j_4, k_4} \alpha_{ij_3k_3, ij_4k_4}}\right) \\
&\quad \times \left(1 + \sum_{j_3, k_3} e^{f_{ij_3k_3}} + \cdots + e^{\sum_{j_3, k_3} f_{ij_3k_3} + \sum_{j_3, k_3} \sum_{j_4, k_4} \alpha_{ij_3k_3, ij_4k_4}}\right)^{-1}, \\
w_{ijk, ijk} &= \frac{\partial^2 b_i}{\partial f_{ijk}^2} = \operatorname{var} Y_{ijk} = \mu_{ijk}(1 - \mu_{ijk}), \\
w_{ijk_1, ijk_2} &= \frac{\partial^2 b_i}{\partial f_{ijk_1} \partial f_{ijk_2}} = \operatorname{cov}(Y_{ijk_1}, Y_{ijk_2}) = E(Y_{ijk_1} Y_{ijk_2}) - EY_{ijk_1} \cdot EY_{ijk_2} = \frac{\partial b_i}{\partial \alpha_{ijk_1, ijk_2}} - \mu_{ijk_1} \mu_{ijk_2} \\
&= \left(e^{f_{ijk_1} + f_{ijk_2} + \alpha_{ijk_1, ijk_2}} + \cdots + e^{\sum_{j_3, k_3} f_{ij_3k_3} + \sum_{j_3, k_3} \sum_{j_4, k_4} \alpha_{ij_3k_3, ij_4k_4}}\right) \\
&\quad \times \left(1 + \sum_{j_3, k_3} e^{f_{ij_3k_3}} + \cdots + e^{\sum_{j_3, k_3} f_{ij_3k_3} + \sum_{j_3, k_3} \sum_{j_4, k_4} \alpha_{ij_3k_3, ij_4k_4}}\right)^{-1} - \mu_{ijk_1} \mu_{ijk_2}, \\
u_{ijk} &= \frac{-\partial l_j(y_{ij}, f_{ij})}{\partial f_{ijk}} = -y_{ijk} + \mu_{ijk}, \\
\mathbf{u}_j &= (u_{1j1}, u_{1j2}, \ldots, u_{1jK_j}, u_{2j1}, \ldots, u_{njK_j})^T,
\end{aligned}$$

$$\mathbf{W}_{ij} = \begin{pmatrix} w_{ij1, ij1} & w_{ij1, ij2} & \cdots & w_{ij1, ijK_j} \\ w_{ij2, ij1} & w_{ij2, ij2} & \cdots & w_{ij2, ijK_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ijK_j, ij1} & w_{ijK_j, ij2} & \cdots & w_{ijK_j, ijK_j} \end{pmatrix}_{K_j \times K_j},$$

and

$$\mathbf{W}_j = \operatorname{diag}(\mathbf{W}_{1j}, \mathbf{W}_{2j}, \ldots, \mathbf{W}_{nj}).$$

Therefore, the one-step updating formula is to solve

$$\begin{pmatrix} \mathbf{Q}_{j,V}^T \mathbf{W}_{j-} \mathbf{Q}_{j,V} + n\mathbf{Q}_{j,V}^* & \mathbf{Q}_{j,V}^T \mathbf{W}_{j-} \mathbf{S}_j \\ \mathbf{S}_j^T \mathbf{W}_{j-} \mathbf{Q}_{j,V} & \mathbf{S}_j^T \mathbf{W}_{j-} \mathbf{S}_j \end{pmatrix} \begin{pmatrix} \mathbf{c}_V^j - \mathbf{c}_{V-}^j \\ \mathbf{d}^j - \mathbf{d}_-^j \end{pmatrix} = \begin{pmatrix} -\mathbf{Q}_{j,V}^T \mathbf{u}_{j-} - n\mathbf{Q}_{j,V}^* \mathbf{c}_{V-}^j \\ -\mathbf{S}_j^T \mathbf{u}_{j-} \end{pmatrix}, \quad (21)$$

where the subscript minus indicates the quantities evaluated at the latest update.

With some abuse of notation, we use $\mathbf{u}_{ij}$ to denote $(u_{ij1}, \ldots, u_{ijK_j})^T$, $\mathbf{f}_{ij}$ to denote $(f_{ij1}, \ldots, f_{ijK_j})^T$, an so on. Here $\tilde{\mathbf{y}}_{ij} = \mathbf{f}_{ij-} - \mathbf{W}_{ij-}^{-1} \mathbf{u}_{ij-}$ are called the *pseudo-data*. It is easy to see that the solution of (21) gives the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{\mathbf{y}}_{ij} - \mathbf{f}_{ij})^T \mathbf{W}_{ij-} (\tilde{\mathbf{y}}_{ij} - \mathbf{f}_{ij}) + \mathbf{c}^{jT} \mathbf{Q}_{j,V}^* \mathbf{c}^j. \quad (22)$$

The block one-step SOR–Newton-Ralphson procedure iteratively reformulates the problem to estimate f_j from the pseudo-data by weighted penalized least squares. The pseudo-data are equivalent to the adjusted dependent vector of Yee and Wild (1996). The pseudo-data can be shown to have the usual (mean and covariance) data structure implicit in (22) if the f_{j-} are not far away from f_j. A theorem is given in Appendix B. We use this result later to construct approximate Bayesian confidence intervals for the f_j.

The preceding discussion assumes no special structure in the design points. The algorithm is specifically designed to handle the unstructured case. However, when special structure is available, the foregoing algorithm can be simplified. One common case is the presence of person-specific covariates only. Hence $\mathbf{x}_{ijk} = \mathbf{x}_{ij}$ for all $k = 1, \ldots, K_j$. Similarly, $f_{ijk} = f_j(\mathbf{x}_{ijk}) = f_j(\mathbf{x}_{ij}) = f_{ij}$. To update f_j, the part of the penalized likelihood that needs to be minimized has the simplified form

$$I_{j,V} = -\sum_{i=1}^{n} \left\{ \left(\sum_{k=1}^{K_j} y_{ijk} \right) f_{ij} - b_i \right\} + \frac{n}{2} \mathbf{c}^{jT} \mathbf{Q}_{j,V}^* \mathbf{c}^j. \quad (23)$$

Now redefine

$$\mathbf{Q}_{j,V} = \begin{pmatrix} \xi_{j,1}(\mathbf{x}_{1j}) & \xi_{j,2}(\mathbf{x}_{1j}) & \cdots & \xi_{j,V}(\mathbf{x}_{1j}) \\ \xi_{j,1}(\mathbf{x}_{2j}) & \xi_{j,2}(\mathbf{x}_{2j}) & \cdots & \xi_{j,V}(\mathbf{x}_{2j}) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{j,1}(\mathbf{x}_{nj}) & \xi_{j,2}(\mathbf{x}_{nj}) & \cdots & \xi_{j,V}(\mathbf{x}_{nj}) \end{pmatrix}_{n \times V}. \quad (24)$$

Denote $y_{ij} = \sum_{k=1}^{K_j} y_{ijk}$, $\mu_{ij} = E(\sum_{k=1}^{K_j} Y_{ijk})$, $W_{ij} = \operatorname{var}(\sum_{k=1}^{K_j} Y_{ijk})$, $u_{ij} = -\sum_{k=1}^{K_j} y_{ijk} + \mu_{ij}$, and $\mathbf{u}_j = (u_{1j}, u_{2j}, \ldots, u_{nj})^T$.

Except for the aforementioned changes, all of the previous formulas and discussions remain true.

Finally, by the SOR–Newton theorem of Ortega and Rheinboldt (1970) and corollary 3.1 of X. Lin (1998), the local convergent property holds if we use block one-step SOR–Newton-Raphson method to find the minimizer of a twice-differentiable convex function. Thus there exists an open ball of the minimizer, and when the starting point is in this open ball, the algorithm will converge to the minimizer for fixed smoothing parameters.

In response to a referee's query, we note that the backfitting algorithm popularized by Hastie and Tibshirani (1990) is a form of block SOR algorithm. A block SOR algorithm was also used by Wahba et al. (1995), who gave a detailed discussion of the relation between the block SOR they used and the backfitting algorithm of Hastie and Tibshirani. However, the way in which we have assigned blocks in this article is different than either of these forms of block SOR.

4. ADAPTIVE CHOICE OF THE SMOOTHING PARAMETERS

So far, all smoothing parameters are considered fixed. Now we consider an automated data-driven method to choose smoothing parameters. The risk function or "target" to be minimized is the comparative Kullback–Leibler (CKL) distance of the estimate from the "truth." Because the truth is not observable, except in simulation studies, the CKL also is not observable. Thus a computable proxy for the CKL whose minimizer is a good estimate of the minimizer of the CKL is desired. The GACV was proposed by Xiang and Wahba (1996) as a method for choosing a smoothing parameter for penalized likelihood estimates when the data are from an exponential family with no nuisance parameters. Those authors showed via Monte Carlo methods with Bernoulli data that its minimizer provides a good estimate of the minimizer of the CKL in relatively small samples. The derivation of the GACV score begins with a leave-out-one argument. Direct calculation of the GACV score becomes unstable in the large-sample case due to the existence of the inverse of a generally ill-conditioned matrix in the formula (see Xiang and Wahba 1996). Later, Lin et al. (2000) and Wahba et al. (1999) demonstrated that the GACV score can be computed and minimized over multiple smoothing parameters with the use of a randomized trace technique, (ranGACV), which provides a stable calculation. In this article we extend the GACV, and then the ranGACV, to the case of correlated Bernoulli observations. Instead of using a leave-out-one-observation to start the derivation, we leave out one person. Following the derivation, in the next section we demonstrate via a small simulation study the properties of the ranGACV for providing an estimate of the minimizer of the CKL in the correlated Bernoulli case.

To choose the smoothing parameters for f_j, we let all of the other conditional logit functions f_ℓ ($\ell \neq j$) and the $\boldsymbol{\alpha}$ be fixed at the estimated values from the previous iteration. Let $f_{ijk}^{[-i]}$ denote the estimated conditional logit function f_j evaluated at $\mathbf{x}_{ijk}$, with the ith subject left out for the estimation procedure. The ordinary leave-out-one-subject cross validation function for choosing Λ_j, $\mathrm{CV}(\Lambda_j)$, is defined as

$$\begin{aligned}\mathrm{CV}(\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk}^{[-i]} + b(\mathbf{f}_{ij})\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(\mathbf{f}_{ij}) \right. \\ &\qquad \left. + \sum_{k=1}^{K_j} y_{ijk}(f_{ijk} - f_{ijk}^{[-i]})\right]. \end{aligned} \tag{25}$$

Because y_{ijk} and $f_{ijk}^{[-i]}$ are independent, and because for large sample sizes $f_{ijk}^{[-i]}$ is expected to be close to f_{ijk}, CV is expected to provide an approximately unbiased estimate for the CKL distance, given $\boldsymbol{\alpha}$ and $f^{(-j)}$ fixed. First, let us assume that the exact solution of the variational problem can be computed. Let $\mathbf{y}_j = (y_{ijk}, 1 \le i \le n, 1 \le k \le K_j)^T$. In each updating step for f_j, the minimizer of

$$I_{\Lambda_j}(f_j, \mathbf{y}_j) = -\sum_{i=1}^{n}\left(-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(\mathbf{f}_{ij})\right) + \frac{n}{2}\mathcal{J}_{\Lambda_j}^{j}(f_j) \tag{26}$$

is obtained. By using several first order Taylor expansions, we derive an approximate cross validation (ACV) score. The GACV is a generalization of the ACV score.

First, we define the *generalized average* of submatrices as a generalization of the trace of a matrix. For any matrix $\mathbf{A}$ with submatrices $\mathbf{A}_{ii}, 1 \le i \le n$ on the diagonal, $\mathbf{A}_{ii} = \left(a_{i,k_1k_2}\right)_{K\times K}, 1 \le k_1, k_2 \le K$, if $K = 1$, the generalized average of $\mathbf{A}_{ii}$ is simply the trace of $\mathbf{A}$ divided by n, $\overline{\mathbf{A}} = \mathrm{tr}(\mathbf{A})/n$. When $K \ge 2$, let

$$\begin{aligned}\delta &= \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K} a_{i,kk} = \mathrm{tr}(\mathbf{A}), \\ \gamma &= \frac{1}{n\cdot K(K-1)}\sum_{i=1}^{n}\sum_{k_1\neq k_2} a_{i,k_1k_2}. \end{aligned} \tag{27}$$

The generalized average of $\mathbf{A}_{ii}$'s is defined as

$$\overline{\mathbf{A}} = (\delta-\gamma)\mathbf{I}_{K\times K} + \gamma\cdot\mathbf{e}\mathbf{e}^T = \begin{pmatrix} \delta & \gamma & \cdots & \gamma \\ \gamma & \delta & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \delta \end{pmatrix}_{K\times K}, \tag{28}$$

where $\mathbf{e} = (1, 1, \cdots, 1)^T$ is the unit vector. When $K \ge 3$, this form is suitable whenever it is reasonable to assume that associations between different pairs of repeated observations are similar. Other generalization forms are possible. The properties of the $K \ge 3$ case are a matter for future research.

Let $\mathbf{H}^j$ denote the inverse Hessian of (26) with respect to f_j. $\mathbf{H}^j$ is an $nK_j \times nK_j$ matrix with $\mathbf{H}_{ii}^j$, $i = 1, \ldots, n$ as $K_j \times K_j$ submatrices on the diagonal. Let $\mathbf{G}_{ii}^j = \mathbf{I} - \mathbf{W}_{ij}\mathbf{H}_{ii}^j$ and let $\overline{\mathbf{H}}^j$ and $\overline{\mathbf{G}}^j$ denote the generalized average of $\mathbf{H}_{ii}^j$ and $\mathbf{G}_{ii}^j$. We now

define the GACV as

$$\begin{aligned}\mathrm{GACV}(\Lambda_j) = \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(f_{ij})\right] \\ + \frac{1}{n}\sum_{i=1}^{n}\left(y_{ij1}\cdots y_{ijK_j}\right) \\ \times \overline{\mathbf{H}}^j(\overline{\mathbf{G}}^j)^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}\end{pmatrix}. \qquad (29)\end{aligned}$$

Details of the derivation of the GACV are given in Appendix C.

Direct computation of the GACV for large sample sizes is slow and tends to be unstable (see Xiang and Wahba 1996, Sec. 3.1). This explicit calculation can be avoided by using a technique in the spirit of the randomized trace method, provided that a solution (either exact or approximate) of the variational problem can be obtained at a lower cost. We propose a one-step randomized estimate of GACV that is fast and cheap to calculate.

Given a square matrix $\mathbf{A}$ with $\mathbf{A}_{ii}$ $(1 \le i \le n)$ being the $K \times K$ submatrices on the diagonal, we discuss how to obtain a randomized estimate of $\mathbf{A}$. First, n iid vectors of K iid random variables distributed as N(0,1) are generated. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_K)^T$ be the ith such vector and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$. Then $\delta = \mathrm{tr}(\mathbf{A})/(nK)$ can be estimated by $(\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon})/(nK)$. On the other hand, if $K > 1$, then $\gamma = (\sum_i \sum_{k_1,k_2} a_{i,k_1k_2} - \mathrm{tr}(\mathbf{A}))/(nK(K-1))$. To estimate $\sum_i \sum_{k_1,k_2} a_{i,k_1k_2}$, let $\bar{\boldsymbol{\epsilon}}_i = (1/\sqrt{K})\sum_{k=1}^K \epsilon_{ik}$ and $\bar{\boldsymbol{\epsilon}} = (\bar{\boldsymbol{\epsilon}}_1, \ldots, \bar{\boldsymbol{\epsilon}}_1, \bar{\boldsymbol{\epsilon}}_2, \ldots, \bar{\boldsymbol{\epsilon}}_n)^T$. Here $\bar{\boldsymbol{\epsilon}}$ is a column vector with K replicates of $\bar{\boldsymbol{\epsilon}}_i$ for each $1 \le i \le n$. Note that $E\bar{\boldsymbol{\epsilon}}^T\mathbf{A}\bar{\boldsymbol{\epsilon}} = \sum_i \sum_{k_1,k_2} a_{i,k_1k_2}$. Thus we can estimate γ by $(\bar{\boldsymbol{\epsilon}}^T\mathbf{A}\bar{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon})/(nK(K-1))$ and obtain a randomized estimate of $\mathbf{A}$.

In practice, the randomized estimate of GACV is calculated by solving the nonlinear system on the perturbed data $\mathbf{Y}_j + \boldsymbol{\epsilon}$ and $\mathbf{Y}_j + \bar{\boldsymbol{\epsilon}}$. Let $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}$ denote the solution of (26) using the original data and let $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon}}$ denote the solution using the perturbed data. If we take $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}$ as the initial value to a Newton-Ralphson calculation of $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon}}$ and run the iteration only once by using all matrix decompositions that have already been performed for calculating $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}$ in the last step, then we obtain the one-step solution $\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon},1}$. Because $(\partial I_{\Lambda_j}/\partial \mathbf{f}_j)(\mathbf{f}_{\Lambda_j}^{Y_j}, \mathbf{Y}_j) = \mathbf{0}$ and $(\partial^2 I_{\Lambda_j}/\partial \mathbf{f}_j^T \partial \mathbf{f}_j)(\mathbf{f}_{\Lambda_j}^{Y_j}, \mathbf{Y}_j) = (\partial^2 I_{\Lambda_j}/\partial \mathbf{f}_j^T \partial \mathbf{f}_j)(\mathbf{f}_{\Lambda_j}^{Y_j}, \mathbf{Y}_j + \boldsymbol{\epsilon})$, we observe the simple relation

$$\begin{aligned}\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon},1} &= \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j} - \left[\frac{\partial^2 I_{\Lambda_j}}{\partial \mathbf{f}_j^T \partial \mathbf{f}_j}\left(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}, \mathbf{Y}_j+\boldsymbol{\epsilon}\right)\right]^{-1}\frac{\partial I_{\Lambda_j}}{\partial \mathbf{f}_j}\left(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}, \mathbf{Y}_j+\boldsymbol{\epsilon}\right) \\ &= \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j} - \left[\frac{\partial^2 I_{\Lambda_j}}{\partial \mathbf{f}_j^T \partial \mathbf{f}_j}\left(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}, \mathbf{Y}_j\right)\right]^{-1}\left(-\boldsymbol{\epsilon} + \frac{\partial I_{\Lambda_j}}{\partial \mathbf{f}_j}\left(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}, \mathbf{Y}_j\right)\right) \\ &= \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j} + \mathbf{H}^j\boldsymbol{\epsilon}. \qquad (30)\end{aligned}$$

Hence we have

$$\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon},1} - \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j} = \mathbf{H}^j\boldsymbol{\epsilon}. \qquad (31)$$

Thus $\boldsymbol{\epsilon}^T(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon},1} - \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}) = \boldsymbol{\epsilon}^T\mathbf{H}^j\boldsymbol{\epsilon}$ and $\bar{\boldsymbol{\epsilon}}^T(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\bar{\boldsymbol{\epsilon}},1} - \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j}) = \bar{\boldsymbol{\epsilon}}^T\mathbf{H}^j\bar{\boldsymbol{\epsilon}}$, and we can obtain a randomized estimate of $\overline{\mathbf{H}}^j$. Similarly, $\boldsymbol{\epsilon}^T\mathbf{G}^j\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{W}_j(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\boldsymbol{\epsilon},1} - \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j})$, and $\bar{\boldsymbol{\epsilon}}^T\mathbf{G}^j\bar{\boldsymbol{\epsilon}} = \bar{\boldsymbol{\epsilon}}^T\bar{\boldsymbol{\epsilon}} - \bar{\boldsymbol{\epsilon}}^T\mathbf{W}_j(\mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j+\bar{\boldsymbol{\epsilon}},1} - \mathbf{f}_{\Lambda_j}^{\mathbf{Y}_j})$. We can thus calculate the randomized estimate of $\overline{\mathbf{G}}^j$. This approach avoids explicit calculation of the inverse Hessian $\mathbf{H}^j$, which is computationally expensive and tends to be unstable if $\mathbf{H}^j$ is ill conditioned. A randomized estimate can always be obtained provided that a cheap and stable "black box" exists for calculating the (approximate) one-step solution for perturbed data. The resulting ranGACV function is

$$\begin{aligned}\mathrm{ranGACV}(\Lambda_j) = \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{j=1}^{K_j} y_{ijk} f_{ijk} + b(f_{ij})\right] \\ + \frac{1}{n}\sum_{i=1}^{n}(y_{ij1}, \ldots, y_{ijK_j}) \\ \times \hat{\overline{\mathbf{H}}}^j\left(\hat{\overline{\mathbf{G}}}^j\right)^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}.\end{pmatrix}, \qquad (32)\end{aligned}$$

where $\hat{\overline{\mathbf{H}}}^j$ and $\hat{\overline{\mathbf{G}}}^j$ denote the randomized estimates. To reduce the variance in the term after the "+" in (32), we may draw R independent random vectors $\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(R)}$, and replace the term after the "+" in (32) by

$$\frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n}(y_{ij1}, \ldots, y_{ijK_j})\hat{\overline{\mathbf{H}}}^{j(r)}\left(\hat{\overline{\mathbf{G}}}^{j(r)}\right)^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}.\end{pmatrix} \qquad (33)$$

to obtain an R-replicated ranGACV function. The GACV and ranGACV functions are derived by assuming that the minimizer of (20) is calculated at each block nonlinear SOR iteration. To speed up the algorithm, however, only a one-step update will be calculated. We remark that all favorable properties of GACV and ranGACV are preserved for the block one-step SOR algorithm and approximate spline estimate. It is very easy to carry out the computation, as no additional matrix decomposition is required. We iteratively minimize ranGACV in each step of the block one-step SOR iteration. This is done repeatedly until some prespecified convergence criteria is met, or the number of iterations exceeds a prespecified limit.

We remark that it is a straightforward generalization to extend the the algorithm and the ranGACV to observations that come from other exponential families with no nuisance parameter. In fact, the original derivation of the GACV did not use the fact that the data were Bernoulli. However, the GACV has not yet been tested on data from other exponential families.

5. MONTE CARLO SIMULATIONS

In this section we demonstrate results from some Monte Carlo simulations to evaluate the performance of the proposed method. The comparative Kullback–Leibler distance (CKL) is used to measure the performance of the estimated values. In our experience, the CKL distance provides a measure that agrees very well with human intuition when eyeballing the fitted surfaces.

In all models presented here and in next section, we used reproducing kernels associated with cubic splines as building blocks for the tensor product space reproducing kernels that

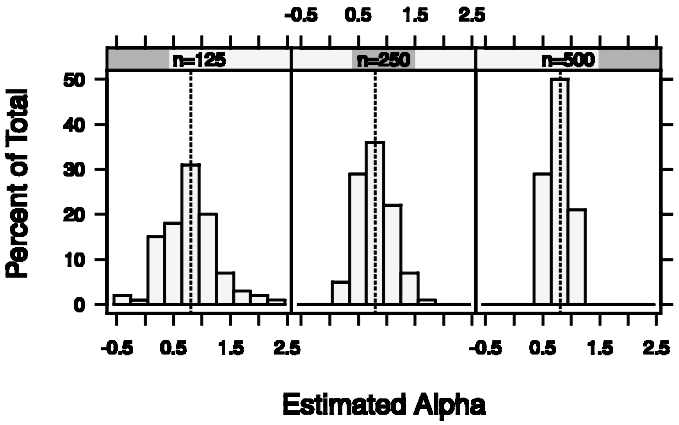


Figure 1. Histogram of $\hat{\alpha}$ for Three Different Sample Sizes. The dotted lines represent the true value of $\alpha = .8$. The means of estimated values are .7926, .7937, and .7686.

generate the ξ's. (This construction was given in Wahba 1990, sec. 10.2, for general SS-ANOVA models.) We briefly note some details. Let $u \in [0, 1]$; the null space of the cubic spline penalty functional $(\int (f'')^2)$ is spanned by the linear functions $\phi_1(u) = 1$ and $\phi_2(u) = u - 1/2$. Define the reproducing kernel $r_1(u, v) = k_2(u)k_2(v) - k_4([u - v])$, where $u, v \in [0, 1]$ and $k_\ell(u) = B_\ell(u)/\ell!$, with $B_\ell(\cdot)$ the ℓth Bernoulli polynomial. In the case where f_j is a function on $[0, 1]$, the ϕ's are the linear functions above and $\xi_{j,\nu}(u) = r_1(x_{j,\nu}, u)$. When f_j is a function on the unit square or higher dimensional unit cube, the ϕ's are built up from tensor sums and products of ϕ_1 and ϕ_2, and reproducing kernels to generate the $\xi_{j,\nu}$ are built up from tensor sums and products of r_1 and $r_0(u, v) = \phi_2(u)\phi_2(v)$, as given by Wahba (1990). Particular details for the model in (37) below were also given by Lin et al. (2000). In the experiments and data analysis that follow, the ranGACV is used to choose the smoothing parameters.

5.1 Repeated Measurements for the Same Endpoint

The first example concerns the single smoothing parameter situation. We try to mimic the characteristic of possible ophthalmology data. Each subject has one endpoint of interest and paired observations. There is one observation-specific covariate, $X_{ik}, (k = 1, 2)$. The X_{i1}'s are assumed to be uniformly distributed on the interval $(.05, .95)$. $X_{i2} = X_{i1} + \epsilon_i$, and the ϵ_i's are uniformly distributed on $(-.05, .05)$. The true conditional logit function is assumed to be $f(x_{ik}) = \text{logit}(P(Y_{ik} = 1|Y_i^{(-k)} = 0, x_{ik})) = 2[\exp(-30(x_{ik} - .25)^2) + \sin(\pi x_{ik}^2)] - 2$, and the conditional log odds ratio $\alpha = \log \text{OR}(Y_{i1}, Y_{i2}|x_i) = .8$. Three different sample sizes are used in this simulation: $n = 125, 250, 500$. For each sample size, 100 independent datasets are randomly generated according to the true joint distribution.

Figure 1 shows histogram plots of the estimated $\hat{\alpha}$ for the three different sample sizes. The dotted lines represent the true value of .8. The fitted values appear to converge to the truth while the sample size increases. The estimator of α appears to be approximately unbiased and normally distributed from the histogram.

Figure 2 plots the true conditional probability function and the estimated curves for each sample size, $P(Y_{ik} = 1|Y_i^{(-k)} = 0, x_{ik}) = e^{f(x_{ik})}/(1 + e^{f(x_{ik})})$. For each sample size, the 100 fitted values are ranked according to the CKL distances between

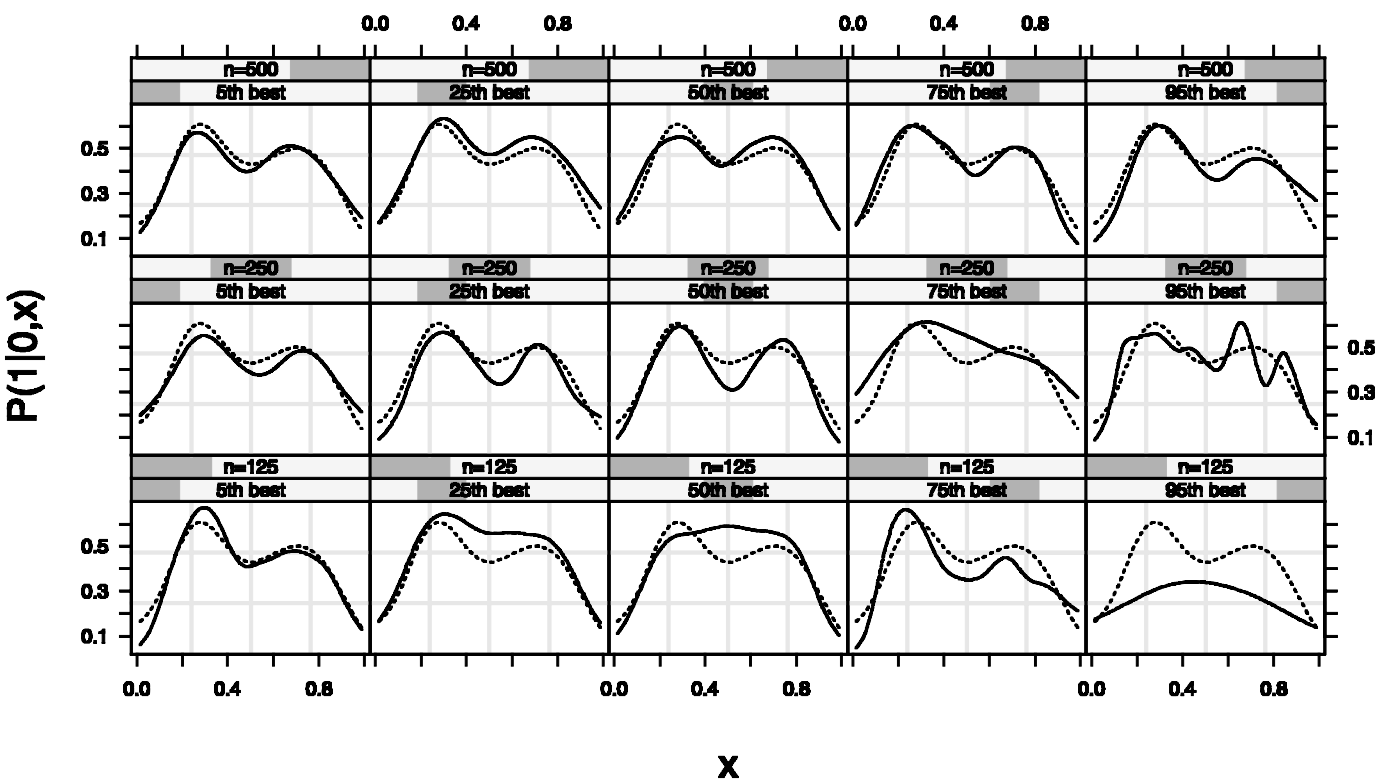


Figure 2. True (Dashed Line) and Estimated (Solid Line) Conditional Probability Functions.

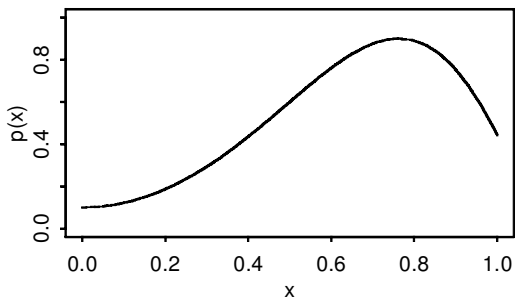


Figure 3. True $p(x_i) = P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i)$ Used for the Simulation Study.

the estimated joint distributions and the truth. The 5th, 25th, 50th, 75th and 95th best fits are plotted for each sample size. The true conditional logit function is a bimodal function. The trend is clear that when the sample size increases, the estimated curves become increasingly accurate. However, for a parametric model, there might be no prior knowledge about the bimodal nature of the truth. Hence a linear or even quadratic form will miss the true curve no matter how large the sample size.

In the next experiment, we compare the proposed new "multivariate" method to the "univariate" fit. In ophthalmology studies, one question of interest is to estimate the probability of at least one eye developing a certain disease given the values of the predictor variables for a person. Assume that there is no eye-specific covariate. The X_i's are uniformly distributed on [0,1], and for each subject, there are paired observations (Y_{i1}, Y_{i2}). We want to estimate the probability $P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i) = (2e^{f_i} + e^{2f_i+\alpha})/(1 + 2e^{f_i} + e^{2f_i+\alpha})$ from the observed data.

For this experiment, we assume that $p(x_i) = P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i) = .8\sin(2.7x_i^2) + .1$. Figure 3 plots the true $p(x)$. Four different values are used for α: 0, .4, .8, and 1.2; $\alpha = 0$ corresponds to the case where Y_{i1} and Y_{i2} are independent. However we pretend that this fact is unknown, and so estimate α by the proposed algorithm. Straightforward calculation yields the following formula to compute f_i for given α and $P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i)$:

$$f_i = \log \frac{(p(x_i) - 1) + \sqrt{(1 - p(x_i))^2 + e^{\alpha} p(x_i)(1 - p(x_i))}}{e^{\alpha}(1 - p(x_i))}. \tag{34}$$

The experiment is conducted as follows. First, for the "univariate" fit, the only information needed is $\check{Y}_i$, which is defined to be 0 when $Y_{i1} = Y_{i2} = 0$ and 1 otherwise. Here $P(\check{Y}_i = 1|x_i) = p(x_i)$. We generate 100 datasets according to the true distribution and fit the data using the "univariate" penalized logistic regression based only on the derived data $\check{Y}_i$'s. For the bivariate fit, we first calculate the true joint distribution of (Y_{i1}, Y_{i2}) according to the previous formula. For each value of α, 100 datasets are randomly generated and the joint distribution is estimated by the proposed multivariate method. Afterward, the probability $P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i)$ can be derived from the estimated joint distribution. For every run, the CKL distance between the estimated $\hat{p}(x_i)$ and $p(x_i)$ is calculated.

The foregoing procedure is performed for three different sample sizes: $n = 100$, 200, 400. Figure 4 shows histograms

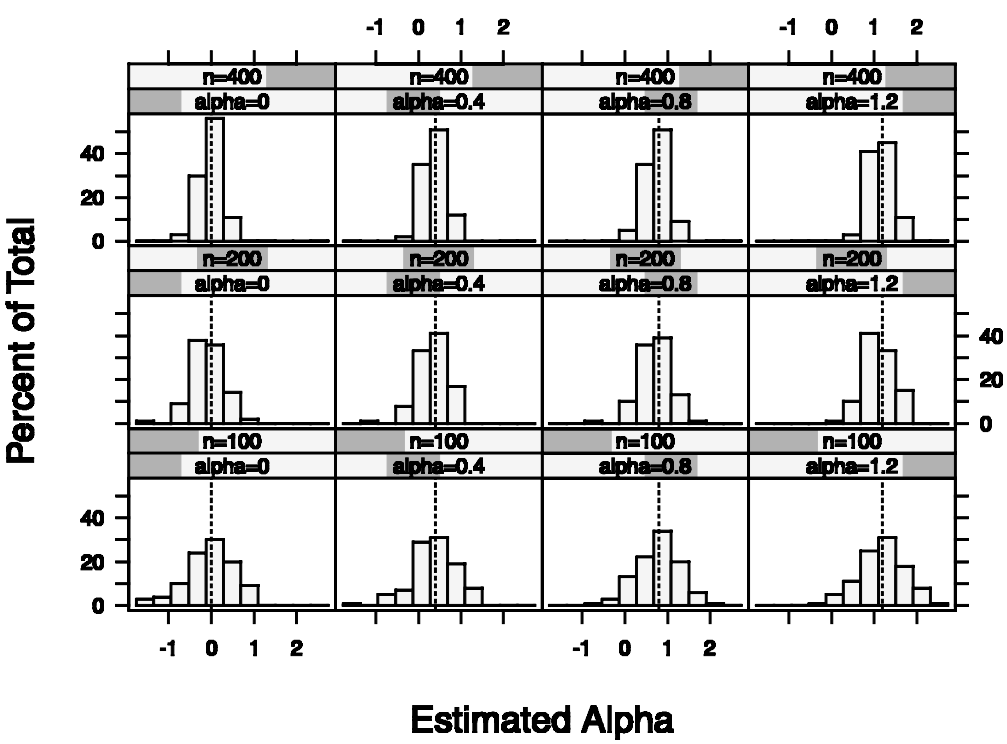


Figure 4. Histograms of Estimated $\hat{\alpha}$'s for n = 100, n = 200, and n = 400. The dotted lines represent the true values of α. From left to right, the means of estimated values are −.0214, .3743, .7596, and 1.1624 for n = 400; −.1028, .3161, .7058, and 1.1025 for n = 200; and −.0619, .3719, .7746, and 1.1775 for n = 100.

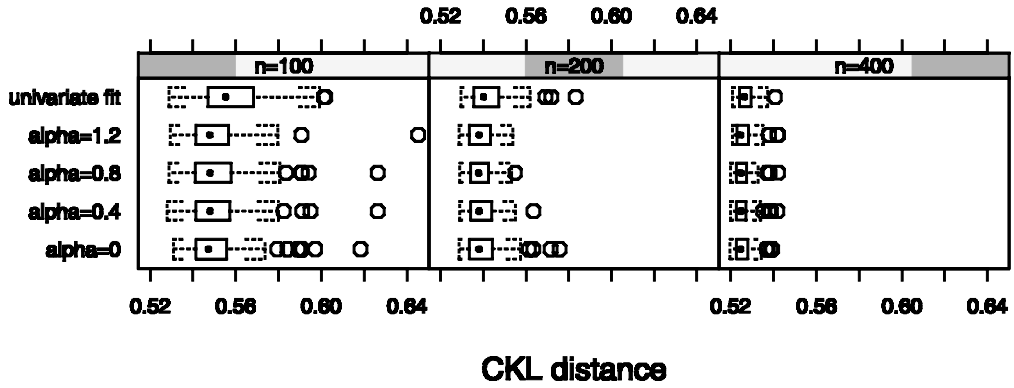


Figure 5. Boxplots of CKL for the Univariate Fit and the Multivariate Fits.

of the estimated $\hat{\alpha}$'s for different sample sizes and true values of α. Dotted lines represent the true values of α. From the plots, the estimated values have an approximate bell-shaped distribution and are approximately unbiased. As the sample size increases, the estimated values become closer to the true value.

Figure 5 compares the CKL distances between the fitted probability and the true probability $p(x_i) = P(Y_{i1} = 1 \text{ or } Y_{i2} = 1|x_i)$ for the two methods. Obviously, for all true values of α, the bivariate fit, which estimates the joint distribution of (Y_{i1}, Y_{i2}), has a better efficiency than the univariate fit, which estimates $P(\check{Y}_i = 1)$ directly. This is not surprising, since the "univariate" fit only needs to know $\check{Y}_i$; hence some information in (Y_{i1}, Y_{i2}) is not used in the estimation procedure.

In this article we omit plots of ranGACV(λ) and CKL(λ) due to lack of space. Plots of these curves (and surfaces in the case of multiple smoothing parameters) in the univariate Bernoulli response case show visually that at the sample sizes considered here, the minimizer of ranGACV is an excellent proxy for CKL. Examples have been given by Lin et al. (1998) and Wahba et al. (1999). Simulations with the same results in the multivariate Bernoulli response case were provided by Gao (1999a).

The next experiment is similar to the previous one but for multiple smoothing parameters. Assume that the (X_{i1}, X_{i2})'s are uniformly distributed on the unit square $(0, 1) \times (0, 1)$. The true conditional logit function is taken to be $f(x_{i1}, x_{i2}) = 2\sin(3x_{i1} - 3x_{i1}x_{i2}) + \cos(2 - 2x_{i2}) - 3(x_{i1} - .35)^2 - 1.5$, which involves both main effects and interaction terms. The conditional log odds ratio α is taken to be a constant 1. Each time, 500 independent pairs of observations (Y_{i1}, Y_{i2})'s are simulated. The proposed penalized multivariate logistic regression is used to estimate the joint distribution. This is repeated 100 times.

We can derive $p(x_{i1}, x_{i2}) = P(\check{Y}_i = 1|x_{i1}, x_{i2})$ from the estimated joint distribution. Figure 6 shows the true $p(x_{i1}, x_{i2})$ and the 5th, 25th, 50th, 75th, and 95th best estimated values

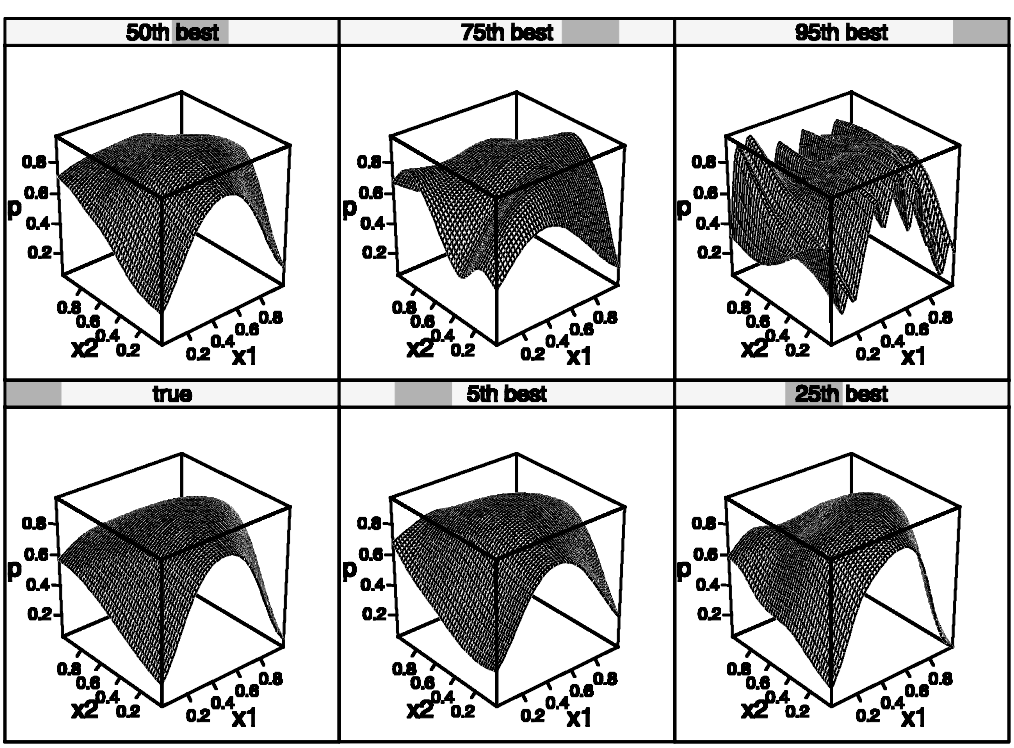


Figure 6. True $p(x_{i1},x_{i2}) = P(\check{Y}_i = 1|x_{i1},x_{i2})$ and Estimated Surfaces.

ranked by the CKL distance. The proposed method gives very good estimates most of the time.

To make the comparison, we also use the "univariate" method to estimate $p(x_{i1}, x_{i2})$ directly for the same 100 datasets. Only the derived outcome variable $\check{Y}_i$ is used in the estimation procedure. Assuming that we are only interested in estimating $P(\check{Y}_i = 1|x_{i1}, x_{i2})$; figure 7 shows the pairwise comparison of CKL distance. About 2/3 of the time, the bivariate fit yields the better estimate.

5.2 Different Endpoints

In this example we assume that there are two correlated endpoints of interest. For each subject, there are two binary outcome variables: Y_{i1} for the first endpoint and Y_{i2} for the second endpoint. The proposed method estimates the conditional joint distribution of $P(Y_{i1}, Y_{i2}|X_i)$. This model is also useful for predicting the outcome of one endpoint, given that the outcome of another endpoint is known. For example, if a person already has one disease, then what is the probability of getting another disease?

The true association factor $\alpha = \log \mathrm{OR}(Y_{i1}, Y_{i2})$ is taken to be 1.5 in this simulation. The true conditional logit functions for the two different endpoints are

$$f_1(x_i) = \mathrm{logit}(P(Y_{i1} = 1|Y_{i2} = 0, x_i)) = 10\cos(2x_i) + 7e^{x_i^2} - 16 \quad (35)$$

and

$$f_2(x_i) = \mathrm{logit}(P(Y_{i2} = 1|Y_{i1} = 0, x_i)) = 2\cos(5x_i + 1.4) + x_i^2. \quad (36)$$

If $f_1 = f_2$, then this is reduced to the one-endpoint case in the previous subsection.

Two sample sizes ($n = 200$ and $n = 500$) are used in this simulation. For each sample size, 100 sets of independent data

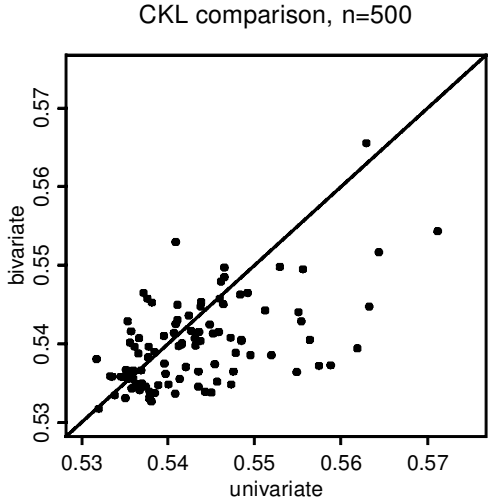


Figure 7. Pairwise Comparison of CKL Distance for the Bivariate Fit and the Univariate Fit.

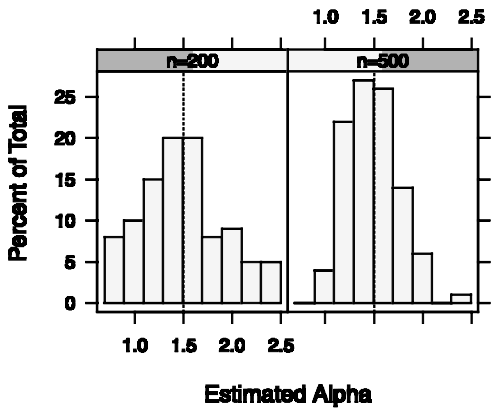


Figure 8. Histograms of estimated $\hat{\alpha}$ for Two Different Sample Sizes. The dotted lines represent the true values of $\alpha = 1.5$. The means of estiamted values are 1.487 and 1.489.

are generated according to the true joint distribution. The predictor variables X_i are assumed to have a uniform distribution over [0, 1]. Only 50 basis functions are selected to generate the approximating subspace for the approximate spline solutions. To compute the randomized version of GACV, we use $R = 20$ replicates to reduce the variance of the estimated values.

Figure 8 presents the histogram plots of the estimated $\hat{\alpha}$ for two different sample sizes. The dotted lines are the true value of $\alpha = 1.5$. The estimated values converge to the truth while sample size increases.

Figure 9 plots the true and estimated conditional probability functions for both endpoints. For each sample size, the 100 fitted values are ranked according to the CKL distance between the estimated joint distribution and the truth. The 5th, 25th, 50th, 75th, and 95th best fits are plotted for both sample sizes. Figure 9(a) shows the conditional probability for the first endpoint, $P(Y_{i1} = 1|Y_{i2} = 0, x_i) = e^{f_1(x_i)}/(1 + e^{f_1(x_i)})$, and Figure 9(b) shows the conditional probability for the second endpoint, $P(Y_{i2} = 1|Y_{i1} = 0, x_i) = e^{f_2(x_i)}/(1 + e^{f_2(x_i)})$.

6. APPLICATION TO THE BEAVER DAM EYE STUDY: PIGMENTARY ABNORMALITIES IN WOMEN

The Beaver Dam Eye Study (BDES) is an ongoing population-based cohort study of age-related eye diseases, cataract and maculopathy. A description of the population and details of the study at baseline were given by Klein, Klein, and Linton (1992). Five-year follow-up data have been collected and analyzed (see, e.g., Klein, Klein, Jensen, and Meuer 1997), and the 10-year follow-up of the cohort is in progress.

A private census of the population of Beaver Dam, Wisconsin was performed from September 15, 1987, to May 4, 1988, to identify the eligible population, which is defined as being age 43–84 years at the time of census. Afterward, the population was examined over a 30-month period. Of the 5,925 eligible people, 4,926 (83.1%) participated in the study. Photographs of each eye were taken and graded; an examination and a standardized questionnaire were administered.

The association of pigmentary abnormalities with six other attributes at the baseline was studied by Lin et al. (2000) using the "univariate" penalized logistic regression. Only the

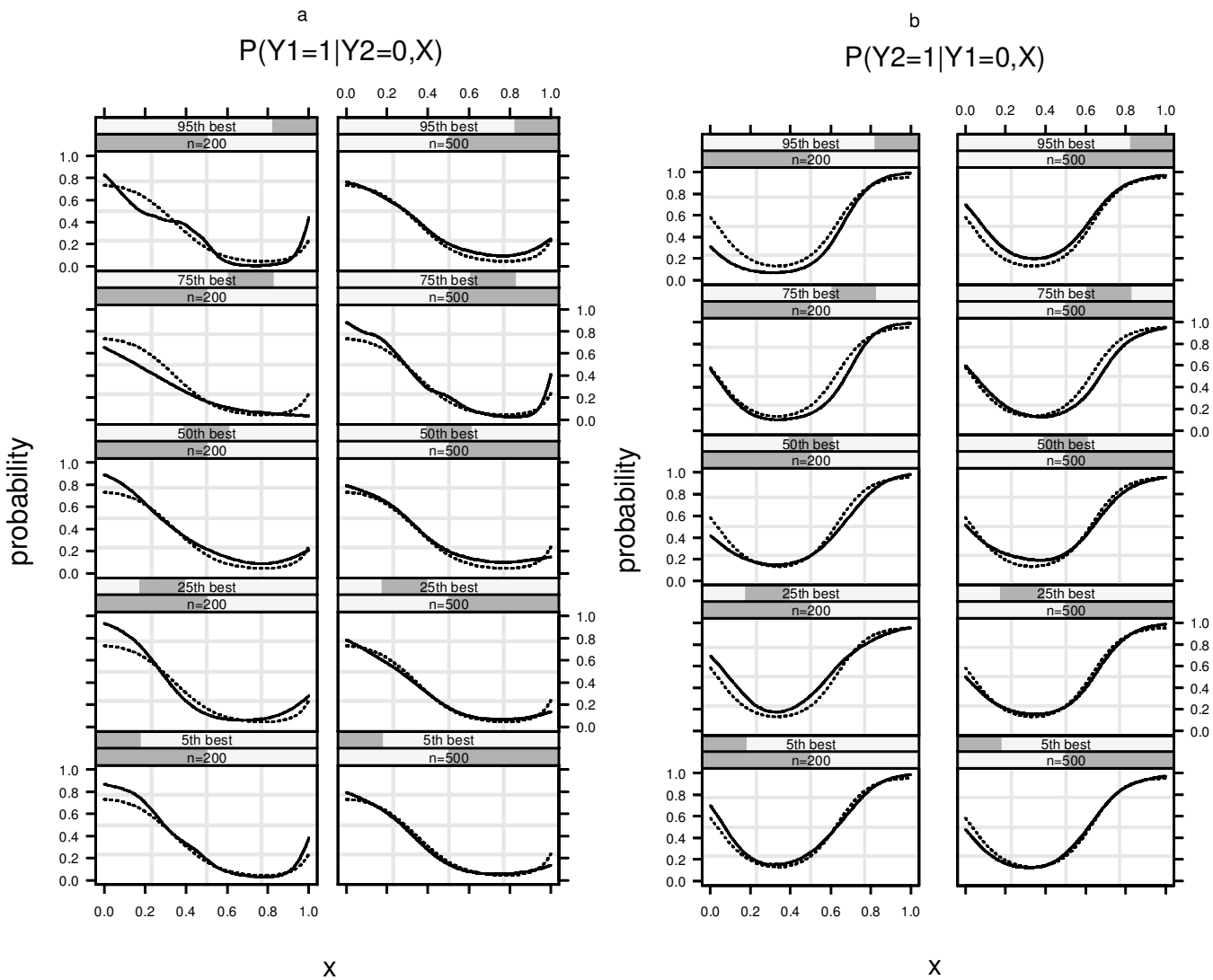


Figure 9. True and Estimated Conditional Probabilities $P(Y_{i1} = 1|Y_{i2} = 0, X_i)$ (a) and $P(Y_{i2} = 1|Y_{i1} = 0, X_i)$(b). Solid lines are the estimated functions, dotted lines represent the true function.

$n = 2{,}585$ women members of the cohort in the baseline with no missing values were considered. Pigmentary abnormalities are an early sign of age-related macular degeneration and are defined by the presence of retinal depigmentation or increased retinal pigmentation in association with retinal drusen. Pigmentary abnormalities were found in 11.88% of the $n = 2{,}585$ cohort studied. Here the question of interest is to estimate the probability of at least one eye developing pigmentary abnormalities given the values of the predictor variables.

Based on previous work, age is known to be a very strong risk factor for the presence of pigmentary abnormalities and other age-related maculopathy in the Beaver Dam Eye Study. The association between cardiovascular disease and its risk factors and the incidence of age-related maculopathy was examined by Klein, Klein, and Jensen (1997). Hormone replacement therapy was associated with a weak protective effect, and a history of heavy alcohol consumption and beer drinking was associated with a deleterious effect for some endpoints (see Klein, Klein, and Ritter 1994; Moss, Klein, Klein, Jensen, and Meuer 1998; Ritter, Klein, Klein, Mares-Perlman, and Jensen 1995). Lin et al. (2000) used multiple linear logistic regression and contingency tables for the preliminary analysis. First, one predictor variable was examined at a time. Only those variables whose p-values are below some threshold (.1) were kept for further analysis. A forward selection procedure was then carried out for the linear logistic regression. Afterward, several candidate nonparametric models were closely examined. If the fitted value of any term had no significant visual effect to the overall fit, then that term was considered to have no practical importance. The six "predictor" variables selected for the final nonparametric model are listed in Table 2.

The model finally fitted is

$$f(x) = C + f_1(\mathsf{sys}) + f_2(\mathsf{chol}) + f_{12}(\mathsf{sys}, \mathsf{chol}) + d_{\mathsf{age}}\mathsf{age} + d_{\mathsf{bmi}}\mathsf{bmi} + d_{\mathsf{horm}} I_1(\mathsf{horm}) + d_{\mathsf{drin}} I_2(\mathsf{drin}). \quad (37)$$

Here I_1 and I_2 are indicator variables. Originally, age and bmi were fitted as smooth main effects, however, visual inspection

Table 2. Predictor Variables for the Beaver Dam Pigmentary Abnormalities Model

Variable	Units	Code
Current usage of hormone replacement therapy	yes/no	horm
History of heavy drinking	yes/no	drin
Body mass index	kg/m^2	bmi
Age	years	age
Systolic blood pressure	mmHg	sys
Serum cholesterol	mg/dL	chol

indicated that they are indistinguishable from linear terms, so that they were set to be linear in the final model. Thus there are five smoothing parameters in the model, one for each of the main effects of sys and chol and another three for the interaction term ($\text{linear}_{\text{sys}} \otimes \text{smooth}_{\text{chol}}$, $\text{smooth}_{\text{sys}} \otimes \text{linear}_{\text{chol}}$ and $\text{smooth}_{\text{sys}} \otimes \text{smooth}_{\text{chol}}$). The results were reported by Lin et al. (1998).

In this section we re-examine the association by using the proposed penalized multivariate logistic regression. $n = 2,495$ women with outcomes available for both eyes are included in the analysis. For reference, the percentiles of the continuous predictor variables are given in Table 3.

We apply the penalized multivariate logistic regression to analyze these data. Here $J = 1$ and $K_1 = 2$. All predictor variables take the same values for both eyes of the same person. The association between fellow eyes is assumed to be a constant, $\alpha = \log[P(1,1|x_i)P(0,0|x_i)/P(1,0|x_i)P(0,1|x_i)]$. The final model takes the same functional form as in (37), although this time on the conditional logit scale. Only 50 basis functions selected by the clustering method are used to fit the final model. To estimate the ranGACV, the number of replicates R is taken to be 20. On convergence, the estimated $\hat{\alpha} = 2.8269$. The naive estimate of odds ratio without adjustment for any covariates is 26.06. The estimated odds ratio from the multivariate model decreases to $\text{OR} = e^{2.8269} = 16.89$. Obviously, the common predictor values for the same person only partly explain the strong association between fellow eyes.

From the estimated joint probability, we can calculate the probability of at least one eye developing the pigmentary abnormalities; $P(Y_{i1} \text{ or } Y_{i2} = 1|X_i) = (2e^{f_i} + e^{2f_i+\alpha})/(1 + 2e^{f_i} + e^{2f_i+\alpha})$. Figures 10 and 11 give the estimated probability of finding pigmentary abnormalities in at least one eye as a function of chol, for various values of sys, age and bmi. In Figure 10, (horm, drin) = (no, no) and in Figure 11, (horm, drin) = (yes, no). A suggestion of a nonlinear protective effect of cholesterol, particularly for those who were older in the horm = no group, may be seen as a result of fitting this model. A protective effect of hormone replacement therapy is still evident from this bivariate model. Figure 12 gives cross-sectional plots of the estimated probabilities along with the 90% Bayesian confidence intervals as a function of chol for both values of horm and four values of age, which are taken to be the middle of the four age groups defined in the Beaver Dam Eye Study. Details of the Bayesian "confidence intervals" are given in Appendix D.

The new analysis basically confirms the result of Lin et al. (2000). The trend of the effect for each predictor variable remains the same. Compared to figures 9–11 of Lin et al. (2000), we notice some small difference between these two models. From the simulation studies, we expect that the new model is closer to the underlying truth. Moreover, we notice that the outcomes for both eyes of the same person are highly correlated ($\text{OR} = e^{2.8269} = 16.89$), even after adjusting for all of the predictor variables in this model. This partly explains why the results from the two models look very similar.

Another merit of this new approach is in estimating the probability $P(Y_k = 1|Y^{(-k)} = 1, \mathbf{X})$. Figure 13 shows this conditional probability as a function of chol. This conditional probability is medically meaningful to a patient who has been diagnosed with a certain disease for one eye. It provides a guideline as to how to reduce the risk of the same disease for the other healthy eye. A referee pointed out that it will also be interesting to estimate the probability of both eyes developing an abnormality, because one might want those patients to have greater priority for receiving health care or prevention measures. We can do so by calculating $P(Y_1 = Y_2 = 1|\mathbf{X}) = e^{2f+\alpha}/(1 + 2e^{f} + e^{2f+\alpha})$.

7. CONCLUSION

We have proposed using penalized multivariate logistic regression with SS-ANOVA models to estimate the joint distribution for multivariate Bernoulli data, given the values of the predictor variables. The estimate is obtained by solving a variational problem involving the penalized likelihood. Numerically, an approximate solution of the minimization problem is obtained by using the block one-step SOR–Newton-Ralphson algorithm. The GACV and ranGACV for multivariate Bernoulli data have been derived, and ranGACV has been used to adaptively select smoothing parameters in every step of the block one-step SOR iteration. The association terms are still kept in simple parametric form in this model. They are estimated iteratively by maximum likelihood in each block one-step SOR updating step. However, we can also generalize to estimate the association term nonparametrically. In principle, once the likelihood is fully specified, we can use the one-step SOR algorithm to fit the model. By leaving out one independent unit at a time, we can also use ranGACV to

Table 3. Percentiles of the Predictor Variables

	Percentile								
	Minimum	*12.5*	*25*	*37.5*	*50*	*62.5*	*75*	*87.5*	*Maximum*
sys(mmHg)	71	108	116	122	129	136	145	157	221
chol(mg/dL)	102	191	210	225	237	252	266.5	290	503
bmi(kg/m^2)	15	22.5	24.25	25.9	27.4	29.5	31.55	35.2	68.4
age(years)	43	48	52	58	62	66	71	76	86

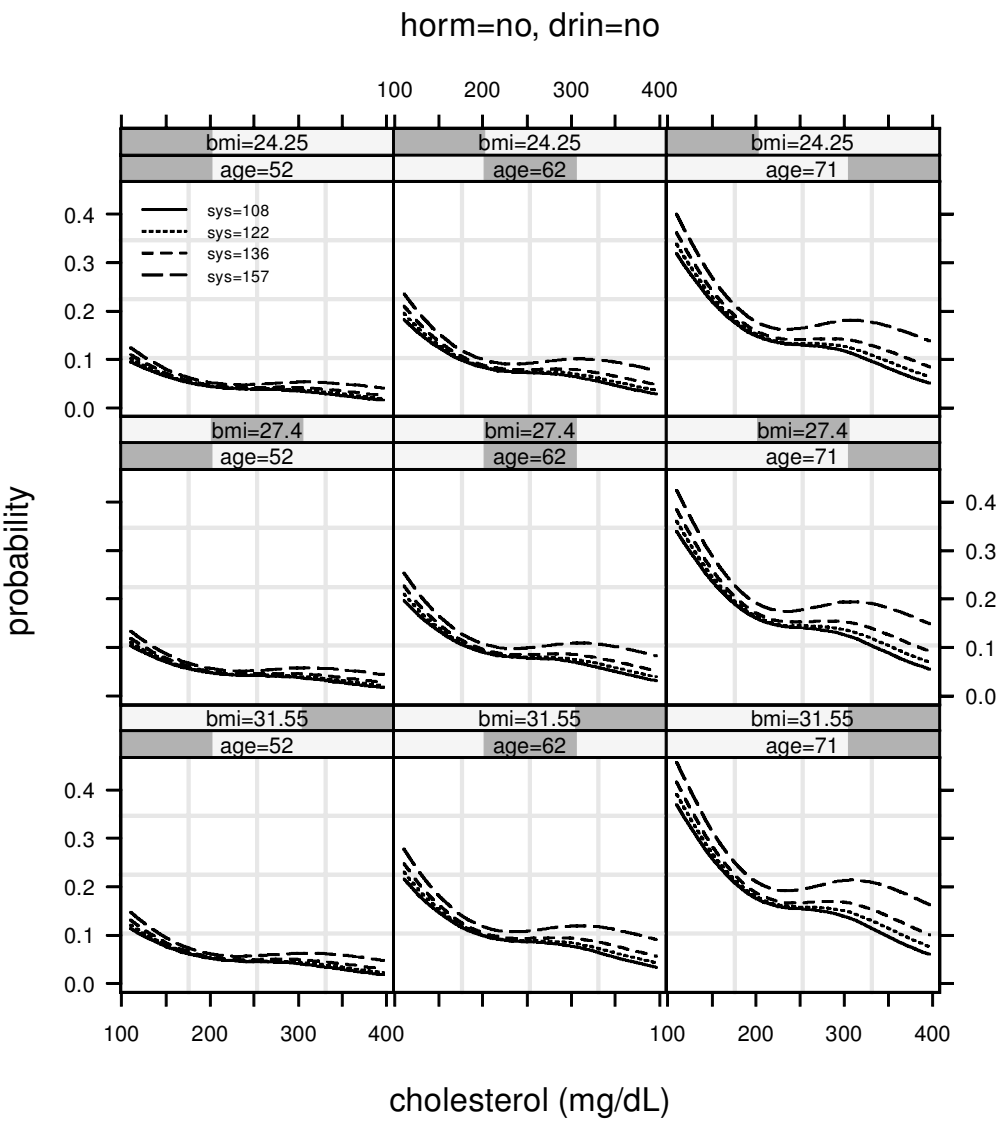


Figure 10. Estimated Probability of at Least One Eye Having Pigmentary Abnormalities as a Function of Cholesterol by Three Levels of age and bmi. horm = no, drin = no.

choose smoothing parameters adaptively. By taking the dependence structure into consideration, we can obtain a partly flexible estimate of the joint probability conditioning on the predictor variables. This approach is particularly useful when the correct form of the function to be estimated is unknown. We successfully applied this method to analyze a medical dataset. Some interesting features of this data set are brought to our attention by the semi parametric model, whereas the more conventional parametric approach is unlikely to reveal such a relationship without further prior knowledge of the data set.

APPENDIX A: PROOF OF THEOREM 1

Before we prove this theorem, we first state two lemmas.

Lemma A.1. Let f_{ijk} denote $f_j(\mathbf{x}_{ijk})$ and let $\alpha_{ij_1k_1,ij_2k_2}$ denote $\alpha_{j_1j_2}(\mathbf{x}_{ij_1k_1}, \mathbf{x}_{ij_2k_2})$. $\mathcal{L}(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha})$ in (8) is a strictly convex function of f_{ijk}'s and $\alpha_{ij_1k_1,ij_2k_2}$'s.

Proof. We need to show that the Hessian is positive definite. To simplify the notation, we relabel $\mathbf{Y}_i = (Y_{ijk})$ to be $(Y_{i1}, \ldots, Y_{iM})$, where $M = \sum_{j=1}^{J} K_j$. We simplify the notation for the $\mathbf{f}$'s and $\boldsymbol{\alpha}$'s similarly. From the property of exponential families, we know that the Hessian with respect to the $\mathbf{f}$'s and $\boldsymbol{\alpha}$'s is $\mathbf{H} = \text{diag}\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_n\}$, where $\mathbf{H}_i$ is the covariance matrix of $\tilde{\mathbf{Y}}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iM}, Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i,M-1}Y_{iM})^T$. Denoting $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iM}, a_{i12}, a_{i13}, \ldots, a_{i,M-1,M})^T$, if $\mathbf{a}_i^T \mathbf{H}_i \mathbf{a}_i = \text{var}(\mathbf{a}_i^T \tilde{\mathbf{Y}}_i) = 0$, then we have $\mathbf{a}_i^T \tilde{\mathbf{Y}}_i = \text{constant}$. We show that $\mathbf{a}_i$ must be a zero vector. First, the constant here must be 0, because we can let all Y_{im}'s be 0. To show $a_{im} = 0$, we let $Y_{im} = 1$ and let the rest of vector $\tilde{\mathbf{Y}}_i$ be 0's. Afterward, to derive $a_{im_1m_2} = 0$, we let the only nonzero elements of the $\tilde{\mathbf{Y}}_i$ vector be $Y_{im_1} = 1$, $Y_{im_2} = 1$, and $Y_{im_1}Y_{im_2} = 1$. This proof also extends to the saturated model.

The following lemma is theorem 4.1 of Gu and Qiu (1993).

Lemma A.2. Suppose that $\mathcal{L}(g)$ is a continuous and strictly convex functional in a Hilbert space, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_1$ has a square norm $\mathcal{J}(g)$ and $\mathcal{H}_0$ is the null space of $\mathcal{J}(g)$ of finite dimension. If $\mathcal{L}(g)$ has a minimizer in $\mathcal{H}_0$, then $\mathcal{L}(g) + \mathcal{J}(g)$ has a unique minimizer in $\mathcal{H}$.

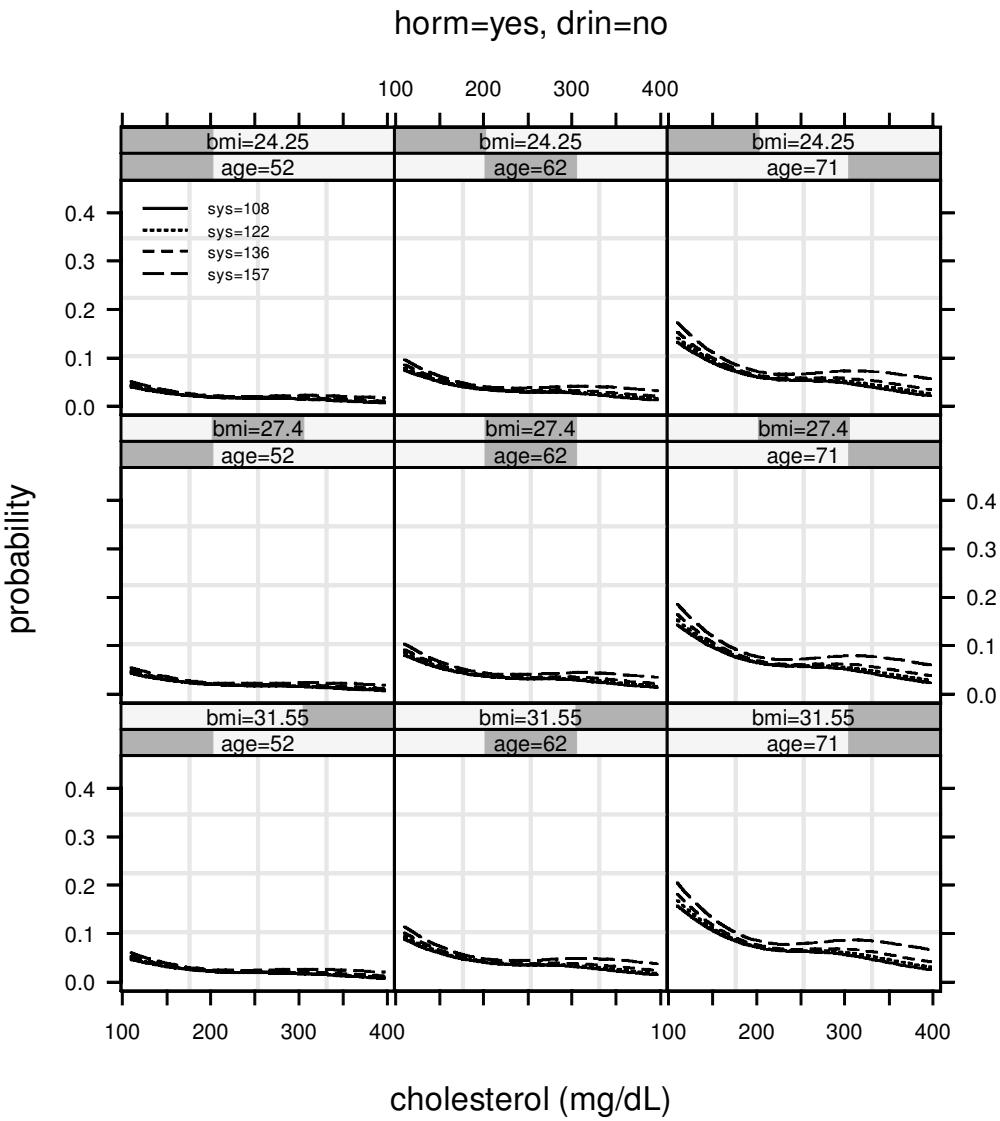


Figure 11. Estimated Probability of at Least One Eye Having Pigmentary Abnormalities as a Function of Cholesterol by Three Levels of age and bmi. horm = yes, drin = no.

Proof of Theorem 1

Define

$$g(x_i, j_1, k_1, j_2, k_2) = \begin{cases} f_j(\mathbf{x}_{ijk}), & 1 \le j = j_1 = j_2 \le J, \\ & 1 \le k = k_1 = k_2 \le K_j, \quad f_j \in \mathcal{H}^j \\ \alpha_{jj}(\mathbf{x}_{ijk_1}, \mathbf{x}_{ijk_2}) & 1 \le j = j_1 = j_2 \le J, \\ & 1 \le k_1 < k_2 \le K_j \\ \alpha_{j_1 j_2}(\mathbf{x}_{ij_1 k_1}, \mathbf{x}_{ij_2 k_2}) & 1 \le j_1 < j_2 \le J, \\ & 1 \le k_1 \le K_{j_1}, \quad 1 \le k_2 \le K_{j_2}. \end{cases}$$

Let $\mathcal{H} = \{g(x_i, j_1, k_1, j_2, k_2) : \mathbf{x}_{ijk} \in \mathcal{X},\ 1 \le j_1 \le j_2 \le J,\ 1 \le k_1 \le K_{j_1},\ 1 \le k_2 \le K_{j_2}\}$. Then $\mathcal{H}$ is a Hilbert space with square seminorm $\mathcal{J}_\Lambda(g) = \mathcal{J}_\Lambda(f_1, \ldots, f_J)$, where $\mathcal{J}_\Lambda$ is defined in (11). Let $\mathcal{L}^*(\mathbf{g}) = \mathcal{L}(\mathbf{y}, \mathbf{f}, \boldsymbol{\alpha})$. By Lemma A.2, it suffices to show that $\mathcal{L}^*(\mathbf{g})$ is continuous and strictly convex in $\mathcal{H}$. Continuity is obvious. Strict convexity follows from Lemma A.1.

APPENDIX B: PROPERTIES OF THE PSEUDODATA

Properties of the pseudo-data of (22), which we invoke in Appendix D to get Bayesian "confidence intervals," are given by the following.

Theorem B.1. For any fixed j, if $|f_{ijk-} - f_{ijk}| = o(1)$ uniformly for $i = 1, 2, \ldots, n$ and $k = 1, \ldots, K_j$, $|\boldsymbol{\alpha}_- - \boldsymbol{\alpha}| = o(\mathbf{1})$ uniformly, $\boldsymbol{\mu}_j(x)$ is uniformly bounded away from 0 and 1, $\boldsymbol{\alpha}$'s are uniformly bounded away from $-\infty$ and ∞. Then

$$\tilde{\mathbf{y}}_{ij} = \mathbf{f}_{ij} + \boldsymbol{\epsilon}_{ij} + o_p(\mathbf{1}),$$

where $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijK_j})^T$ has mean 0 and covariance matrix $\mathbf{W}_{ij}^{-1}$ and $\boldsymbol{\epsilon}_{1j}, \ldots, \boldsymbol{\epsilon}_{nj}$ are independent.

Proof. The proof is given as theorem 3.3 of Gao (1999b), generalizing a similar argument in the univariate case of Gu (1990), but is omitted here for lack of space.

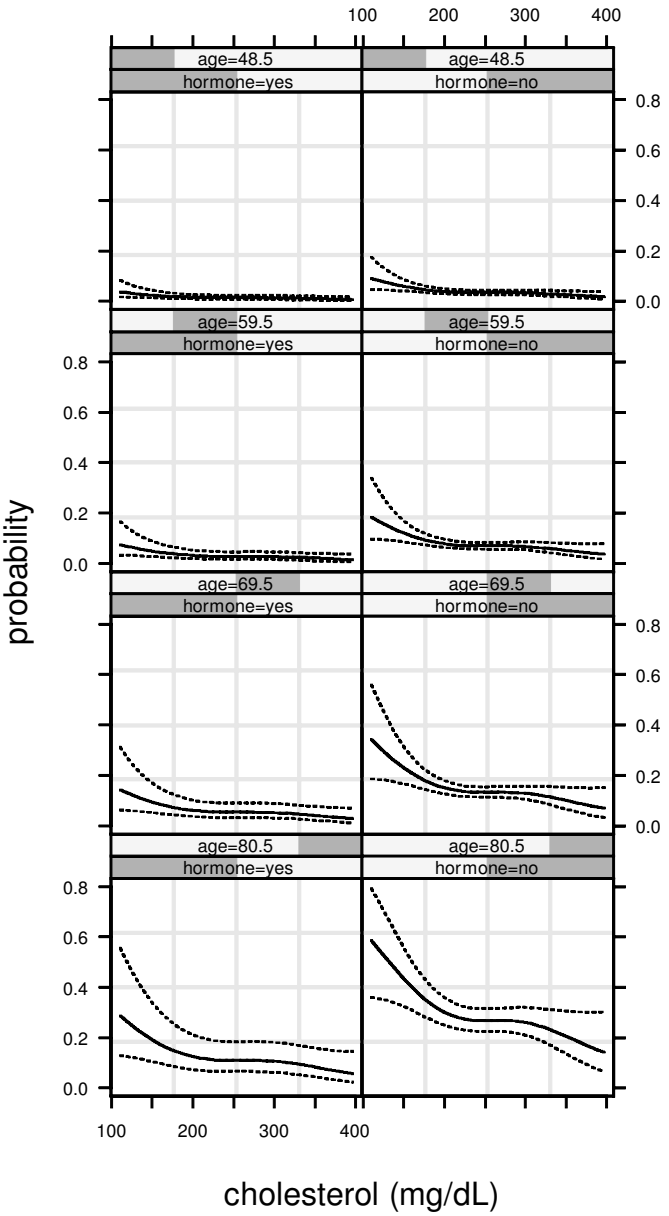


Figure 12. Bayesian Confidence Intervals for the Probability of at Least One Eye Having Pigmentary Abnormalities. bmi and sys are fixed at their Median drin = no.

APPENDIX C: DERIVATION OF GENERALIZED APPROXIMATE CROSS VALIDATION

Before proceeding, we need to generalize the leave-out-one lemma of Craven and Wahba (1979) first. This time, we need to leave out one independent unit at a time, that is, one independent subject.

Lemma C.1 (Leave-Out-One-Subject Lemma). Let $-l_j(\mathbf{y}_{ij}, \mathbf{f}_{ij}) = -\sum_k y_{ijk} f_{ijk} + b(\mathbf{f}_{ij})$ be the part of the likelihood function related to the jth endpoint. All other parts of the likelihood function are considered as fixed. $I_{\Lambda_j}(f_j, \mathbf{Y}_j) = -\sum_i l_j(\mathbf{y}_{ij}, \mathbf{f}_{ij}) + (n/2)\mathcal{J}_{\Lambda_j}(f_j)$. Suppose that $h(i, \mathbf{z}, \cdot)$ is the minimizer of $I_{\Lambda_j}(\mathbf{f}_j, \mathbf{Z})$, where $\mathbf{Z} = (\mathbf{y}_{1j}^T, \ldots, \mathbf{y}_{i-1,j}^T, \mathbf{z}^T, \mathbf{y}_{i+1,j}^T, \ldots, \mathbf{y}_{nj}^T)^T$; then

$$h(i, \boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), \cdot) = f_{\Lambda_j}^{[-i]}(\cdot),$$

where $f_{\Lambda_j}^{[-i]}$ is the minimizer of $\sum_{i_1 \neq i} l(\mathbf{y}_{i_1 j}, \mathbf{f}_{i_1, j}) + (n/2)\mathcal{J}_{\Lambda_j}(f_j)$ and $\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}) = (\mu^{[-i]}(x_{ij1}), \ldots, \mu^{[-i]}(x_{ijK_j}))^T$ is the vector of means corresponding to $f_{\Lambda_j}^{[-i]}(\cdot)$.

Proof. We have

$$-l_j(\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij})) \leq -l_j(\boldsymbol{\mu}^{[-i]}(x_{ij}), f_j(\mathbf{x}_{ij})). \tag{C.1}$$

This follows because setting $\{[-\partial l_j(\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), \boldsymbol{\tau})]/\partial\tau_k\} = -\mu^{[-i]}(\mathbf{x}_{ijk}) + \{[\partial b(\tau)]/\partial\tau_k\} = 0$ and using the fact $\{[\partial^2 b(\tau)]/\partial\tau^T\partial\tau\} > 0$, the foregoing equation implies that $-l_j(\mu^{[-i]}(x_{ij}), \boldsymbol{\tau})$ achieves its unique minimum for $\{[\partial b(\tau)]/\partial\tau_k\} = \mu^{[-i]}(\mathbf{x}_{ijk})$; hence $\tau_k = f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ijk})$. Therefore, for any f_j,

$$\begin{aligned} I_{\Lambda_j}(f_j, Z) &= -l_j(\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), \mathbf{f}_{ij}) - \sum_{i_1 \neq i} l_j(\mathbf{y}_{i_1 j}, \mathbf{f}_{i_1 j}) + \frac{n}{2}\mathcal{J}_{\Lambda_j}(f_j) \\ &\geq -l_j(\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij})) \\ &\quad - \sum_{i_1 \neq i} l_j(\mathbf{y}_{i_1 j}, \mathbf{f}_{i_1 j}) + \frac{n}{2}\mathcal{J}_{\Lambda_j}(f_j) \\ &\geq -l_j(\boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij})) \\ &\quad - \sum_{i_1 \neq i} l_j(\mathbf{y}_{i_1 j}, f_{\Lambda_j}^{[-i]}(\mathbf{x}_{i_1 j})) + \frac{n}{2}\mathcal{J}_{\Lambda_j}(f_{\Lambda_j}^{[-i]}) \end{aligned}$$

The first inequality is due to (C.1); the second one is due to the fact that $f_{\Lambda_j}^{[-i]}(\cdot)$ is the minimizer of $-\sum_{i_1 \neq i} l(\mathbf{y}_{i_1 j}, \mathbf{f}_{i_1, j}) + (n/2)\mathbf{J}_{\Lambda_j}(f_j)$. Therefore, we have $h(i, \boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij}), \cdot) = f_{\Lambda_j}^{[-i]}(\cdot)$.

Let $\mathbf{Y}_j^{[-i]} = (\mathbf{y}_{1j}^T, \ldots, \mathbf{y}_{i-1,j}^T, \boldsymbol{\mu}^{[-i]}(\mathbf{x}_{ij})^T, \mathbf{y}_{i+1,j}^T, \ldots, \mathbf{y}_{nj}^T)^T$. With abuse of notation, let $\mathbf{f}_j$ denote $(f_{ijk}, 1 \leq i \leq n, 1 \leq k \leq K_j)^T$. Because $(\mathbf{f}_{\Lambda_j}, \mathbf{Y}_j)$ and $(\mathbf{f}_{\Lambda_j}^{[-i]}, \mathbf{Y}_j^{[-i]})$ are two local minimizers of $I_{\Lambda_j}(f, \mathbf{Z})$, $\partial I_{\Lambda_j}/\partial\mathbf{f}_j$ is equal to 0 on those two points. Thus, $(\partial I_{\Lambda_j}/\partial\mathbf{f}_j)(\mathbf{f}_{\Lambda_j}, \mathbf{Y}_j) = 0$, and $(\partial I_{\Lambda_j}/\partial\mathbf{f}_j)(\mathbf{f}_{\Lambda_j}^{[-i]}, \mathbf{Y}_j^{[-i]}) = 0$. It is also easy to verify that $(\partial^2 I_{\Lambda_j}/\partial\mathbf{f}_j\partial\mathbf{f}_j^T) = \mathbf{W}_j(f_j) + n\boldsymbol{\Sigma}_{\Lambda_j}$, $(\partial^2 I_{\Lambda_j}/\partial\mathbf{Y}_j\partial\mathbf{f}_j^T) = -\mathbf{I}$, where $\mathbf{W}_j(f_j) = \mathrm{diag}(\mathbf{W}_{1j}, \mathbf{W}_{2j}, \ldots, \mathbf{W}_{nj})$ is as defined earlier and $\boldsymbol{\Sigma}_{\Lambda_j}$ is the positive semidefinite matrix satisfying $\mathcal{J}_{\Lambda_j}(f_j) = \mathbf{f}_j^T\boldsymbol{\Sigma}_{\Lambda_j}\mathbf{f}_j$.

Using a first order Taylor expansion, we have the equation

$$\begin{aligned} \mathbf{0} &= \frac{\partial I_{\Lambda_j}}{\partial\mathbf{f}_j}\left(\mathbf{f}_{\Lambda_j}^{[-i]}, \mathbf{Y}_j^{[-i]}\right) \\ &= \frac{\partial I_{\Lambda_j}}{\partial\mathbf{f}_j}(\mathbf{f}_{\Lambda_j}, \mathbf{Y}_j) + \frac{\partial^2 I_{\Lambda_j}}{\partial\mathbf{f}_j\partial\mathbf{f}_j^T}(f_j^*, \mathbf{Y}_j^*)\left(\mathbf{f}_{\Lambda_j}^{[-i]} - \mathbf{f}_{\Lambda_j}\right) \\ &\quad + \frac{\partial^2 I_{\Lambda_j}}{\partial\mathbf{Y}_j\partial\mathbf{f}_j^T}(\mathbf{f}_j^*, \mathbf{Y}_j^*)\left(\mathbf{Y}_j^{[-i]} - \mathbf{Y}_j\right) \\ &= \frac{\partial^2 I_{\Lambda_j}}{\partial\mathbf{f}_j\partial\mathbf{f}_j^T}(\mathbf{f}_j^*, \mathbf{Y}_j^*)\left(\mathbf{f}_{\Lambda_j}^{[-i]} - \mathbf{f}_{\Lambda_j}\right) \\ &\quad + \frac{\partial^2 I_{\Lambda_j}}{\partial\mathbf{Y}_j\partial\mathbf{f}_j^T}(\mathbf{f}_j^*, \mathbf{Y}_j^*)\left(\mathbf{Y}_j^{[-i]} - \mathbf{Y}_j\right), \end{aligned} \tag{C.2}$$

where $(\mathbf{f}_j^*, \mathbf{Y}_j^*)$ is a point somewhere between $(f_{\Lambda_j}, \mathbf{Y}_j)$ and $(\mathbf{f}_{\Lambda_j}^{[-i]}, \mathbf{Y}_j^{[-i]})$. Equivalently, this is $(\mathbf{f}_{\Lambda_j} - \mathbf{f}_{\Lambda_j}^{[-i]}) = (\mathbf{W}_j(f_j^*) + n\boldsymbol{\Sigma}_{\Lambda_j})^{-1}(\mathbf{Y}_j - \mathbf{Y}_j^{[-i]})$. Approximate $\mathbf{W}_j(f_j^*)$ by $\mathbf{W}_j(f_{\Lambda_j})$ and note that

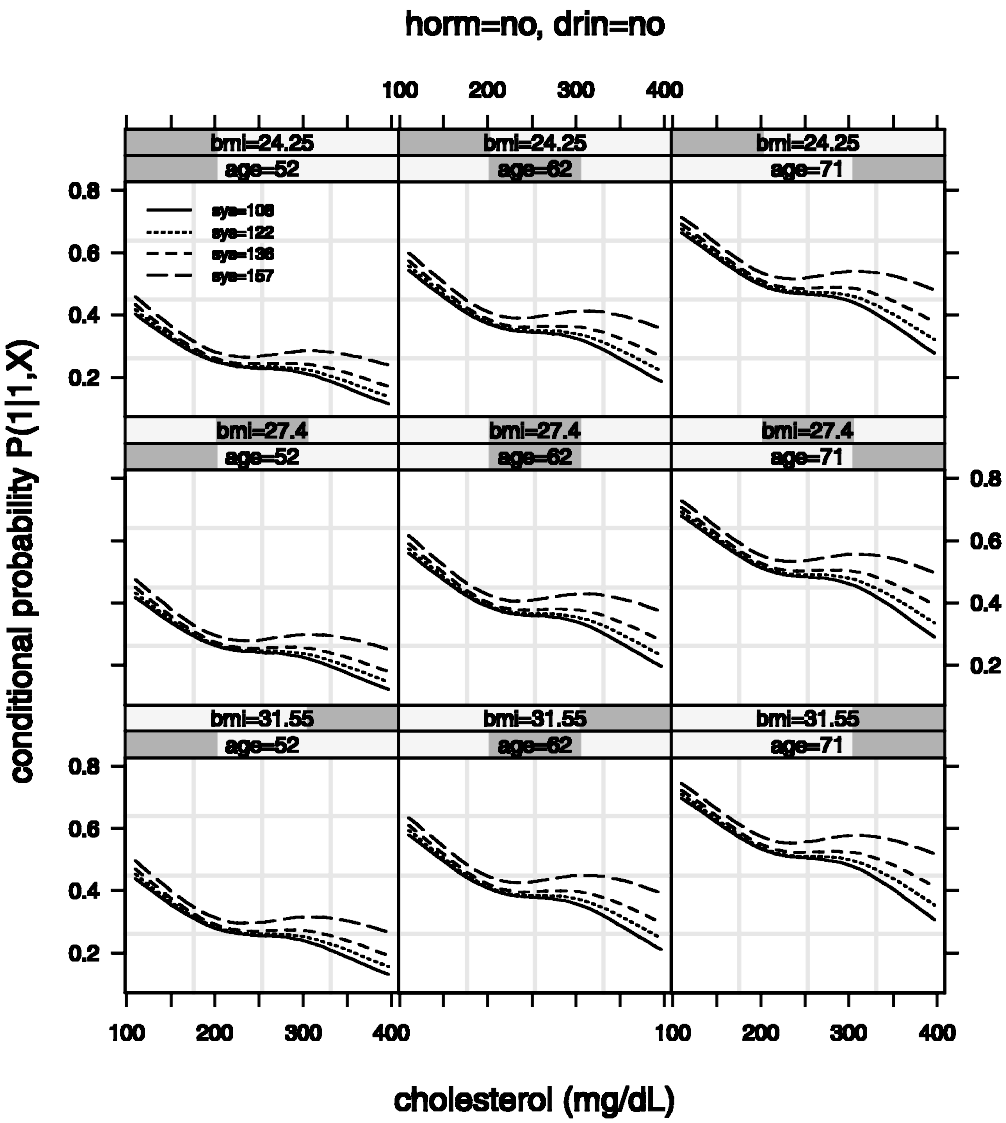


Figure 13. Estimated Probability of One Eye Developing Pigmentary Abnormalities Conditioning on the Other Eye Already Having This Disease as a Function of Cholesterol by Three Levels of age and bmi; horm = no, drin = no.

$\mathbf{Y}_j - \mathbf{Y}_j^{[-i]} = (0, \ldots, 0, (\mathbf{y}_{ij} - \boldsymbol{\mu}^{[-i]}(x_{ij}))^T, 0, \ldots, 0)^T$. We have

$$\begin{pmatrix} f_{\Lambda_j}(\mathbf{x}_{1j1}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{1j1}) \\ \vdots \\ f_{\Lambda_j}(\mathbf{x}_{ij1}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij1}) \\ \vdots \\ f_{\Lambda_j}(\mathbf{x}_{ijK_j}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ijK_j}) \\ \vdots \\ f_{\Lambda_j}(\mathbf{x}_{njK_j}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{njK_j}) \end{pmatrix}_{nK_j \times 1}$$

$$\approx (\mathbf{W}_j(f_{\Lambda_j}) + n\boldsymbol{\Sigma}_{\Lambda_j})^{-1} \begin{pmatrix} 0 \\ \vdots \\ y_{ij1} - \mu^{[-i]}(\mathbf{x}_{ij1}) \\ \vdots \\ y_{ijK_j} - \mu^{[-i]}(\mathbf{x}_{ijK_j}) \\ \vdots \\ 0 \end{pmatrix}_{nK_j \times 1} . \quad \text{(C.3)}$$

Denote $\mathbf{H}^j = [\mathbf{W}_j(f_{\Lambda_j}) + n\boldsymbol{\Sigma}_{\Lambda_j}]^{-1}$, which is the inverse Hessian of $I_{\Lambda_j}(\mathbf{f}_j, \mathbf{Y}_j)$ with respect to $\mathbf{f}_j$ evaluated at $\mathbf{f}_{\Lambda_j}$. $\mathbf{H}^j$ has the structure

$$\mathbf{H}^j = \begin{pmatrix} \mathbf{H}_{11}^j & & & * \\ & \mathbf{H}_{22}^j & & \\ & & \ddots & \\ * & & & \mathbf{H}_{22}^j \end{pmatrix}_{nK_j \times nK_j} , \quad \text{(C.4)}$$

where $\mathbf{H}_{ii}^j$ is the $K_j \times K_j$ submatrix on the diagonal. Hence we have

$$\begin{pmatrix} f_{\Lambda_j}(\mathbf{x}_{ij1}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij1}) \\ \vdots \\ f_{\Lambda_j}(\mathbf{x}_{ijK_j}) - f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ijK_j}) \end{pmatrix} \approx \mathbf{H}_{ii}^j \begin{pmatrix} y_{ij1} - \mu^{[-i]}(\mathbf{x}_{ij1}) \\ \vdots \\ y_{ijK_j} - \mu^{[-i]}(\mathbf{x}_{ijK_j}) \end{pmatrix} . \quad \text{(C.5)}$$

Starting with the ordinary leave-out-one cross validation function $\mathrm{CV}(\Lambda_j)$, we use the foregoing relation and several first order Taylor

expansions in our derivation:

$$\begin{aligned}\mathrm{CV}(\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk}^{[-i]} + b(\mathbf{f}_{ij})\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(\mathbf{f}_{ij}) + \sum_{k=1}^{K_j} y_{ijk}\left(f_{ijk} - f_{ijk}^{[-i]}\right)\right] \\ &= \mathrm{OBS}(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n}(y_{ij1},\ldots,y_{ijK_j}) \\ &\quad\times\left(f_{ij1} - f_{ij1}^{[-i]},\ldots,f_{ijK_j} - f_{ijK_j}^{[-i]}\right)^T. \end{aligned} \tag{C.6}$$

Next, we need to show that the following relation is true. The first approximation is due to a Taylor expansion for a function with vector responses:

$$\begin{aligned}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}\end{pmatrix} &= \begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} + \begin{pmatrix} \mu_{ij1}^{[-i]}-\mu_{ij1} \\ \vdots \\ \mu_{ijK_j}^{[-i]}-\mu_{ijK_j}\end{pmatrix} \\ &= \begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} \\ &\quad + \begin{pmatrix} \frac{\partial b}{\partial f_{ij1}}(f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij})) - \frac{\partial b}{\partial f_{ij1}}(f_{\Lambda_j}(\mathbf{x}_{ij})) \\ \vdots \\ \frac{\partial b}{\partial f_{ijK_j}}(f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij})) - \frac{\partial b}{\partial f_{ijK_j}}(f_{\Lambda_j}(\mathbf{x}_{ij}))\end{pmatrix} \\ &\approx \begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} + \mathbf{W}_{ij}\begin{pmatrix} f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ij1}) - f_{\Lambda_j}(\mathbf{x}_{ij1}) \\ \vdots \\ f_{\Lambda_j}^{[-i]}(\mathbf{x}_{ijK_j}) - f_{\Lambda_j}(\mathbf{x}_{ijK_j})\end{pmatrix} \\ &\approx \begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} - \mathbf{W}_{ij}\mathbf{H}_{ii}^{j}\begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} \\ &= (\mathbf{I}-\mathbf{W}_{ij}\mathbf{H}_{ii}^{j})\begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix}. \end{aligned} \tag{C.7}$$

Hence we have the following approximate relation, which we use to define the ACV function:

$$\begin{aligned}\mathrm{CV}(\Lambda_j) &\approx \mathrm{OBS}(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n}(y_{ij1},\ldots,y_{ijK_j})\mathbf{H}_{ii}^{j}\begin{pmatrix} y_{ij1}-\mu_{ij1}^{[-i]} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}^{[-i]}\end{pmatrix} \\ &\approx \mathrm{OBS}(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n}(y_{ij1},\ldots,y_{ijK_j}) \\ &\quad\times\mathbf{H}_{ii}^{j}(\mathbf{I}-\mathbf{W}_{ij}\mathbf{H}_{ii}^{j})^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}\end{pmatrix} \\ &\equiv \mathrm{ACV}(\Lambda_j). \end{aligned} \tag{C.8}$$

Now define $\mathbf{G}_{ii}^{j} = (\mathbf{I}-\mathbf{W}_{ij}\mathbf{H}_{ii}^{j})$. In a step reminiscent of that used to get from leave-out-one cross validation to GCV in the Gaussian case, we will obtain a generalized form of the ACV. There the diagonal elements of certain matrix were replaced by $1/n$ times its trace. Here, for any matrices $\mathbf{A}_{ii}$, $1 \le i \le n$, $\mathbf{A}_{ii} = \left(a_{i,k_1k_2}\right)_{K\times K}$, $1 \le k_1, k_2 \le K$, we define the *generalized average* of $\mathbf{A}_{ii}$'s in (28). Because $\mathbf{A}$ has a very special structure, it is very easy to obtain the closed form of its inverse,

$$\begin{aligned}\overline{\mathbf{A}}^{-1} &= \frac{1}{\delta-\gamma}\mathbf{I}_{K\times K} - \frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)}\mathbf{e}\mathbf{e}^T \\ &= \begin{pmatrix} \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} \\ -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)}\end{pmatrix}. \end{aligned} \tag{C.9}$$

Hence we define the GACV for a multivariate Bernoulli distribution as

$$\begin{aligned}\mathrm{GACV}(\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(\mathbf{f}_{ij})\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(y_{ij1},\ldots,y_{ijK_j}) \\ &\quad\times\overline{\mathbf{H}}^{j}(\overline{\mathbf{G}}^{j})^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}\end{pmatrix}. \end{aligned} \tag{C.10}$$

We remark that the above formula is reduced to (2.9) of Xiang and Wahba (1996) when $J = 1$ and $K_j = 1$. In practice, we will iteratively choose smoothing parameters in each block nonlinear SOR iteration to minimize GACV.

When only person-specific covariates exist, following the notation defined at the end of Section 3, we can rewrite the foregoing formula to a simpler form, which has a similar form to the original GACV of Xiang and Wahba (1996)

$$\begin{aligned}\mathrm{GACV}(\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}\left[-y_{ij}f_{ij} + b(f_{ij})\right] \\ &\quad + \frac{\mathrm{tr}(\mathbf{H}^j)/n\cdot\sum_{i=1}^{n} y_{ij}(y_{ij}-\mu_{ij})}{n-\mathrm{tr}(\mathbf{W}_j^{1/2}\mathbf{H}^j\mathbf{W}_j^{1/2})}. \end{aligned} \tag{C.11}$$

However, by the new definition, $y_{ij} = \sum_{k=1}^{K_j} y_{ijk}$, and $\mathbf{W}_j$ is an estimated covariance matrix for $(y_{1j},\ldots,y_{nj})^T$.

APPENDIX D: BAYESIAN INFERENCE

To construct the approximate Bayesian "confidence interval" for f_j, we let all other estimated values f_ℓ $(\ell \ne j)$ and $\boldsymbol{\alpha}$ fixed at their estimated values. Theorem B.1 shows that the pseudo-data defined in Section 3 have approximately the usual data structure. We use this observation to construct the approximate Bayesian confidence interval. The following argument is a straightforward extension of work of Lin et al. (1998) to multivariate observations. An approach similar to that taken by Silverman (1985) is adapted for the approximate solution to the variational problem. In the case where all basis functions are used to solve the variational problem, the Bayesian confidence intervals are known from theoretical and simulation results to

have the across-the-function property; this means that the expected number of the true points that will be covered by a 95% confidence interval is approximately $.95n$ (see Wahba, et al. 1995, Sec. 4). The arguments there apply to each f_j here, if all other f_ℓ $(\ell \neq j)$ and $\boldsymbol{\alpha}$'s are considered fixed. To the extent that the solutions computed from an appropriate subset of the representers are a good approximation to the exact solution, Bayesian confidence intervals computed via the same representers will be a good approximation to the exact confidence intervals. Simulation results of Lin et al. (1998) demonstrate that the Bayesian confidence intervals so computed have this across-the-function property.

First, we consider the Bayesian formulation of the variational problem associated with correlated Gaussian observations. For fixed smoothing parameter(s), we will identify the variational problem with a Bayesian problem. Assume that there is only one endpoint. $j = 1$. On domain $\mathcal{X}$, $y_{ik} = f_1(\mathbf{x}_{ik}) + \epsilon_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, K$, where $(\epsilon_{i1}, \ldots, \epsilon_{iK})$, $i = 1, \ldots, n$ are iid distributed as $N(0, \mathbf{W}_{i1}^{-1})$, with $\mathbf{W}_{i1}$ a known positive-definite matrix, $\mathbf{W}_1 = \text{diag}(\mathbf{W}_{11}, \mathbf{W}_{21}, \ldots, \mathbf{W}_{n1})$. The approximate solution of $f_1(\cdot)$ is a combination of the selected basis functions

$$f_1(\cdot) = \boldsymbol{\phi}^1(\cdot)^T \mathbf{d}^1 + \boldsymbol{\xi}_V^1(\cdot)^T \mathbf{c}_V^1, \tag{D.1}$$

where $\boldsymbol{\phi}^1(\cdot)^T$ and $\boldsymbol{\xi}_V^1(\cdot)^T$ are as in (19) on setting $j = 1$. Recall the definition of $\mathbf{Q}_{j,V}$ and $\mathbf{Q}_{j,V}^*$ in (18). By assuming an improper prior distribution on the coefficient $\mathbf{d}^1$ and assuming a normal distribution for $\mathbf{c}_V^1$, we let their log-density function take the form

$$l_{\text{prior}}(\mathbf{c}_V^1, \mathbf{d}^1) \stackrel{c}{=} -\frac{n\lambda}{2} \mathbf{c}_V^{1\,T} \mathbf{Q}_{1,V}^* \mathbf{c}_V^1, \tag{D.2}$$

where the notation "$\stackrel{c}{=}$" means "equals up to a constant." Following some standard Bayesian manipulation, the posterior log-likelihood has the form

$$l_{\text{post}}(\mathbf{c}_V^1, \mathbf{d}^1) \stackrel{c}{=} -\frac{n\lambda}{2} \mathbf{c}_V^{1\,T} \mathbf{Q}_{1,V}^* \mathbf{c}_V^1 - \frac{1}{2}(\mathbf{y} - \mathbf{Q}_{1,V}\mathbf{c}_V^1 - \mathbf{S}_1\mathbf{d}^1)^T \times \mathbf{W}_1(\mathbf{y} - \mathbf{Q}_{1,V}\mathbf{c}_V^1 - \mathbf{S}_1\mathbf{d}^1). \tag{D.3}$$

Hence by minimizing the posterior negative log-likelihood of $(\mathbf{c}_V^1, \mathbf{d}^1)$ with this prior, we obtain exactly the same solution as solving the variational problem in the approximating subspace.

From (D.3), $(\mathbf{c}_V^1, \mathbf{d}^1)$ in fact has a proper posterior distribution as a multivariate normal with mean $(\hat{\mathbf{c}}_V^1, \hat{\mathbf{d}}^1)$ and covariance matrix $\mathbf{M}^{-1}$, where

$$\mathbf{M} = \begin{pmatrix} \mathbf{Q}_{1,V}^T \mathbf{W}_1 \mathbf{Q}_{1,V} + n\lambda \mathbf{Q}_{1,V}^* & \mathbf{Q}_{1,V}^T \mathbf{W}_1 \mathbf{S}_1 \\ \mathbf{S}_1^T \mathbf{W}_1 \mathbf{Q}_{1,V} & \mathbf{S}_1^T \mathbf{W}_1 \mathbf{S}_1 \end{pmatrix} \tag{D.4}$$

and

$$\begin{pmatrix} \hat{\mathbf{c}}_V^1 \\ \hat{\mathbf{d}}_V^1 \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} \mathbf{Q}_{1,V}^T \\ \mathbf{S}_1^T \end{pmatrix} \mathbf{W}_1 \mathbf{Y}. \tag{D.5}$$

Hence the posterior distribution of $f_1(\cdot) = \boldsymbol{\phi}^1(\cdot)^T \mathbf{d}^1 + \boldsymbol{\xi}_V^1(\cdot)^T \mathbf{c}_V^1$ is as

$$E(f_1(x)|\mathbf{Y}, \lambda) = (\boldsymbol{\phi}^1(x)^T \quad \boldsymbol{\xi}_V^1(x)^T)\mathbf{M}^{-1} \begin{pmatrix} \mathbf{Q}_{1,V}^T \\ \mathbf{S}_1^T \end{pmatrix} \mathbf{W}_1 \mathbf{Y}, \tag{D.6}$$

$$\text{cov}(f_1(x), f_1(x')|\mathbf{Y}, \lambda) = (\boldsymbol{\phi}^1(x)^T \quad \boldsymbol{\xi}_V^1(x)^T)\mathbf{M}^{-1} \begin{pmatrix} \boldsymbol{\phi}^1(x') \\ \boldsymbol{\xi}_V^1(x') \end{pmatrix}. \tag{D.7}$$

For multivariate Bernoulli data, the block one-step SOR–Newton-Ralphson algorithm iteratively reformulates the optimization problem as a penalized weighted least squares problem (22). By dealing with the pseudo-data, the approximate posterior variance of $f_j(\cdot)$ is at hand. The confidence intervals in Figure 12 are computed based on the foregoing argument, whereas α is considered as fixed.

One of our future research areas is to construct confidence interval for $\boldsymbol{\alpha}$. It may be reasonable to extend this Bayesian argument to calculate the posterior variance of $\boldsymbol{\alpha}$. More empirical evidence or theoritical justification is needed for taking this approach.

[Received July 1999. Revised March 2000.]

REFERENCES

Aronszajn, N. (1950), "Theory of Reproducing Kernels," *American Mathematical Society*, 68, 337–404.

Berhane, K., and Tibshirani, R. (1998), "Generalized Additive Models for Longitudinal Data," *The Canadian Journal of Statistics*, 26, 517–535.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analyses: Theory and Practice*, Cambridge, MA: MIT Press.

Brumback, B. A., and Rice, J. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves," *Journal of the American Statistical Association*, 93, 961–975.

Carey, V., Zeger, S. L., and Diggle, P. (1993), "Modelling Multivariate Binary Data With Alternating Logistic Rregressions," *Biometrika*, 80, 517–526.

Cessie, S. L., and Houwelingen, J. C. V. (1994), "Logistic Regression for Correlated Binary Data," *Applied Statistics*, 43, 95–108.

Chambers, J., and Hastie, T. (1992), *Statistical Models in S*, Wadsworth and Brooks, Pacific Grove, CA.

Cox, D. R. (1972), "The Analysis of Multivariate Binary Data," *Applied Statistics*, 21, 113–120.

Craven, P., and Wahba, G. (1979)," Smoothing Noisy Data With Spline Functions," *Numer. Math.* 31, 377-403.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, New York: Oxford University Press.

Fitzmaurice, G. M., and Laird, N. M. (1993), "A Likelihood-Based Method for Analysis Longitudinal Binary Responses," *Biometrika*, 80, 141–151.

Gao, F. (1999a), "Iterated *ranGACV*: A Computational Proxy for the Comparative Kullback–Leibler Distance," Technical Report 1011, University of Wisconsin-Madison, Dept. of Statistics. WI.

——— (1999b), "Penalized Multivariate Logistic Regression With a Large Data Set," unpublished doctoral thesis, University of Wisconsin-Madison, Dept. of Statistics.

Glonek, G. F. V., and McCullagh, P. (1995), "Multivariate Logistic Models," *Journal of the Royal Statistical Society*, Ser. B, 57, 533-546.

Gu, C. (1990), "Adaptive Spline Smoothing in Non-Gaussian Regression Models," *Journal of American Statistical Assocation*, 85, 801–807.

Gu, C., and Qiu, C. (1993), "Smoothing Spline Density Estimation: Theory," *The Annals of Statistics*, 21, 217–234.

Gu, C., and Wahba, G. (1993), "Smoothing Spline ANOVA With Component-Wise Bayesian Confidence Intervals," *Journal of Computational and Graphical Statistics*, 2, 97–117.

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

Heagerty, P. J., and Zeger, S. L. (1996), "Marginal Regression Models for Clustered Ordinal Measurements," *Journal of the American Statistical Association*, 91, 1024–1036.

——— (1998), "Lorelogram: A Regression Approach to Exploring Dependence in Longitudinal Categorical Responses," *Journal of the American Statistical Association*, 93, 150–162.

Jones, B., and Kenward, M. G. (1989), *Design and Analysis of Cross-Over Trials*, London: Chapman Hall.

Jordan, M. (1998), *Learning in Graphical Models*, Kulwer.

Katz, J., Zeger, S., and Liang, K.-Y. (1994), "Approporiate Statistical Methods to Account for Similarities in Binary Outcomes Between Fellow Eyes," *Investigate Ophthalmology and Visual Science*, 35, 2461–2465.

Kimeldorf, G., and Wahba, G. (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis Applications*, 33, 82–95.

Klein, R., and Klein, B. E., and Jensen, S. (1997), "The Relation of Cardiovascular Disease and Its Risk Factors to the 5-Year Incidence of Age-Related Maculopathy: The Beaver Dam Eye Study," *Ophthalmology*, 104, 1804–1812.
Klein, R., Klein, B. E., Jensen, S., and Meuer, S. (1997), "The Five-Year Incidence Progression of Age-Related Maculopathy: The Beaver Dam Eye Study," *Ophthalmology*, 104, 7–21.
Klein, R., Klein, B. E., and Linton, K. (1992), "Prevalence of Age-Related Maculopathy: The Beaver Dam Eye Study," *Ophthalmology*, 99, 933–943.
Klein, R., Klein, B. E., and Ritter, L. (1994), "Is There Evidence of an Estrogen Effect on Age-Related Lens Opacities? The Beaver Dam Eye Study," *Arch. Ophthamology*, 112, 85–91.
Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analyses for Categorical Data (disc: P24-40),"*Journal of the Royal Statistical Society*, Ser. B, 54, 3–24.
Lin, X. (1998), "Smoothing Spline ANOVA for Polychotomous Response Data," Technical Report 1003, University of Wisconsin-Madison, Dept. of Statistics.
Lin, X., Wahba, G. Xiang, D., Gao, F., Klein, R., and Klein, B. (2000), "Smoothing Spline ANOVA Models for Large Data Sets with Bernoulli Observations and the Randomized GACV," unpublished manuscript submitted to *The Annals of Statistics.*
Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Model Using Smoothing Splines," *Jounral of the Royal Statistical Society,* Ser, B, 61, 381–400.
Lin, Y. (2000), "Tensor Product Space ANOVA Models," *The Annals of Statistics* 28, 734–755.
Lisitz, S. R., Laird, N. M., and Harrington, D. P. (1991), "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association," *Biomtrika*, 78, 153–160.
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear models* (2nd ed.), London: Chapman and Hall.
Molenberghs, G., and Ritter, L. L. (1996), "Methods for Analyzing Multivariate Binary Data, With Associaiton Between Outcomes of Interest," *Biometrics*, 52, 1121–1133.
Moss, S., Klein, R., Klein, B. E., Jensen, S., and Meuer, S. (1998), "Alcohol Consumption and the 5-Year Incidence of Age-Related Maculopathy: The Beaver Dam Eye Study," *Ophthalmology* 105, 789–794.
Ortega, J., and Rheinboldt, W. (1970), *Iteration Solution of Nonlinear Equations in Several Variables*, New York: Academic Press.
O'Sullivan, F. (1983), "The Analysis of Some Penalized Likelihood Estimation Schemes," Technical Report 726, University of Wisonsin-Madison, Dept. of Statistics.
Qu, Y., Williams, G. W., Beck, G. J., and Goormastic, M. (1987), "A Generalized Model of Logistic Regression for Clustered Data," *CommStA* 16, 3447–3476.
Ritter, L., Klein, R., Klein, B. E., Mares-Perlman, J., and Jensen, S. (1995), "Alcohol Use and Agerelated Maculopathy in the Beaver Dam Eye Study," *American Journal of Ophthalmology*, 120, 190–196.
Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, 47, 1–21.
Wahba, G., (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.
Wahba, G. Lin, X., Gao, F., Xiang, D., Klein, R., and Klein, B. (1999), "The Bias-Variance trade-off and the Randomized GACV," in *Advances in Information Processing Systems 11*, eds. M. Kearns, S. Solla, and D. Cohn, Cambridge, MA: MIT Press, pp. 620–626.
Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," *The Anals of Statistics*, 23, 1865–1895.
Wang, Y. (1998a), "Mixed-Effects Smoothing Spline ANOVA," *Journal of the Royal Statistical Society*, Ser. B, 60, 159–174.
——— (1998b), "Smoothing Spline Models With Correlated Random Errors," *Journal of the American Statistical Association*, 93, 341–348.
Wang, Y., and Brown, M. B. (1996), "A Flexible Model for human Circadian Rhythms," *Biometrics*, 52, 588–596.
Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics*, New York: Wiley.
Wild, C. J., and Yee, T. W. (1996), "Additive Extension to generalized Estimating Equation Methods," *Journal of the Royal Statistical Society*, Ser. B, 58, 711–725.
Williamson, J. M., Kim, K., and Lipsitz, S. R. (1995), "Analyzing Bivariate Ordinal Data Using a Global Odds Ratio," *Journal of the American Statistical Association*, 90, 1432–1437.
Xiang, D., and Wahba, G. (1996), "A Genralized Approximate Croos-Validation for Smoothing Splines with non-Gaussian dtata," *Statistica Sinica*, 6, 675–692.
Yee, T. W., and Wild, C. J. (1996), "Vector Generalized Additive Models," *Journal of the Royal Statistical Society*, Ser. B, 58, 481–493.
Zhao, L., and Prentice, R. L. (1990), "Correlated Binary Regression Using a Quadratic Exponential Model," *Biometrika*, 77, 642–648.
Zhao, L. P., Prentice, R. L., and Self, S. G. (1992), "Multivariate Mean Parameter Estimation by Using a Partly Exponential Model," *Journal of the Royal Statistical Society*, Ser. B, 805–811.

Comment

T. W. YEE and C. J. WILD

Over the last two decades, much attention has been directed toward the analysis of correlated data and the field of nonparametric regression. The article is an important contribution to the cross-fertilization of these two areas.

1. EXTENSIONS

Our main comment is that we believe the class of models presented in the article can be extended in several important directions beyond log-linear models for correlated binary data. Wahba et al. (1995) gave a general framework for applying SS-ANOVA to models in exponential families. Their methodology handled a univariate response, and it was shown that, with fixed smoothing parameters, backfitting was an alternative method of solution. The present article extends SS-ANOVA to vector responses but only within a log-linear binary data setting. We believe that the SS-ANOVA idea can be further extended to a multivariate exponential family, or at least something similar to the vector generalized additive model (VGAM) class; see Section 2.

There are advantages in embedding the SS-ANOVA models in a larger class for which similar models fall out easily as special cases; for example, log-linear models for correlated Poisson counts and other methods for correlated binary data discussed by Liang et al. (1992). A large framework facilitates the construction of modular software and gives the data analyst a flexible but coherent modeling environment in which to work with a large array of models. Generalized linear models (GLMs) are a prime example. Most of our suggestions for

T. W. Yee is Lecturer and C. J. Wild is Professor, Department of Statistics, University of Auckland, New Zealand.

generalizations concern using the authors' SS-ANOVA ideas to extend the VGAM class. We conclude this section with a few suggestions that are not of that form.

One variant of SS-ANOVA is suggested by recent work of T. W. Yee with T. Hastie. Consider the decomposition

$$\beta_0 + \sum_{r=1}^{R} f_r\left(\mathbf{c}_r^T\mathbf{x}\right) + \sum_{r<s} f_{rs}\left(\mathbf{c}_r^T\mathbf{x}, \mathbf{c}_s^T\mathbf{x}\right) + \cdots . \tag{1}$$

Here R is much smaller than the number of predictors D, and the $\mathbf{c}_r^T\mathbf{x}$ are like latent variables. These help overcome the curse of dimensionality inherent in SS-ANOVA models when D is large and there are interaction terms of degree 2 or higher. Even when D is only moderately large but there are many interaction terms $f_{k\ell}$, some reduction in the number of parameters via (1) may be advantageous. We note that the main effects of (1) are those of projection pursuit regression and have been applied to models in the exponential family by Rosen and Hastie (1993).

In recent years, some members of the local regression school of smoothing have started using splines. There may be benefits from members of the spline school also considering other smoothing methods—in particular, local regression methods (Fan and Gijbels 1996), which are also widely applied. Moreover, much flexibility is gained when local scoring and smoothing are separated in the estimation process (cf. Sec. 2). Consideration might also be given to local likelihood estimation (Loader 1999). This is certainly feasible with only a single covariate, and the software described in Section 3 can be used to help compute such estimates.

2. VECTOR GENERALIZED LINEAR MODELS AND VECTOR GENERALIZED ADDITIVE MODELS

VGAMs are a sufficiently large class of models to accommodate many data types encountered in practice. There are many points of contact between the theory of the article and that of VGAMs. We originally defined the VGAM class as any model for which the conditional distribution of $\mathbf{y}$ (which may be multivariate) given $\mathbf{x}$ is of the form

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{y}, \eta_1, \ldots, \eta_M), \tag{2}$$

where $h(\cdot)$ is some known function and

$$\eta_j(\mathbf{x}) = \beta_{(j)0} + \sum_{k=1}^{D} f_{(j)k}x(k) \tag{3}$$

are additive predictors. For example, if $Y \sim t_\nu$, then we may want to model the (positive) degrees of freedom ν in terms of covariates using $\eta(\mathbf{x}) = \log\nu(\mathbf{x})$. When the component functions $f_{(j)k}$ are constrained to be linear, the result is a vector generalized linear model (VGLM), with $\eta_j = \boldsymbol{\beta}_j^T\mathbf{x}$, say. These are a superset of the GLM class.

If the contribution to the log-likelihood of the ith "individual" or cluster ℓ_i is strictly concave in each η_j, then one can apply Newton-Raphson or Fisher scoring to maximize the log-likelihood $\ell = \sum\ell_i$. For the VGAM class, this results in iteratively reweighted least squares (IRLS). In particular, the adjusted dependent vectors

$$\mathbf{z}_i = \boldsymbol{\eta}_i + \mathbf{W}_i^{-1}(\partial\ell_i/\partial\boldsymbol{\eta}) \tag{4}$$

are regressed on the $\mathbf{x}_i$ at each iteration. (In fact, the likelihood need not even be specified; e.g., generalized estimating equation modeling, Wild and Yee 1996). A huge advantage of estimating VGAMs by separating the local scoring and smoothing procedures, (e.g., by backfitting) is that one can easily make extensions. Some examples follow.

Example 1: Reduced-Rank Vector Generalized Linear Models. Yee and Hastie (2000) proposed the class of reduced-rank VGLMs (RR-VGLMs) where the matrix $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ is approximated by a lower-rank matrix via the product of two low-rank matrices ($\mathbf{A}$ and $\mathbf{C}$, say). RR-VGAMs can be estimated by minimizing over $\boldsymbol{\eta}_0$, $\mathbf{A}$, and $\mathbf{C}$ the quantity

$$\sum_{i=1}^{n}\{\mathbf{z}_i - \boldsymbol{\eta}_0 - \mathbf{A}\mathbf{C}^T\mathbf{x}_i\}^T\mathbf{W}_i\{\mathbf{z}_i - \boldsymbol{\eta}_0 - \mathbf{A}\mathbf{C}^T\mathbf{x}_i\}$$

at each local scoring algorithm. Certain independently proposed statistical models have been identified as belonging to the RR-VGLM subclass.

Example 2: Vector Local Regression. Welsh and Yee (2000) proposed local regression estimators for vector responses and derived some of their asymptotic properties. The practical outcome of this is that it could allow one to use LOESS-type code as a smoothing option for fitting VGAMs.

Example 3: Vector Generalized Analysis of Multivariate Models. Prompted by the article, we now see that it would be useful to extend the VGAM class to allow ANOVA-type terms. That is, for $j = 1, \ldots, M$,

$$\eta_j(\mathbf{x}) = \beta_{(j)0} + \sum_k f_{(j)k}(x_k) + \sum_{k<\ell} f_{(j)k\ell}(x_k, x_\ell) + \cdots . \tag{5}$$

This could be likened to functional MANOVA applied to regression models outside the exponential family. We could call this class *vector generalized MANOVA models (VGMMs)*. Estimation of VGMMs would simply involve fitting a MANOVA model to the $\mathbf{z}_i$ against $\mathbf{x}_i$ with weights

$$\mathbf{W}_i = -\frac{\partial^2\ell_i}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}$$

[or $E(\mathbf{W}_i)$] at each local scoring iteration. For $M = 1$, the interaction terms could be estimated by thin-plate splines (Wahba 1990), and because of the separation between local scoring and backfitting, other multivariate smoothers such as the bivariate local linear regression estimator (Fan and Gijbels 1996, sec. 7.8; Ruppert and Wand 1994) could be used. For $M > 1$, vector versions of these could be proposed. One advantage of splines over other smoothers, however, is the justification by penalized likelihood. Indeed, VGAMs using vector splines can be justified by this means (Yee and Wild 1994); the derivation is a natural extension to section 6.5.2 of Hastie and Tibshirani (1990). VGAMs estimated in this way result in an exact, rather than approximate, solution to a penalized likelihood problem.

It can be noted that the *vector linear model* (VLM) is the central model behind VGLMs and VGAMs. The crux of the algorithm is to minimize the quantity.

$$\sum_{i=1}^{n}\left(\mathbf{z}_i-\sum_{k=1}^{D}\mathbf{B}_k\boldsymbol{\beta}_k^*x_{ik}\right)^T\mathbf{W}_i\left(\mathbf{z}_i-\sum_{k=1}^{D}\mathbf{B}_k\boldsymbol{\beta}_k^*x_{ik}\right),$$

where the $\mathbf{B}_k$ are known constraint matrices of full column rank. With no constraints ($\mathbf{B}_k=\mathbf{I}_M$), that is equivalent to fitting the model

$$\mathbf{z}_i=\begin{pmatrix}\mathbf{x}_i^T\boldsymbol{\beta}_1\\ \vdots\\ \mathbf{x}_i^T\boldsymbol{\beta}_M\end{pmatrix}+\boldsymbol{\varepsilon}_i,\qquad \boldsymbol{\varepsilon}_i\sim(\mathbf{0},\mathbf{W}_i^{-1}),\qquad i=1,\ldots,n.$$

When $D=1$, the nonparametric extension has the properties described in Appendix B of the article.

The authors are to be commended for tackling the problem of automatic selection of smoothing parameters. Having good automatic data-based selection of smoothing parameters available is clearly superior to needing to rely on fixed values (although automatic selections cannot always be trusted, even in much simpler problems). In our work we have considered only the fixed case (by assigning some effective degrees of freedom to each component function). The reason for this is that the problem of smoothing parameter selection is very difficult for general multivariate responses and predictors. We would dearly love the authors to apply their theory and some of their analytical firepower to VGAMs! Fixed smoothing parameters are often adequate for exploratory data analysis where we typically want to try out a large number of candidate models. In such situations, it is important that each set of estimates be obtained promptly, so we must sacrifice some quality of estimation for speed. When optimally choosing about seven smoothing parameters, Wahba et al. (1995) found that $n=800$ was about the size limit of datasets that could be handled when fitting certain SS-ANOVA models on a fast workstation (and taking up to 8 hours).

One advantage of VGAMs is the availability of residuals for diagnostics. For example, the working residuals

$$\mathbf{r}_i^W=\boldsymbol{W}_i^{-1}(\partial\ell_i\partial/\boldsymbol{\eta})$$

and the Pearson residuals

$$\mathbf{r}_i^P=\boldsymbol{W}_i^{1/2}\mathbf{r}_i^W=\boldsymbol{W}_i^{-1/2}(\partial\ell_i/\partial\boldsymbol{\eta})$$

are always defined for the VGAM class.

2.1 Constraints on the Functions

In the first experiment of Section 5.1 of the article, the same smoothing function is used for both $k=1,2$; that is, $f_k(x_{ik})\equiv f(x_{ik})$. Here x_{ik} could be the ocular pressure of the kth eye of the ith person, for example. Before we wrote this comment, our description of VGAMs had not allowed the same smooth to be applied in different $\eta_j s$ to variables x_{ik} that varied within an individual. Variables that were constant within an individual could be constrained to have the same effect for different η_j's in a straightforward way using linear constraints (Yee and Wild 1996). It is easy to handle the extension in the parametric case; we now do so for the nonparametric case.

2.2 Solving the $f(x_{ik})$ Smoothing Problem

Suppose that we wish to fit

$$y_{ik}=f(x_{ik})+\varepsilon_{ik},\qquad \varepsilon_i\sim(\mathbf{0},\boldsymbol{W}_i^{-1})\quad\text{independently,}\quad(6)$$

with splines, $i=1,\ldots,n,\ k=1,\ldots,M$. We can do so by minimizing

$$\sum_{i=1}^{n}\left(\mathbf{y}_i-\begin{pmatrix}f(x_{i1})\\ \vdots\\ f(x_{iM})\end{pmatrix}\right)^T\boldsymbol{W}_i\left(\mathbf{y}_i-\begin{pmatrix}f(x_{i1})\\ \vdots\\ f(x_{iM})\end{pmatrix}\right)+\lambda\int_a^b\{f''(x)\}^2dx\quad(7)$$

for some a and b, where λ is a nonnegative smoothing parameter. Write

$$f(x)=\sum_{j=1}^{n^*+2}\theta_jB_j(x),$$

where n^*+6 is the number of knots, θ_j are B-spline coefficients, and $B_j(x)$ are B-spline basis functions. Defining the $nM\times(n^*+2)$ matrix $\mathbf{B}_*=(\mathbf{B}_1^T,\ldots,\mathbf{B}_n^T)^T$ and the $(n^*+2)\times(n^*+2)$ penalty matrix $\boldsymbol{\Omega}$ by

$$\left[(\mathbf{B}_i)_{kj}\right]=B_j(x_{ik})$$

and

$$\left[(\boldsymbol{\Omega}_{jk})\right]=\int_a^b B_j''(x)B_k''(x)dx,$$

we can rewrite the objective function (7) as

$$(\mathbf{y}-\mathbf{B}_*\boldsymbol{\theta})^T\boldsymbol{W}(\mathbf{y}-\mathbf{B}_*\boldsymbol{\theta})+\lambda\boldsymbol{\theta}^T\boldsymbol{\Omega}\boldsymbol{\theta},\quad(8)$$

where $\mathbf{y}=(\mathbf{y}_1^T,\ldots,\mathbf{y}_n^T)^T$ and $\mathbf{W}=\operatorname{diag}(\boldsymbol{W}_1,\ldots,\boldsymbol{W}_n)$. The penalty matrix $\boldsymbol{\Omega}$ is symmetric and has a half-bandwidth of 4.

Setting the derivative of (8) with respect to $\boldsymbol{\theta}$ to $\mathbf{0}$ gives the solution

$$\left(\mathbf{B}_*^T\mathbf{W}\mathbf{B}_*+\lambda\boldsymbol{\Omega}\right)\hat{\boldsymbol{\theta}}=\mathbf{B}_*^T\mathbf{W}\mathbf{y}.\quad(9)$$

The matrix $\mathbf{B}_*^T\mathbf{W}\mathbf{B}_*+\boldsymbol{\Omega}$ is symmetric but is not generally banded for this problem. One can solve (9) by Cholesky decomposition in $O(n^{*3})$ operations.

The smoothing problem (6) can also be solved by local regression methods—(see Welsh and Yee 2000). Neither solution has yet been implemented in software.

3. SOFTWARE

Software implementing methods as complex as those in the article are difficult and time-consuming to write. We hope that

the authors will soon be able to make their software available to interested users. T. W. Yee has almost completed writing an S-PLUS/R library for fitting VGAMs, called VGAM. Currently available freely at http://www.stat.auckland.ac.nz/~yee.VGAM already fits a very wide range of models, including some of those described in the article. The primary function vgam() can be thought of as a generalization of the gam() function for generalized additive models. Family functions called loglinb2() and loglinb3() have been written for bivariate/trivariate binary responses. We restrict our discussion here to loglinb2().

We have

$$\log P(Y_1 = y_1, Y_2 = y_2|\mathbf{x}) = u_0(\mathbf{x}) + u_1(\mathbf{x})y_1 + u_2(\mathbf{x})y_2 + u_{12}(\mathbf{x})y_1y_2$$

as the log-linear model. We can fit this as a VGAM by specifying

$$\begin{pmatrix} u_1(\mathbf{x}) \\ u_2(\mathbf{x}) \\ u_{12}(\mathbf{x}) \end{pmatrix} = \boldsymbol{\eta}(\mathbf{x}) = \begin{pmatrix} \eta_1(\mathbf{x}) \\ \eta_2(\mathbf{x}) \\ \eta_3(\mathbf{x}) \end{pmatrix},$$

where the η_j are additive predictors (3).

As an example, if ymatrix is a $n \times 2$ matrix of 1's and 0's then

```
fit <- vgam(ymatrix ~ s(x, df=c(4,2)),
                        loglinb2(exchangeable=T))
```

would fit

$$\log P(Y_1 = y_1, Y_2 = y_2|x) = u_0(x) + u_1(x)y_1 + u_2(x)y_2 + u_{12}(x)y_1y_2,$$

subject to $u_1 = u_2$, using vector (smoothing) splines. VGAM implements a newly developed algorithm using B-splines as the basis functions. Here u_1 and u_{12} are assigned 4 and 2 degrees of freedom (1= linear fit). A number of methods functions support objects of class "vgam"; for example `fitted(fit)` returns the $n \times 4$ matrix of joint probabilities, and `resid(fit, ``working)` returns the working residuals.

In the article, the authors fit (37) to the BDES data as their final model. It would be interesting to replace the term f_{12} (sys, chol) by f_{12}(sys * chol) (i.e., a nonparametric function of their product) and fit the model with VGAM.

3.1 Computational Details

Suppose generally that the data are $(y_{i1}, \ldots, y_{iS}, \mathbf{x}_i)$, $i = 1, \ldots, n$, where each y_j is a binary response. For their log-linear model, the authors force $u_{jkl} \equiv 0$ and similarly for other higher-order associations. We follow suit. Such assumptions are often necessary because unless the data contain all fitted combinations, the estimates become unbounded. Furthermore, higher-order associations become increasingly more difficult to interpret.

Consequently, we fit the log-linear model

$$\log P(Y_1 = y_1, \ldots, Y_S = y_S|x) = u_0(\mathbf{x}) + \sum_{j=1}^{S} u_j(\mathbf{x})y_j + \sum_{j<k} u_{jk}(\mathbf{x})y_jy_k.$$

The normalizing parameter u_0 satisfies

$$e^{-u_0} = 1 + \sum_{j=1}^{S} e^{u_j} + \sum_{j<k} e^{u_j+u_k+u_{jk}} + \sum_{j<k<\ell} e^{u_j+u_k+u_\ell+u_{jk}+u_{j\ell}+u_{k\ell}+\cdots} + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right).$$

One has $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T = (u_1, \ldots, u_S, u_{12}, \ldots, u_{S-1,S})^T$, where $M = S(S+1)/2$. (An identity link for each of the u's is chosen because the parameter space is unconstrained). Then $\partial\ell_i/\partial\boldsymbol{\eta} = \partial u_{0i}/\partial\boldsymbol{\eta} + (y_{i1}, \ldots, y_{iS}, y_{i1}y_{i2}, \ldots, y_{i,S-1}y_{iS})^T$.

In the following formulas, a, b, c, and d are distinct indices of $\{1, \ldots, S\}$. It may be verified that

$$\begin{aligned}
\frac{\partial u_0}{\partial u_a} &= -e^{u_0}\left\{e^{u_a} + \sum_{j=1, j\neq a}^{S} e^{u_a+u_j+u_{aj}} + \sum_{j<k} e^{u_a+u_j+u_k+u_{aj}+u_{ak}+u_{jk}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right)\right\}, \\
&= -e^{u_0}A_a, \quad \text{say}, \\
\frac{\partial u_0}{\partial u_{ab}} &= -e^{u_0}\left\{e^{u_a+u_b+u_{ab}} + \sum_{j=1, j\neq a, j\neq b}^{S} e^{u_a+u_b+u_j+u_{ab}+u_{aj}+u_{bj}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right)\right\} \\
&= -e^{u_0}A_{ab}, \quad \text{say}, \\
-\frac{\partial^2 u_0}{\partial u_a^2} &= e^{u_0}A_a(1 - e^{u_0}A_a), \\
-\frac{\partial^2 u_0}{\partial u_{ab}^2} &= e^{u_0}A_{ab}(1 - e^{u_0}A_{ab}), \\
-\frac{\partial^2 u_0}{\partial u_a \partial u_{ab}} &= e^{u_0}A_{ab}(1 - e^{u_0}A_a), \\
-\frac{\partial^2 u_0}{\partial u_a \partial u_b} &= e^{u_0}\left(\frac{\partial A_b}{\partial u_a} - e^{u_0}A_aA_b\right), \\
-\frac{\partial^2 u_0}{\partial u_{ab} \partial u_{ac}} &= e^{u_0}\left(\frac{\partial A_{ac}}{\partial u_{ab}} - e^{u_0}A_{ab}A_{ac}\right), \\
-\frac{\partial^2 u_0}{\partial u_a \partial u_{bc}} &= e^{u_0}\left(\frac{\partial A_{bc}}{\partial u_a} - e^{u_0}A_aA_{bc}\right),
\end{aligned}$$

and

$$-\frac{\partial^2 u_0}{\partial u_{ab} \partial u_{cd}} = e^{u_0}\left(\frac{\partial A_{cd}}{\partial u_{ab}} - e^{u_0}A_{ab}A_{cd}\right),$$

where

$$\frac{\partial A_b}{\partial u_a} = e^{u_a+u_b+u_{ab}} + \sum_{j=1,\, j\neq a,\, j\neq b}^{S} e^{u_j+u_a+u_b+u_{ab}+u_{aj}+u_{bj}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right),$$

$$\frac{\partial A_{bc}}{\partial u_a} = e^{u_a+u_b+u_c+u_{ab}+u_{ac}+u_{bc}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right),$$

$$\frac{\partial A_{ac}}{\partial u_{ab}} = e^{u_a+u_b+u_c+u_{ab}+u_{ac}+u_{bc}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right),$$

and

$$\frac{\partial A_{cd}}{\partial u_{ab}} = e^{u_a+u_b+u_c+u_d+u_{ab}+u_{ac}+u_{ad}+u_{bc}+u_{bd}+u_{cd}} + \cdots + \exp\left(\sum_{j=1}^{S} u_j + \sum_{j<k} u_{jk}\right).$$

Thus Newton-Raphson coincides with Fisher scoring. When estimating VGAMs, modified backfitting (Hastie and Tibshirani 1990) is used. Cholesky decomposition is used to invert the $\mathbf{W}_i$ and the QR method is used to solve for the linear parts of the component functions.

For $S = 2$, the information matrix simplifies to

$$\mathbf{W}_i = e^{2u_0} \begin{pmatrix} (1+e^{u_2})(e^{u_1}+e^{u_1+u_2+u_{12}}) & e^{u_1+u_2+u_{12}} - e^{u_1+u_2} & (1+e^{u_2})e^{u_1+u_2+u_{12}} \\ e^{u_1+u_2+u_{12}} - e^{u_1+u_2} & (1+e^{u_1})(e^{u_2}+e^{u_1+u_2+u_{12}}) & (1+e^{u_1})e^{u_1+u_2+u_{12}} \\ (1+e^{u_2})e^{u_1+u_2+u_{12}} & (1+e^{u_1})e^{u_1+u_2+u_{12}} & (1+e^{u_1}+e^{u_2})e^{u_1+u_2+u_{12}} \end{pmatrix}$$

Although the iterative proportional fitting procedure (IPFP; Bishop et al. 1975) is the usual algorithm for log-linear models, our limited experience has indicated that IRLS works well.

ADDITIONAL REFERENCES

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

Roosen C. B., and Hastie, T. J. (1993), "Logistic Response Projection Pursuit" Technical Report, BL011214-930806-09TM, AT&T Bell Laboratories.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Welsh, A. H., and Yee, T. W. (2000), "Local Regression for Vector Responses," Unpublished manuscript.

Yee, T. W., and Hastie, T. J. (2000), "Reduced-Rank Vector Generalized Linear Models," Unpublished manuscript.

Yee, T. W., and Wild, C. J. (1994), *Vector Generalized Additive Models*, Technical Report STAT04 University of Auckland, Dept. of Statistics.

Comment

Sally WOOD

This article presents an innovative and practical solution to the problem of modeling and nonparametrically estimating the conditional probability of two correlated binary outcomes. The work is an excellent blend of innovative methodology and a great application, and I congratulate the authors on making such a significant contribution to the literature.

In this discussion I present an alternative Bayesian approach to modeling the probability of correlated Bernoulli outcomes, and I compare this approach to that presented in the article. My approach is based on work by Albert and Chib (1993), who introduced latent variables to turn the correlated binary regression problem into a correlated Gaussian regression problem. Specifically, define the latent variables $\mathbf{w}_i = (w_{i1}, w_{i2})$ as

$$w_{1i} = g_1(x_i) + e_{i1}$$

and

$$w_{2i} = g_2(x_i) + e_{i2},$$

where $\mathbf{e}_i = (e_{1i}, e_{2i}) \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

To specify the prior on g_j, write g_j as

$$g_j(x) = \alpha_0 + \alpha_1 x + f_j(x) \tag{1}$$

and place the smoothing spline prior from Wahba (1978) on the function f_1 and f_2. To complete the prior specification on the functions g_j, flat priors are assumed for $\boldsymbol{\alpha}_j = (\alpha_{j0}, \alpha_{j1})'$ and τ_j^2. The prior on the correlation coefficient ρ is $p(\rho) \sim U[-1, 1]$.

To estimate the bivariate probabilities $P(y_{1i} = 1, y_{2i} = 1|X)$, place the following constraints on the latent variables:

$$\begin{aligned} w_{1i} &> 0 \quad \text{if } y_{1i} = 1 \\ &< 0 \quad \text{if } y_{1i} = 0 \end{aligned}$$

and

$$\begin{aligned} w_{2i} &> 0 \quad \text{if } y_{2i} = 1 \\ &< 0 \quad \text{if } y_{2i} = 0. \end{aligned}$$

Sally Wood is a visiting Assistant Professor, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208.

Journal of the American Statistical Association
March 2001, Vol. 96, No. 453, Theory and Methods

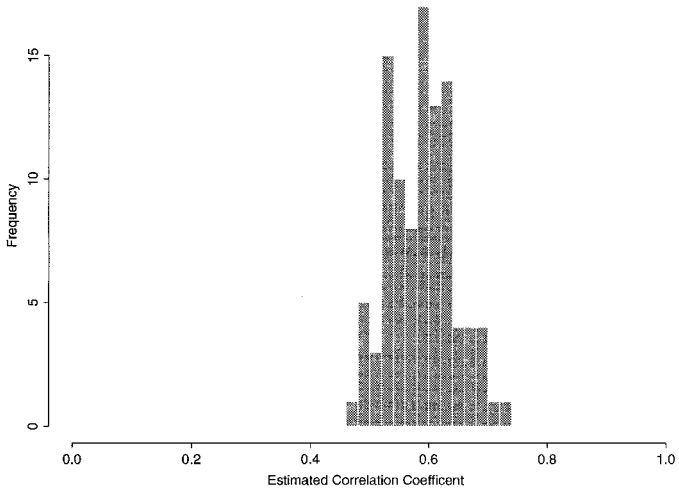


Figure 1. Histogram of the Estimated Value, $\hat{\rho}$ for the 100 Realizations When $P(y_{1i} = 1 * bvX) = \Phi(f_1(x_i)$ and $f_1(x_i) = 2[\exp - 30(x_{1i} - .25)^2 \sin(\pi x_{1i}^2)] - 2$.

This specification leads to the following probabilities:

$$\begin{aligned} P(y_{1i} = 1|X) &= P(w_{1i} > 0|X) \\ &= \Phi\{f_1(x_i)\}, \\ P(y_{2i} = 1|X) &= P(w_{2i} > 0|X) \\ &= \Phi\{f_2(x_i)\}, \end{aligned}$$

and

$$P(y_{2i} = 1, y_{1i} = 1|X) = P(w_{1i} > 0, w_{2i} > 0|X), \qquad (2)$$

where Φ is the standard normal cdf. One drawback of this model is that the joint probability $P(y_{1i}, y_{2i})$ does not have a closed-form expression. However, obtaining these probabilities is straightforward either by numerical integration or by simulation. In this model, ρ is a measure of the dependence between y_{1i} and y_{2i}. If $\rho = 0$, then y_{1i} and y_{2i} are independent; if $\rho = 1$ and we know y_{1i}, then we also know y_{2i}. The correlation coefficient ρ is analogous to the authors' measure of pairwise association which is the log-odds ratio, however, it is not directly comparable to this parameter. For example, in the logistic regression model, the log odds ratio is constant over the domain for x, whereas in the model given by (2), the log odds ratio varies over x.

As before, the problem is now to estimate $f_1(x)$ and $f_2(x)$ nonparametrically. This can be done by using the method outlined by Wood and Kohn (1998), with a few minor modifications. In that article, the functions $f_1(x)$ and $f_2(x)$ were estimated by their posterior means using Monte Carlo simulation to integrate out the smoothing parameters, τ_1^2 and τ_2^2. This involves obtaining draws of τ_1^2 and τ_2^2 from their posterior distributions. The method outlined here uses the same idea with the additional step of drawing the correlation coefficient ρ from its posterior distribution. This is done by performing the following transformation. Define $\mathbf{f}_j = (f_j(x_1), \ldots, f_j(x_n))$ for $j = 1, 2$; $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)$, where $\mathbf{w}_j = (w_{j1}, \ldots, w_{jn})$; $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, where $\mathbf{y}_j = (y_{j1}, \ldots, y_{j1})$; and $z_i = (1, x_i)$. Let $\boldsymbol{\Sigma} = \text{var}(\mathbf{e}_i)$ have the Cholesky decomposition $\mathbf{LDL}'$, where

$$\mathbf{L} = \begin{pmatrix} 1 & 0 \\ \rho & 1 \end{pmatrix}$$

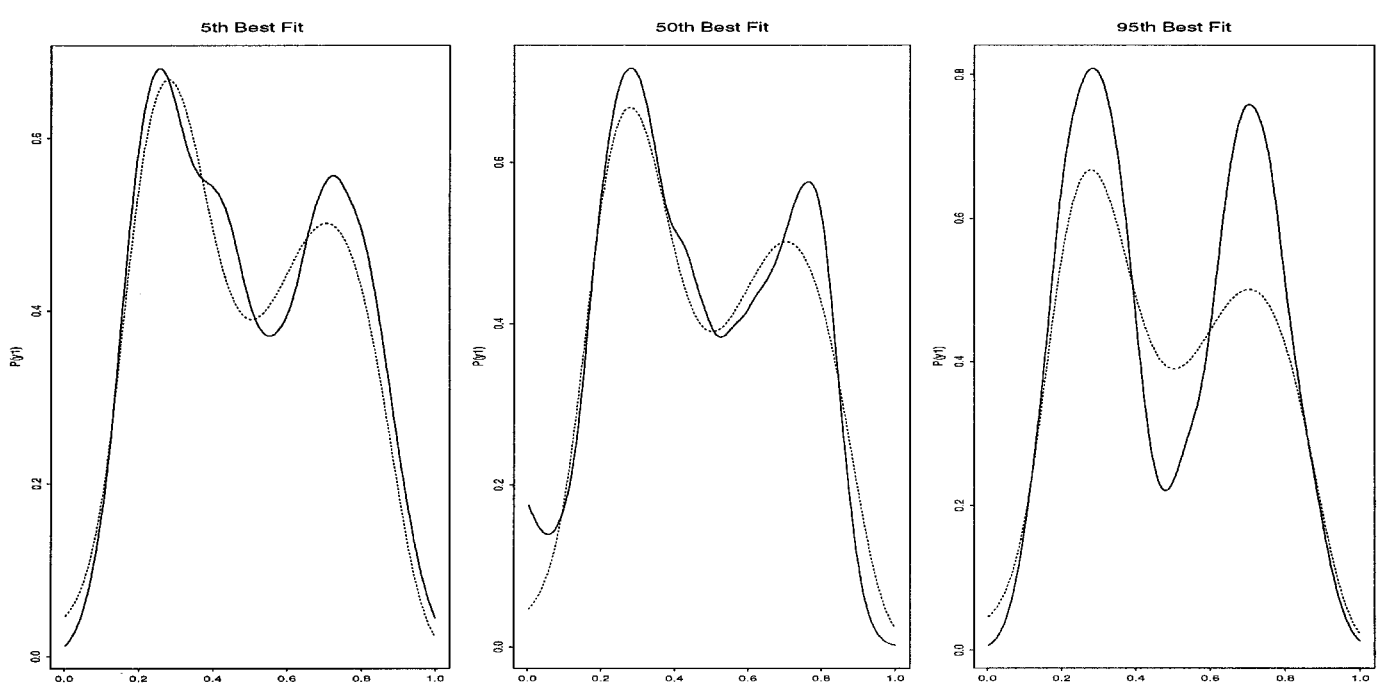


Figure 2. True and Estimated Probability $\Phi(f_1(x_{1i}))$ for the 5th(a), 50th(b), and 95th(c) Best Fits, Where $f_1(x_i) = 2[\exp - 3(x_{1i} - .25)^2 + \sin(\pi x_{1i}^2)] - 2$..

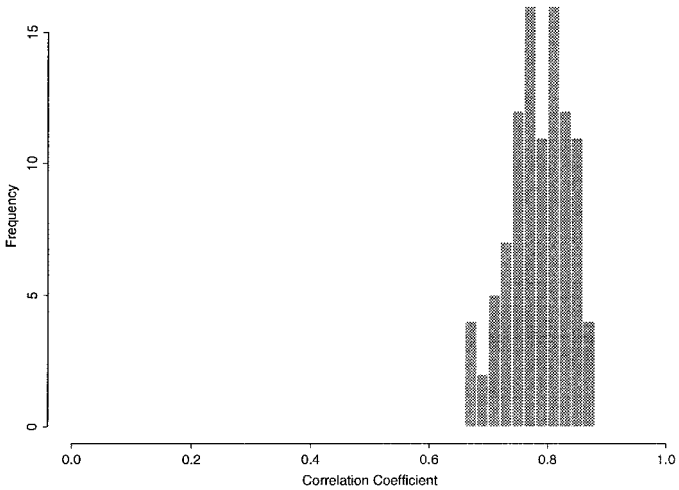


Figure 3. Histogram of the Estimated Value, $\hat{\rho}$, for the 100 Realizations Where $f_1(x_i) = 10\cos(2x_i) + 7e^{x_i^2} - 16$ and $f_2(x_i) = 2\cos(5x_i + 1.4) + x_i^2$.

and

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 1-\rho^2 \end{pmatrix},$$

and let

$$w_i^* = \mathbf{L}^{-1} w_i,$$
$$f_i^* = \mathbf{L}^{-1} f_i,$$
$$\boldsymbol{\alpha}^* = \mathbf{L}^{-1} \boldsymbol{\alpha},$$

and

$$e_i^* = \mathbf{L}^{-1} e_i.$$

Draws from the posterior distribution of the parameters of interest are obtained using the following sampling scheme:

GIBBS SAMPLER

0. Initialize $\mathbf{f}_1, \mathbf{f}_2$ and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ as $\mathbf{f}_1^{[0]} = 0, \mathbf{f}_2^{[0]} = 0$ and $\boldsymbol{\alpha}_1^{[0]} = 0$, $\boldsymbol{\alpha}_2^{[0]} = 0$ and draw $\rho^{[0]}$ from $U[-1, 1]$.

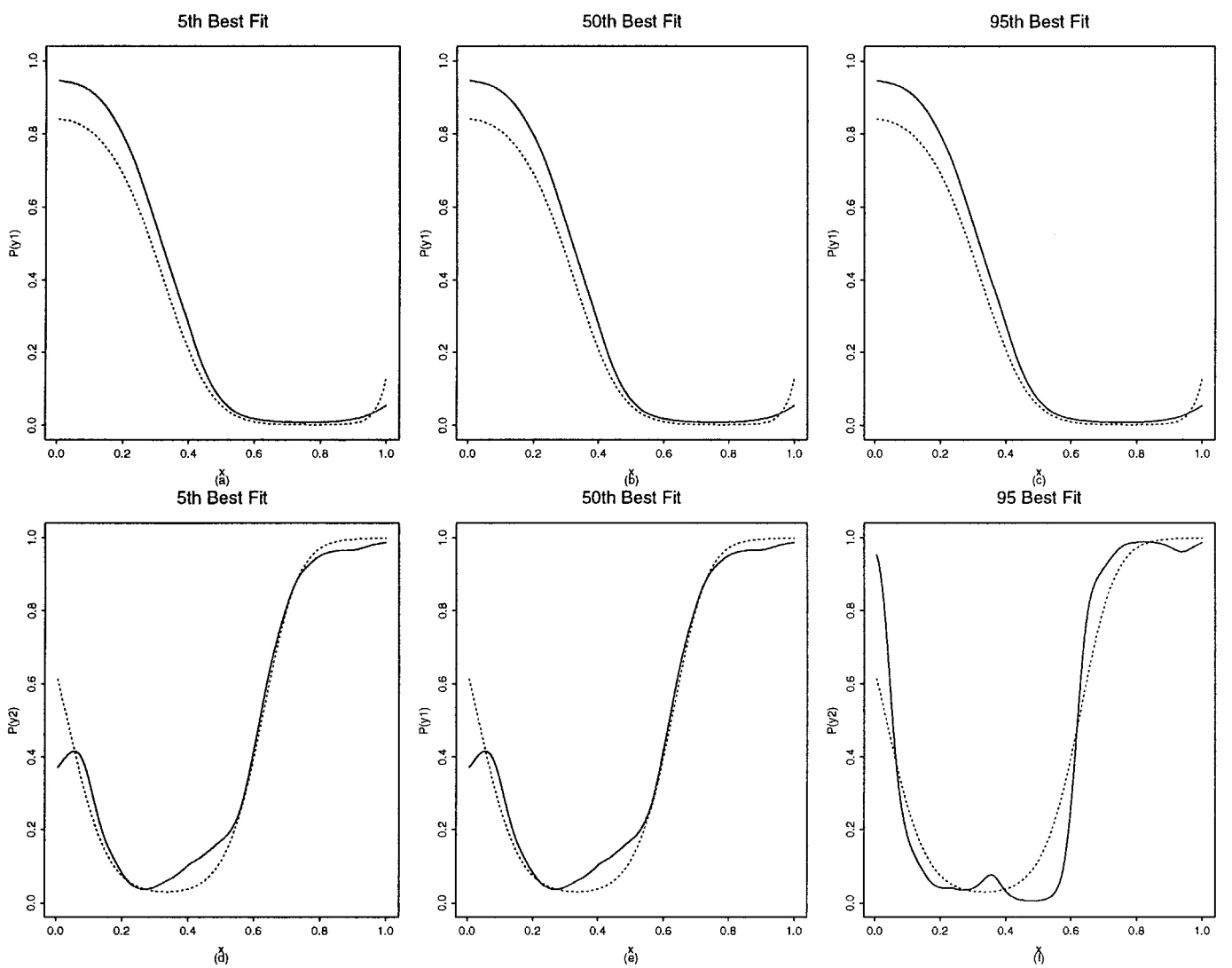


Figure 4. The 5th, 50th, and 95th Best Fits for True and Estimated Probabilities $\Phi(f_1(x_i))$ [(a), (b), and (c)] and $\Phi(f_2(x_i))$ [(d), (e), and (f)], Where $f_1(x_i) = 10\cos(2x_i) + 7e^{x_i^2} - 16$ and $f_2(x_i) = 2\cos(5x_i + 1.4) + x_i^2$.

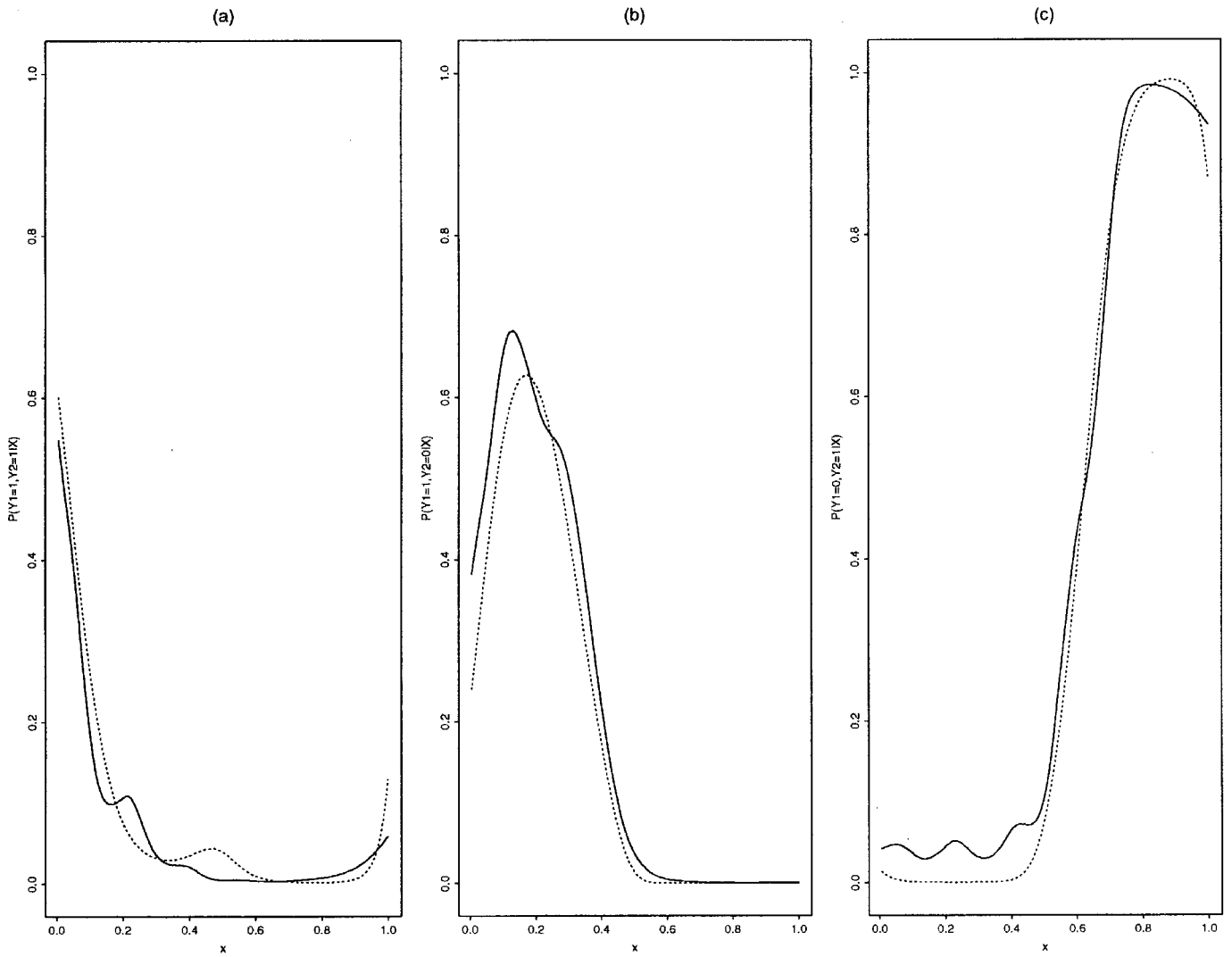


Figure 5. True and Estimated Probabilities $P(y_1 = 1, y_2 = 1|x)$ (a), $P(y_1 = 1, y_2 = 0|x)$ (b), and $P(y_1 = 0, y_2 = 1|x)$ (c).

1. Draw $\mathbf{w}_1$ from $p(\mathbf{w}_1|\mathbf{w}_2, \mathbf{f}_1, \mathbf{f}_2, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_1, \mathbf{y}_1, \mathbf{y}_2, \rho) = p(\mathbf{w}_1|\mathbf{w}_2, \mathbf{f}_1\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_1, \mathbf{y}_1, \mathbf{f}_2, \rho)$ and $p(\mathbf{w}_1|\mathbf{f}_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \mathbf{w}_2, \mathbf{y}_1, \mathbf{f}_2, \rho) = \prod_{i=1}^{n} p(w_{1i}|w_{2i}, f_2(x_i), f_1(x_i), \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, y_{1i}, \rho)$. Generate w_{1i} from a normal distribution with mean $f_1(x_i) + z_i\boldsymbol{\alpha}_1 + \rho(y_{2i} - f_2(x_i) - z_i\boldsymbol{\alpha}_2)$ and variance $1 - \rho^2$. If

$$y_{1i} = 1, \quad \text{then constrain } w_{1i} \text{ to be positive; if}$$

$$y_{1i} = 0, \quad \text{then constrain } w_{1i} \text{ to be negative.}$$

Similarly, generate w_{2i} from a normal distribution with mean $f_2(x_i) + z_i\boldsymbol{\alpha}_2 + \rho(y_{1i} - f_1(x_i) - z_i\boldsymbol{\alpha}_1)$ and variance $1 - \rho^2$. If

$$y_{2i} = 1, \quad \text{then constrain } w_{2i} \text{ to be positive; if}$$

$$y_{2i} = 0, \quad \text{then constrain } w_{2i} \text{ to be negative.}$$

Calculate $w_i^* = \mathbf{L}^{-1}w_i$, where $w_i = (w_{1i}, w_{2i})'$.

2. Generate $\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \mathbf{f}_2^*$, and $\mathbf{f}_1^*$ as a block from $p(\mathbf{f}_1^*, \mathbf{f}_2^*, \boldsymbol{\alpha}_2^*, \boldsymbol{\alpha}_1^*|\mathbf{w}^*, \tau_1^2, \tau_2^2, \rho) = p(\mathbf{f}_1^*, \boldsymbol{\alpha}_1^*|\mathbf{w}_1^*, \tau_1^2)p(\mathbf{f}_2^*, \boldsymbol{\alpha}_2^*|\mathbf{w}_2^*, \tau_2^2, \rho)$ by generating $\boldsymbol{\alpha}_j^*$ from $p(\boldsymbol{\alpha}_j^*|\mathbf{w}_j^*, \tau_j^2, \rho)$ and then generating $\mathbf{f}_j^*$ from $p(\mathbf{f}_j^*|\mathbf{w}_j^*, \tau^2, \boldsymbol{\alpha}_j^*, \rho)$, conditional on the generated value of $\boldsymbol{\alpha}_j^*$, for $j = 1, 2$. Calculate $f_i = \mathbf{L}f_i^*$, where $f_i = (f_1(x_i), f_2(x_i))'$.

3. Generate τ_1^2 and τ_2^2 from $p(\tau_2^2, \tau_1^2|\mathbf{f}_1^*, \mathbf{f}_2^*) = p(\tau_2^2|\mathbf{f}_2^*) \times p(\tau_1^2|\mathbf{f}_1^*)$. The conditional density $p(\tau_j^2|\mathbf{f}_j^*) \propto p(\mathbf{f}_j^*|\tau_j^2)p(\tau_j^2)$ and is inverse gamma.

4. Generate ρ^2 in the following manner. Let $\nu = 1 - \rho^2$, then $p(\nu|\mathbf{w}_1^*, \mathbf{w}_2^*, \boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \mathbf{f}_1^*, \mathbf{f}_2^*) = p(\nu|\mathbf{w}_2^*, \mathbf{f}_2^*, \boldsymbol{\alpha}_2^*) \propto p(\mathbf{w}_2^*|\nu, \mathbf{f}_2^*, \boldsymbol{\alpha}_2^*)p(\nu)$ and is the inverse gamma (u, v) with parameters $u = (n-1)/2$ and $v = .5\sum_{i=1}^{n}(w_{2i}^* - f_{2i}^* - \alpha_2^* z_i^*)^2$ and constrained to be less than 1. Then $\rho^2 = 1 - \nu$, and the sign of ρ is given by the sign of $\tilde{\mathbf{w}}_1'\tilde{\mathbf{w}}_2'$, where $\tilde{\mathbf{w}}_j = \mathbf{w}_j - \mathbf{f}_j - z\boldsymbol{\alpha}_j$. Alternatively, ρ can be drawn directly using a Metropolis–Hastings step.

This procedure was tested using the first simulated example in the article where there is only one observation-specific covariate, X_{ki} $(k = 1, 2)$, and X_{1i} is uniformly distributed on the interval [0,1] and $X_{2i} = X_{1i} + \epsilon_i; \epsilon_i \sim U[-.05, .05]$. The regression function used to generate the observations in this example is

$$f_j(x_{ji}) = 2[\exp -30(x_{ji} - .25)^2 + \sin(\pi x_{ji}^2)] - 2.$$

These f_j's $(j = 1, 2)$ were used for f_1 and f_2 in (2). One hundred realizations, each containing 250 pairs of correlated Bernoulli observations, were generated from the model given

by (2), with ρ set to .6. Following Wood and Kohn, the criteria used to measure performance is the ISKL distance. Figure 1 is a histogram of the estimated correlation coefficient ρ for the 100 realizations. The figure shows that the estimates of $\hat{\rho}$ are very close to the true value of .6. To test whether this technique is recovering the regression functions, Figure 2 plots the 5th, 50th, and 95th best fits for the probability $P(y_{1i} = 1|X)$. Note that given the probabilities $P(y_1|X)$, $P(y_2|X)$, and ρ, it is straightforward to obtain $P(y_1|X, y_2)$. These compare favorably with the fits produced by the authors in their Figure 2. The plots suggest that the estimates obtained via the Bayesian method are less variable than those obtained by the authors. One possible explanation for this is that in the article, GACV is used to select the smoothing parameters, and estimates of the regression function given this smoothing parameter are calculated. When the choice of the smoothing parameter is good, the procedure produces excellent curve estimates; however, when a poor smoothing parameter is chosen, the curve estimate is correspondingly poor. The Bayesian technique avoids choosing a specific smoothing parameter by averaging the curve over a number of smoothing parameters drawn from their posterior distributions. Hence there is less variability in the curve estimates.

A second set of simulations were run when there were different endpoints of interest. Again, the functions used for generating the data were those in the used in the article. These functions are

$$f_1(x_i) = 10\cos(2x_i) + 7e^{x_i^2} - 16$$

and

$$f_2(x_i) = 2\cos(5x_i + 1.4) + x_i^2$$

The true correlation co-efficient, ρ is set to be 0.8 in this simulation. Figure 3 is a histogram of the estimates $\hat{\rho}$. As in the single endpoint case, the data are tightly centered around the true value of .8. Figure 4 plots the true and estimated probabilities $\Phi\{f_1(x)\}$ and $\Phi\{f_2(x)\}$ for the 5th, 50th, and 95th best estimates. These estimates compare favorably to the results in the authors Figure 9. Figure 5 shows the true and estimated joint probabilities probabilities $P(y_{1i} = 1, y_{2i} = 1|X)$, $P(y_{1i} = 1, y_{2i} = 0|X)$, and $P(y_{1i} = 0, y_{2i} = 1|X)$ for the 5th best fit. This figure shows that, given the marginal distributions $P(y_1)$ and $P(y_2)$ and the correlation coefficient ρ, it is straightforward to calculate the joint and hence the conditional distributions of y_1 and y_2.

A more rigorous test of the robustness of both techniques would be to generate data from a model different from that used for estimation and then compare the performance of the two models.

Again, I thank the authors for their contribution and welcome their thoughts on the present Bayesian approach.

Comment

Marc AERTS and Geert MOLENBERGHS

The authors make a serious effort to present a comprehensive modeling strategy for multivariate binary observations. All details, both technical and computational, are worked out in detail. They are to be congratulated for this.

Much work has been done in the area of repeated and/or multivariate binary observations. The authors acknowledge only part of this work. For example, an important review was provided by Pendergast et al. (1996).

Cox (1972) provided a multivariate log-linear type model that has formed the basis for many modeling strategies, including the one proposed by the auhtors. Similar models were proposed by Rosner (1984) and Liang and Zeger (1989). Because it falls within the class of (multivariate) exponential family models, it enjoys all of that class's desirable mathematical and statistical properties. However, it has some serious drawbacks (Liang et al. 1992). The parameters have a conditional interpretation, and the models are not upward compatible. This means that they cannot adequately handle sequences of unequal length. One must be careful with a 0/1 coding in the event that sequences are of unequal length (Cox and Wermuth 1996), and calculation of the normalizing constant can be formidable for long sequences. Further, the research questions at hand often beg a marginal answer. Precisely such a log-linear representation has been chosen by the authors as the basis for their modeling strategy, but we believe that a fully marginal model or a random-effects representation may be of greater value.

Important early models of the marginal type are the beta-binomial model (Kleinman 1973; Skellman 1948), the Bahadur model (Bahadur 1961; Cox 1972), and the probit model (Ashford and Sowden 1970). The beta-binomial model has had some success in various applications, whereas the Bahadur model, because of a very restrictive parameter space, has seen few applications (Declerck, Aerts, and Molenberghs 1998). The probit model is very popular in econometric applications but has also seen some sucess in biometry in various forms (Lesaffre and Molenberghs 1991; Ochi and Prentice 1984). Recently, more work has been done in this same area. A few attempts have been made to produce marginal models that combine logistic regressions for the margins with odds ratios to describe the associations

Marc Aerts and Geert Molenberghs are Professors of Statistics, Center for Statistics, Limburgs Universitair Centrum, B-3590 Diepenbeck, Belgium.

Journal of the American Statistical Association
March 2001, Vol. 96, No. 453, Theory and Methods

and several variations thereof (Glonek and McCullagh 1995; Lang and Agresti 1994; Molenberghs and Lesaffre 1994). (A review and unification of several model was given by Molenberghs and Lesaffre 1999.) When full likelihood methods are deemed cumbersome, GEE-based alternatives, or pseudolikelihood models (Geys, Molenberghs, and Ryan 1999), can be considered. Random-effects approaches have been studied by Stiratelli, Laird, and Ware (1984), Zeger, Liang, and Albert (1988), Breslow and Clayton (1993), and Wolfinger and O'Connell (1993). A flexible random-effects model was given by Hedeker and Gibbons (1994, 1996). Heagerty (1999) constructed a model that combines the advantages of a marginal specification and the flexibility of random effects. A thorough account of many methods was given by Fahrmeir and Tutz (1994).

Having selected a model for the (clustered) binary response data and an appropriate estimation method, it is important to allow flexible functional forms to describe the dependence of parameters like the

(conditional) probability parameter as a function of predictor variables like age. Different smoothing techniques can be used for these purposes. The authors mention some related work based on smoothing splines. An interesting alternative that received much of attention is local polynomial estimation (Fan and Gijbels 1996). Aerts and Claeskens (1997) studied local polynomial estimators in multiparameter likelihood models, illustrated on multivariate binary response data. In further work, Claeskens and Aerts (2000a, b) examined using a bootstrap method to construct confidence bands and some backfitting algorithms for additive models. Carroll, Ruppert, and Welsh (1998) proposed local versions of GEEs. A comparison starting from the multivariate log-linear type model between both smoothing methods using the Beaver Dam Eye Study would be very interesting. Of course, both methods must deal with delicate issues, such as adaptive choice of the smoothing parameter and the curse of dimensionality. Whatever method chosen, the access to flexible models revealing interesting nonlinear relationships must be gained by computational complexity. In many cases, analyzing complex multivariate data necessitates making some minimal assumptions to simplify and make the method feasible. Nonparametric methods are only partly able to fill the gap between restrictive and in principle incorrect parametric models and the "truth." As an illustration, consider the author's ophthalmology application. They simplify the general model and its associated likelihood (8) by assuming a constant pairwise association α. In clusters larger than 2, it is also often assumed that all higher-order association are 0. It is hard to predict the effect of misspecifying such particular aspects. Forcing the association to be constant might affect the estimation of the other parameters and result in misleading smooth curves. Related to that, Aerts, Claeskens, and Hart (1999) indicated, for a similar kind of application—macular edema for younger onset diabetic persons (see Klein et al. 1984)—the necessity of modeling the intraperson correlation as a nonconstant function of systolic blood pressure (at least in a model only containing blood pressure as predictor variable). They developed a formal lack-of-fit test in a general likelihood-based framework using orthogonal series estimators and modifications of the Akaike information criterion. As the authors mention in their conclusion, their method can also be generalized to get a nonparametric estimate for the association term, which also could be used as a diagnostic tool for checking constant association. Finally, next to other smoothing methods (e.g., penalized splines, series estimators), another useful approach are so-called fractional polynomials (Royston and Altman 1994; Sauerbrei and Royston 1999).

The authors state that an adequate fit is necessary to "understand the cause of certain outcomes." This is somewhat misleading because even though a badly fitting model will not enhance such understanding, causality requires much deeper reflection than merely a good fit. Indeed, causation goes well beyond correlation.

Equations (35) and (36) clearly show the counterintuitive nature of a conditional specification. Indeed, it would be much more natural to just specify the marginal logits for each eyes separately.

ADDITIONAL REFERENCES

Aerts, M., and Claeskens, G. (1997), "Local Polynomial Estimators in Multiparameter Likelihood Models," *Journal of the American Statistical Association*, 92, 1536–1545.

Aerts, M., Claeskens, G., and Hart, J. D. (1999), "Testing the Fit of a Parametric Function," *Journal of the American Statistical Association*, 94, 869–879.

Ashford, J. R., and Sowden, R. R. (1970), "Multivariate Probit Analysis," *Biometrics*, 26, 535–546.

Bahadur, R. R. (1961), "A Representation of The Joint Distribution of Responses to n Dichotomous items," in *Studies in Item Analysis and Prediction*, H. Solomon, Stanford CA: Stanford University Press.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Carroll, R. J., Ruppert, D., and Welsh, A. H. (1998), "Local Estimating Equations," *Journal of the American Statistical Association*, 93, 214–227.

Claeskens, G., and Aerts, M. (2000a), "Bootstrapping Local Polynomial Estimators in Likelihood Based Models," *Journal of Statistical Planning and Inference*, 86, 63–80.

——— (2000b), "On Local Estimating Equations in Additive Multiparameter Models," *Statistics and Probability Letters*, to appear.

Cox, D. R. (1972), "The Analysis of Multivariate Binary Data," *Applied Statistics*, 21, 113–120.

Cox, D. R., and Wermuth, N. (1996), *Multivariate Dependencies*, London: Chapman and Hall.

Declerck, L., Aerts, M., and Molenberghs, G. (1998), "Behavior of the Likelihood Ratio Test Statistic Under a Bahadur Model for Exchangeable Binary Data," *Journal of Statistical Computation and Simulation*, 61, 15–38.

Fahrmeir, L., and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Heidelberg: Springer-Verlag.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Geys, H., and Molenberghs, G., and Ryan, L. (1999), "Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology," *Journal of the American Statistical Association*, 94, 734–745.

Heagerty, P. (1999), "Marginally Specified Logistic-Normal Models for Longitudinal Binary Data," *Biometrics*, 55, 688–698.

Hedeker, D., and Gibbons, R. D. (1994), "A Random-Effects Ordinal Regression Model For Multilevel Analysis," *Biometrics*, 50, 933–944.

——— (1996), "MIXOR: A Computer Program for Mixed-Effects Ordinal Regression Analysis," *Computer Methods and Programs in Biomedicine*, 49, 157–176.

Kleinman, J. C. (1973), "Proportions With Extraneous Variance: Single and Independent Samples," *Journal of the American Statistical Association*, 68, 46–54.

Klein, R., Kelin, B. E. K., Moss, S. E., Davis, M. D., and DeMets, D. L. (1984), "The Wisconsin Epidemiologic Study of Diabetic Retinopaty: II. Prevalence and Risk of Diabetic Retinopathy When Age at Diagnosis is Less Than 30 Years," *Archives of Ophthalmology*, 102, 520–526.

Lang, J. B., and Agresti, A. (1994), "Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses," *Journal of the American Statistical Association*, 89, 625–632.

Lesaffre, E., and Molenberghs, G. (1991), "Multivariate Probit Analysis: A Neglected Procedure in Medical Statistics," *Statistics in Medicine*, 10, 1391–1403.
Liang, K.-Y., and Zeger, S. L. (1989), "A Class of Logistic Regression Models for Multivariate Binary Time Series," *Journal of the American Statistical Association*, 84, 447–451.
Molenberghs, G., and Lesaffre, E. (1994), "Marginal Modelling of Correlated Ordinal Data Using a Multivariate Plackett Distribution," *Journal of the American Statistical Association*, 89, 633–644.
——— (1999), "Marginal Modelling of Multivariate Categorical Data," *Statistics in Medicine*, 18, 2237–2255.
Ochi, Y., and Prentice, R. L. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, 71, 531–543.
Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Plata, M., and Fisher, M. R. (1996), "A Survery of Methods for Analyzing Clustered Binary Response Data," *International Statistical Review*, 64, 89–118.
Rosner, B. (1984), "Multivariate Methods in Ophtalmology With Applications to Other Paired-Data Situations," *Biometrics*, 40, 1025–1035.
Royston, P., and Altman, D. G. (1994), "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Applied Statistics*, 43, 429–468.
Sauerbrei, W., and Royston, P. (1999), "Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials," *Journal of the Royal Statistical Society*, Ser. A, 162, 71–94.
Skellam, J. G. (1948), "A Probability Distribution Derived From the Binomial Distribution by Regarding the Probability of a Success as Variable Between the Sets of Trials," *Journal of the Royal Statistical Society*, Ser. B, 10, 257–261.
Stiratelli, R., Laird, N., and Ware, J. (1984), "Random Effects Models for Serial Observations With Dichotomous Response," *Biometrics*, 40, 961–972.
Wolfinger, R., and O'Connell, M. (1993), "Generalized Linear Mixed Models; A Pseudo-Likelihood Approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.
Zeger, S. C., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060.

Comment

Joe WHITTAKER

First, I congratulate the authors for a stimulating article. I wish to illustrate the structure of the independence graph associated with certain discrete approximations to the SS-ANOVA models considered here. The underlying theory of graphical models was considered by Lauritzen (1996), and we use similar ideas for representation as did Spiegelhalter (1998). Hoped-for benefits include increased insight where the direct visualization may suggest common patterns in rather disparate models, which in turn may suggest borrowing algorithms from one research area and applying them to another.

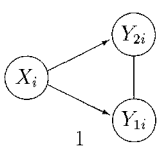
Rather than tackle the authors' general models here, I give a possible graph that relates specifically to the model suggested by the authors for the Beaver Eye Dam study. They consider the distribution of two binary variables $Y = (Y_1, Y_2)$, indicating pigment abnormalities in the left and right eyes, conditionally on the values of six covariates $X = (X_1, X_2, \ldots, X_6)$. The model for a generic individual i is specified by

$$\log p(y_{1i}, y_{2i}|x_i) = f(x_i)[y_{1i} + y_{2i}] + \alpha y_{1i} y_{2i} + b_i, \qquad (1)$$

where the main effect f is a smooth function of the covariates and the pairwise interaction parameter α is constant, although potentially also a function of the covariates, and b_i is a normalizing constant depending on the values of $f(x_i)$ and α.

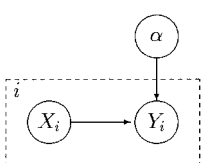
The argument for identical main effects is symmetry between the eyes.

The generic independence graph of (1) for the random variables is the chain graph



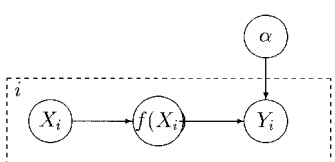
in which the edge between the two response variables is undirected. Time series models might have a mixture of directed and undirected edges in this part of the model.

The parameters and the sampled individuals may be represented within the same graph. By supposing that the parameter α has a prior distribution [proper, but diffuse; e.g., $\alpha \sim \mathrm{N}(0, \lambda_\alpha^{-1})$ where λ_α is small] leads to the graph



in which the rectangular dashed box indexed by i denotes the replication of the diagram for each sample individual and, consequently, the independence of the random variables Y_i conditioned on the parameter α. The separate elements of Y_i have
been suppressed for clarity. The graph indicates that X_i and α are independent when Y_i is unobserved.

The covariates affect Y only through f, which can be made explicit by using a logical link



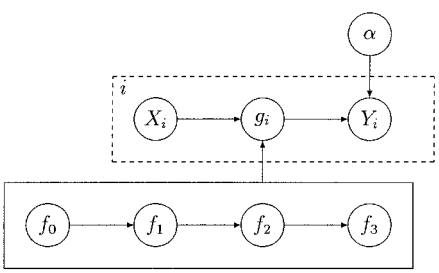
Joe Whittaker is Senior Lecturer, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA14YF, U.K. (E-mail: *joe.whittaker@lancaster.ac.uk*).

The independence graph requires a full distribution for the unknown parameters, as well as for the observables, so we interpret the penalty term in the penalized likelihood as a Bayesian prior. Representing the model when f is unknown by a finite number of random variables necessitates a measure of approximation.

Consider a continous covariate x, say. At first pass, one might divide the range into $R+1$ intervals with center points x_r with $r = 0, 1, \ldots, R$. If R is sufficiently large, then the continuity of f implies that $f_r = f(x_r)$ and $f_{r+1} = f(x_{r+1})$ are close. A stochastic model for this is that the increments are independent Gaussian random variables so that $\{f_r\}$ is a realization of a random walk in discrete time. The term in the prior is proportional to

$$\exp\left\{-\frac{1}{2}\lambda_f \sum_{r=0}^{R-1}(f_{r+1}-f_r)^2 - \frac{1}{2}\lambda_0 f_0^2\right\}.$$

The graph to represent this now includes the Markov chain



where $R = 3$ for clarity and g_i is the value of the function that delivers f_r if X_i is nearest x_r and the node in the graph gives a deterministic outcome.

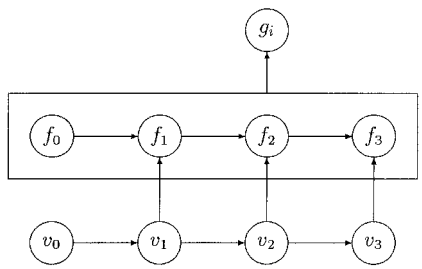
The sample paths of the random walk are not differentiable. As the standard penalty for cubic splines is based on second derivates, the natural discrete time approximation is the integrated random walk

$$f_{r+1} - f_r = v_r$$

and

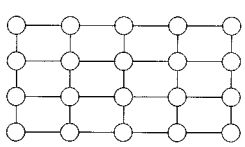
$$v_{r+1} - v_r \sim \mathrm{N}\left(0, \lambda_v^{-1}\right),$$

which has independent second forward differences. The graph of the Markov chain is replaced by the graph



This is an instance of a graph of a hidden Markov model (see, e.g., MacDonald and Zucchini 1997). The graph suggests that intermediate state-space models, such as the smoothed integrated random walk, may also be of interest, and also highlights that the algorithmic procedures associated with state-space models, such as Kalman filtering or recursive fixed interval smoothing, can be applied when the covariates are deliberately ordered.

The Markov chain graph of the one-dimensional random walk is Markov equivalent to the undirected Gibbs chain where each of the arrows are replaced by undirected edges. Presumably the graph of the two-dimensional smoother, $f(x_1, x_2)$, that corresponds to this is the Markov random field, illustrated here by



ADDITIONAL REFERENCES

Lauritzen, S. (1996), *Graphical Models*, Oxford, U.K.: Oxford University Press.

MacDonald, I., and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, London: Chapman and Hall.

Spiegelhalter, D. J. (1998), "Bayesian Graphical Modelling: A Case–Study in Monitoring Health Outcomes," *Applied Statistics*, 47, 115–133.

Rejoinder

Fangyu GAO, Grace WAHBA, Ronald KLEIN, and Barbara KLEIN

We heartily thank all of the discussants for their interesting comments.

1. YEE AND WILD: EXTENSIONS AND COMPUTATION

We believe that the current model can be extended in numerous ways. We thank Yee and Wild for providing some excellent examples of possible extensions for the modeling of vectors of smooth functions. These extensions and variations will form a rich family of models. As Yee and Wild mention, one advantage of the SS-ANOVA setup is to provide a unified theoretical framework. Under this framework, we pose a

variational problem and solve it. By separating the fitting procedure into natural "blocks," these models enjoy the property of easy extension.

Nonparametric function estimation always needs more computing power. The good news is that advances in computer speed and memory made it possible to perform tasks once thought impossible. On the other hand, the birth of the internet and advances in database technology have provided an explosion of data waiting to be analyzed. We imagine two research areas that might be interesting for statisticians in the future. The first is parallel computing algorithms. Instead of waiting for the first step to be completed before starting the second step, several tasks can be executed at the same time on different CPUs. We believe that the block SOR algorithm or backfitting (which could be viewed as a special case of block SOR) can be modified to serve as a general building block for this purpose. The second area is on-line real time model updating. When new data come in sequentially, a fast and efficient updating algorithm could be valuable. There will be a lot of interesting problems for nonparametric modeling. In more recent work here we have been able to use the Condor system run by the University of Wisconsin-Madison Computer Sciences Department, in which multiple jobs are directed to any of the literally hundreds of machines in the Condor system that have unused cycles, thus allowing model evaluation for many values of the smoothing parameters simultaneously.

We mention that in a real application, what type of model to use may depend strongly on the application itself. As for the Beaver Dam Eye Study, we tried to avoid using any derived variables from the beginning, unless they are widely accepted, such as body mass index. However, the product of systolic blood pressure and cholesterol level lacks an accepted medical interpretation. But we do believe that model forms like Yee and Wild's (1) are useful and provide an important addition to the literature.

2. WOOD: BAYESIAN INTERPRETATIONS AND MORE

One property of smoothing spline models is that they can be identified with Bayesian and other regularization models. This helps us understand the models from other perspectives, and opens the door to more possibilities. A now-classical application is to construct Bayesian "confidence intervals" for smoothing spline models. Also, the connection between smoothing spline models and recently popular machine learning algorithms such as support vector machines (SVMs) and Gaussian process learning (GPI) has been noted (see, e.g., Seeger 2000, Wahba, Lui, and Zhang 2000; the website http://www.kernel-machines.org).

We are glad to see Wood's stimulating comments. Tanner and Wong (1987) introduced data augmentation in a Bayesian context. Chib and Greenberg (1998) provided a Bayesian analysis for a multivariate probit model. Wood and Kohn (1998) used a Bayesian approach and a Gibbs sampler for fitting smoothing splines for binary data. Given a fixed smoothing parameter, Wood's approach is equivalent to solving the variational problem by using a Gibbs sampler. However, in Wood's work, she also puts a flat prior on the smoothing parameters, as opposed to the approach in our article, where we used GACV to "select" a specific smoothing parameter. We remark that the plots in Wood's comments are not directly comparable to plots in our work due to the different model setting, but nevertheless they are interesting. We also make an interesting observation here. Given that the latent normal random variables are not observed and the "flexible" means of the normal distributions are unknown, Wood's estimate of the correlation ρ has a very small variance. To examine this more closely, we did a simple simulation study by generating 250 pairs of correlated normal random variables with mean 0, variance 1, and $\rho = .6$, estimating ρ by maximum likelihood and repeating this 100 times. Figure 1 gives a histogram for the 100 repetitions of the estimated ρ's. It is almost the same as Wood's Figure 1. We need to understand how Wood's flat prior affects the final estimate.

3. WHITTAKER: CONNECTION WITH GRAPHICAL MODELS

We are pleased that Joe Whittaker has provided comments. One of us (GW) became aware some time ago of the interesting relation between the terms in a (smoothing spline-type) ANOVA decomposition in the log-likelihood for exponential families, and graphical models. Indeed, this relationship (and Whittaker 1990) were mentioned by Gu (1993). In some sense, the model selection problem (i.e. which terms in the ANOVA decomposition to keep in the model) are equivalent to the choice of a graphical model and important to the study of conditional dependencies. We think that there are a number of interesting problems to be solved here in the nonparametric context, whose solution would also contribute to the literature on graphical models.

4. AERTS AND MOLENBERGHS: STRATEGIES FOR MODELING MULTIVARIATE RESPONSES

We acknowledge that numerous approaches exist to model multivariate response variables, with most of the literature involving parametric models. In our article we were not attempting to give a comprehensive review of the literature. Indeed, with so many existing techniques, determining which one to use will depend on the application. It is hard to see that one technique will universally outperform other methods under all circumstances.

We thank Aerts and Molenberghs for listing more references about modeling correlated responses. Although we believe that a well-fitted and interpretable model will be helpful in understanding the correlation between independent and dependent variables, we did not mention that this is "sufficient" to understand causality.

As far as the "base" model to use, we chose a log-linear model in this article. Our concern is focused mainly on "genuine" multivariate Bernoulli outcomes, which have the same number of repetitions for every independent cluster. This setup provides all the desirable properties of the exponential family model. It is clear that this setup can serve as a basis for many different generalizations. One extension is to model marginal probabilities. This can be achieved by reparameterization within the penalized likelihood framework, given that

we can write down the likelihood function. As previously noted, the smoothing spline setup here has connections with several other areas of research.

Certain compromises must be made to model higher-order associations among multivariate categorical responses. In our article, for simplicity, we assumed all higher-order associations to be 0, although in principle we can fit them from data. The GEE model was developed when the full likelihood function is too complicated. Because the mean vector and covariance matrix fully specify a multivariate normal distribution, latent variable models like the multivariate probit model are also useful. Lin and Zhang (1999) and Ke and Wang (2000) developed nonlinear mixed effects models by using smoothing splines. We always need to balance among flexibility, simplicity, interpretability, and efficiency. For a real application, the situation is more likely to be, quoting from George Box, that "no model is correct, but some are more useful than others."

ADDITIONAL REFERENCES

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrica*, 85, 347–361.

Gu, C. (1993), "Smoothing Spline Density Estimation: A Dimension Automatic Algorithm," *Journal of the American Statistical Association*, 88, 495–504.

Ke, C., and Wang, Y. (2000), "Semi-Parametric Nonlinear Mixed Effects Models and Their Applications," unpublished manuscript, available at http://www.pstat.ucsb.edu/yuedong.

Seeger, M. (2000), "Bayesian Model Selection for Support Vector Machine Classifiers, in *Advances in Neural Information Processing Systems*, eds. T. L. S. Solla and K. Muller, Cambridge, MA: MIT Press, p. 12.

Tanner, M. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Wahba, G., Lin, Y., and Zhang, H. (2000), "Generalized Approximate Cross-Validation for Support Vector Machines, or Another Way to Look at Margin-Like Quantities," in *Advances in Large Margin Classifiers*, eds. B. S. A. Smola, P. Bartlett, and D. Schuurmans, Cambridge, MA: MIT Press, pp. 297–311.

Wood, S. and Kohn, R. (1998), "A Bayesian Approach to Robust Binary Nonparametric Regression," *J. Amer. Statist. Assoc.*, 93, 203–213.