The Canadian Journal of Statistics
Vol. 39, No. 4, 2011, Pages 591–609
La revue canadienne de statistique

Nonparametric regression with cross-classified responses

Chong GU[1]* and Ping MA[2]*

[1]*Department of Statistics, Purdue University, West Lafayette, IN 47906, USA*
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*

Key words and phrases: Conditional density; cross-validation; eyetracking experiments; penalized likelihood; random effects.

MSC 2010: Primary 62G08; secondary 62G05, 62G20, 62H12, 41A15.

Abstract: In this article, we develop regression models with cross-classified responses. Conditional independence structures can be explored/exploited through the selective inclusion/exclusion of terms in a certain functional ANOVA decomposition, and the estimation is done nonparametrically via the penalized likelihood method. A cohort of computational and data analytical tools are presented, which include cross-validation for smoothing parameter selection, Kullback–Leibler projection for model selection, and Bayesian confidence intervals for odds ratios. Random effects are introduced to model possible correlations such as those found in longitudinal and clustered data. Empirical performances of the methods are explored in simulation studies of limited scales, and a real data example is presented using some eyetracking data from linguistic studies. The techniques are implemented in a suite of R functions, whose usage is briefly described in the appendix. *The Canadian Journal of Statistics* 39: 591–609; 2011 © 2011 Statistical Society of Canada

Résumé: Dans cet article, nous développons des modèles de régression avec variables réponses provenant de classifications croisées. Les structures d'indépendance conditionnelle peuvent être explorées/exploitées grâce à l'inclusion/exclusion de termes dans une décomposition de type anova fonctionnelle et l'estimation se fait de façon non paramétrique en utilisant la méthode de vraisemblance pénalisée. Un ensemble d'outils informatiques et d'analyse de données sont présentés et incluent la validation croisée pour la sélection du paramètre de lissage, la projection de Kullback-Leibler pour la sélection de modèles et les intervalles de crédibilité bayésiens pour les rapports de cotes. Des effets aléatoires sont inclus dans le modèle pour prendre en compte les possibles corrélations telles que celles trouvées dans les données longitudinales et en grappes. Des études de simulations limitées montrent les performances empiriques de ces méthodes. Un vrai jeu de données sur l'oculométrie dans des études linguistiques est aussi traité. Ces techniques sont implantées dans un ensemble de fonctions R et nous en présentons brièvement l'utilisation en annexe. *La revue canadienne de statistique* 39: 591–609; 2011 © 2011 Société statistique du Canada

1. INTRODUCTION

To motivate the models under development, consider some eyetracking experiments in linguistic studies, in which participants in front of computer monitors listen to instructions such as "click on the purple bottle," and their eye fixation on the "target" (i.e., purple bottle), on some "colour competitor" (e.g., purple pencil), on some "object competitor" (e.g., yellow bottle), or on something else is monitored during the trial on a fine time grid. The purpose of such studies is to explore how linguistic variables may affect the ease with which the listeners can select a visually available referred-to item; more details can be found in Section 7. Our task in this article is to develop

** Author to whom correspondence may be addressed.*
E-mail: chong@purdue.edu, pingma@illinois.edu

modeling tools that can be used to estimate the probabilities of the eye fixation on items of the four different categories, as functions of time. For the ultimate goal of the linguists, the response is the eye fixation profile along time and the covariates are qualitative linguistic variables. For the quantitative estimation of the eye fixation profile under fixed linguistic conditions, the task we are concerned with here, the response is the eye fixation category and the covariate is time.

If one reduces the number of categories to two, say by combining categories, the estimation can be performed via logistic regression with time as the covariate (or x), so the tools to be developed generalize logistic regression. The response (or y) in the aforementioned eyetracking data appears as univariate with four categories, but may also be taken as bivariate of a pair of binary variables of colour (purple or not) and object (bottle or not); our general treatment allows for multivariate y with marginals not necessarily binary. The eyetracking data are longitudinal, so observations are generally correlated at least within each trial if not also within the same human participant in different trials, and random effects will be incorporated into the setting to accommodate such correlation.

Write $\mathcal{Y} = \prod_{k=1}^{K} \mathcal{Y}_k$, where $\mathcal{Y}_k = \{1, \ldots, N_k\}$ are discrete, and consider regression problem with response $y \in \mathcal{Y}$ and covariate $x \in \mathcal{X}$, where $\mathcal{X}$ is a generic domain. We shall employ the logistic conditional density transform

$$f(y|x) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{Y}} e^{\eta(x,y)}}, \tag{1}$$

where the integral over $\mathcal{Y}$ is summation, and estimate the conditional density $f(y|x)$ via $\eta(x, y)$; one may decompose

$$\eta(x, y) = \eta_{\emptyset} + \eta_x(x) + \eta_y(y) + \eta_{xy}(x, y), \tag{2}$$

where the identifiability of $\eta_x(x)$, $\eta_y(y)$, $\eta_{xy}(x, y)$ depends on appropriate side conditions, and to ensure a one-to-one transform in (1), one may set $\eta_{\emptyset} + \eta_x = 0$. Observing (x_i, y_i), $i = 1, \ldots, n$, one may estimate $\eta(x, y) = \eta_y(y) + \eta_{xy}(x, y)$ via the penalized minus log likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(x_i, y_i) - \log \int_{\mathcal{Y}} e^{\eta(x_i,y)}\right\} + \frac{\lambda}{2}J(\eta), \tag{3}$$

where $J(\eta)$ is a roughness penalty and the smoothing parameter λ controls the tradeoff between the goodness-of-fit and the smoothness of the estimate; this is a special case of penalized likelihood conditional density estimation of Gu (1995), with $\mathcal{Y}$ discrete. The response y is generally multivariate and so is the covariate x, so η_y and η_{xy} can be further decomposed similar to (2); conditional independence structures among the components of $y = (y_{\langle 1\rangle}, \ldots, y_{\langle K\rangle})$ can be explored/exploited via selective term elimination in such functional ANOVA decompositions, with details to be filled in later. For $K = 1$ and $N_1 = 2$, this reduces to the standard penalized likelihood logistic regression; see, for example, Gu (2002,sect. 6.7.3). Other special cases of the formulation include regression with multinomial responses ($K = 1$, $N_1 > 2$) treated in Lin (1998) and regression with multivariate Bernoulli responses ($N_1 = \cdots = N_K = 2$) treated in Gao et al. (2001).

As a special case of an even more general formulation, some existing results can be inherited, such as the asymptotic convergence and the numerical computation, but the many data analytical tools to be discussed in this article were not available in Gu (1995). Some of these tools, such as the Kullback–Leibler projection for model selection, will be adapted from later developments in other settings, and some of these tools, such as the Bayesian confidence intervals for log odds ratios and random effects for correlated data, are designed only for a discrete $\mathcal{Y}$.

When the covariate x is absent, the data are to be aggregated into contingency tables, for which log-linear models are among standard analytical tools. From this perspective, the models under development effectively add a nonparametric "x-axis" to the log-linear models that "disaggregates" contingency tables.

The rest of the article is organized as follows. Some details of the model structure involving functional ANOVA decompositions are spelled out in Section 2. In Section 3, pertinent technical details concerning penalized likelihood estimation is reviewed. Model selection and inferential tools are discussed in Section 4, and mixed-effect models for correlated data are introduced in Section 5. Simulations of limited scales are conducted in Section 6, and an analysis of some eyetracking data is presented in Section 7. Section 8 collects a few remarks. To facilitate the practical application of the tools being developed in this article, a suite of R functions have been developed, whose usage is briefly described in an appendix.

2. CONDITIONAL DENSITY MODELS

We now fill in some details concerning functional ANOVA decomposition and log-linear models, which are largely repackaged from materials in the literature. See, for example, Gu (2002,sect. 1.3).

2.1. Functional ANOVA Decomposition

Consider a bivariate function $\eta(x) = \eta(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ on a domain $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$; subscripts in brackets are used in this article to denote coordinates of a point on a multi-dimensional domain while ordinary subscripts are reserved for multiple points. One may decompose the function through a functional ANOVA decomposition

$$\begin{aligned}\eta(x_{\langle 1\rangle}, x_{\langle 2\rangle}) &= (I - A_1 + A_1)(I - A_2 + A_2)\eta \\ &= A_1A_2\eta + (I - A_1)A_2\eta + A_1(I - A_2)\eta + (I - A_1)(I - A_2)\eta \\ &= \eta_\emptyset + \eta_1(x_{\langle 1\rangle}) + \eta_2(x_{\langle 2\rangle}) + \eta_{12}(x_{\langle 1\rangle}, x_{\langle 2\rangle}),\end{aligned} \tag{4}$$

where I is the identity operator, A_1 and A_2 are averaging operators acting on arguments $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$, respectively, that satisfy $A_1 1 = 1$ and $A_2 1 = 1$, with the main effects η_1, η_2 and the interaction η_{12} satisfying side conditions $A_1\eta_1 = A_1\eta_{12} = 0$ and $A_2\eta_2 = A_2\eta_{12} = 0$; examples of averaging operators include $A\eta = \int_a^b \eta(x)\mathrm{d}x/(b-a)$, $A\eta = \eta(x_0)$, and $A\eta = \sum_{i=1}^m \eta(x_i)/m$. The averaging operators on different axes are independent of each other, and (4) has a one-way ANOVA on $\mathcal{X}$ built-in, $\eta(x) = \eta_\emptyset + \eta_x(x)$, with $A\eta_x = A_1A_2\eta_x = 0$ for $\eta_x = \eta_1 + \eta_2 + \eta_{12}$.

For $\mathcal{X}_1 \times \mathcal{X}_2$ discrete, $\eta(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ is a matrix of "treatment means" μ_{ij} in the standard ANOVA model notation, with (4) in the form of

$$\mu_{ij} = \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}) + (\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where $\mu_{i\cdot} = \sum_j c_j\mu_{ij}$ for $\sum_j c_j = 1$, $\mu_{\cdot j} = \sum_i d_i\mu_{ij}$ for $\sum_i d_i = 1$, and $\mu_{\cdot\cdot} = \sum_{i,j} c_jd_i\mu_{ij}$.

Note that the marginal domains $\mathcal{X}_1$ and $\mathcal{X}_2$ are generic so can be product domains themselves, and (2) is simply (4) in slightly different notation. Similar constructions in more than two dimensions can be done recursively, or directly via $\prod_k(I - A_k + A_k)\eta$.

2.2. Log-Linear Models

A standard log-linear model for an $N_1 \times \cdots \times N_K$ table is a surrogate Poisson regression model on $\mathcal{Y} = \prod_{k=1}^K \mathcal{Y}_k$ for $\mathcal{Y}_k = \{1, \ldots, N_k\}$, which is equivalent to density estimation on $\mathcal{Y}$ (see Lindsey, 1997, chap. 3). For a one-to-one logistic density transform $f(y) = e^{\eta(y)}/\int_{\mathcal{Y}} e^{\eta(y)}$, one sets

$\eta_\emptyset = 0$ in a one-way ANOVA decomposition $\eta(y) = \eta_\emptyset + \eta_y(y)$ of the log density, and conditional independence structures among the marginals of $y = (y_{\langle 1\rangle}, \dots, y_{\langle K\rangle})$ can be characterized via the selective elimination of interaction terms in an ANOVA decomposition of $\eta(y) = \eta_y(y)$.

For an example, consider $K = 3$ with $y = (y_{\langle 1\rangle}, y_{\langle 2\rangle}, y_{\langle 3\rangle})$. An ANOVA decomposition yields

$$\begin{aligned}\eta_y(y) = {} & \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_3(y_{\langle 3\rangle}) + \eta_{12}(y_{\langle 1\rangle}, y_{\langle 2\rangle}) \\ & + \eta_{13}(y_{\langle 1\rangle}, y_{\langle 3\rangle}) + \eta_{23}(y_{\langle 2\rangle}, y_{\langle 3\rangle}) + \eta_{123}(y_{\langle 1\rangle}, y_{\langle 2\rangle}, y_{\langle 3\rangle}). \end{aligned} \tag{5}$$

Setting $\eta_{23} + \eta_{123} = 0$, one has $f(y_{\langle 1\rangle}, y_{\langle 2\rangle}, y_{\langle 3\rangle}) \propto \exp\{\eta_1 + \eta_2 + \eta_3 + \eta_{12} + \eta_{13}\}$, so $f(y_{\langle 2\rangle}, y_{\langle 3\rangle}|y_{\langle 1\rangle}) = f(y_{\langle 2\rangle}|y_{\langle 1\rangle})f(y_{\langle 3\rangle}|y_{\langle 1\rangle})$, or $y_{\langle 2\rangle} \perp y_{\langle 3\rangle}|y_{\langle 1\rangle}$. Setting all the interaction terms to 0, an additive model $\eta_y = \eta_1 + \eta_2 + \eta_3$ implies the mutual independence of the marginals.

2.3. Log-Linear Regression Models

Adding an x-axis, of interest are models for $f(y|x) = e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$, where $\eta(x, y) = \eta_y(y) + \eta_{xy}(x, y)$. Again taking $K = 3$ for an example, $\eta_y(y)$ can still be decomposed as in (5), and $\eta_{xy}(x, y)$ can be decomposed in a parallel manner, with x added to the argument lists and to the subscripts of each term in (5). Setting $\eta_{23} + \eta_{123} + \eta_{x23} + \eta_{x123} = 0$ implies $y_{\langle 2\rangle} \perp y_{\langle 3\rangle}|(x, y_{\langle 1\rangle})$, and the conditional independence of the y marginals given x can be obtained by setting all interactions involving more than one y marginals to 0.

Association between the marginals of contingency tables are often characterized via the odds ratios. Consider $K = 2$ with $y = (y_{\langle 1\rangle}, y_{\langle 2\rangle})$. The log conditional density is given by

$$\eta(x, y) = \eta_y + \eta_{xy} = \eta_1 + \eta_2 + \eta_{12} + \eta_{x1} + \eta_{x2} + \eta_{x12},$$

and the log odds ratio depends only on $\eta_{12} + \eta_{x12}$,

$$\begin{aligned}\log \frac{f(y_{\langle 1\rangle}, y_{\langle 2\rangle}|x) f(y'_{\langle 1\rangle}, y'_{\langle 2\rangle}|x)}{f(y_{\langle 1\rangle}, y'_{\langle 2\rangle}|x) f(y'_{\langle 1\rangle}, y_{\langle 2\rangle}|x)} = {} & \eta_{12}(y_{\langle 1\rangle}, y_{\langle 2\rangle}) + \eta_{x12}(x, y_{\langle 1\rangle}, y_{\langle 2\rangle}) \\ & + \eta_{12}(y'_{\langle 1\rangle}, y'_{\langle 2\rangle}) + \eta_{x12}(x, y'_{\langle 1\rangle}, y'_{\langle 2\rangle}) - \eta_{12}(y_{\langle 1\rangle}, y'_{\langle 2\rangle}) \\ & - \eta_{x12}(x, y_{\langle 1\rangle}, y'_{\langle 2\rangle}) - \eta_{12}(y'_{\langle 1\rangle}, y_{\langle 2\rangle}) - \eta_{x12}(x, y'_{\langle 1\rangle}, y_{\langle 2\rangle}). \end{aligned} \tag{6}$$

When $\eta_{x12} = 0$, the odds ratio is independent of x, with the model sitting in between the "saturated" model and the independence model $y_{\langle 1\rangle} \perp y_{\langle 2\rangle}|x$ (with $\eta_{12} + \eta_{x12} = 0$); the models developed in Gao et al. (2001) effectively set $\eta_{x12} = 0$, not allowing the "saturated" model.

Note that x can also be multidimensional, in which case η_{xy} can be further decomposed, allowing for further model choices.

3. PENALIZED LIKELIHOOD ESTIMATION

As noted earlier, the models being considered are special cases of the conditional density estimation formulated in Gu (1995), with $\mathcal{Y}$ discrete, and we will review the basics of the approach in this section; it is assumed that at least part of x is continuous, for otherwise the models would be parametric. Settings with continuous $\mathcal{Y}$ needs alternative treatment due to the cost of numerical integration, which will be studied elsewhere.

3.1. Reproducing Kernel Hilbert Spaces

The minimization of (3) is implicitly conducted in a Hilbert space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square semi norm with a finite dimensional null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$. A Hilbert space has a metric and a geometry that facilitate analysis and computation, and a finite dimensional $\mathcal{N}_J$ prevents interpolation. Function evaluations appear in (3), so for (3) to be continuous in η, one also needs the evaluation functional $[u]\eta = \eta(u)$ to be continuous in $\eta \in \mathcal{H}$, $\forall u \in \mathcal{U}$, where $u = (x, y) \in \mathcal{U} = \mathcal{X} \times \mathcal{Y}$.

A Hilbert space in which evaluation functional is continuous is a reproducing kernel Hilbert space with a reproducing kernel $R(\cdot, \cdot)$, a non-negative definite bivariate function on $\mathcal{U}$ such that $R(u, \cdot) = R(\cdot, u) \in \mathcal{H}$, $\forall u \in \mathcal{U}$, and $\langle R(u, \cdot), \eta(\cdot)\rangle = \eta(u)$ (the reproducing property), $\forall \eta \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. A reproducing kernel Hilbert space can be generated from its reproducing kernel R, for which any non-negative definite function qualifies, as the "column space" span$\{R(u, \cdot), u \in \mathcal{U}\}$. A general theory can be found in Aronszajn (1950).

In the setting of (3), one may write $\langle \cdot, \cdot \rangle = J(\cdot, \cdot) + \tilde{J}(\cdot, \cdot)$, where $J(\cdot, \cdot)$ is the semi inner product associated with $J(\eta)$ and $\tilde{J}(\cdot, \cdot)$ is an inner product in $\mathcal{N}_J$. One has a tensor-sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, with $J(\eta)$ being a square full norm in $\mathcal{H}_J$. For computation, one needs a basis of $\mathcal{N}_J$ and the reproducing kernel R_J in $\mathcal{H}_J$ satisfying $J(R_J(u, \cdot), \eta(\cdot)) = \eta(u)$, $\forall \eta \in \mathcal{H}_J$, $\forall u \in \mathcal{U}$.

The discussion above applies to a generic domain $\mathcal{U}$, not necessarily a product domain, and in fact reproducing kernel Hilbert spaces on product domains are typically constructed via tensor products of spaces on their marginal domains, as illustrated in the examples below.

3.2. Examples of Marginal and Tensor-Product Spaces

The examples are extracted from Gu (2002,chap. 2), where further details can be found.

On $\mathcal{X} = [0, 1]$, one may set $J(\eta) = \int_0^1 (\eta''(x))^2 \mathrm{d}x$ and $\tilde{J}(\eta, \eta) = (\int_0^1 f(x)\mathrm{d}x)^2 + (\int_0^1 f'(x)\mathrm{d}x)^2$. The reproducing kernel in $\mathcal{H}_J = \{f : J(f) < \infty, \int_0^1 f(x)\mathrm{d}x = \int_0^1 f'(x)\mathrm{d}x = 0\}$ is given by $R_J(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. Combining with $\mathcal{N}_J = \{1\} \oplus \{k_1(x)\}$, where $k_1(x) = x - 0.5$, one has a one-way ANOVA decomposition $\eta = \eta_\emptyset + \eta_x$, with $\eta_x \in \{k_1(x)\} \oplus \mathcal{H}_J$ satisfying the side condition $\int_0^1 \eta_x(x)\mathrm{d}x = 0$. The construction provides building blocks for tensor-product spaces, to be discussed below, for which one writes $\mathcal{H} = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1$ with reproducing kernels $R_{00}(x_1, x_2) = 1$, $R_{01}(x_1, x_2) = k_1(x_1)k_1(x_2)$, and $R_1(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$.

On $\mathcal{Y} = \{1, \ldots, N\}$, one may set $J(\eta) = \sum_{y=1}^N (\eta(y) - \bar{\eta})^2$ for y nominal, where $\bar{\eta} = \sum_{y=1}^N \eta(y)/N$, or set $J(\eta) = \sum_{y=1}^{N-1} (\eta(y + 1) - \eta(y))^2$ for y ordinal; in both cases, $\mathcal{N}_J = \{1\}$ contains the constant functions (i.e., length-N vectors) and one may simply set $\tilde{J}(\eta, \eta) = \bar{\eta}^2$. This defines a one-way ANOVA decomposition $\eta = \eta_\emptyset + \eta_y \in \mathcal{H}_0 \oplus \mathcal{H}_1$ with $\bar{\eta}_y = 0$. The reproducing kernels can be written as $N \times N$ matrices, with $R_0(y_1, y_2) = 1$ (i.e., the matrix $\mathbf{1}\mathbf{1}^T$) and $R_1(y_1, y_2)$ given by $(I - \mathbf{1}\mathbf{1}^T/N)^+ = I - \mathbf{1}\mathbf{1}^T/N$ for y nominal or by $B^+ = (C^T C)^+$ for y ordinal, where $(\cdot)^+$ denotes the Moore–Penrose inverse and

$$C = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & \ldots & -1 & 1 \end{pmatrix}.$$

Note that there is no distinction between nominal or ordinal for $N = 2$.

On $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$, one may construct tensor-product spaces from marginal constructions. Given reproducing kernels $R^{(x)}(x_1, x_2)$ of $\mathcal{H}^{(x)}$ on $\mathcal{X}$ and $R^{(y)}(y_1, y_2)$ of $\mathcal{H}^{(y)}$ on $\mathcal{Y}$, the non-negative definite function $R(u_1, u_2) = R^{(x)}(x_1, x_2)R^{(y)}(y_1, y_2)$ on $\mathcal{U}$, where $u_1 = (x_1, y_1)$, $u_2 = (x_2, y_2)$, generates the tensor-product space $\mathcal{H}^{(x)} \otimes \mathcal{H}^{(y)}$. From the above $\mathcal{H}_{00}^{(x)} \oplus \mathcal{H}_{01}^{(x)} \oplus \mathcal{H}_1^{(x)}$ constructed on $\mathcal{X} = [0, 1]$ and $\mathcal{H}_0^{(y)} \oplus \mathcal{H}_1^{(y)}$ constructed on $\mathcal{Y} = \{1, \ldots, N\}$, one has 6 tensor-product spaces

$\mathcal{H}_{\mu.\nu} = \mathcal{H}_{\mu}^{(x)} \otimes \mathcal{H}_{\nu}^{(y)}$, $\mu = 00, 01, 1$, $\nu = 0, 1$, with reproducing kernels $R_{\mu.\nu} = R_{\mu}^{(x)} R_{\nu}^{(y)}$. An ANOVA decomposition is built in with $\eta_{\emptyset} \in \mathcal{H}_{00,0}$, $\eta_x \in \mathcal{H}_{01,0} \oplus \mathcal{H}_{1,0}$, $\eta_y \in \mathcal{H}_{00,1}$, and $\eta_{xy} \in \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,1}$; only η_y and η_{xy} are needed here. $\mathcal{H}_{1,1}$ is infinite-dimensional so has to be part of the penalized space $\mathcal{H}_J$, but $\mathcal{H}_{00,1}$ and $\mathcal{H}_{01,1}$ are both $(N-1)$-dimensional (remember that $\bar{\eta}_y = 0$) and can be left in $\mathcal{N}_J$. In the software tools to be illustrated in the appendix, we chose to set $\mathcal{H}_J = \mathcal{H}_{1,1}$ and $\mathcal{N}_J = \mathcal{H}_{00,1} \oplus \mathcal{H}_{01,1}$ for $N = 2$, and set $\mathcal{H}_J = \mathcal{H}_{00,1} \oplus \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,1}$ with an empty $\mathcal{N}_J$ for $N > 2$; in the latter case,

$$R_J = \theta_{00,1} R_{00,1} + \theta_{01,1} R_{01,1} + \theta_{1,1} R_{1,1}, \tag{7}$$

where $\theta_{\mu,\nu}$ are a set of extra smoothing parameters adjusting the relative weights of the roughness of different components.

Similar constructions on higher dimensional domains can be done recursively. For an example, consider $y = (y_{\langle 1 \rangle}, y_{\langle 2 \rangle})$, $N_1 = 2$, $N_2 = 3$, and $x \in [0, 1]$. In the place of the above $\mathcal{H}_1^{(y)}$ one has $\mathcal{H}_{0,1}^{(y)} \oplus \mathcal{H}_{1,0}^{(y)} \oplus \mathcal{H}_{1,1}^{(y)}$, where $\mathcal{H}_{\mu,\nu}^{(y)} = \mathcal{H}_{\mu}^{(y_{\langle 1 \rangle})} \otimes \mathcal{H}_{\nu}^{(y_{\langle 2 \rangle})}$, $\mu, \nu = 0, 1$, and taking tensor-products with $\mathcal{H}_{00}^{(x)} \oplus \mathcal{H}_{01}^{(x)} \oplus \mathcal{H}_{1}^{(x)}$ yields 9 spaces. In obvious notation, one has $\eta_y \in \mathcal{H}_{00,0,1} \oplus \mathcal{H}_{00,1,0} \oplus \mathcal{H}_{00,1,1}$ and $\eta_{xy} \in \mathcal{H}_{01,0,1} \oplus \mathcal{H}_{01,1,0} \oplus \mathcal{H}_{01,1,1} \oplus \mathcal{H}_{1,0,1} \oplus \mathcal{H}_{1,1,0} \oplus \mathcal{H}_{1,1,1}$, and the "saturated" model can be fitted with $\mathcal{H}_J = \mathcal{H}_{00,0,1} \oplus \mathcal{H}_{00,1,1} \oplus \mathcal{H}_{01,0,1} \oplus \mathcal{H}_{01,1,1} \oplus \mathcal{H}_{1,0,1} \oplus \mathcal{H}_{1,1,0} \oplus \mathcal{H}_{1,1,1}$ and $\mathcal{N}_J = \mathcal{H}_{00,1,0} \oplus \mathcal{H}_{01,1,0}$; note that both subspaces of $\mathcal{N}_J$ are finite-dimensional and none of those of $\mathcal{H}_J$ are. To eliminate certain ANOVA terms from the model, the corresponding subspaces can be taken out of $\mathcal{H}_J$ or $\mathcal{N}_J$; a model with $\eta_{x12} = 0$ can be fitted using $\mathcal{H}_J = \mathcal{H}_{00,0,1} \oplus \mathcal{H}_{00,1,1} \oplus \mathcal{H}_{01,0,1} \oplus \mathcal{H}_{1,0,1} \oplus \mathcal{H}_{1,1,0}$ and $\mathcal{N}_J = \mathcal{H}_{00,1,0} \oplus \mathcal{H}_{01,1,0}$, with $\mathcal{H}_{01,1,1}$ and $\mathcal{H}_{1,1,1}$ taken out of $\mathcal{H}_J$.

3.3. Asymptotic Convergence and Approximate Solutions

Fixing x, the estimation precision of $e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$ by $e^{\hat{\eta}(x,y)} / \int_{\mathcal{Y}} e^{\hat{\eta}(x,y)}$ can be assessed via the symmetrized Kullback–Leibler discrepancy $\mathrm{SKL}(\eta, \hat{\eta}|x) = \mu_{\eta}(\eta - \hat{\eta}|x) - \mu_{\hat{\eta}}(\hat{\eta} - \eta|x)$, where $\mu_g(h|x) = \int_{\mathcal{Y}} h(x, y) e^{g(x,y)} / \int_{\mathcal{Y}} e^{g(x,y)}$. A proxy of $\mathrm{SKL}(\eta, \hat{\eta}|x)$ is given by the square error

$$v(\eta - \hat{\eta}|x) = \mu_{\eta}((\eta - \hat{\eta})^2|x) - \mu_{\eta}^2(\eta - \hat{\eta}|x),$$

and the normed distance

$$V(\eta - \hat{\eta}) = \int_{\mathcal{X}} v(\eta - \hat{\eta}|x) f(x) \tag{8}$$

makes an adequate performance measure in the setting, where $f(x)$ is the limiting density of x_i.

Under regularity conditions, the minimizer $\hat{\eta}$ of (3) in $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ has a convergence rate $V(\eta - \hat{\eta}) = O_p(\lambda^p + n^{-1}\lambda^{-1/r})$, where $r > 1$ characterizes the growth rate of the eigenvalues of $J(\eta)$ with respect to $V(\eta)$ and $p \in [1, 2]$ depending on how smooth the true η is; for the examples in Section 3.2, $r = 4$, and $p = 2$ when $d^4\eta/\mathrm{d}x^4$ is square integrable on $\mathcal{X} = [0, 1]$.

For practical applications, one may calculate the minimizer $\hat{\eta}^*$ of (3) in a space $\mathcal{H}^* = \mathcal{N}_J \oplus \mathrm{span}\{R_J(v_j, \cdot), j = 1, \ldots, q\}$, where $\{v_j\}$ is a random subset of the observations $\{u_i = (x_i, y_i)\}$. Under conditions, $V(\eta - \hat{\eta}^*) = O_p(\lambda^p + n^{-1}\lambda^{-1/r})$ as $q\lambda^{2/r} \to \infty$, so $\hat{\eta}^*$ is as good as $\hat{\eta}$ for q sufficiently large. The optimal rate is achieved at $\lambda \asymp n^{-r/(rp+1)}$, so it is sufficient to have $q \asymp n^{2/(rp+1)+\epsilon}$, $\forall \epsilon > 0$. In the software implementation, we set $\epsilon = 0$, $r = 4$, and $p = 2$, and use $q = 10n^{2/9}$, which was shown to be adequate through numerical simulations in similar settings. Due to the random selection of $\{v_j\}$, $\hat{\eta}^*$ is not unique unless $q = n$; larger q would bring more "stability" but the computation time is of the order $O(nq^2)$.

Further details concerning the asymptotic analysis can be found in Gu (1995), Gu & Qiu (1993), and Gu & Wang (2003).

3.4. Computation

Write $\xi_j = R_J(v_j, \cdot)$ and $\mathcal{N}_J = \{\phi_\nu\}_{\nu=1}^{m}$. A function in $\mathcal{H}^*$ has an expression

$$\eta(u) = \sum_{j=1}^{q} c_j \xi_j(u) + \sum_{\nu=1}^{m} d_\nu \phi_\nu(u) = \boldsymbol{\xi}^T \mathbf{c} + \boldsymbol{\phi}^T \mathbf{d}, \tag{9}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions and $\mathbf{c}$ and $\mathbf{d}$ are vectors of coefficients. Fixing smoothing parameters, $\hat{\eta}^*$ can be calculated by minimizing

$$-\frac{1}{n}\mathbf{1}^T(R\mathbf{c} + S\mathbf{d}) + \frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{Y}} \exp\{\boldsymbol{\xi}_i^T \mathbf{c} + \boldsymbol{\phi}_i^T \mathbf{d}\} + \frac{\lambda}{2}\mathbf{c}^T Q \mathbf{c} \tag{10}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where R is $n \times q$ with the (i, j)th entry $\xi_j(u_i)$, S is $n \times m$ with the (i, ν)th entry $\phi_\nu(u_i)$, Q is $q \times q$ with the (j, k)th entry $R_J(v_j, v_k) = J(\xi_j, \xi_k)$, $\boldsymbol{\xi}_i$ is $q \times 1$ with the jth entry $\xi_j(x_i, y)$, and $\boldsymbol{\phi}_i$ is $m \times 1$ with the νth entry $\phi_\nu(x_i, y)$.

Substituting the empirical distribution for $f(x)$, one may write $\mu_g(h) = (1/n)\sum_{i=1}^{n} \mu_g(h|x_i)$ and $V_g(h, h') = (1/n)\sum_{i=1}^{n} v_g(h, h'|x_i)$, where

$$v_g(h, h'|x) = \mu_g(hh'|x) - \mu_g(h|x)\mu_g(h'|x).$$

From an iterate $\tilde{\eta} = \boldsymbol{\xi}^T \tilde{\mathbf{c}} + \boldsymbol{\phi}^T \tilde{\mathbf{d}}$, the Newton updating equation for the minimization of (10) is given by

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \mathbf{c} - \tilde{\mathbf{c}} \\ \mathbf{d} - \tilde{\mathbf{d}} \end{pmatrix} = \begin{pmatrix} R^T \mathbf{1}/n - \boldsymbol{\mu}_\xi - \lambda Q \tilde{\mathbf{c}} \\ S^T \mathbf{1}/n - \boldsymbol{\mu}_\phi \end{pmatrix} \tag{11}$$

where $\boldsymbol{\mu}_\xi = \mu_{\tilde{\eta}}(\boldsymbol{\xi})$, $\boldsymbol{\mu}_\phi = \mu_{\tilde{\eta}}(\boldsymbol{\phi})$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$, $V_{\xi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\phi}^T)$, $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$, $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \tilde{\eta})$, and $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \tilde{\eta})$.

4. MODEL SELECTION AND INFERENCE

We now present tools for model selection and inference, which are essential for practical data analysis using the models being developed.

4.1. Cross-Validation

With varying smoothing parameters, the minimizer of (3) defines a family of estimates, from which one needs to pick one that performs well. For the purpose, one may use as performance measure the aggregated Kullback–Leibler discrepancy

$$\mathrm{KL}(\eta, \hat{\eta}) = \int_{\mathcal{X}} f(x) \left\{ \mu_\eta(\eta - \hat{\eta}|x) - \log \int_{\mathcal{Y}} e^{\eta(x,y)} + \log \int_{\mathcal{Y}} e^{\hat{\eta}(x,y)} \right\}, \tag{12}$$

which is a proxy of $V(\eta - \hat{\eta})$ in (8); $\mathrm{KL}(\eta, \hat{\eta})$ and $V(\eta - \hat{\eta})$ are equivalent measures and the different choices in different contexts are dictated by technical convenience.

Write $\hat{\eta}$ as η_λ to spell out its dependence on the smoothing parameters λ in (3) and θ's hidden in $J(\eta)$. Dropping terms from (12) that do not involve η_λ, one has the relative Kullback–Leibler discrepancy

$$\mathrm{RKL}(\eta, \eta_\lambda) = \int_{\mathcal{X}} f(x) \log \int_{\mathcal{Y}} e^{\eta_\lambda(x,y)} - \int_{\mathcal{X}} f(x) \mu_\eta(\eta_\lambda | x). \tag{13}$$

The first term of (13) only involves η_λ and can be estimated by $n^{-1} \sum_{i=1}^n \log \int_{\mathcal{Y}} e^{\eta_\lambda(x_i,y)}$, but the second term involves both η and η_λ and has to be estimated by the cross-validated sample mean $n^{-1} \sum_{i=1}^n \eta_\lambda^{[i]}(x_i, y_i)$, where $\eta_\lambda^{[i]}$ minimizes a delete-one version of the quadratic approximation of (3) at $\tilde{\eta} = \eta_\lambda$,

$$-\frac{1}{n-1} \sum_{j \neq i} \eta(x_j, y_j) + \mu_{\tilde{\eta}}(\eta) + \frac{1}{2} V_{\tilde{\eta}}(\eta - \tilde{\eta}, \eta - \tilde{\eta}) + \frac{\lambda}{2} J(\eta),$$

for $\mu_g(h)$ and $V_g(h, h')$ as in Section 3.4. Such RKL estimate yields the cross-validation score,

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \eta_\lambda(x_i, y_i) - \log \int_{\mathcal{Y}} e^{\eta_\lambda(x_i,y)} \right\} + \alpha \frac{\mathrm{trace}(P_{\mathbf{1}}^\perp \tilde{R} H^{-1} \tilde{R}^T P_{\mathbf{1}}^\perp)}{n(n-1)}, \tag{14}$$

where H is the matrix on the left-hand side of (11) for $\tilde{\eta} = \eta_\lambda$, $\tilde{R} = (R, S)$, $P_{\mathbf{1}}^\perp = I - \mathbf{1}\mathbf{1}^T/n$, and $\alpha = 1$; the derivation parallels that in Gu (2002,sect. 6.3) for density estimation, where a fudge factor $\alpha = 1.4$ was shown to deliver more robust performances in empirical studies.

For the minimization of (14) as a function of smoothing parameters, the λ in (3) and the θ's in the likes of (7), one may use quasi-Newton methods with numerical derivatives. When the number of θ's is large, which is the case for models containing many ANOVA terms, the process could be prohibitively time-consuming. Using Algorithm 3.2 in Gu & Wahba (1991), one may perform two steps of fixed-θ λ-selection to obtain good starting values, which could leave the subsequent quasi-Newton iteration to pick up only the "last 20%" performance with extra effort many times over the initial one. It is often a good practice to skip the quasi-Newton iteration.

4.2. Kullback–Leibler Projection

Lacking a sampling distribution in settings with infinite-dimensional nulls, the classical testing approach is of little help for the assessment of practically negligible ANOVA terms. A geometric approach was developed in Gu (2004), in which for the "testing" of the null $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, say, one calculates an estimate $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, obtains its Kullback–Leibler projection $\tilde{\eta} \in \mathcal{H}_0$ by minimizing $\mathrm{KL}(\hat{\eta}, \eta)$ over $\eta \in \mathcal{H}_0$, then checks the "entropy decomposition" $\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c)$, where η_c is a degenerate fit such as a constant regression function or an uniform density. When $\mathrm{KL}(\hat{\eta}, \tilde{\eta})$ is only a small portion of $\mathrm{KL}(\hat{\eta}, \eta_c)$, say no more than 2–3%, one loses little by cutting out $\mathcal{H}_1$.

In the current context, one may use the empirical version of (12),

$$\mathrm{KL}(\hat{\eta}, \eta) = \mu_{\hat{\eta}}(\hat{\eta} - \eta) - \frac{1}{n} \sum_{i=1}^n \left\{ \log \int_{\mathcal{Y}} e^{\hat{\eta}(x_i,y)} - \log \int_{\mathcal{Y}} e^{\eta(x_i,y)} \right\}. \tag{15}$$

Taking derivative of $\mathrm{KL}(\hat{\eta}, \tilde{\eta} + \alpha h)$ with respect to α, where $\tilde{\eta}$ minimizes $\mathrm{KL}(\hat{\eta}, \eta)$ for $\eta \in \mathcal{H}_0$ and $h \in \mathcal{H}_0$, and setting $\alpha = 0$, one can verify that $\mu_{\hat{\eta}}(h) = \mu_{\tilde{\eta}}(h)$. It then follows that $\mathrm{KL}(\hat{\eta}, \eta_c) =$

$\mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c)$ for $\eta_c \in \mathcal{H}_0$. In the software implementation, η_c is taken as an additive model involving only y variables, that is, flat on the x-axis with y variables mutually independent.

4.3. Bayesian Confidence Intervals for Log Odds Ratios

Penalized least squares regression with quadratic penalties is equivalent to Bayesian estimation with Gaussian process priors (Wahba, 1978), based on which Bayesian confidence intervals were derived (Wahba, 1983). The utility can be extended to penalized likelihood regression using the quadratic approximation of the penalized likelihood functional at its minimizer (Gu, 1992). Modifications of the arguments for estimation in a finite-dimensional space $\mathcal{H}^*$ can be found in Kim & Gu (2004) and Du & Gu (2006), which we adopt here.

Write $\eta = \boldsymbol{\xi}^T\mathbf{c} + \boldsymbol{\phi}^T\mathbf{d} = \boldsymbol{\psi}^T\mathbf{a}$ as in (9), where $(\boldsymbol{\xi}^T, \boldsymbol{\phi}^T) = \boldsymbol{\psi}^T$ and $(\mathbf{c}^T, \mathbf{d}^T) = \mathbf{a}^T$, and refer η and $\mathbf{a}$ interchangeably. The quadratic approximation of (3) at $\tilde{\eta} = \eta_\lambda$ is seen to be

$$\frac{1}{2n}(\mathbf{a} - \tilde{\mathbf{a}})^T(nH)(\mathbf{a} - \tilde{\mathbf{a}}) + C,$$

where H is as in (14), $\tilde{\eta} = \boldsymbol{\psi}^T\tilde{\mathbf{a}}$, and C is a constant; (3) is the posterior likelihood of the data divided by n, so the posterior of $\mathbf{a}$ is approximately normal with mean $\tilde{\mathbf{a}}$ and covariance H^+/n, where H^+ is the Moore–Penrose inverse of H. The posterior of $\eta(u)$ is thus approximately normal with mean $\tilde{\eta}(u) = \boldsymbol{\psi}^T(u)\tilde{\mathbf{a}}$ and variance $\boldsymbol{\psi}^T(u)H^+\boldsymbol{\psi}(u)/n$. This however is of little practical use as the conditional density $f(y|x)$ involves a normalizing constant that varies with x.

Given $y_1, \ldots, y_p \in \mathcal{Y}$, for any $x \in \mathcal{X}$, one may consider contrasts of the form

$$\kappa(x) = \beta_1\eta(u_1) + \cdots + \beta_p\eta(u_p),$$

where $u_j = (x, y_j)$ and $\beta_1 + \cdots + \beta_p = 0$; the log odds ratios of $f(y|x)$ can be expressed as such contrasts with the normalizing constant cancelling out. The posterior of $\kappa(x)$ is seen to have a mean $\tilde{\kappa}(x) = \tilde{\boldsymbol{\psi}}^T(x)\tilde{\mathbf{a}}$ and a variance $s^2(x) = \tilde{\boldsymbol{\psi}}^T(x)H^+\tilde{\boldsymbol{\psi}}(x)/n$, where $\tilde{\boldsymbol{\psi}}(x) = \beta_1\boldsymbol{\psi}(u_1) + \cdots + \beta_p\boldsymbol{\psi}(u_p)$. Bayesian confidence intervals of $\kappa(x)$ are given by $\tilde{\kappa}(x) \pm z_{1-\alpha/2}\, s(x)$.

5. MIXED-EFFECT MODELS FOR CORRELATED DATA

When the responses are correlated, one may model the correlation via random effects, yielding mixed-effect models. Mixed-effect models for non-Gaussian regression with parametric fixed-effects were studied in, for example, Zeger & Karim (1991), Breslow & Clayton (1993), and McCulloch (1997), and those with nonparametric fixed-effects by Lin & Zhang (1999), Karcher & Wang (2001), and Gu & Ma (2005).

We now introduce mixed-effect models for cross-classified responses. To our knowledge, this is the first attempt on random effects for multivariate responses, which include the case of multinomial responses with $K = 1$ and $N_1 > 2$.

5.1. Random Effects

For regression with univariate responses, a mixed-effect model is of the form $\zeta = \eta(x) + \mathbf{z}^T\mathbf{b}$, where ζ is the modeling parameter such as the logit for logistic regression, $\eta(x)$ is the fixed-effect, and $\mathbf{z}^T\mathbf{b}$ is the random effect with $\mathbf{b} \sim N(0, B)$. The covariance matrix B of $\mathbf{b}$ is often structured but with unknown parameters; see Section 5.2 below.

In the current setting, one may replace (1) by

$$f(y|x) = \frac{e^{\eta_y + \eta_{xy} + \mathbf{z}^T\mathbf{b}_y}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{xy} + \mathbf{z}^T\mathbf{b}_y}}, \tag{16}$$

where $\eta(x, y) = \eta_y + \eta_{xy}$ is spelled out explicitly and $\mathbf{b}_y \sim N(0, B)$ varies with y. Note that $\int_{\mathcal{Y}} \eta(x, y) = 0$, and we shall specify the correlations among $\mathbf{b}_y$ to ensure $\int_{\mathcal{Y}} \mathbf{z}^T\mathbf{b}_y = 0$.

For $K = 1$ and $\mathcal{Y} = \{1, \ldots, N\}$, write $\tilde{\mathbf{b}} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_N^T)^T$. We shall specify

$$\tilde{\mathbf{b}} \sim N(0, c(I_N - \mathbf{1}_N\mathbf{1}_N^T/N) \otimes B), \tag{17}$$

where $\otimes$ denotes the Kronecker product of matrices and c is a constant. For $K > 1$, we consider an additive model $\mathbf{b}_y = \mathbf{b}_{y_{\langle 1\rangle}} + \cdots + \mathbf{b}_{y_{\langle K\rangle}}$, with independent components $\mathbf{b}_{y_{\langle k\rangle}}$ specified as above; the structure of B should remain the same for all the components $\mathbf{b}_{y_{\langle k\rangle}}$ but the unknown parameters may differ. For $K = 1$ and $N_1 = 2$, this reduces to the mixed-effect logistic regression model of Gu & Ma (2005).

5.2. Examples

The formulation through (16) and (17) propagates a random effect $\mathbf{z}^T\mathbf{b}$ for univariate responses to cross-classified responses, and we shall review a couple of examples of commonly used $\mathbf{z}^T\mathbf{b}$.

First consider a longitudinal study involving p subjects, where y_i is taken from subject s_i with covariate x_i. Observations from different subjects are independent, while observations from the same subject are naturally correlated. The intra-subject correlation may be modeled by $\mathbf{z}_i^T\mathbf{b} = b_{s_i}$, where $\mathbf{b} \sim N(0, \sigma_s^2 I)$ and $\mathbf{z}_i$ is the s_i-th unit vector. The $p \times p$ matrix $B = \sigma_s^2 I$ involves only one tunable parameter. The random effects b_s can be interpreted as the subject effects.

Now consider observations from p clusters, such as in multi-centre studies, where y_i is taken from cluster c_i with covariate x_i. Observations from different clusters are independent, while observations from the same cluster may be correlated to various degrees. The intra-cluster correlation may be modeled by $\mathbf{z}_i^T\mathbf{b} = b_{c_i}$, where $\mathbf{b} \sim N(0, B)$ with $B = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{z}_i$ is the c_i-th unit vector. The $p \times p$ matrix B involves p tunable parameters on the diagonal. The random effects b_c are not quite interpretable in the setting.

5.3. Modeling With Random Effects

Note that $\mathbf{b}_1 + \cdots + \mathbf{b}_N = 0$ given (17), so one only needs the first $(N - 1)$ $\mathbf{b}_y$'s. Rewriting $\tilde{\mathbf{b}} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_{N-1}^T)^T$, the minus log likelihood of random effect is seen to be proportional to $\tilde{\mathbf{b}}^T\Sigma\tilde{\mathbf{b}}$ for

$$\Sigma = c^{-1}(I_{N-1} + \mathbf{1}_{N-1}\mathbf{1}_{N-1}^T) \otimes B^{-1}, \tag{18}$$

where $(I_{N-1} + \mathbf{1}_{N-1}\mathbf{1}_{N-1}) = (I_{N-1} - \mathbf{1}_{N-1}\mathbf{1}_{N-1}/N)^{-1}$. For $K > 1$, one may concatenate all the independent components of $\mathbf{b}_y$ in $\tilde{\mathbf{b}}$ with Σ block-diagonal with blocks of the form (18). The estimation can then be performed via the minimization of the penalized joint likelihood of $(\eta, \tilde{\mathbf{b}})$,

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(x_i, y_i) + \mathbf{z}_i^T\mathbf{b}_{y_i} - \log\int_{\mathcal{Y}} e^{\eta(x_i,y)+\mathbf{z}_i^T\mathbf{b}_y}\right\} + \frac{1}{2n}\tilde{\mathbf{b}}^T\Sigma\tilde{\mathbf{b}} + \frac{\lambda}{2}J(\eta). \tag{19}$$

The computation follows straightforward adaptation of Section 3.4 and the cross-validation of Section 4.1 can be used to select the tuning parameters consisting of the smoothing parameters in $\lambda J(\eta)$ and the correlation parameters in $\tilde{\mathbf{b}}^T\Sigma\tilde{\mathbf{b}}$. The Kullback–Leibler projection of Section 4.2 can be computed with the random effect $\mathbf{z}^T\mathbf{b}_y$ treated as an offset, and the Bayesian confidence intervals of Section 4.3 for odds ratios remain available following straightforward adaptation.

6. SIMULATION STUDIES

We conduct simulation studies of limited scales to explore various aspects of the modeling tools being developed. For the test function on $\mathcal{Y} = \{0, 1\} \times \{0, 1\}$ and $\mathcal{X} = [0, 1]$, we start with $p_1(x)$, $p_2(x)$, and $p_3(x)$ given by

$$\log \frac{p_1(x)}{1 - p_1(x)} = 400x^5(1 - x)^3 - 1, \tag{20}$$

$$\log \frac{p_2(x)}{1 - p_2(x)} = 500x^7(1 - x)^3 + 250x^2(1 - x)^{10} - 1, \tag{21}$$

$$\log \frac{p_3(x)}{1 - p_3(x)} = 50x^2(1 - x)^4. \tag{22}$$

A setting with $y_{\langle 1\rangle} \perp y_{\langle 2\rangle} | x$ would have

$$(f(0, 0), f(0, 1), f(1, 0), f(1, 1)) = (q_1q_2, q_1p_2, p_1q_2, p_1p_2),$$

where $q_k = 1 - p_k$, but we modify it by

$$(f(0, 0), f(0, 1), f(1, 0), f(1, 1)) \propto (q_1q_2p_3, q_1p_2q_3, p_1q_2q_3, p_1p_2p_3);$$

note that $p_1(x)$ and $p_2(x)$ are no longer the marginal probabilities $P(y_{\langle 1\rangle} = 1|x)$ and $P(y_{\langle 2\rangle} = 1|x)$ after the modification, but the log odds ratio is given by

$$\log \frac{f(0, 0|x)f(1, 1|x)}{f(1, 0|x)f(0, 1|x)} = 2 \log \frac{p_3(x)}{1 - p_3(x)} = 100x^2(1 - x)^4.$$

Samples of size $n = 200$ were generated, for $x_i \sim U(0, 1)$, with and without random effects. For samples with random effects, $\mathbf{z}^T\mathbf{b}_i = b_1(s_i, y_{\langle 1\rangle}) + b_2(s_i, y_{\langle 2\rangle})$, where $s_i \in \{1, \ldots, 20\}$, 10 each, $b_1(s, 1) = -b_1(s, 0) \sim N(0, 0.5^2)$, and $b_2(s, 1) = -b_2(s, 0) \sim N(0, 0.5^2)$. Models of the form

$$\eta(x, y) = \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{12}(y_{\langle 1\rangle}, y_{\langle 2\rangle}) + \eta_{x1}(x, y_{\langle 1\rangle}) + \eta_{x2}(x, y_{\langle 2\rangle}) + \eta_{x12}(x, y_{\langle 1\rangle}, y_{\langle 2\rangle})$$

were fitted to the data.

6.1. Effectiveness of Cross-Validation

To assess the performance of $\hat{f}(y|x)$ as an estimate of $f(y|x)$, one may use the Kullback–Leibler discrepancy

$$\mathrm{KL}(f, \hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Y}} \log \left\{ \frac{f(y|x_i)}{\hat{f}_\lambda(y|x_i)} \right\} f(y|x_i),$$

where the dependence of $\hat{f}_\lambda(y|x)$ on the tuning parameters is made explicit, with the subscript λ representing the λ in (3) or (19), the θ's hidden in $J(\eta)$, and also the Σ in (19) for mixed-effect models. The conditional density $f(y|x)$ may be given by (1), for which we denote $L(\lambda) = \mathrm{KL}(f, \hat{f}_\lambda)$, or may be given by (16), for which we write $L_w(\lambda) = \mathrm{KL}(f, \hat{f}_\lambda)$; for fixed-effect fits via (3), $L(\lambda) = L_w(\lambda)$ as $\mathbf{z}^T\mathbf{b} = \mathbf{z}^T\hat{\mathbf{b}} = 0$. While $L(\lambda)$ is of more practical interest, only $L_w(\lambda)$ is "chaseable" via cross-validation.

For both the fixed-effect (without random effects) and mixed-effect (with random effects) simulations, one hundred replicates were generated, and tuning parameters were selected using the cross-validation score of (14) with $\alpha = 1, 1.4$. $L_w(\lambda_v)$ was evaluated for each of the cross-validated fits, where λ_v denotes the cross-validation choices of the tuning parameters. Also calculated for each of the replicates were the "optimal" fits minimizing $L_w(\lambda)$, with the resulting minimum value written as $L_w(\lambda_o)$. The left and centre frames of Figure 1 plots $L_w(\lambda_v)$ versus $L_w(\lambda_o)$ in the fixed-effect and mixed-effect simulations, respectively, for $\alpha = 1$ (solid) and $\alpha = 1.4$ (faded); the relative efficacy $L_w(\lambda_o)/L_w(\lambda_v)$ is shown in the right frame in box-plots. The results suggest that the fudge factor $\alpha = 1.4$ hurts the performance of cross-validation in the setting by an appreciable margin, especially with mixed-effect models; in fact, it was outperformed by $\alpha = 1$ 64-to-36 in the fixed-effect simulation and 100-to-0 in the mixed-effect simulation.

The cross-validation of (14) appears highly effective, but the findings concerning the fudge factor $\alpha = 1.4$ contrast empirical findings in numerous parallel settings, where it at least does no harm. Undersmoothing however occurs much less often with multiple smoothing parameters, a given in the current setting, thus stable performances of cross-validation may still be expected without a fudge factor.

Throughout the rest of the article, we will be using $\alpha = 1$ in cross-validation.

6.2. Fixed-Effect Fits Versus Mixed-Effect Fits

The results shown in Figure 1 were obtained with $q = n$ in (9) to eliminate any effect due to the choice of $\{v_j\}$. For a quick check on the adequacy of the default $q = 10n^{2/9}$, we selected one random subset $\{v_j\} \subset \{(x_i, y_i)\}$ of size $q = 10(200)^{2/9} \approx 33$ for each of the replicates and recalculated the cross-validated fits. The $L_w(\lambda_v)$ of the $q = 33$ fits are plotted against that of the $q = 200$ fits in the left frame of Figure 2, with their ratio shown in boxplots in the left half of the right frame. The same-data fits with different $\{v_j\}$ were mostly "duplicates" of each other. The one hundred $q = 200$ fits took a total of 4853 CPU seconds to compute on a laptop with Core2 duo 2.6 GHz and 4 Gb RAM running Linux Mint 9 and R 2.12.1, while the $q = 33$ fits took 383 CPU seconds.

We now compare fixed-effect fits and mixed-effect fits to the same data using the one hundred independent samples in the fixed-effect simulation and the one hundred correlated samples in the mixed-effect simulation. For each of the two hundreds replicates, a pair of cross-validated fits were calculated via (3) and (19) using the same $\{v_j\}$ of size $q = 33$. The respective $L(\lambda_v)$ of the fits are plotted in the centre frame of Figure 2, with those of independent samples in solid circles and those of correlated samples in faded; the ratio of $L(\lambda_v)$ for fixed-effect fits over that for mixed-effect fits is shown in the right half of the right frame. For the independent samples, the mixed-effect model is correct though with unnecessary complications, and the two sets of fits largely share the same performance. The correlated samples paint a different picture, however,

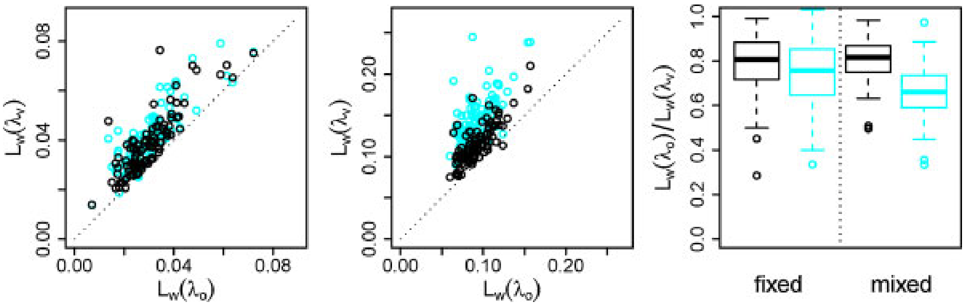

FIGURE 1: Performance of cross-validation. Left: fixed-effect simulation, for $\alpha = 1$ (solid) and $\alpha = 1.4$ (faded). Centre: mixed-effect simulation, for $\alpha = 1$ (solid) and $\alpha = 1.4$ (faded). Right: relative efficacy, for $\alpha = 1$ (solid) and $\alpha = 1.4$ (faded). [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

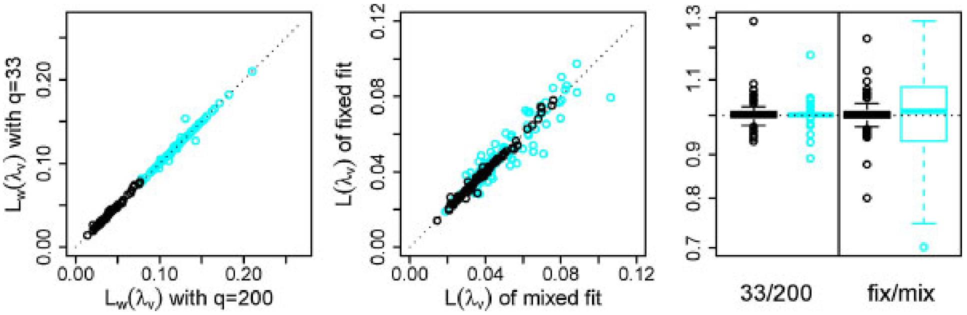

FIGURE 2: Some comparisons concerning $L_w(\lambda_v)$ and $L(\lambda_v)$. Left: $L_w(\lambda_v)$ of $q = 200$ fits versus that of $q = 33$ fits, with fixed-effect simulation in solid and mixed-effect simulation in faded. Centre: fixed-effect fits versus mixed-effect fits to same data, with independent samples in solid circles and correlated samples in faded. Right: $L_w(\lambda_v)$ of $q = 33$ fits over that of $q = 200$ fits on the left; $L(\lambda_v)$ of fixed-effect fits over that of mixed-effect fits on the right. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

with greater discrepancies in the respective performances of the two sets of fits but the overall preference a tossup. Intuitively, the mixed-effect fits should do better on correlated samples, but the added model components $\mathbf{z}^T\mathbf{b}$ and the extra tuning parameters in Σ make the task harder plus one has no direct handle on $L(\lambda)$ in the estimation process. Still, one would expect some benefits in using the mixed-effect models for correlated data, but perhaps in the presence of stronger correlation.

To look further into the comparison of fixed-effect fits and mixed-effect fits to correlated data, we repeated the experiments using four more sets of samples, each consisting of one hundred replicates. The sample size remained at $n = 200$, but to make correlation stronger, we took $s_i \in \{1, \ldots, 10\}$, 20 each, and $b_1(s, 1) \sim N(0, 1)$, $b_2(s, 1) \sim N(0, 1)$. One set of samples were generated using the same test function used earlier, with the resulting $L(\lambda_v)$ plotted in the left frame of Figure 3 in solid circles; for the second test function we use the same $p_1(x)$ and $p_2(x)$ but a constant $p_3(x) = 2/3$, with results shown in the same frame in faded circles. Using the function of (20) for both $p_1(x)$ and $p_2(x)$ and that of (22) for $p_3(x)$ yields the third test function, and changing out $p_3(x)$ to $2/3$ in the third yields the fourth; the centre frame of Figure 3 demonstrates the results from the third (solid) and the fourth (faded) sets of samples. The ratio of $L(\lambda_v)$ of fixed-effect fits over that of mixed-effect fits is shown in boxplots in the right frame of Figure 3

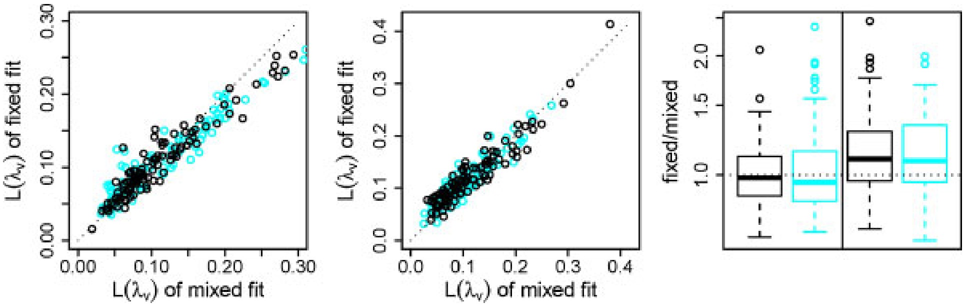

FIGURE 3: Comparison of fixed-effect and mixed-effect fits to correlated data. Left: $p_1(x)$ as in (20), $p_2(x)$ as in (21), $p_3(x)$ as in (22) (solid) or $p_3(x) = 2/3$ (faded). Centre: $p_1(x)$ and $p_2(x)$ both as in (20), $p_3(x)$ as in (22) (solid) or $p_3(x) = 2/3$ (faded). Right: $L(\lambda_v)$ of fixed-effect fits over that of mixed-effect fits, for the four sets of samples in order. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

for the four sets of samples. The results do seem to favour the mixed-effect fits overall, though the absolute values of $L(\lambda_v)$ are not all that different. As can be seen in the left and centre frames of Figure 3, the mixed-effect fits tend to do better in "easier" cases with smaller $L(\lambda_v)$ but do not fare as well in the "hard" cases; this is confirmed by plots, not shown here, of the ratio in the right frame versus the absolute $L(\lambda_v)$ in the left and centre frames.

6.3. Performances of Estimates

One case each were selected from the fixed-effect and random-effect simulations of Section 6.1 for further study, with $q = 33$, that delivered the median performances of $L_w(\lambda_v) = 0.037$,

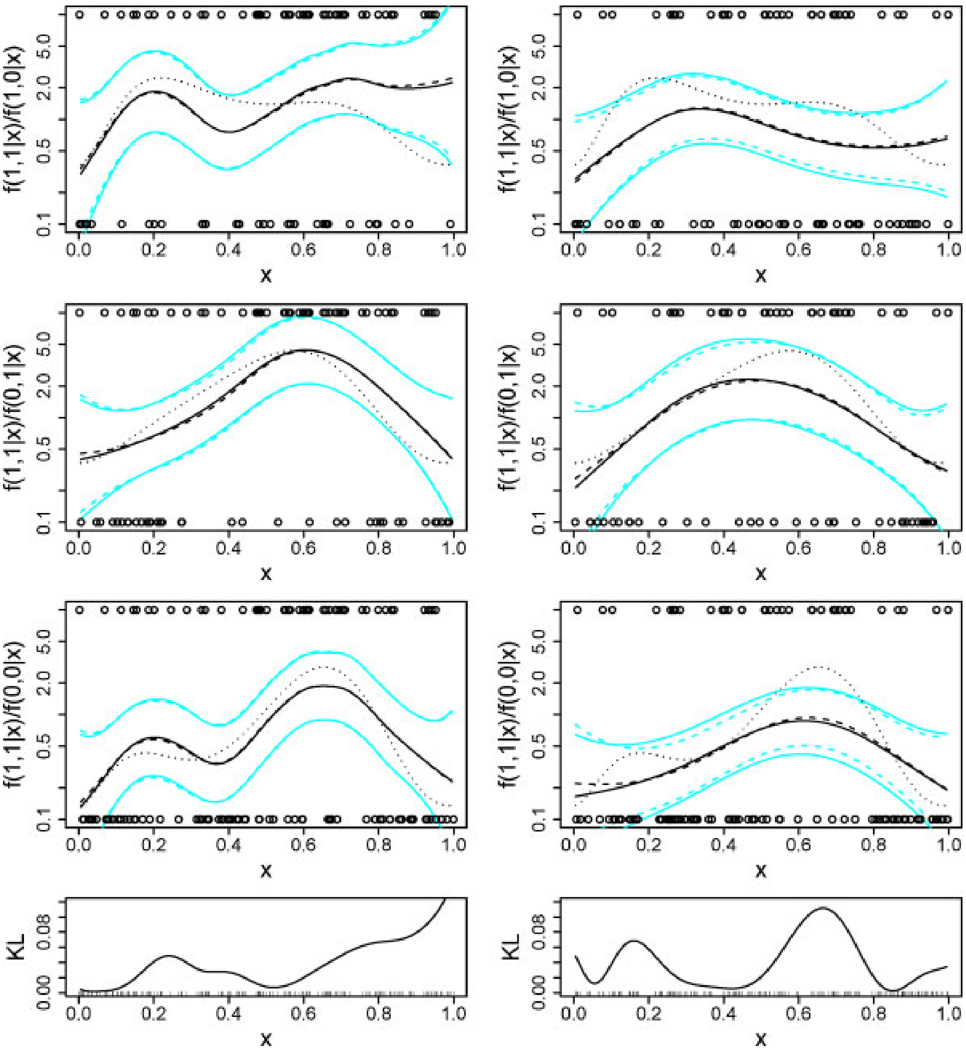

FIGURE 4: Some fits with median performances. Left: a fixed-effect fit. Right: a mixed-effect fit. Top rows: estimated $f(1, 1|x)/f(1, 0|x)$, $f(1, 1|x)/f(0, 1|x)$, and $f(1, 1|x)/f(0, 0|x)$ (solid) with 95% Bayesian confidence intervals (faded); true functions (dotted) and data (circles) are superimposed. Fits by the wrong models are also superimposed in dashed lines. Bottom: $\mathrm{KL}(f(x), \hat{f}(x)) = \int_{\mathcal{Y}} \log\{f(y|x)/\hat{f}(y|x)\} f(y|x)$ as a function of x. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

0.109, respectively. The left frames of Figure 4 are from the fixed-effect fit, where the estimated $f(1,1|x)/f(1,0|x)$, $f(1,1|x)/f(0,1|x)$, $f(1,1|x)/f(0,0|x)$ (solid) and the respective 95% Bayesian confidence intervals (faded) are plotted in the top three frames, with the true functions superimposed (dotted); the data are also superimposed, with $(x_i, 1, 1)$ marked on the top and $(x_i, 1, 0)$, $(x_i, 0, 1)$, or $(x_i, 0, 0)$ marked on the bottom in their respective frames. The bottom frame depicts $\mathrm{KL}(f(x), \hat{f}(x)) = \int_{\mathcal{Y}} \log\{f(y|x)/\hat{f}(y|x)\} f(y|x)$ as a function of x, with the rug showing the locations of x_i. The right frames are parallel results from the mixed-effect fit but evaluated with the random-effect $\mathbf{z}^T\mathbf{b}$ set to zero; note that $\mathrm{KL}(f(x), \hat{f}(x))$ in the bottom right frame does not average to $L_w(\lambda_v) = 0.109$ but to $L(\lambda_v)$, which was 0.036. As nonparametric fits based on samples of size $n = 200$, the estimation precision appears reasonable, and the Bayesian confidence intervals seem to have the adequate width though not necessarily the nominal coverage; some of the miscues, such as the rise of $f(1,1|x)/f(1.0|x)$ towards $x = 1$ in the fixed-effect fit (see the top left frame of Figure 4), are apparently responding to patterns in the data. The estimation precision is certainly not uniform on the $\mathcal{X}$ domain.

For comparison, the fits using the "wrong" models are also superimposed in the top three rows of Figure 4 in dashed lines, which are the unnecessary mixed-effect model fit in the left frames and the incorrect fixed-effect model fit in the right frames. The fits are almost visually indistinguishable.

For the mixed-effect fit in Figure 4, cross-validation selected (0.195, 0.251) as the variances of $b_1(s, 1)$ and $b_2(s, 1)$ used in Σ of (19). The sample variances of the generated $b_1(s, 1)$ and $b_2(s, 1)$ were (0.164, 0.170), and the sample variances of the fitted $\hat{b}_1(s, 1)$ and $\hat{b}_2(s, 1)$ were (0.051, 0.080). Cross-validation is designed to minimize the Kullback–Leibler discrepancy $L_w(\lambda)$ and Σ in (19) only forms part of the tuning parameters but *not* part of the model parameters to be estimated, so the cross-validation choice of Σ should not be taken as estimate.

7. ANALYSIS OF EYETRACKING DATA

We now analyze some eyetracking data collected by Dr. Anouschka Foltz during her dissertation research at The Ohio State University. The data we use is a subset consisting of 288 trials of human participants' eye movements monitored on a time grid (−867)(17)(1428) ms. In each run of the experiments, the participant in front of a computer monitor listened to three consecutive instructions, and we are looking at the time segment associated with the second instruction. For the selected subset, the first instruction given to the participant was something like "click on the YELLOW pencil" with emphasis on the adjective, and the second instruction was something like "click on the PURPLE bottle" with neither the adjective nor the noun being repeated. Upon hearing the emphasized adjective "PURPLE" but before the noun "bottle," one usually expects a noun repetition ("pencil") and starts to look for purple pencil on the monitor, and of interest is how long it takes for the participant to recover from the trap to focus on the target, the purple bottle. The participants' eye fixation on the target (purple bottle), the colour competitor (purple pencil), and the object competitor (yellow bottle) were recorded as binary indicators every 17 ms; the time 0 is the noun onset of the second instruction, that is, between the adjective (purple) and the noun (bottle). The common linguistic condition in the 288 trials was the accented (i.e., emphasized) adjectives and the change in both the adjective and the noun. The particular word choice in the discussion above (i.e., "yellow pencil" followed by "purple bottle") was taken from one of the six instruction lists used in the trials. The six lists were actually carefully designed using Latin squares to balance out certain linguistic characteristics, but further details are not needed for our task at hand. There were 48 human participants involved in 6 trials each, for a total of 288 trials. The raw data were reformatted into $288 \times 136 = 39168$ observations of time (x), matching colour indicator ($y_{\langle 1\rangle}$, eye fixation on target or colour competitor), matching object indicator ($y_{\langle 2\rangle}$, eye fixation on target or object competitor), and subject identification (s).

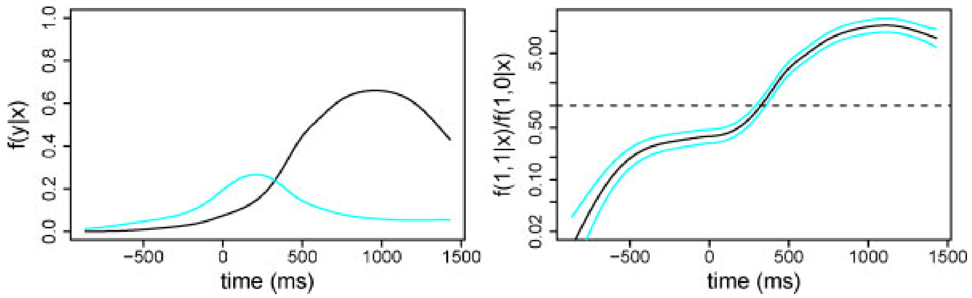

FIGURE 5: Fitted probabilities along time. Left: $f(1, 1|x)$ (solid) versus $f(1, 0|x)$ (faded). Right: $f(1, 1|x)/f(1, 0|x)$ along with 95% Bayesian confidence intervals. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

A model of the form (16) was fitted to the data, with $\eta_y = \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{12}(y_{\langle 1\rangle}, y_{\langle 2\rangle})$, $\eta_{xy} = \eta_{x1}(x, y_{\langle 1\rangle}) + \eta_{x2}(x, y_{\langle 2\rangle}) + \eta_{x12}(x, y_{\langle 1\rangle}, y_{\langle 2\rangle})$, and $\mathbf{z}^T\mathbf{b}_y = b_1(s, y_{\langle 1\rangle}) + b_2(s, y_{\langle 2\rangle})$, where $b_1(s, 1) = -b_1(s, 0) \sim N(0, \sigma_1^2)$ and $b_2(s, 1) = -b_2(s, 0) \sim N(0, \sigma_2^2)$ are independent. Since x is on a regular grid here, we took the first 136 observations as the v_j's discussed in Section 3.3 instead of a random subset, though this could not ensure "even coverage" on the "y-axis." Cross-validation selected $(\sigma_1^2, \sigma_2^2) = (0.163, 0.195)$ along with the smoothing parameters in $\lambda J(\eta)$. The sample variances of the fitted $b_1(s, 1)$ and $b_2(s, 1)$ are 0.0791 and 0.1076, respectively. Setting the random effects $\mathbf{b}_y$ to 0, the estimated $f(y|x) = e^{\eta(x,y)}/\int_{\mathcal{Y}} e^{\eta(x,y)}$ for $y = (1, 1)$ (target) and $y = (1, 0)$ (colour competitor) are plotted in the left frame of Figure 5, and the ratio $f(1, 1|x)/f(1, 0|x) = \exp\{\eta(x, 1, 1) - \eta(x, 1, 0)\}$ is plotted in the right frame along with 95% Bayesian confidence intervals; the transition time at which the target overtook the colour competitor was around 310 ms.

Of further interest is how such profile along time changes with experimental conditions, for which fits also need to be calculated for other subsets of the data. Different experimental conditions could be noun repetition (e.g., "purple pencil" following "yellow pencil"), adjective repetition (e.g.,"yellow bottle" following "yellow pencil"), and/or numerous accent patterns on the adjectives and the nouns.

Projecting the fit $\hat{\eta}$ into a model space of structure $\eta(x, y) = \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{x1}(x, y_{\langle 1\rangle}) + \eta_{x2}(x, y_{\langle 2\rangle})$, one has $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) = 36.7\%$, so $y_{\langle 1\rangle}$ and $y_{\langle 2\rangle}$ are not independent given x. Projecting $\hat{\eta}$ into a space of structure $\eta(x, y) = \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{12}(y_{\langle 1\rangle}, y_{\langle 2\rangle}) + \eta_{x1}(x, y_{\langle 1\rangle}) + \eta_{x2}(x, y_{\langle 2\rangle})$, one has $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) = 3.2\%$, so the dependence on x of the log odds ratio $\log\{f(0, 0|x)f(1, 1|x)/f(1, 0|x)f(0, 1|x)\}$ appears rather weak. The association between $y_{\langle 1\rangle}$ and $y_{\langle 2\rangle}$ however is not of primary interest in the current application.

8. SUMMARY AND DISCUSSION

In this article, we have presented a cohort of modeling tools for nonparametric regression with cross-classified responses. The general model formulation includes many settings in the literature as special cases, and allows one to study the associations among marginals of contingency tables as functions of covariates. The techniques are implemented in a suite of R functions, and are demonstrated through the analysis of some eyetracking data found in linguistic studies.

To our knowledge, the only existing approach to accommodating an "x-axis" in the analysis of cross-classified responses is through surrogate log-linear models (see Venables & Ripley, 2002, chap. 7), but the method works naturally only with discrete x variables. For continuous x's, the specification/justification of parametric models poses a real challenge, due mainly to the less intuitive meaning of $\eta(x, y)$, which yields the conditional probability $f(y|x) = e^{\eta(x,y)}/\int_{\mathcal{Y}} e^{\eta(x,y)}$

only after the normalization. The nonparametric treatment in this setting is thus less of an "extending parametric model" flavour but more for its own practical convenience. Also, mixed-effect models for correlated data in this setting do not seem to exist in current literature, parametric or nonparametric.

For the simulations of Section 6 and the eyetracking data of Section 7, one has a univariate x and only three θ's hidden in $J(\eta)$ associated with η_{x1}, η_{x2}, and η_{x12}, so the quasi-Newton optimization of (14) is still feasible. For multivariate x one could have many more θ's, in which case the quasi-Newton iteration will have to be skipped. On the same note, the simulations of Section 6.1 quickly become infeasible as the number of θ's increases.

Traditionally, eyetracking data are often aggregated over coarser time intervals then analyzed using ANOVA models for repeated measures; see, for example, Ito & Speer (2008). The loss of information in data aggregation was recognized by Barr (2008), who suggested the use of parametric logistic regression in the setting. The modeling tools developed here provide yet another approach to the analysis of eyetracking data, one that is flexible and preserves the fine-scale dynamics in the experiments.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Anouschka Foltz who kindly provided the eyetracking data and pertinent explanations of the experiments. Chong Gu's research was supported by U.S. National Science Foundation under grant DMS-0705961. Ping Ma's research was partially supported by U.S. National Science Foundation under grants DMS-0800631 and DMS-0723759 and the Office of Science (BER), U.S. Department of Energy.

APPENDIX

In this appendix, we illustrate the user-interface of some open-source R code that implements the techniques presented in this article, using the eyetracking data of Section 7. The code is in the `ssllrm` suite of the `gss` package by the first author, as of version 1.1-7.

R resources are archived at `http://cran.r-project.org`, where the source code of base R and that of over 2000 add-on packages can be found along with installation instructions. Assuming that base R and the `gss` package have been installed, the following line load the `gss` package and the `eyetrack` data frame at the R prompt,

```
library(gss); data(eyetrack)
```

where `eyetrack` consists of components `time`, `colour`, `object`, `id`, and `cnt`; duplicated data points are merged and the multiplicity counts are recorded in `cnt`, and there are 13891 distinctive records.

To fit the model as discussed in Section 7, one uses

```
fit <- ssllrm(~time*colour*object, ~colour+object, data=eyetrack,
              weight=cnt, id.basis=1:136, random=~1|id)
```

where the first model formula specifies model terms and the second lists response variables which are necessarily factors; terms not involving response variables are removed internally. The `weight` entry is only necessary for data with multiplicities. The `id.basis` specifies the v_j's that determine the basis functions ξ_j as in (9); a random subset will be used if `id.basis` is unspecified. The `random` entry specifies a univariate random effect $\mathbf{z}^T\mathbf{b}$ that will be propagated into a multivariate one following the lines of Section 5.1. If one would like to skip the quasi-Newton iteration for the optimization of (14), simply add an entry `skip.iter=TRUE`. A run of the fit on a laptop, with Core2 duo 2.6 GHz and 4 Gb RAM running Linux Mint 9 and R 2.12.1, took 5647 CPU seconds, and a run with `skip.iter=TRUE` took 4959 CPU seconds.

To evaluate the fit at time points in `tt`, use

```
predict(fit, data.frame(time=tt))
```

which returns a matrix of $f(y|x)$ with the columns adding up to one; the columns are ordered according to `fit$qd.pt`, which is `(colour,object)=(0,0)`, `(0,1)`, `(1,0)`, and `(1,1)` here. To calculate $\log\{f(1,1|x)/f(1,0|x)\} = \eta(x,1,1) - \eta(x,1.0)$ along with standard errors as discussed in Section 4.3, use

```
predict(fit, data.frame(time=tt), odds=c(0,0,-1,1), se=TRUE)
```

which return a list with components `fit` and `se.fit`.

The six terms of the fit η_1, η_2, η_{12}, η_{x1}, η_{x2}, and η_{x12} are coded as `"colour"`, `"object"`, `"colour:object"`, `"time:colour"`, `"time:object"`, and `"time:colour:object"`. To project the fit into a space of structure $\eta = \eta_1 + \eta_2 + \eta_{x1} + \eta_{x2}$, use

```
project(fit, include=c("colour","object","time:colour",
"time:colour"))
```

which returns a list with components `ratio` $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ and `kl` $\mathrm{KL}(\hat{\eta}, \tilde{\eta})$.

BIBLIOGRAPHY

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 337–404.

Barr, D. J. (2008). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59, 457–474.

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

Du, P. & Gu, C. (2006). Penalized likelihood hazard estimation: Efficient approximation and Bayesian confidence intervals. *Statistics & Probability Letters*, 76, 244–254.

Gao, F., Wahba, G., Klein, R., & Klein, B. E. (2001). Smoothing spline ANOVA for multivariate bernoulli observations, with application to ophthalmology data (with discussion). *Journal of the American Statistical Association*, 96, 127–160.

Gu, C. (1992). Penalized likelihood regression: A Bayesian analysis. *Statistica Sinica*, 2, 255–264.

Gu, C. (1995). Smoothing spline density estimation: Conditional distribution. *Statistica Sinica*, 5, 709–726.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer-Verlag, New York.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Canadian Journal of Statistics*, 32, 347–358.

Gu, C. & Ma, P. (2005). Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. *Journal of Computational and Graphical Statistics*, 14, 485–504.

Gu, C. & Qiu, C. (1993). Smoothing spline density estimation: Theory. *Annals of Statistics*, 21, 217–234.

Gu, C. & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific Computing*, 12, 383–398.

Gu, C. & Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, 13, 811–826.

Ito, K. & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58, 541–573.

Karcher, P. & Wang, Y. (2001). Generalized nonparametric mixed effects models. *Journal of Computational and Graphical Statistics*, 10, 641–655.

Kim, Y.-J. & Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B*, 66, 337–356.

Lin, X. (1998). *Smoothing Spline Analysis of Variance for Polychotomous Response Data*. Ph. D. thesis, University of Wisconsin—Madison.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B*, 61, 381–400.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*, Springer-Verlag, New York.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162–170.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics With S-PLUS*, 4th ed., Springer, New York.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B*, 40, 364–372.

Wahba, G. (1983). Bayesian “confidence intervals” for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B*, 45, 133–150.

Zeger, S. L. & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Received 10 February 2011
Accepted 12 June 2011