Lorelogram: A Regression Approach to Exploring Dependence in Longitudinal Categorical Responses

Patrick J. HEAGERTY and Scott L. ZEGER

We propose flexible regression estimators of the marginal pairwise log-odds ratio measure of association for longitudinal categorical responses. The function that we estimate is the log-odds ratio analog of the correlogram; hence we name the function the *lorelogram.* Measuring the association of categorical responses on the log-odds scale allows ease of interpretation and allows pairwise association to remain unconstrained by the marginal means, a feature not shared by correlations with binary or multinomial responses. Estimation of the function is achieved through the use of standard parametric estimating equations or through an extension of generalized additive models that allows nonparametric estimation of dependence functions for fixed smoothing parameters. We apply the methodology to binary longitudinal data where scientific interest focuses on the dependence structure.

KEY WORDS: Correlogram; Estimating equation; Variogram.

1. INTRODUCTION

In the analysis of longitudinal data, the dependence structure can be of direct scientific interest, useful for the efficient estimation of mean parameters, or simply a nuisance. This article is concerned with the description of the dependence structure when it is either of direct interest or of interest as an exploratory prelude to further regression modeling.

Section 2 describes the variogram as used with continuous longitudinal responses. For categorical longitudinal responses, we propose an alternative measure of dependence based on the marginal pairwise log-odds ratio that Heagerty (1995) termed the "lorelogram." The advantages of this measure include familiarity with the scale of the measure and the lack of mean constraints. In Section 3 we develop three estimation strategies for obtaining a fitted lorelogram: a parametric approach based on estimating equations (Liang and Zeger 1986), a related regularization estimator (O'Sullivan 1986), and a nonparametric estimator that naturally extends the generalized additive models of Hastie and Tibsirani (1990) to the estimation of dependence functions.

We apply the methodology to six longitudinal binary responses collected monthly for 90 schizophrenia patients. The estimated lorelograms display the components of covariance due to serial dependence and patient heterogeneity and clearly distinguish two classes of symptoms.

Finally, we suggest other uses for the estimated lorelogram, including assessment of hierarchical model assumptions by using covariates other than time and use for general dependent data structures, including spatially correlated categorical responses.

2. LORELOGRAM: PROPERTIES

For categorical data, we propose an alternative to the variogram or correlogram (Cressie 1993; Diggle 1990) that measures dependence in terms of the marginal pairwise log-odds ratio. First, we briefly review the properties of the variogram and correlogram; then, we describe the proposed function, the lorelogram. We specifically consider isotropic models in which the dependence between observations is assumed to depend only on the distance between them. We avoid the stronger assumption of stationarity (Cinlar 1975), because we wish to address situations in which the mean may depend on covariates including time.

2.1 Variogram and Correlogram

The variogram (Cressie 1993; Diggle 1990) is a functional description of the isotropic dependence structure in time series, spatial observations, or longitudinal responses. Without loss of generality, we review the variogram and its estimation specifically for continuous longitudinal responses. Consider measurements $Y_{ij}, j = 1, \ldots, n_i$ on subjects $i = 1, \ldots, N$, taken at times t_{ij}. Define the residuals, $R_{ij} = Y_{ij} - E[Y_{ij}]$. For $j \neq k$, the variogram is defined as

$$\gamma(|t_{ij} - t_{ik}|) = \frac{1}{2}E[(R_{ij} - R_{ik})^2].$$

An alternative function that is used is the covariogram,

$$C(|t_{ij} - t_{ik}|) = \text{cov}(R_{ij}, R_{ik}).$$

Finally, the correlogram is a rescaled version of the covariogram, $r(\Delta t) = C(\Delta t)/V$, where $V = \text{var}(R_{ij}) \geq C(0)$. Each of these functions captures the isotropic dependence equivalently if the marginal variance exists. The major differences lie in the interpretation of the function and the statistical properties of their estimators.

For longitudinal data, the correlogram features $1 - r(0)$, the difference $r(0) - r(+\infty)$, and $r(+\infty)$ have useful interpretations in terms of proportions of variability due to measurement error, serial dependence, and subject heterogeneity (Diggle 1988). In a similar fashion we can use the variogram features $\gamma(0), \gamma(+\infty)$, and V to characterize the variance components under Diggle's model (see Diggle, Liang, and Zeger 1994 for examples and discussion).

Patrick J. Heagerty is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195. Scott L. Zeger is Professor and Chairman, Department of Biostatistics, Johns Hopkins University School of Public Health, Baltimore, MD 21205. The data on schizophrenia originate from the Madras Longitudinal Study, conducted by R. Thara. The Madras Longitudinal Study was supported by the National Institute of Mental Health grant 44653 (W. Eaton, Principal Investigator), the India Council on Medical Research, and the Schizophrenia Research Foundation (India).

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

For continuous responses, the variogram is estimated nonparametrically by the empirical variogram, which is defined as a smoothing of the squared residual differences, $\frac{1}{2}(R_{ij} - R_{ik})^2$, plotted against the time separation, $|t_{ij} - t_{ik}|$ (Diggle et al. 1994). Conversion to the correlogram is achieved by $\hat{r}(u) = 1 - \{[\hat{\gamma}(u)]/\hat{V}\}$.

The empirical variogram or correlogram provides a graphical summary of the isotropic dependence structure. This summary may then be used to assess the relative contributions of serial dependence and subject heterogeneity and to guide selection of appropriate parametric covariance models for continuous longitudinal responses.

2.2 Lorelogram

For binary or multinomial data, the marginal variance is a function of the mean, and hence the variogram or covariogram are not appropriate summaries of the dependence. The correlogram may be used, but for binary responses it is known that the correlation is constrained by the means due to the Frechet inequality, $P[Y_1 = Y_2 = 1] \leq \min(P[Y_1 = 1], P[Y_2 = 1])$ (Lipsitz, Laird, and Harrington 1991). For example, let $\mu_1 = E[Y_{i1}], \mu_2 = E[Y_{i2}]$ and assume $\mu_1 \leq \mu_2$. Then

$$E[Y_{i1}Y_{i2}] \leq \mu_1$$

and

$$\text{corr}(Y_{i1}, Y_{i2}) \leq \left[\frac{\mu_1(1-\mu_2)}{(1-\mu_1)\mu_2}\right]^{1/2}.$$

The bound is recognized as the square root of the odds ratio. For equal means, there is no constraint; however, if the two means differ, then the allowable range for the correlation can be severely constrained. For $(\mu_1, \mu_2) = (.1, .3)$, the upper bound on the correlation is .51. Thus even "moderately" different means can greatly restrict the range for the correlation.

To illustrate the potential impact of mean constraints, we considered a generalized linear mixed model,

$$g(P[Y_{ij} = 1|\mathbf{b}_i]) = \mathbf{X}_{ij}\boldsymbol{\beta} + b_{ij}$$

$$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}_i),$$

where the (j, k) element of the covariance matrix $\mathbf{D}_i$ is given by

$$D_i(j,k) = \tau^2 + \sigma^2\rho^{|t_{ij}-t_{ik}|}.$$

We further assume that the Y_{ij} are conditionally independent given the vector $\mathbf{b}_i$. Thus under this model, dependence among the observed binary data is induced by a latent autocorrelated Gaussian process. Figure 1 shows the induced pairwise correlation for a vector of 12 equally spaced binary observations where the mean function is linear in time, $\mathbf{X}_{ij}\boldsymbol{\beta} = \beta_0 + \beta_1 t_{ij}; g = \Phi^{-1}$ the probit link function; and the dependence parameters are fixed at $\tau^2 = 1.5, \sigma^2 = 2.5$, and $\rho = .5$. Using this model yields a marginal prevalence function $P[Y_{ij} = 1] = E[\Phi(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{ij})] = \Phi\{[1/(\sqrt{1 + \tau^2 + \sigma^2})]\mathbf{X}_{ij}\boldsymbol{\beta}) = \Phi[(1/\sqrt{5})\mathbf{X}_{ij}\boldsymbol{\beta}]$. In Figure 1a, $\boldsymbol{\beta}^T = \sqrt{5} \times (.5, 0)$, yielding stationarity and an autocorrelation function that directly reflects the latent process. (See Keenan 1982 for a discussion of stationary binary processes induced in this fashion.) But in Figure 1, where $\boldsymbol{\beta}^T = \sqrt{5} \times (.5, -.2)$, the process is no longer stationary,

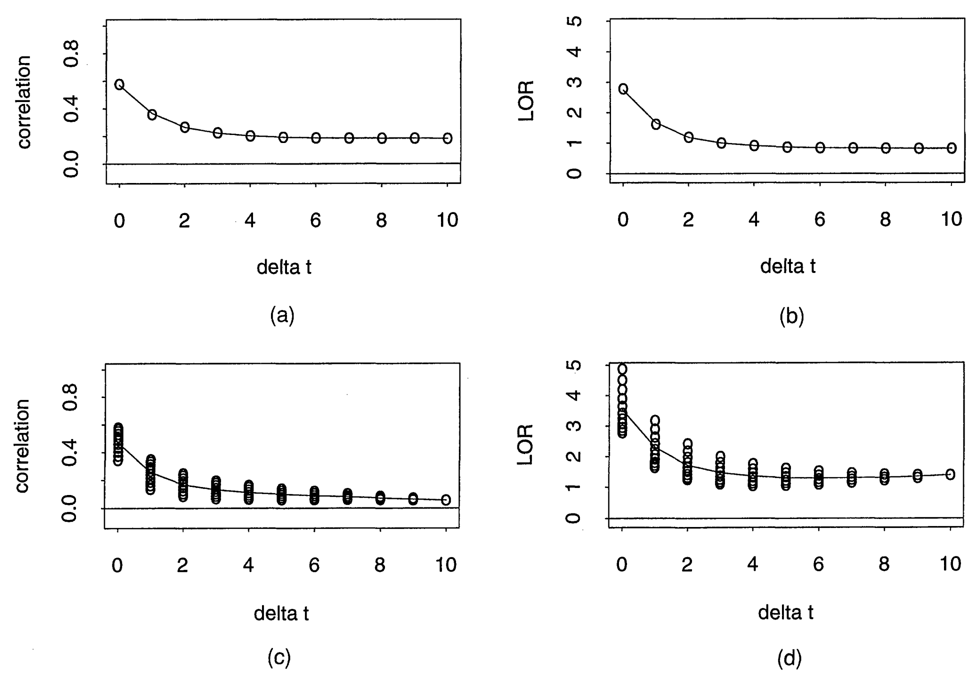

*Figure 1. Correlogram and Lorelogram for Binary Data Assuming a Latent Autocorrelated Gaussian Process. (a) COR: $\beta = \sqrt{(5)}$*c(.5, 0); (b) COR: $\beta = \sqrt{(5)}$*c(.5, −.2); (c) LOR: $\beta = \sqrt{(5)}$*c(.5, 0); (d) LOR: $\beta = \sqrt{(5)}$*c(.5, −.2). The model assumes that the binary responses are conditionally independent given the latent process with the mean a function of time, given by $\Phi^{-1}(P[Y_{ij} = 1|\mathbf{b}_i]) = \beta_0 + \beta_1 \times t_{ij} + b_{ij}$. The first row shows data with a constant mean function; the second row shows data with a linearly declining conditional mean function. The latent Gaussian process is mean 0 with covariance given by $\text{cov}(b_{ij}, b_{ik}) = 1.5 + 2.5 \times .5^{|t_{ij}-t_{ik}|}$. ○, exact value of the pairwise correlation (log-odds ratio) of Y_{ij} and Y_{ik} versus $|j - k|$; the lines give the average correlation (log-odds ratio).*

and the average correlation continues to decline over the time range. Thus naive inspection of the marginal correlations may lead to the erroneous conclusion that the latent dependence structure is purely serial.

In contrast to correlations, the marginal pairwise odds ratio is unconstrained by the marginal prevalence. The pairwise odds ratio is given by

$$\Psi(Y_{i1}, Y_{i2}) = \frac{P[Y_{i1} = 1, Y_{i2} = 1]P[Y_{i1} = 0, Y_{i2} = 0]}{P[Y_{i1} = 1, Y_{i2} = 0]P[Y_{i1} = 0, Y_{i2} = 1]}.$$

In the exponential family representation of a pair of binary responses, the likelihood can be written as $L(Y_1, Y_2; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 Y_1 + \theta_2 Y_2 + \theta_{12} Y_1 Y_2)$. In this representation the unconstrained canonical parameter θ_{12} is the pairwise log-odds ratio.

Following Lipsitz et al. (1991), we propose using the marginal pairwise log-odds ratio to describe the serial dependence for binary responses. Define the lorelogram as

$$\text{LOR}(t_{ij}, t_{ik}) = \log \Psi(Y_{ij}, Y_{ik}).$$

As a special case, we define the isotropic lorelogram as

$$\text{LOR}(|t_{ij} - t_{ik}|) = \log \Psi(Y_{ij}, Y_{ik}).$$

The odds ratio is strictly positive and unbounded, so taking its logarithm yields the entire real line as the range of allowable values. Furthermore, for binary responses the log-odds ratio is the scale on which regression parameters are measured, assuming a logistic link function. Therefore, using a log-odds ratio dependence function permits meaningful and familiar interpretation of the scale.

Returning to the earlier example of a binary series induced via a latent autocorrelated Gaussian process, we find that the lorelogram retains the basic features of the latent process's autocorrelation function. Figure 1c shows the lorelogram for the stationary binary series used to construct the plot in 1a. Figure 1d plots the average log-odds ratio for the binary longitudinal series used to construct 1b, where the conditional mean function declines linearly on the probit scale. In this case the average correlation is constrained by the changing means. However, the pairwise log-odds ratio is unconstrained, and the average log-odds ratio clearly reflects both the serial correlation and the long-range, or subject-specific, dependence.

Odds ratios also naturally describe dependence for ordinal or nominal responses. For ordinal responses, we represent a datum in terms of cumulative indicators. Define $O_{ij} \in [1, 2, \ldots, C]$ as the ordinal response and consider $\mathbf{Y}_{ij} = \text{vec}(Y_{ijc})$, where $Y_{ijc} = \mathbf{1}(O_{ij} > c)$ $c \in [1, 2, \ldots, C-1]$. Note that $E[Y_{ijc}]$ is used in the cumulative link regression models such as the discrete proportional hazards model and the proportional odds model (McCullagh and Nelder 1989). The global odds ratio can be used to describe the dependence between two ordinal responses (Dale 1986). For a given pair of ordinal responses there are $(C-1)^2$ global odds ratios defined as

$$\Psi_{i(j,k)(c_1,c_2)} = \Psi(Y_{ijc_1}, Y_{ikc_2}) \quad \text{for} \quad c_1 \in [1, \ldots, C-1]$$
$$\text{and} \quad c_2 \in [1, \ldots, C-1].$$

One may adopt simpler models that assume that for a given pair there exists a single common global odds ratio yielding a single association parameter for a pair of ordinal responses:

$$\Psi_{i(j,k)(c_1,c_2)} = \Psi_{i(j,k)} \quad \text{independent of cutpoints.}$$

These global odds ratios do not require any assumptions regarding the distance between categories nor any assignment of scores. Thus the global odds ratio is able to reflect category ordering and hence collapsibility of ordinal measures without imposing an arbitrary distance or score.

Odds ratios can be used similarly to measure dependence for nominal responses. But reasonable assumptions to reduce the number of parameters are generally unavailable. This is analogous to the parameter dimension issue for polytomous logistic regression for independent responses (McCullagh and Nelder 1989).

Finally, we note that the lorelogram as a function, though unconstrained for any given pair of times, is not globally unconstrained. The dependence function combined with the mean must correspond to a valid covariance structure. This condition is similarly required of a valid correlation function (Cressie 1993) and is difficult to ensure for flexible empirical estimators.

To summarize, we propose modeling the marginal pairwise log-odds ratio as a function of the observation times (t_{ij}, t_{ik}) as an exploratory method for assessing the dependence structure in categorical longitudinal responses. The advantage of regression modeling is that estimation of $\text{LOR}(t_{ij}, t_{ik})$ is possible even when observation times vary widely among subjects. The principal advantages of the log-odds ratio summary are as follows:

- It is pairwise unconstrained as opposed to the correlation of binary responses, which may have severe constraints imposed by differing marginal means.
- The log-odds ratio is on the same scale as the linear predictor assuming a logit link, and thus is on a familiar and interpretable scale.
- It extends naturally to ordinal or nominal responses.

3. ESTIMATION

In this section we outline estimation of the lorelogram. For parametric models we use a pair of estimating equations that allow modeling of the marginal mean and pairwise log-odds ratio (Heagerty and Zeger 1996; Lipsitz et al. 1991). We introduce a regularization estimator useful for stabilizing log-odds ratio estimates influenced by sparse or zero cells in local pairwise 2×2 association tables and suggest a simple data-driven selection method for the penalty parameter. For a nonparametric approach, we adopt a generalized additive model (Hastie and Tibshirani 1990) for the lorelogram. Pointwise standard errors are obtained by jackknifing the estimating equations (Lele 1991).

3.1 Estimating Equations

3.1.1 Paired Regression Models. For continuous responses, Y_{ij}, the dependence structure, is described through an empirical variogram after removing any mean trend. This is accomplished by considering the dependence among the residuals, $R_{ij} = Y_{ij} - \mu_{ij}$. Typically a flexible or saturated mean model is used (Diggle et al. 1994) to avoid bias in the dependence estimates. We propose using a marginal generalized linear model to adjust for mean trends. Let $E[Y_{ij}] = \mu_{ij}$. The mean model is given by specifying a link function, g_1, that relates the expected value to covariates: $g_1(\mu_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$. For binary or ordinal data, common link functions include the logit, probit, and complementary log-log functions.

Given a model for the marginal mean, we now consider a model for the association structure. We choose to model the pairwise odds ratios through a second generalized linear model. Let $\Psi_{i(j,k)}$ be the pairwise odds ratio measuring the association between Y_{ij} and Y_{ik}. We consider both a parametric model and a nonparametric additive model that can be represented jointly as

$$g_2(\Psi_{i(j,k)}) = \mathbf{Z}_{1,ijk}\boldsymbol{\alpha} + \sum_{m=1}^{r} s_m(Z_{2m,ijk}).$$

Here $\mathbf{Z}_1$ represents covariates used parametrically and $s_m(Z_{2m})$ represents smooth functions of covariates Z_{2m} used additively. Common choices for the second link function g_2 would include the log and identity functions. In subsequent sections we restrict our attention to the log link.

3.1.2 Paired Estimating Equations. First, we consider the fully parametric model ($r = 0$). For estimating the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, we use a pair of estimating equations. By defining models for the mean and association we are identifying only the first two moments of the joint distribution of a cluster of observations. A likelihood formulation requires a complete model for the joint distribution and is generally computationally intense (Fitmaurice, Laird, and Rotnitzky 1993; Heagerty and Zeger 1996). But given only the first two moments, a semiparametric approach is feasible by defining the estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ as the root of the equations

$$\mathbf{0} = \mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \boldsymbol{\mu}_i}{\partial \beta}\right]^T \mathbf{V}_{1i}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

and

$$\mathbf{0} = \mathbf{U}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^T \mathbf{V}_{2i}^{-1}(\mathbf{S}_i - \boldsymbol{\sigma}_i),$$

where $\mathbf{Y}_i = \text{vec}(Y_{ij}), \mu_i = E[\mathbf{Y}_i], \mathbf{S}_i = \text{vec}((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik}))$, and $\boldsymbol{\sigma}_i = E[\mathbf{S}_i]$. The weight matrix $\mathbf{V}_{1i} = \text{cov}(\mathbf{Y}_i)$ is completely determined by the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, because $\sigma_{ijk} = \text{cov}(Y_{ij}, Y_{ik})$ is uniquely defined by the odds ratio Ψ_{ijk} and the marginal means (μ_{ij}, μ_{ik}) via the relationship

$$\Psi_{ijk} = \frac{[\sigma_{ijk} + \mu_{ij}\mu_{ik}][\sigma_{ijk} + (1 - \mu_{ij})(1 - \mu_{ik})]}{[\sigma_{ijk} - \mu_{ij}(1 - \mu_{ik})][\sigma_{ijk} - \mu_{ik}(1 - \mu_{ij})]}.$$

Thus, given $g_1(\mu) = \mathbf{X}\boldsymbol{\beta}$ and $g_2(\Psi) = \mathbf{Z}\boldsymbol{\alpha}$, we can solve the foregoing expression to obtain σ_{ijk} as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ (Mardia 1967). The matrix $\mathbf{V}_{2i}$ approximates the covariance of $\mathbf{S}_i$ using only the first- and second-moment assumptions. The estimators defined by these conditions are known to be consistent and asymptotically Gaussian under mild regularity conditions (Liang and Zeger 1986; Prentice 1988; Prentice and Zhao 1991). Lipsitz et al. (1991) first proposed using a log-odds ratio model to specify the pairwise covariance structure for binary longitudinal responses and used similar paired estimating equations to obtain model estimates.

Finally, we note that the valid use of estimating equations for lorelogram estimation with time dependent covariates requires careful consideration that both the marginal expectation and covariance be specified correctly conditional on the entire covariate vector $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in_i})$. Implicit in our notation is the sufficient condition that $E[Y_{ij}|\mathbf{X}_i] = E[Y_{ij}|\mathbf{X}_{ij}]$. Pepe and Anderson (1994) have presented a detailed discussion of issues surrounding the use of estimating equations with time-dependent covariates.

3.2 Parametric Estimation: Spline Regression

We consider the parametric model by setting $r = 0$, thereby reducing estimation of the lorelogram to a parametric regression problem. For description of a stationary dependence structure, we define the covariate $z_{ijk} = |t_{ij} - t_{ik}|$, the time separation between the pair of observations, and use this variable to create a knotted cubic spline basis matrix that imposes the natural boundary constraints of linearity beyond the range of the data. For K specified knots, we require $K+2$ basis vectors. Estimation of the lorelogram is done by estimation of the parameter $\boldsymbol{\alpha}$, the coefficients of the natural spline basis elements.

The estimating equation strategy avoids assumptions regarding the third- and higher-order moments and as a result there is no model based estimate of the variance of the second moment parameter estimate $\hat{\boldsymbol{\alpha}}$. An empirical estimate of the variance of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ is obtained through an "information sandwich" (Liang and Zeger 1986; Royall 1986; White 1982) given by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{N} \mathbf{D}_{1i}^T \mathbf{W}_i \mathbf{D}_{2i}\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{U}_i \mathbf{U}_i^T\right) \times \left(\sum_{i=1}^{N} \mathbf{D}_{2i}^T \mathbf{W}_i \mathbf{D}_{1i}\right)^{-1}$$

where $\mathbf{D}_{2i} = \{[\partial(\mu_i, \sigma_i)]/[\partial(\beta, \alpha)]\}, \mathbf{D}_{1i}$ sets $\partial\sigma_i/\partial\beta$ to 0, $\mathbf{W}_i$ is the block diagonal matrix with $\mathbf{V}_{1i}^{-1}$, and $\mathbf{V}_{2i}^{-1}$ on the diagonal, and $\mathbf{U}_i = (\mathbf{U}_{1i}^T, \mathbf{U}_{2i}^T)^T$. We use this variance estimate and the asymptotic normality of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ to place pointwise confidence bands around the estimated lorelogram.

3.3 Zero Cell Correction: A Regularization Estimator

Zero cells and sparse cells can cause problems with the existence of estimates in general log-linear models. Many researchers routinely add a small constant to every cell to stabilize estimates (Agresti 1991). In examples, Goodman (1970) recommended adding 1/2 to every cell for estimation in saturated log-linear models. Similarly, our estimation algorithm runs the risk of diverging in highly flexible models similar to saturated models in classical log-linear analysis. We propose a solution that is qualitatively similar to adding a small constant to cells of contingency tables but is formally motivated in the regression context via a Bayesian argument. Our approach is related to Bayesian estimation for contingency tables (Laird 1978; Leonard 1975). Similar to Laird (1978), we adopt a flat prior on the parameters that describe the marginal means but adopt a Gaussian prior on the pairwise association parameters.

The approach that we propose is derived from putting a weak prior on the association regression parameters. For the lorelogram using natural splines, we modify the second estimating function to be

$$\mathbf{U}_2^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^T \mathbf{V}_{2i}^{-1}(\mathbf{S}_i - \boldsymbol{\sigma}_i) - \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\alpha}.$$

This equation can be derived as an approximation to the derivative of the log-posterior based on a quadratic exponential family likelihood and a Gaussian prior on α with mean 0 and covariance $\boldsymbol{\Lambda}_2$.

More generally, consider a quadratic exponential family model with independent Gaussian priors on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The posterior distribution under this model has the form

$$f(\boldsymbol{\beta}, \boldsymbol{\alpha} | \mathbf{Y}_i, i = 1, \ldots, N)) \propto \prod_{i=1}^{N} \exp(\theta_{0i} + \theta_{1i}^T \mathbf{Y}_i + \theta_{2i}^T \mathbf{T}_i) \times \pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}),$$

where $\boldsymbol{\theta}_i = (\theta_{0i}, \boldsymbol{\theta}_{1i}^T, \boldsymbol{\theta}_{2i}^T)^T$ represents the canonical parameters as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}, \mathbf{Y}_i = \text{vec}(Y_{ij})$, and $\mathbf{T}_i = \text{vec}(Y_{ij}Y_{ik})$, the $\binom{n_i}{2}$ vector of pairwise products. The priors are assumed to be independent and are given as $\pi(\boldsymbol{\beta}) = \text{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Lambda}_1)$ and $\pi(\boldsymbol{\alpha}) = \text{MVN}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Lambda}_2)$. The posterior mode is given by setting the derivative of the log posterior to 0:

$$\sum_{i=1}^{N} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & \mathbf{0} \\ \frac{\partial \sigma_i}{\partial \beta} & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}^T \begin{bmatrix} \mathbf{V}_{11i} & \mathbf{V}_{12i} \\ \mathbf{V}_{21i} & \mathbf{V}_{22i} \end{bmatrix}^{-1} \times \begin{pmatrix} \mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \\ \mathbf{S}_i - \boldsymbol{\sigma}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{pmatrix} - \begin{bmatrix} \boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) \\ \boldsymbol{\Lambda}_2^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha) \end{bmatrix} = \mathbf{0},$$

where $\mathbf{S}_i = \text{vec}((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik}))$ and $\mathbf{V}_{11i} = \text{cov}(\mathbf{Y}_i), \mathbf{V}_{12i} = \text{cov}(\mathbf{Y}_i, \mathbf{S}_i)$, and $\mathbf{V}_{22i} = \text{cov}(\mathbf{S}_i)$. Direct solution of this equation is difficult due to the need to recover the third and fourth moments to construct the covariance matrices. An estimating equation approximation is obtained by setting $\mathbf{V}_{12i} = \mathbf{0}$ and $\partial\sigma_i/\partial\beta = \mathbf{0}$ (Liang, Zeger, and Qaqish 1992; Prentice and Zhao 1991). This results in the pair of estimating functions

$$\mathbf{U}_1^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \mu_i}{\partial \beta}\right]^T \mathbf{V}_{11i}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) - \boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)$$

and

$$\mathbf{U}_2^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^T \mathbf{V}_{22i}^{-1}(\mathbf{S}_i - \boldsymbol{\sigma}_i) - \boldsymbol{\Lambda}_2^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha).$$

Although we motivate the equations $\mathbf{U}_1^* = \mathbf{0}, \mathbf{U}_2^* = \mathbf{0}$ via a Bayesian formulation, such equations have a long history in the non-Bayesian literature as well. Penalized likelihood methods (Green 1987) and ridge regression Hoerl and Kennard 1970) also yield estimating equations of this form. We now direct attention to the sampling properties of these equations, deriving asymptotic properties of the resulting estimators $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\alpha}}^*)$.

Proposition 1. For fixed $(\boldsymbol{\mu}_\beta, \boldsymbol{\Lambda}_1)$ and $(\boldsymbol{\mu}_\alpha, \boldsymbol{\Lambda}_2)$, the estimator $(\hat{\boldsymbol{\beta}}_N^*, \hat{\boldsymbol{\alpha}}_N^*)$ obtained as the solution to $\mathbf{U}_1^* = \mathbf{0}, \mathbf{U}_2^* = \mathbf{0}$ is weakly consistent as $N \to \infty$ given standard regularity conditions.

The proof of this proposition follows the results of Crowder (1986), giving general conditions on the consistency of estimators resulting from the solution of estimating equations. Note that asymptotically, the estimating functions $\mathbf{U}_1$ and $\mathbf{U}_1^*$ are equivalent in the sense that

$$\frac{1}{N}\mathbf{U}_1 - \frac{1}{N}\mathbf{U}_1^* = \frac{1}{N}\boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta).$$

Thus although $\mathbf{U}_1^*$ is not unbiased, the bias is $O(1/N)$ for fixed $\boldsymbol{\beta}$, resulting in a consistent estimator. The equations $\mathbf{U}_2$, and $\mathbf{U}_2^*$ are similarly asymptotically equivalent. Thus the following result also obtains:

Proposition 2. For the population parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$, the estimators $(\hat{\boldsymbol{\beta}}_N^*, \hat{\boldsymbol{\alpha}}_N^*)$ are asymptotically Gaussian with mean $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ and variance $\mathbf{V}_\infty$.

The asymptotic variance $\mathbf{V}_\infty$ is given as the limit $\lim N \times \mathbf{V}_N$ as $N \to \infty$ where

$$\mathbf{V}_N(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\alpha}}^*) = \left(\sum_{i=1}^{N} \mathbf{D}_{1i}^T \mathbf{W}_i \mathbf{D}_{2i} + \boldsymbol{\Lambda}^{-1}\right)^{-1} \times \sum_{i=1}^{N} E[\mathbf{U}_i \mathbf{U}_i^T] \left(\sum_{i=1}^{N} \mathbf{D}_{2i}^T \mathbf{W}_i \mathbf{D}_{1i} + \boldsymbol{\Lambda}^{-1}\right)^{-1},$$

where $\mathbf{D}_{1i}, \mathbf{D}_{2i}$, and $\mathbf{W}_i$ are as defined in Section 3.2 and $\boldsymbol{\Lambda}$ is a block diagonal matrix with $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ on the diagonal. We obtain a consistent estimate of the variance by using the empirical information, $\sum_{i=1}^N \mathbf{U}_i \mathbf{U}_i^T$, in place of $\sum_{i=1}^N E[\mathbf{U}_i \mathbf{U}_i^T]$ (Royall 1986). A sketch of the proof of this proposition can be found in the Appendix.

In practice, specification of the hyperparameters $(\boldsymbol{\mu}_\beta, \boldsymbol{\Lambda}_1)$ and $(\boldsymbol{\mu}_\alpha, \boldsymbol{\Lambda}_2)$ is somewhat arbitrary. In the example that follows we have chosen to set $\boldsymbol{\Lambda}_1^{-1} = 0$, thereby eliminating any shrinkage for the estimation of $\boldsymbol{\beta}$. For estimation of the lorelogram, one choice is to set $\boldsymbol{\Lambda}_2^{-1} = \lambda \mathbf{I}$, where λ is a small constant to mimic the zero cell correction method of adding 1/2. Heuristically, adding 1/2 to the cells of a 2×2 table is like combining the observed data with weak prior data that yield an estimated log odds ratio of 0 with large variance—a prior precision of 1/8 if we use the variance formula $\sum 1/(\text{cell count})$. We see in our example that this approach yields an estimated lorelogram that tracks closely the crude log-odds ratios based on the 1/2 adjustment to each cell in the pairwise tables.

Alternatively, we can consider the mean squared error (MSE) for the estimator $\hat{\boldsymbol{\delta}}_N(\boldsymbol{\Lambda}, \boldsymbol{\mu}_\delta) = (\hat{\boldsymbol{\beta}}^*_N, \hat{\boldsymbol{\alpha}}^*_N)$. Given $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}_\delta = (\boldsymbol{\mu}_\beta, \boldsymbol{\mu}_\alpha)$, the MSE can be approximated as

$$\begin{aligned}\text{MSE}[\hat{\boldsymbol{\delta}}_N(\boldsymbol{\Lambda}, \boldsymbol{\mu}_\delta)] &= E[(\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta})(\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta})^T] \\ &\approx [\mathcal{I}^*_N]^{-1}[\boldsymbol{\Lambda}^{-1}(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta)(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta)^T\boldsymbol{\Lambda}^{-1} + N\boldsymbol{\Sigma}][\mathcal{I}^{*^T}_N]^{-1},\end{aligned}$$

where $\mathcal{I}^*_N = \sum_{i=1}^N \mathbf{D}^T_{1i}\mathbf{W}_i\mathbf{D}_{2i} + \boldsymbol{\Lambda}^{-1}$ and $\boldsymbol{\Sigma} = \lim(1/N) \sum_{i=1}^N E[\mathbf{U}_i\mathbf{U}_i^T]$. Details are presented in the Appendix. If we choose $\boldsymbol{\Lambda}_2^{-1} = \lambda\mathbf{J}$ for a known matrix $\mathbf{J}$, and fix $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\mu}_\delta$ (often at $\mathbf{0}$), then we can minimize the approximate MSE as a function of λ. But evaluating the MSE requires knowledge of both $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$. In practice, we substitute a consistent estimate of $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ obtained from a preliminary regression fit with λ small. Thus we obtain adaptive penalization where the estimator is chosen to minimize the approximate MSE. In general, $\text{MSE}(\hat{\boldsymbol{\delta}}(\lambda))$ is a matrix, so an appropriate summary is minimized such as the maximum eigenvalue that bounds the MSE of all normed linear functions of $\hat{\boldsymbol{\delta}}_N$, for all $\mathbf{a} \in \Re^q, q = \dim(\boldsymbol{\delta})$, $\text{MSE}(\mathbf{a}^T\hat{\boldsymbol{\delta}}_N) \le \lambda^*_{(q)}\|\mathbf{a}\|_2$, where $\lambda^*_{(q)}$ is the maximum eigenvalue of $\text{MSE}(\hat{\delta}_N)$ (Horn and Johnson 1992). More generally, we consider selection of λ to minimize the maximum eigenvalue of $\text{MSE}(\mathbf{B}\hat{\boldsymbol{\delta}}_N(\lambda))$, for a square matrix $\mathbf{B}$, acknowledging that in applications we may wish to choose $\mathbf{B}$ such that $\mathbf{B}\boldsymbol{\delta} = \boldsymbol{\alpha}$.

3.4 Nonparametric Estimation: Generalized Additive Models

We first restrict attention to the nonparametric estimation of a single smooth log-odds ratio function. Our interest is in the function

$$\log(\Psi_{i(j,k)}) = s(z_{ijk}).$$

Our approach to estimation of the lorelogram is to use a local scoring procedure (Hastie and Tibshirani 1990) in conjunction with a smoothing spline with specified degrees of freedom. For the rth iteration, we define the following:

$$\eta^{(r)}_{ijk} = \log(\Psi^{(r-1)}_{i(j,k)}).$$

$$d^{(r)}_{ijk} = \eta^{(r)}_{ijk} + \left(\frac{\partial\eta_{ijk}}{\partial\sigma_{ijk}}\right)^{(r)} \times (S_{ijk} - \sigma_{ijk}(\boldsymbol{\beta}^{(r)}, \Psi^{(r-1)}_{ijk})).$$

$$w^{(r)}_{ijk} = \left(\frac{\partial\eta_{ijk}}{\partial\sigma_{ijk}}\right)^{(r)} V^{-1}(S_{ijk}) \left(\frac{\partial\eta_{ijk}}{\partial\sigma_{ijk}}\right)^{(r)}.$$

The local scoring algorithm alternates between a Newton step for $\boldsymbol{\beta}$ and updating the association function using the linearized data $d^{(r)}$. The algorithm is summarized as follows.

Algorithm.

1. Initialize: $\boldsymbol{\eta}^{(0)}, \boldsymbol{\beta}^{(0)}, r = 1$.
2. Update: Take a Newton step for $\boldsymbol{\beta}$: $\boldsymbol{\beta}^{(r-1)} \to \boldsymbol{\beta}^{(r)}$.
3. Create: Construct $\mathbf{d}^{(r)}$ and $\mathbf{w}^{(r)}$.
4. Update: Fit a smooth of $d^{(r)}_{ijk}$ on z_{ijk} using weights $w^{(r)}_{ijk}$: $\boldsymbol{\Psi}^{(r-1)} \to \boldsymbol{\Psi}^{(r)}$.
5. Iterate: Repeat steps 2–4, $r \leftarrow r + 1$, until the convergence criterion is satisfied, $\|\boldsymbol{\beta}^{(r)} - \boldsymbol{\beta}^{(r-1)}\|_\infty < \varepsilon_1$ and $\|\boldsymbol{\Psi}^{(r)} - \boldsymbol{\Psi}^{(r-1)}\|_\infty < \varepsilon_2$.

Hastie and Tibshirani (1990) stated results regarding the convergence and uniqueness of solutions obtained via the local scoring algorithm for generalized linear models. Our implementation is an extension from linear exponential family models to quadratic exponential family models using a "working independence" weight matrix for the second moment-estimating equation. When more than one term is included in the linear predictor for the log-odds ratio regression, backfitting (Hastie and Tibshirani 1990) can be used to obtain estimates.

Although a kernel smoother or other smoother may be used, we restrict attention to the use of smoothing splines with the degree of smoothing fixed by prior specification of the degrees of freedom. The effective degrees of freedom for a smoothing spline is defined as the trace of the smoother matrix $\mathbf{S}$ such that $\hat{\mathbf{Y}} = \mathbf{SY}$. This allows us to compare the parametric and nonparametric fits more easily.

In the independent data setting pointwise standard errors are obtained using the smoothing matrix $\mathbf{S}$. For example, $\text{var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{SS}^T$ for independent homoscedastic Gaussian responses. For dependent data, the covariance of the vector of responses has a block diagonal structure and additional computational complexity arises (Berhane and Tibshirani 1995). We choose to use the jackknife estimate of variance for linear estimating equations as outlined by Lele (1991). By dropping a single estimating equation, $\mathbf{U}_j$, the resulting pseudo-estimates $(\hat{\boldsymbol{\beta}}^{(-j)}, \hat{\boldsymbol{\alpha}}^{(-j)})$ can be used to obtain a consistent estimate of the variance of the complete sample estimator.

Using the jackknife variance estimator is feasible, because parameter estimation is computationally easy. We alternate between GEE1 estimation of $\boldsymbol{\beta}$ and working independence weighted smoothing of the linearized covariance pairwise products. In situations where N is large, alternative resampling methods such as the drop-m jackknife (Shao and Wu 1989; Wu 1987) or bootstrapping the estimating equations (Moulton and Zeger 1989) may be computationally more economical.

For initial values, we have used $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}$ obtained from a preliminary fit assuming a constant log-odds ratio α and have used $\eta^{(0)}_{ijk} = \hat{\alpha}$, the estimated common log-odds ratio. We have experienced excellent convergence using these starting values.

Finally, we have not considered data-driven methods for the selection of the smoothing parameter used in the nonparametric empirical lorelogram estimation. We assume throughout that this parameter has been fixed. A sensitivity analysis can be performed to investigate the influence of this parameter. Further research into objective criterion for the selection of the degree of smoothness is warranted.

4. EXAMPLE

In this section we use the lorelogram methodology to address scientific interest in the dependence structure of repeated measures of disease activity. The data are monthly symptom measures for 90 first-episode schizophrenia patients from Madras India (Thara, Henrietta, Joseph, Rajkumar, and Eaton 1994). After hospitalization each patient had the presence or absence of six symptoms recorded monthly. The six symptoms are grouped into two categories: "positive symptoms" include hallucinations, delusions, and thought disorders; "negative symptoms" include flat affect, apathy, and withdrawal. Investigators suspect that schizophrenia is a heterogeneous disorder and that positive symptoms reflect a temporary disruption in the patient's mental health and thus reflect a "state," whereas negative symptoms reflect a patient characteristic or "trait," potentially due to structural differences in the brain. One primary goal of the data analysis is to characterize the longitudinal binary series and summarize the evidence supporting the state–trait hypothesis.

We analyze the first 12 months of data on each of the six symptoms separately. In addition to time (month), the patient's age and gender were recorded. Some observations are missing, and the validity of the estimating equation approach depends on the assumption that the data are missing completely at random.

Table 1. Parametric Lorelogram Models Fit to the Six Schizophrenia Symptoms

	Estimates of positive symptoms		
Covariate	Hallucinations	Delusions	Thoughts
	Mean regression		
Intercept	1.106 (.362)	1.579 (.324)	1.012 (.369)
Time ($k1$)	−2.399 (.428)	−3.358 (.423)	−2.976 (.419)
Time ($k2$)	−6.113 (.679)	−4.083 (.568)	−3.000 (.423)
Time ($k3$)	−2.091 (.352)	−2.589 (.346)	−2.150 (.358)
Age	.651 (.299)	.378 (.256)	−.184 (.306)
Gender	−.058 (.305)	−.594 (.254)	−.634 (.290)
	Association regression		
Intercept	3.640 (.359)	3.030 (.248)	4.422 (.382)
Δt ($k1$)	−1.696 (.469)	−1.814 (.414)	−3.046 (.483)
Δt ($k2$)	−2.638 (.707)	−2.776 (.571)	−3.995 (.563)
Δt ($k3$)	−4.508 (.961)	−5.259 (.500)	−6.278 (.844)
Δt ($k4$)	−2.646 (.962)	−2.414 (.671)	−1.315 (.774)
	Estimates of negative symptoms		
Covariate	Flat affect	Apathy	Withdrawal
	Mean regression		
Intercept	−.389 (.368)	−.739 (.411)	−.408 (.374)
Time ($k1$)	−1.665 (.376)	−1.362 (.474)	−1.434 (.409)
Time ($k2$)	−2.910 (.448)	−3.370 (.523)	−3.186 (.494)
Time ($k3$)	−1.546 (.304)	−1.547 (.364)	−1.045 (.326)
Age	.642 (.338)	.595 (.390)	.442 (.352)
Gender	.137 (.349)	.178 (.365)	−.112 (.353)
	Association regression		
Intercept	3.765 (.396)	3.219 (.418)	3.391 (.345)
Δt ($k1$)	−.934 (.583)	−1.448 (.586)	−.961 (.746)
Δt ($k2$)	−1.245 (.904)	−1.335 (.697)	−1.986 (.655)
Δt ($k3$)	−3.699 (.904)	−.836 (1.095)	−3.043 (1.036)
Δt ($k4$)	−.872 (.742)	1.828 (1.070)	.619 (1.302)

NOTE: The mean model is a marginal logistic regression using the estimated lorelogram for covariance weighting. Values within parentheses represent standard error.

Because the patients were enrolled after a psychotic episode that leads to hospitalization, the frequency of all symptoms was much higher initially and declined over the first year. Positive symptoms occurred in approximately 70% of patients in the first month and declined to a prevalence of 20% by the twelfth month. Similarly, negative symptoms had an initial prevalence of 40% that declined to 10% at the end of the year.

For each symptom ($s = 1, 2, 3, 4, 5, 6$), we use a marginal logistic regression

$$\text{logit}(P[Y_{it}^s = 1]) = \beta_0^s + \sum_{r=1}^{3} \beta_r^s t_r^* + \beta_5^s I(\text{age}_i < 21) + \beta_6^s I(\text{female}_i).$$

Here $Y_{it}^s = 1$ if patient i has symptom s at time t and 0 otherwise. The covariates t_r^* are natural spline basis elements using knots at 3 and 7 months, allowing the prevalence to change smoothly over time. Inclusion of the other covariates allows adjustment of the marginal mean for other potentially prognostic factors.

The state–trait hypothesis is not addressed by the marginal mean but rather by the nature of the dependence structure. Symptoms that represent states should exhibit strong temporal association that declines as observation times separate. Alternatively, symptoms that represent traits are expected to exhibit long-range dependence indicating within-subject association independent of the time lag in observations. We use the estimated lorelogram to characterize the strength of association as a function of the time lag, $|t_{ij} - t_{ik}|$, to assess the state–trait hypothesis.

4.1 Schizophrenia Data—Spline Regression

As outlined in Section 3.2, a flexible parametric approach proceeds by constructing a spline basis matrix in the covariate of interest and then performing a parametric regression fit. We used the covariate $z_{ijk} = |t_{ij} - t_{ik}|$ to generate a natural spline basis given knots at $\Delta t = (2, 5, 8)$ months. The log-odds ratio regression thus is given by

$$\log \Psi(Y_{ij}^s, Y_{ik}^s) = \mathbf{Z}_{ijk}^* \boldsymbol{\alpha}^s,$$

where $\mathbf{Z}_{ijk}^* = [1, z_{ijk}^{(1)}, \ldots, z_{ijk}^{(4)}]$, the basis elements evaluated for the pair of times t_{ij} and t_{ik}. Again we assume a separate function for each of the six symptoms.

Table 1 gives the mean and association parameter estimates for each of the six symptoms. Figure 2 shows the fitted lorelograms for the six symptoms with asymptotic pointwise 95% standard error bands. It also shows the crude pairwise log-odds ratios obtained from creating 2×2 tables for each of the $\binom{12}{2}$ combinations of observation times with each point's asymptotic 95% confidence limits. Note that for all three positive symptoms, the estimated lorelogram decays to 0, supporting the hypothesis that these symptoms represent "states." However, none of the lorelograms for the negative symptoms decays to 0, nor is 0 included in any pointwise 95% confidence band. These functions support the "trait" hypothesis for the negative symptoms, because they exhibit strong long-range dependence.

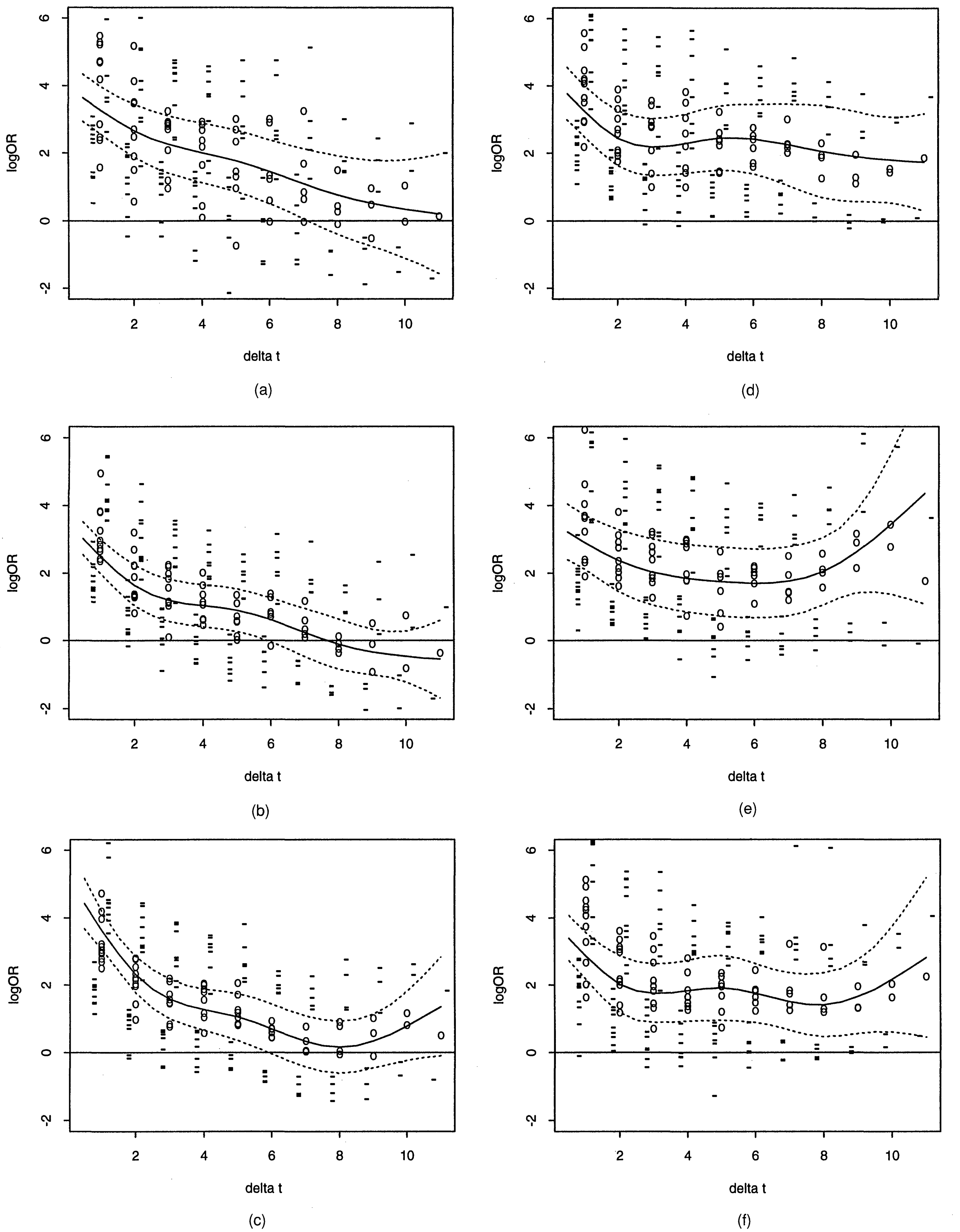

Figure 2. Parametric Lorelograms Fit to the Six Schizophrenia Symptoms Using a Natural Spline Basis With Knots at Δt = 3, 5, 7 Months. (a) Hallucinations; (b) delusions; (c) thoughts; (d) flat affect; (e) apathy; (f) withdrawal The dashed lines are pointwise 95% confidence bands. Also shown are the crude pairwise log-odds ratios and corresponding asymptotic 95% confidence limits. ○, the crude zero cell corrected log-odds ratio computed from a 2 × 2 table of Y_{ij} versus Y_{ik}; −, the confidence limits for these point estimates.

Table 2. Maximum Eigenvalue for the Mean Squared Error Matrix as a Function of λ, Where $\boldsymbol{\Lambda}_1^{-1} = \mathbf{0}$, and $\boldsymbol{\Lambda}_2^{-1} = \lambda\mathbf{J}$ Such That $\mathbf{J} = \mathbf{I}_{5\times 5}$ With (1, 1) Entry Set to 0

λ	Hallucinations	Delusions	Thoughts	Flat affect	Apathy	Withdrawal
0	2.116	.637	1.762	1.538	2.719	2.559
.05	1.849	.588	.751*	1.355	1.328	2.149
.10	1.758*	.577*	1.128	1.269	.851	1.844
.15	1.797	.607	2.267	1.254*	.663	1.616
.20	1.943	.682	3.785	1.293	.609*	1.451
.25	2.179	.804	5.473	1.379	.616	1.343
.30	2.487	.967	7.214	1.503	.648	1.296*
.35	2.850	1.164	8.944	1.658	.693	1.314
.40	3.251	1.388	10.626	1.834	.746	1.386
.45	3.679	1.635	12.241	2.027	.802	1.496
.50	4.125	1.900	13.780	2.230	.861	1.627

* Minimum value.

4.2 Schizophrenia Data—Regularization Estimator

One feature of Figure 2 is the wide confidence bands for $\Delta t = 11$ for both apathy and withdrawal. This is due to the sparsity of discordant pairs at the long time lags for these symptoms. Thus we consider the zero cell correction methodology of Section 3.3 to stabilize these estimators. We choose $\boldsymbol{\Lambda}_1^{-1} = \mathbf{0}$, yielding no shrinkage of $\hat{\boldsymbol{\beta}}$, and investigate $\boldsymbol{\Lambda}_2^{-1} = \lambda\mathbf{J}$. For simplicity, and because we used approximately orthonormal basis elements, we let $\mathbf{J}$ be the identity matrix with the upper left diagonal element set to 0. Using this $\mathbf{J}$ results in $(\hat{\alpha}_2, \ldots, \hat{\alpha}_5)$ shrinking to 0 as $\lambda \to \infty$. Maximal λ reduces the association model to an intercept model, or an exchangeable association model.

We then considered the approximate MSE of the estimator $\hat{\boldsymbol{\alpha}}(\lambda)$ as described in Section 3.3 using an estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ obtained from the regression fit with $\lambda = 0$. Because the $\text{MSE}(\hat{\boldsymbol{\alpha}}_N(\lambda))$ is a matrix, we selected the maximum eigenvalue as the quantity to minimize. Table 2 displays selected values of λ and the corresponding MSE for each of the six symptoms. Note that shrinkage toward the exchangeable model does not provide a substantial reduction in MSE for the first two positive symptoms. But $\lambda = .05$ does reduce the MSE for thoughts by $(1.76 - .75)/1.76 = 57\%$, whereas for negative symptoms $\lambda = .20$ yields a substantial reduction in the MSE for apathy, $(2.72 - .61)/2.72 = 67\%$, and $\lambda = .30$ reduces the MSE for withdrawal by 49%. Figure 3 displays the lorelograms obtained using the optimal λ for apathy and withdrawal. Note that the point estimates remain virtually unaffected for Δt less than 6 but are stabilized for the longer lags, as evidenced by both the reduction in the confidence bands for both symptoms and the dampened trajectory for apathy.

4.3 Schizophrenia Data—Nonparametric Estimator

We also fit nonparametric lorelograms to the six symptoms. This estimator is desirable, because we do not need to select the number or position of knots as in the spline estimator. We used the algorithm described in Section 3.4 using smoothing splines with the degrees of freedom fixed at 5. Figure 4 shows the fitted nonparametric functions and pointwise 95% confidence bands obtained from jackknifing the estimating equations (Lele 1991). These functions are

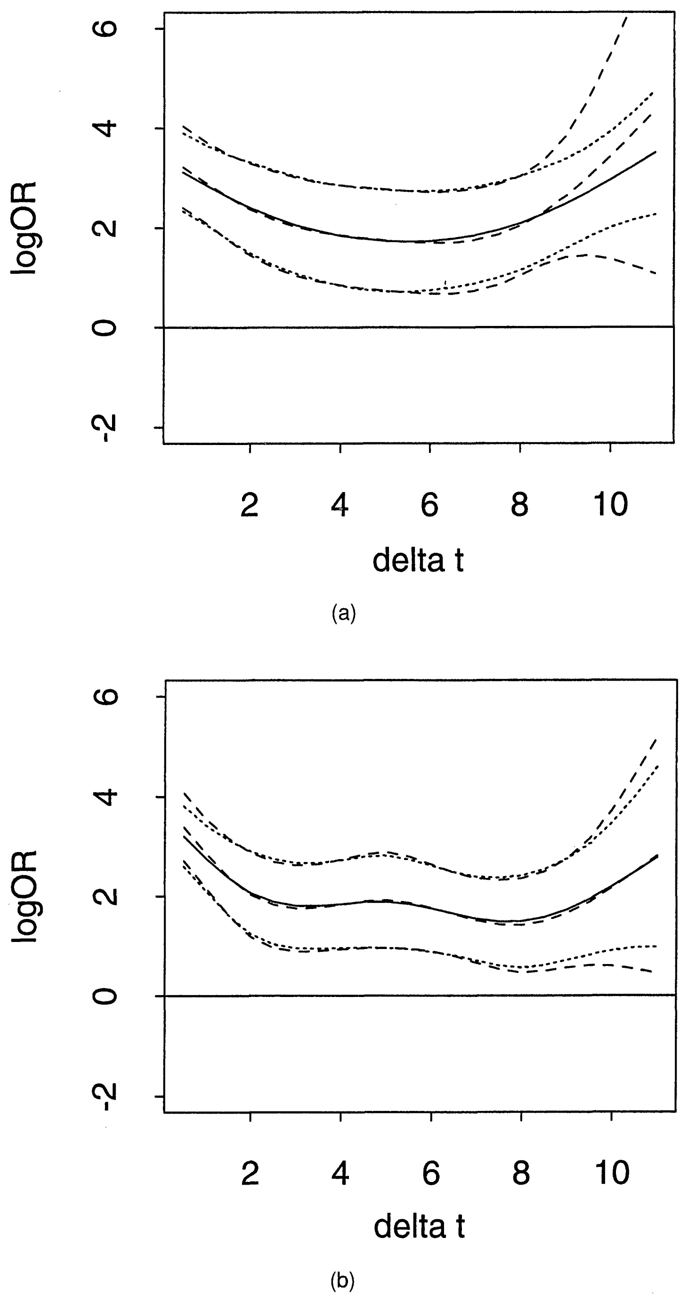

Figure 3. Regularization Estimator for (a) Apathy (λ = .20) and (b) Withdrawal (λ = .30), with Pointwise 95% Confidence Limits Shown Over the Crude Log-Odds Ratios. Also displayed (– – –) is the point estimate and confidence bands obtained using λ = 0.

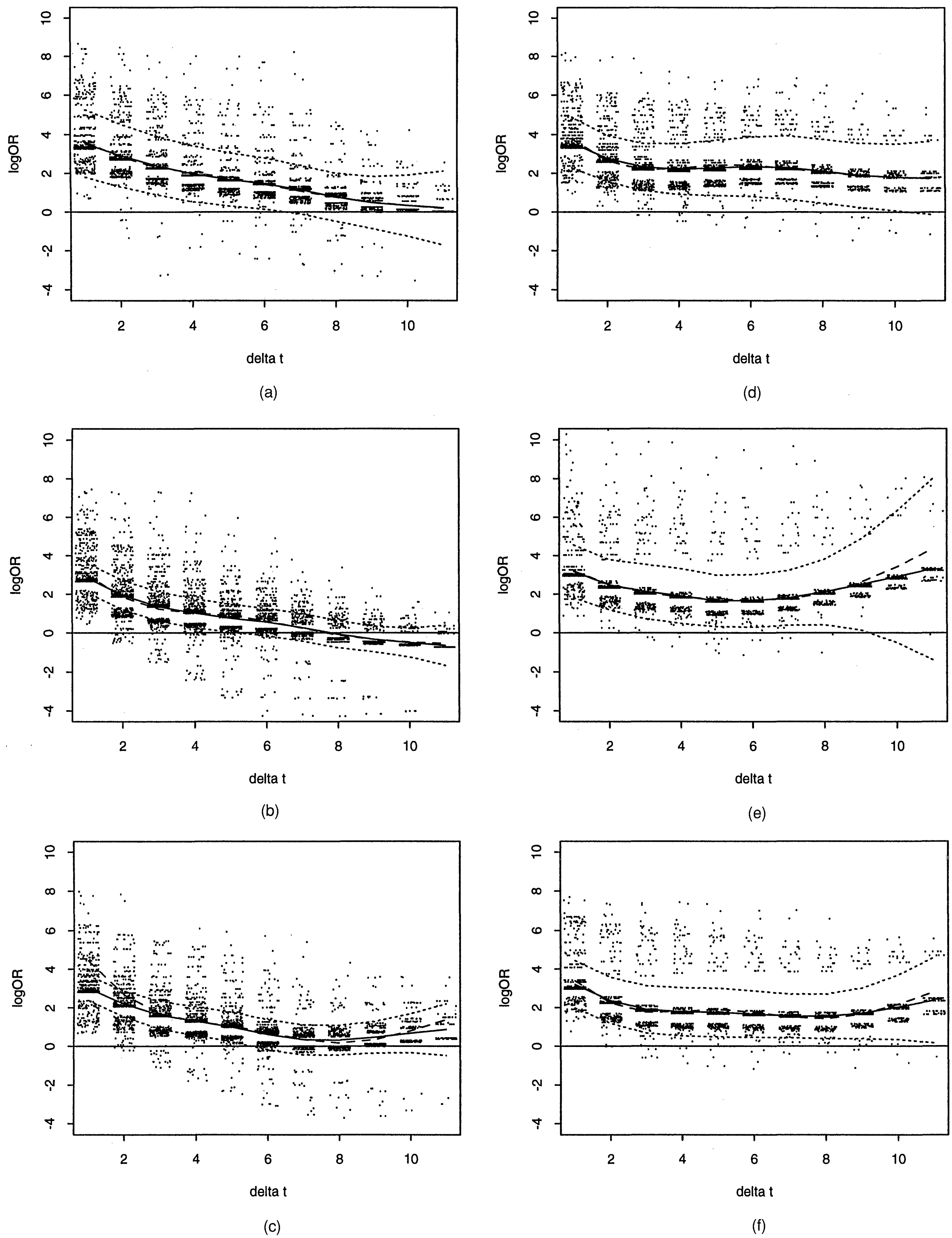

Figure 4. Nonparametric Lorelograms Fit to the Six Schizophrenia Symptoms Using a Smoothing Spline with df = 5. (a) Hallucinations; (b) delusions; (c) thoughts; (d) flat affect; (e) apathy; (f) withdrawal. Pointwise 95% jackknife based confidence bands are shown. The standardized adjusted covariance residuals, $\eta_{ijk} + \sqrt{w_{ijk}}(d_{ijk} - \eta_{ijk})$, *are also plotted (with* Δt *jittered). (– – –), the point estimate obtained using natural splines.*

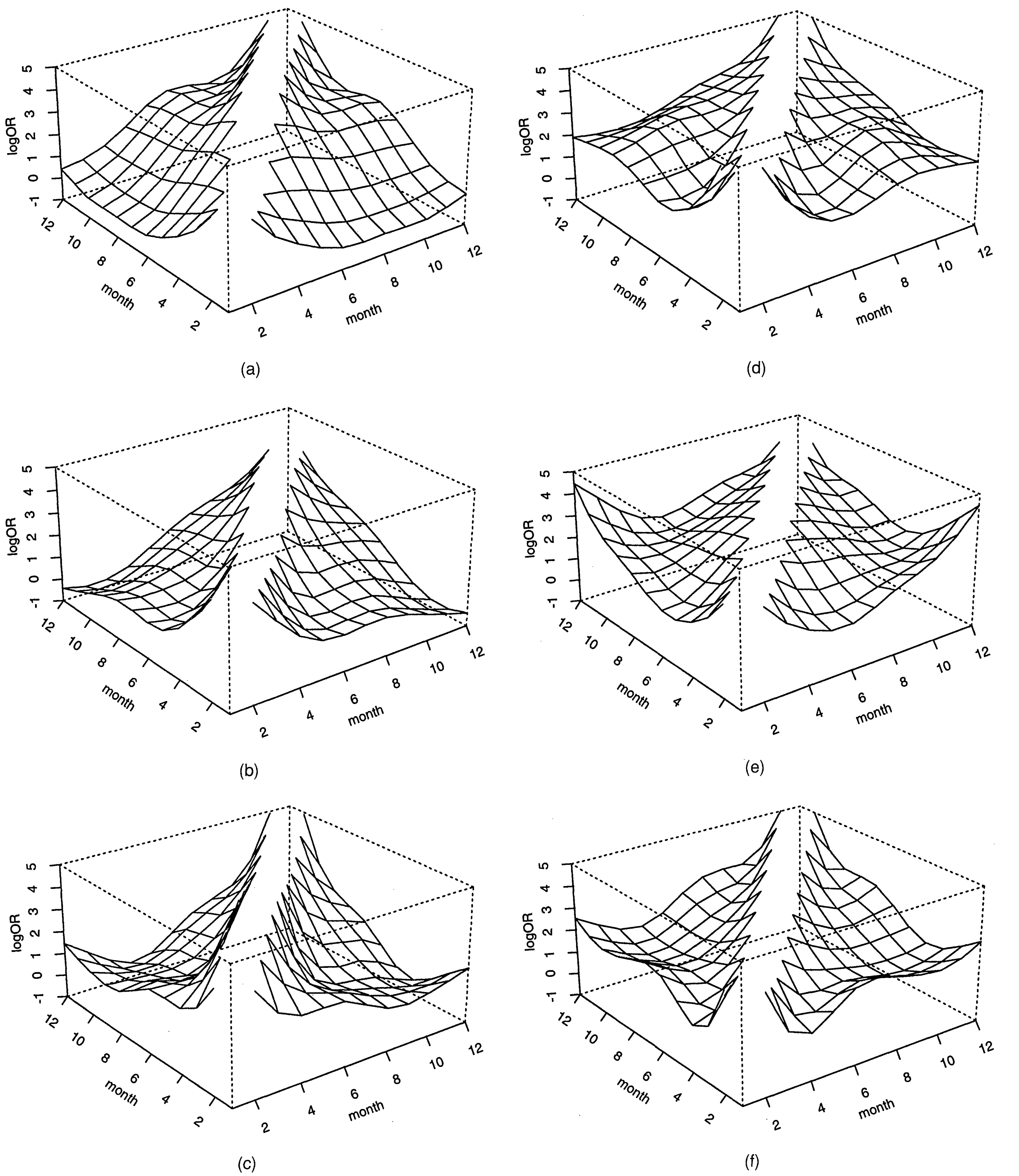

Figure 5. Nonisotropic Lorelograms Fit to the Six Schizophrenia Symptoms Using Natural Splines, Working Independence for V_{11i}, and λ = .05. (a) Hallucinations; (b) delusions; (c) thoughts; (d) flat affect; (e) apathy; (f) withdrawal.

quite similar to the parametric fits, with the exception of slightly wider confidence bands. Interestingly, the nonparametric fits appear to have dampened trajectories for $\Delta t \geq 9$ for thoughts, apathy, and withdrawal, those symptoms for which the regularization estimator yields the greatest reduction in MSE. Again, the fitted lorelograms clearly support the state–trait hypothesis.

4.4 Schizophrenia Data—Nonisotropic Model

Finally, Figure 5 displays nonisotropic association models fit to the six symptoms. We used a natural spline basis in $|t_{ij} - t_{ik}|$ with knots at 2, 5, and 8 months; a natural spline basis in $(t_{ij} + t_{ik})$ with knots at 7, 12, and 17 months; and their tensor product. We selected these bases for the bivariate surface because the log-odds ratio is symmetric in the two times. These association surfaces suggest that the isotropy assumption is plausible for these binary series. The greatest violation of isotropy appears for hallucinations, where a trend of increasing association with later times is evident.

5. DISCUSSION

We have proposed methods for the graphical exploration of the dependence structure in categorical longitudinal data. Two regression models are used: a generalized linear model for the marginal mean; and a log linear model for the pairwise log-odds ratio. A parametric estimator for the association function uses a flexible cubic spline basis with natural boundary constraints, whereas a nonparametric estimator extends generalized additive models to the estimation of covariance components. Finally, we have also introduced a shrinkage parameter for parametric models that stabilizes sparse or zero cells in the association regression.

The flexible association regression can be used to characterize components of covariance due to serial dependence or cluster heterogeneity. For the schizophrenia data, we used the fitted lorelograms to contrast the chronicity of two classes of symptoms.

Although we have focused specifically on models for longitudinal data, the methods that we have developed are potentially useful for general dependent data structures. For example, given spatial categorical responses, Y_i measured at locations s_i, we may use the model $\log \Psi(Y_i, Y_j) = f(\|s_i - s_j\|)$ to characterize isotropic dependence as measured on the log-odds scale. In addition, random-effects models assume that the observed dependence is due to shared latent effects. Using the empirical lorelogram with covariates to be modeled as random effects permits direct characterization of the marginal pairwise dependence and assessment of the parametric covariance assumptions.

Finally, an open problem is the selection of the smoothing parameter for the nonparametric lorelogram estimator. Cross-validation of the covariance pairwise products by dropping a cluster at a time (see Rice and Silverman 1991 and Zeger and Diggle 1994 for mean function cross-validation) may provide data-driven methods for the selection of the penalty parameter and warrant exploration.

APPENDIX: PENALIZED ESTIMATING FUNCTIONS

Define $(\hat{\beta}_N^*, \hat{\alpha}_N^*)$ as the solution to the estimating equations

$$\mathbf{0} = \mathbf{U}_1^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \mu_i}{\partial \beta}\right]^T \mathbf{V}_{11i}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) - \boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)$$

and

$$\mathbf{0} = \mathbf{U}_2^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^T \mathbf{V}_{22i}^{-1}(\mathbf{S}_i - \boldsymbol{\sigma}_i) - \boldsymbol{\Lambda}_2^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha),$$

where $(\mu_\beta, \boldsymbol{\Lambda}_1)$ and $(\boldsymbol{\mu}_\alpha, \boldsymbol{\Lambda}_2)$ are fixed and $\mathbf{Y}_i = \text{vec}(Y_{ij})$, $E[\mathbf{Y}_i] = \boldsymbol{\mu}_i, \mathbf{S}_i = \text{vec}((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})), \mathbf{V}_{11i} = \text{cov}(\mathbf{Y}_i)$, and $\mathbf{V}_{22i} \approx \text{cov}(\mathbf{S}_i)$.

For ease of notation, we combine the two functions to form a single function:

$$\mathbf{U}_N^*(\boldsymbol{\delta}) = \sum_{i=1}^{N} \mathbf{D}_{1i}^T \mathbf{W}_i \mathbf{R}_i - \boldsymbol{\Lambda}^{-1}(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta),$$

where $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)$, $\boldsymbol{\mu}_\delta^T = (\boldsymbol{\mu}_\beta^T, \boldsymbol{\mu}_\alpha^T)$, $\mathbf{D}_{1i}$ is a block diagonal matrix with $[\partial\mu_i/\partial\beta]$, and $[\partial\sigma_i/\partial\alpha]$ on the diagonal, similarly $\mathbf{W}_i$ has $\mathbf{V}_{11i}$ and $\mathbf{V}_{22i}$ on the diagonal, $\mathbf{R}_i^T = ((\mathbf{Y}_i - \boldsymbol{\mu}_i)^T, (\mathbf{S}_i - \boldsymbol{\sigma}_i)^T)$, and $\boldsymbol{\Lambda}$ is block diagonal with $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ on the diagonal. Define $\mathbf{D}_{2i} = \{[\partial(\mu_i, \sigma_i)]/[\partial(\beta, \alpha)]\}$.

The estimator $\hat{\boldsymbol{\delta}}_N$ is defined by

$$\mathbf{0} = \mathbf{U}_N^*(\hat{\boldsymbol{\delta}}_N).$$

Using a multivariate Taylor series expansion, we obtain

$$\mathbf{0} = \mathbf{U}_N^*(\boldsymbol{\delta}) - [\mathcal{I}_N^*](\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) + o_p(\|\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta}\|^2),$$

where $\mathcal{I}_N^* = -E[(\partial/\partial\delta)\mathbf{U}_N^*] = \sum_{i=1}^N \mathbf{D}_{1i}^T \mathbf{W}_i \mathbf{D}_{2i} + \boldsymbol{\Lambda}^{-1}$. Under mild regularity conditions, we can apply the central limit theorem for independent but nonidentically distributed random vectors (Serfling 1980, p. 30):

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{D}_{1i}^T \mathbf{W}_i \mathbf{R}_i \to \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{\Sigma} = \lim \frac{1}{N} \sum_{i=1}^{N} \mathbf{D}_{1i}^T \mathbf{W}_i \text{cov}(\mathbf{R}_i) \mathbf{W}_i \mathbf{D}_{1i}.$$

This result, together with the Taylor series approximation, yields the large-sample approximation

$$(\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) \mathrel{\dot\sim} \text{MVN}(-[\mathcal{I}_N^*]^{-1}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta), [\mathcal{I}_N^*]^{-1}[N\boldsymbol{\Sigma}][\mathcal{I}_N^{*T}]^{-1}).$$

Note that under regularity conditions, $N\mathcal{I}_{\mathbf{N}}^*$ converges to a positive definite matrix, implying that the bias of the estimator $\hat{\boldsymbol{\delta}}_N$ is $O(1/N)$.

Therefore, the MSE of $\hat{\delta}_N$ can be approximated by

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\delta}}_N) &= E[(\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta})(\hat{\boldsymbol{\delta}}_N - \boldsymbol{\delta})^T] \\ &\approx [\mathcal{I}_N^*]^{-1}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta)(\boldsymbol{\delta} - \boldsymbol{\mu}_\delta)^T\boldsymbol{\Lambda}^{-1}[\mathcal{I}_N^{*T}]^{-1} \\ &\quad + [\mathcal{I}_N^*]^{-1}[N\boldsymbol{\Sigma}][\mathcal{I}_N^{*T}]^{-1}. \end{aligned}$$

[Received November 1995. Revised August 1996.]

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.
Berhane, K., and Tibshirani, R. (1995), "Generalized Additive Models for Longitudinal Data," technical report, University of Toronto, Dept. of Preventive Medicine and Biostatistics.
Cinlar, E. (1975), *Introduction to Stochastic Processes*, Englewood Cliffs, NJ: Prentice-Hall.
Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.
Crowder, M. J. (1986), "On Consistency and Inconsistency of Estimating Equations," *Econometric Theory*, 2, 305–330.
Dale, J. R. (1986), "Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses," *Biometrics*, 42, 909–917.
Diggle, P. J. (1988), "An Approach to the Analysis of Repeated Measures," *Biometrics*, 44, 959–971.
——— (1990), *Time Series: A Biostatistical Introduction*, New York: Oxford University Press.
Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, New York: Oxford University Press.

Efron, B., and Stein, C. (1981), "The Jackknife Estimate of Variance," *The Annals of Statistics*, 9, 586–596.

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

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993), "Regression Models for Discrete Longitudinal Responses," *Statistical Science*, 8, 284–309.

Goodman, L. A. (1970), "The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications," *Journal of the American Statistical Association*, 65, 226–256.

Green, P. (1987), "Penalized Likelihood for General Semi-Parametric Regression Models," *International Statistical Reviews*, 55, 245–260.

Hastie, T. R., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Heagerty, P. J. (1995), "Multivariate Multinomial Marginal Models," unpublished Ph.D. thesis, Johns Hopkins University, Dept. of Biostatistics.

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

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Non-Orthogonal Problems," *Technometrics*, 12, 55–67.

Horn, R. A., and Johnson, C. R. (1992), *Matrix Analysis*, New York: Cambridge University Press.

Keenan, B. (1982), "A Time Series Analysis of Binary Data," *Journal of the American Statistical Society*, 77, 816–821.

Laird, N. M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables," *Biometrika*, 65, 581–590.

Lele, S. (1991), "Jackknifing Linear Estimating Equations: Asymptotic Theory and Applications in Stochastic Processes," *Journal of the Royal Statistical Society*, Ser. B, 53, 253–267.

Leonard, T. (1975), "Bayesian Estimation Methods for Two-Way Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 37, 23–37.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analyses for Categorical Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 3–40.

Lindsay, B. G. (1985), "Using Empirical Bayes Inference for Increased Efficiency," *The Annals of Statistics*, 13, 914–931.

Lipsitz, 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," *Biometrika*, 78, 153–160.

Mardia, K. V. (1967), "Some Contributions to the Contingency-Type Bivariate Distribution," *Biometrika*, 54, 235–249.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Moulton, L., and Zeger, S. L. (1989), "Analyzing Repeated Measures on Generalized Linear Models via the Bootstrap," *Biometrics*, 45, 381–394.

O'Sullivan, F. (1986), "A Statistical Perspective on Ill-Posed Inverse Problems," *Statistical Science*, 4, 502–527.

Pepe, M. S., and Anderson, G. L. (1994), "A Cautionary Note on Inference for Marginal Regression Models with Longitudinal Data and General Correlated Response Data," *Communications in Statistics*, 23, 939–951.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Response," *Biometrics*, 44, 1033–1048.

Prentice, R. L., and Zhao, L. P. (1991), "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses," *Biometrics*, 47, 825–839.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When Data are Curves," *Journal of the Royal Statistical Society*, Ser. B, 53, 233–234.

Royall, R. M. (1986), "Model Robust Confidence Intervals Using Maximum Likelihood Estimators," *International Statistical Review*, 54, 221–226.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Shao, J., and Wu, C. F. J. (1989), "A General Theory for Jackknife Variance Estimation," *The Annals of Statistics*, 17, 1176–1197.

Thara, R., Henrietta, M., Joseph, A., Rajkumar, S., and Eaton, W. (1994), "Ten-Year Course of Schizophrenia—The Madras Longitudinal Study," *Acta Psychiatrica Scandinavica*, 90, 329–336.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–24.

Wu, C. F. J. (1987), "On the Asymptotic Properties of the Jackknife Histogram," Technical Report #817, University of Wisconsin-Madison, Dept. of Statistics.

Zeger, S. L., and Diggle, P. J. (1994), "Semi-Parametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. A. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060.

Zhao, L. P., and Prentice, R. L. (1990), "Correlated Binary Regression Using a Quadratic Exponential Model," *Biometrika*, 77, 642–648.