Statistica Sinica **14**(2004), 695-711

SELF-MODELING REGRESSION FOR MULTIVARIATE CURVE DATA

Brent A. Coull and John Staudenmayer

Harvard University and University of Massachusetts

Abstract: We present self-modeling regression models for flexible nonparametric modeling of multiple outcomes measured longitudinally. Based on penalized regression splines, the models borrow strength across multiple outcomes by specifying a global time profile, thereby yielding a means of dimension reduction and estimates of trend more precise than those based on univariate regressions. The proposed models represent nonparametric regression extensions to existing factor analytic models for a multivariate response recorded at a single timepoint, and are easily generalized to incorporate serial correlation above that captured by nonlinear effects over time. We illustrate the methods by applying them to data on the respiratory effects of residual oil fly ash inhalation in humans.

Key words and phrases: Correlated curves, multiple outcomes, nonparametric regression, penalized regression spline, respiratory health, time series.

1. Introduction

Investigations of the effects of environmental exposures on human health often record measurements on multiple outcomes with interest focusing on the overall effect of the exposure on the set of outcomes. When each outcome is measured once, it is well-known that multivariate models that account for correlation among outcomes hold several advantages over the simpler approach that fits a univariate model to each outcome. For instance, multivariate models allow one to reduce the effective dimension of the multivariate response, which in turn yields greater power to detect effects that are consistent across outcomes (Lefkopoulou and Ryan (1993) and Pocock (1997)).

In many instances, researchers monitor multiple outcomes over time, resulting in data taking the form of multivariate time profiles, or multivariate curves. A simple approach to analyzing such data is to fit a separate nonparametric smoother to data on each outcome. As in the univariate multiple outcome setting, disadvantages of this approach include loss of power from ignoring the multivariate structure of the data (Lefkopoulou and Ryan (1993) and Pocock (1997)) and difficulty in synthesizing a common effect. Alternatively, one could

fit a parametric multiple outcome model, but this approach can be susceptible to misspecification.

Most existing work on smoothing correlated data has focused on the longitudinal data setting (e.g., Müller (1988), Altman and Casella (1995), Staniswalis and Lee (1998), Zeger and Diggle (1994), Wang (1998), Zhang, Lin, Raz and Sowers (1998), Lin and Zhang (1999), Zhang (1999), Lin and Carroll (2001) and Coull, Schwartz and Wand (2001)), with different approaches relating to different ways of accounting for correlation among longitudinal observations measured on the same subject. Zeger and Diggle (1994) used semiparametric regression methods that estimate a common nonlinear trend over time in the presence of serial correlation. Altman and Casella (1995), Staniswalis and Lee (1998) and Zhang (1999) took nonparametric approaches to growth curve analysis. Others have formulated generalized additive mixed models (e.g., Zhang et al. (1998), Lin and Zhang (1999) and Coull, Schwartz and Wand (2001)) in which correlation among repeated measures is modeled using random effects.

In this article, we propose self-modeling regression (SEMOR, Lawton and Sylvestre (1971), Lawton, Sylvestre and Maggio (1972), Kneip and Gasser (1988), Kneip and Engel (1995), Lindstrom (1995), Wang and Brown (1996), Ladd and Lindstrom (2000) and Ke and Wang (2001)) as a latent curve formulation for multivariate curve data. The model combines the strengths of simpler approaches in that it provides flexibility in modeling the time profiles while pooling information on exposure across outcomes. Existing applications of SEMOR models to multiple outcomes (1) specify that the outcome-specific curves from multiple subjects depend on a latent curve common to all subjects, and (2) account for correlation among outcomes by specifying a general covariance structure for the residuals (Wang, Guo and Brown (2000)). In contrast, our chosen SEMOR formulation models correlation among multivariate curves by specifying a latent function common to curves arising from the same subject.

We propose estimating the form of the latent curves using penalized regression splines (e.g., Eilers and Marx (1996)). This approach results in a mixed model representation of SEMOR (Altman (2001) and Altman and Villarreal (2001)), which allows one to use standard likelihood-based methods for both estimation and inference as well as automatic smoothing parameter selection. A simulation study presented in Section 5 shows that this likelihood-based approach can yield large efficiency gains over existing estimation methods for SEMOR models. Our formulation also shows that SEMOR models generalize factor-analytic models for outcomes measured at a single timepoint (Sammel and Ryan (1996)) to the functional data setting.

We use the proposed models to analyze data from an experiment that studies the respiratory effects of residual oil fly ash (ROFA), a surrogate for ambient

air pollution, inhalation in humans. Epidemiological studies have repeatedly shown associations between air particulate matter and increased morbidity and mortality in human populations, particularly subjects with pre-existing respiratory or cardiac vulnerability (e.g., Dockery, Pope, Xu, Spengler, Ware, Fay, Ferris and Speizer (1993) and Samet, Dominici, Curriero, Coursac and Zeger (2000)). Current laboratory research focuses on the physiological mechanisms behind these effects by subjecting test subjects to air pollution. In the study motivating this research, investigators recorded 16 respiratory outcomes on eight human subjects during one hour exposures, with each subject monitored during both filtered air and ROFA exposures. The top row of Figure 1 shows minutely averages of two components of respiratory frequency, log(time to inspiration), denoted here as TI, and log(time to exhalation), denoted TE, for one subject under filtered air conditions. The y-axis represents an adjusted response formed by centering minutely averages of each outcome around a mean value observed during a pre-conditioning phase conducted before the start of exposure. Fitted values and pointwise 95% confidence bands from scatterplot smoothing (as discussed in Section 2) show that, as expected, these outcomes are highly correlated at any given time point. The bottom row of Figure 1 shows the analogous plot for the same subject during ROFA exposure. The outcomes are again highly correlated, yet note that the observed relation is in the opposite direction. Two goals of this study were to investigate the effect of ROFA exposure on (1) trends in respiration times over the one hour exposure period, and (2) the relationship between the inhalation and exhalation components of respiration.

The remainder of this article is organized as follows. Section 2 briefly reviews penalized regression splines for a single curve. Section 3 presents SEMOR models for multivariate curve data. Section 4 outlines model fitting and inference for the penalized spline formulation of SEMOR, and is followed by a small simulation study in Section 5. Sections 6 and 7 contain an analysis of the human respiratory data and further discussion, respectively.

2. Penalized Regression Splines for Nonparametric Regression

Consider the case in which repeated measures are recorded on a single outcome for a single individual. Let y_j be the response at time t_j, $j = 1, \ldots, T$. A simple nonparametric model that specifies an arbitrary, smooth time profile for y is

$$y_j = f(t_j) + \varepsilon_j, \tag{1}$$

where the $\{\varepsilon_j\}$ are normally distributed with error variance σ_ε^2. We first consider the case in which these errors are independent, but discuss serial correlation momentarily.

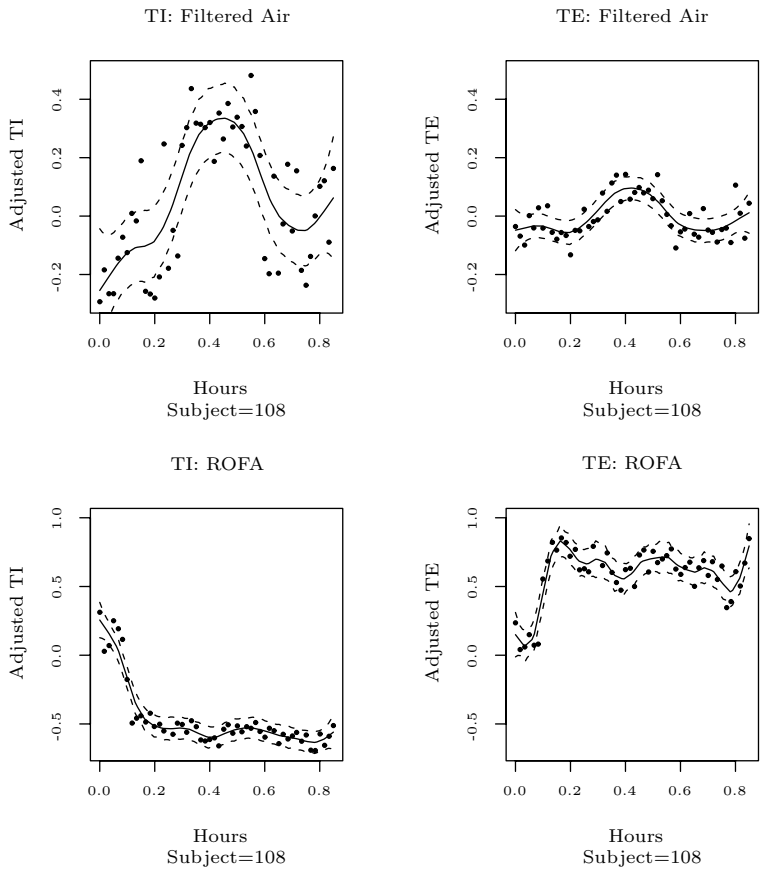

Figure 1. Data from a representative subject in the human respiratory study under two experimental conditions, univariate penalized spline fits, and pointwise confidence intervals. The top row displays log(time to inspiration) (TI) and log(time to expiration) (TE) when exposed to filtered air. The bottom row displays the same responses when the subject is exposed to residual oil fly ash (ROFA).

Let $\kappa_1, \ldots, \kappa_K$ be a set of distinct numbers, or *knots*, inside the range of the time points $1, \ldots, T$, and let $x_+ = \max(0, x)$. As we discuss below, the model adaptively smoothes by penalizing a least squares fit on a vector space that has an excess of basis elements. As a result, the precise locations of the knots does not matter as long as they are relatively "dense" (e.g., one knot for every 3-4 unique covariate values) among the covariate observations in order to allow f to have enough curvature (Ruppert (2002)). In our application and simulations, we use twenty-five equally spaced knots over the range of the covariates. The fits did not appreciably change when the number of knots was changed by plus or minus ten. We have had similar experiences in other projects (Coull, Schwartz

and Wand (2001)), but the number and location of the knots could matter if there are gaps in the distribution of the covariate.

A linear mixed model formulation of a penalized spline model (Eilers and Marx (1996) and Brumback, Ruppert and Wand (1999)) for (1) is

$$y_j = \beta_0 + \beta_1 t_j + \sum_{k=1}^{K} u_k (t_j - \kappa_k)_+ + \varepsilon_j, \tag{2}$$

where $u_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2)$. The mixed model fit is the solution to a ridge regression (Ruppert, Wand and Carroll (2003)), with $\sigma_\varepsilon^2/\sigma_u^2$ acting as the smoothing parameter. For $\sigma_\varepsilon^2/\sigma_u^2$ close to zero, $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated by least squares, and larger $\sigma_\varepsilon^2/\sigma_u^2$ causes $\mathbf{u}$ to adaptively shrink closer to zero elementwise. The truncated linear basis in (2) has the benefit of interpretability, but can be numerically unstable in some settings (Hansen and Kooperberg (2002)). We did not experience any such numerical problems with this basis in either our simulations or application. However, one could implement penalized splines with an alternative numerically stable basis, such as B-splines (Eilers and Marx (1996)) or radial basis functions (French, Kammann and Wand (2001)).

In matrix notation, (2) takes the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a 2×1 vector of unknown, fixed parameters, $\mathbf{u} = (u_1, \ldots, u_K)^\mathsf{T} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_K)$ is a $K \times 1$ vector of random effects corresponding to the random truncated line coefficients, $\mathbf{X}$ and $\mathbf{Z}$ are appropriate covariate matrices, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$. Here, $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. Thus, (2) falls within the linear mixed model framework, and we can rely on the well-developed body of methodology for this broad class of models. In particular, the *best linear unbiased predictor* (BLUP) of $\mathbf{y}$ is $\widehat{\mathbf{y}} = \widehat{\mathbf{f}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}}$, where expressions for estimates $\widehat{\boldsymbol{\beta}}$ and predictions $\widehat{\mathbf{u}}$ are provided in Robinson (1991).

So far we have assumed that the residual errors ε_j are independent. However, it is often the case that these measures are correlated over time. In theory, the mixed model representation of (1) extends naturally to accommodate more general correlation structures. That is, it is straightforward in SAS PROC MIXED or R / Splus lme() to fit the model assuming a first order autoregressive process (AR(1)) $\varepsilon_j = \rho \varepsilon_{j-1} + \zeta_j$, where $\varepsilon_1 \sim N(0, \sigma_\varepsilon^2)$ and $\zeta_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2(1-\rho))$. Several authors have shown that, for other smoothers using automatic smoothing parameter selection, there can be difficulties in estimating serial correlation when estimating nonparametric trend (Simonoff (1996, Section 5.5.2) and Opsomer, Wang and Yang (2001)). To our knowledge, however, this issue has not been formally investigated in the context of mixed model regression splines. In the

next section we turn to a latent curve formulation that allows us to place more structure on the trend and correlation components of the model.

3. Self-Modeling Regression for Multiple Curves

We use SEMOR to extend penalized spline regression to multiple outcome data. A simple SEMOR model that specifies an underlying global time profile specifies the multivariate model in two stages. Again for a fixed subject, let y_{mj} be the response on the mth outcome, $m = 1, \ldots, M$, taken at time t_{mj}, $j = 1, \ldots, T_m$. Consider the model

$$y_{mj} = \lambda_{0,m} + \lambda_{1,m} f(t_{mj}) + \varepsilon_{mj} \tag{3}$$

and $f(t_{mj}) = \beta t_{mj} + \sum_{k=1}^{K} u_k (t_{mj} - \kappa_k)_+$, where $u_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma_u^2)$. Here, we assume $\lambda_{1,1} = 1.0$ to ensure identifiability. This constraint corresponds to the usual errors-in-variables parameterization in a standard factor-analysis model (Yalcin and Amemiya (2001)). We select the knots $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_K)^{\mathsf{T}}$ and check the sensitivity of the model fits to this choice using the strategies outlined in Section 2. Results of this sensitivity analysis for the human respiratory data suggest that the results are insensitive to this choice in the SEMOR setting as well.

We use SEMOR to model multiple curves for each subject as a parametric transformation of a single subject-specific latent curve. This represents a generalization of existing factor-analytic models for outcomes measured at a single timepoint (Sammel and Ryan (1996)) to the functional data setting. The latent curve assumption induces correlation among the multiple curves within a given subject. In particular, the shared random effects $\mathbf{u}$ induce correlation among the M curves, with $\text{Cov}\,(y_{mj}, y_{m'j}) = \sigma_u^2 \lambda_{1,m} \lambda_{1,m'} \mathbf{z}_{mj} \mathbf{z}_{m'j}^{\mathsf{T}}$. Here, $\mathbf{z}_{mj}$ is the row vector of truncated polynomial basis functions for $\mathbf{u}$ associated with observation j on outcome m. Because this basis and σ_u^2 are strictly non-negative, the sign of $\lambda_{1,m}\lambda_{1,m'}$ determines the direction of correlation between curves m and m'. In the special case of $M = 2$, this product simplifies to $\lambda_{1,2}$ due to the identifiability constraint on $\lambda_{1,1}$.

We assume that the residual errors $\boldsymbol{\varepsilon}_m = (\varepsilon_{m1}, \ldots, \varepsilon_{mT_m})^{\mathsf{T}}$ for outcome m are normally distributed with variance covariance matrix $\boldsymbol{\Sigma}_m = \sigma_{\varepsilon,m}^2 \mathbf{R}$, where the (jj')th element of $\mathbf{R}$ is $\rho^{|t_j - t_{j'}|}$. That is, we assume that the serial correlation, but not the residual error, is homogeneous across outcome.

4. Estimation and Inference

In this section we describe an Expectation-Conditional Maximization (ECM) algorithm (Meng and Rubin (1993)) for model fitting. The algorithm yields both maximum likelihood estimates of the fixed effects and variance components, as

well as predictions of the random effects given the data. In particular, we treat the random effects $\mathbf{u}$ as missing data, and iterate between calculating the expectation of complete data log-likelihood given the observed data and maximizing this expectation with respect to the fixed effects and variance components. In the case of nonlinear mixed model (3), the ECM algorithm yields closed-form estimators for all fixed effects and residual variances at each maximization step (Sammel and Ryan (1996)), allowing for easy and quick implementation of the model.

Let $\boldsymbol{\psi} = (\boldsymbol{\lambda}_0^\mathsf{T}, \boldsymbol{\lambda}_1^\mathsf{T}, \beta, \sigma_u, \boldsymbol{\sigma}_\varepsilon^\mathsf{T}, \rho)^\mathsf{T}$ denote the vector of fixed parameters and variance components, where the mth elements of $\boldsymbol{\lambda}_0$ and $\boldsymbol{\sigma}_\varepsilon$ are $\lambda_{0,m}$ and $\sigma_{\varepsilon,m}$, respectively, and the mth element of $\boldsymbol{\lambda}_1$ is $\lambda_{1,m+1}$. In addition, let $\boldsymbol{\psi}^{(p)}$ and $\mathbf{R}^{(p)}$ denote the current values of $\boldsymbol{\psi}$ and $\mathbf{R}$, respectively, after ECM iteration p. Iteration $p+1$ of the ECM algorithm consists of

- E-step: Calculate
  1. $E(\mathbf{u}|\mathbf{y}, \boldsymbol{\psi}^{(p)})$
  2. $E(\mathbf{u}\mathbf{u}^\mathsf{T}|\mathbf{y}, \boldsymbol{\psi}^{(p)})$
- M-step: Update parameter estimates using the following steps:
  1. Fixing $\rho^{(p)}$, $\boldsymbol{\sigma}_\varepsilon^{(p)}$, $\mathbf{R}^{(p)}$, $\sigma_u^{(p)}$, $\boldsymbol{\lambda}_0^{(p)}$ and $\boldsymbol{\lambda}_1^{(p)}$ update β using weighted least squares.
  2. Fixing $\rho^{(p)}$, $\boldsymbol{\sigma}_\varepsilon^{(p)}$, $\mathbf{R}^{(p)}$, $\sigma_u^{(p)}$, $\beta^{(p+1)}$, update $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ using weighted least squares.
  3. Fixing $\beta^{(p+1)}$, $\boldsymbol{\lambda}_0^{(p+1)}$ and $\boldsymbol{\lambda}_1^{(p+1)}$, update σ_u^2, ρ and $\boldsymbol{\sigma}_\varepsilon^2$ using the expected (REML) log likelihood.

See the appendix for the necessary expectations for the E-step and detail on the M-steps. An R implementation of the algorithm is very fast, converging in under a minute when applied to the bivariate response (TI, TE) for a given subject in the human respiratory experiment.

Ruppert, Wand and Carroll (2002) showed that the relevant quantity for standard error estimation in penalized spline models is with respect to the joint distribution of $\mathbf{y}$ and $\mathbf{u}$. In the case of the linear mixed model, Henderson (1975) showed that this quantity is the expected negative inverse Hessian, or equivalently the negative inverse Fisher information, of the joint log density of $\mathbf{y}$ and $\mathbf{u}$ with respect to the fixed and random effects.

We extend this strategy and use the negative inverse Fisher information under (3). Let $\mathbf{t}_m$ and $\mathbf{Z}_m$ be the vector of times and the matrix of basis functions associated with $\mathbf{u}$ for outcome m, respectively, let $\mathbf{K}_m = [\mathbf{1} \quad \beta\mathbf{t}_m + \mathbf{Z}_m\mathbf{u}]$, $m =$

$2, \ldots, M$, and let

$$\mathbf{K} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{t}_1 & \mathbf{Z}_1 \\ \mathbf{0} & \mathbf{K}_2 & \mathbf{0} & \ldots & \mathbf{0} & \lambda_{1,2}\mathbf{t}_2 & \lambda_{1,2}\mathbf{Z}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_3 & \ldots & \mathbf{0} & \lambda_{1,3}\mathbf{t}_3 & \lambda_{1,3}\mathbf{Z}_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{K}_M & \lambda_{1,M}\mathbf{t}_M & \lambda_{1,M}\mathbf{Z}_M \end{bmatrix}.$$

Further, let $\boldsymbol{\lambda} = (\lambda_{0,1}, \lambda_{0,2}, \lambda_{1,2}, \ldots, \lambda_{0,M}, \lambda_{1,M})^{\mathsf{T}}$. Then the negative inverse Fisher information is

$$\text{Cov} \begin{pmatrix} \boldsymbol{\lambda} \\ \beta \\ \widehat{\mathbf{u}} - \mathbf{u} \end{pmatrix} = \left(\mathbf{K}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{K} + \mathbf{D} \right)^{-1}, \tag{4}$$

where $\mathbf{D} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \mathbf{I}_K \end{bmatrix}$. We calculate standard errors by plugging in maximum likelihood estimates $\widehat{\boldsymbol{\psi}} = (\widehat{\boldsymbol{\lambda}}_0^{\mathsf{T}}, \widehat{\boldsymbol{\lambda}}_1^{\mathsf{T}}, \widehat{\beta}, \widehat{\sigma}_u, \widehat{\boldsymbol{\sigma}}_\varepsilon^{\mathsf{T}}, \widehat{\rho})^{\mathsf{T}}$ and predictions $\widehat{\mathbf{u}}$ into (4).

5. Simulation Study

In this section we report the results of a small simulation study designed to compare the finite sample performance of our proposed likelihood-based approach to SEMOR model fitting (denoted by MM for mixed model) to that from (1) existing SEMOR fitting strategies, and (2) univariate penalized splines fit to each outcome separately. The existing approach we consider is the two-stage procedure of Kneip and Engel (1995). At the first stage, this approach estimates the cross-sectional mean over time by smoothing all of the observations $(t_{mj}, y_{mj})_{m=1,\ldots,M;j=1,\ldots,T_m}$ together. The second stage then entails estimating the parameters $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_1$ and the variance components by least squares and moment estimators, respectively. Kneip and Engel (1995) showed that the estimates from this simple scheme are asymptotically as efficient as those that would be available if the cross-sectional mean function were known. For the univariate penalized spline fits, we base estimation on restricted maximum likelihood (REML) estimation of the variance components.

In view of the human respiratory data, our experiment uses $M = 2$, $T_1 = T_2 = 60$ with t_js evenly spaced between zero and one, $\boldsymbol{\lambda}_0 = (0.27, 0.13)$ and $\boldsymbol{\lambda}_1 = (1.00, 1.14)$. We use a factorial combination based on functional forms $f(t) = \{\sin(4\pi t), 3 - 4(t - 0.5)^2\}$, autoregressive parameter values $\rho = \{0.0, 0.2\}$, and residual variance values $\boldsymbol{\sigma}_\varepsilon^2 = \left\{ \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 0.1 \\ 0.6 \end{pmatrix} \right\}$. The Monte Carlo sample

size is 500 and we re-used the simulated errors from case to case as a variance reduction technique. We used lme() in R to perform the REML penalized spline fits and to conduct the smoothing step in the KE method. All fits are computed assuming first-order serial correlation ρ among observations taken on the same outcome.

Tables 1 and 2 show results for the latent sine curve and the latent quadratic curve, respectively. In both cases, we see that the differences in both the root mean squared error and the absolute bias of the three curve estimators are small when both outcomes have equal residual variances. However, there exist noticeable differences among the three strategies when one outcome is noisier than the other. In this case, both SEMOR estimators outperform their univariate counterpart substantially. These SEMOR estimators achieve this large efficiency gain by borrowing strength from the low noise observations to help fit the curve observations. Further, our proposed MM method outperforms the KE method by using a likelihood to more efficiently borrow this strength. The gains come in terms of both root mean squared error and absolute bias. We also note that this difference between the MM method and the others grows as the autocorrelation increases. Thus, the MM method uses a single likelihood to combine the information from both outcomes, estimate the autocorrelation, and adjust the amount of smoothing accordingly.

Table 1. Average RMSE and average absolute bias of the SEMOR-based mixed model (MM) estimates and two-stage estimates of Kneip and Engel (KE), as well as univariate penalized splines based on REML when the latent curve is $f(t) = \sin(4\pi t)$.

Average Root Mean Squared Error X 1000

		Outcome 1 SEMOR			Outcome 2 SEMOR		
ρ	$(\sigma_{e,1}, \sigma_{e,2})$	MM	KE	REML	MM	KE	REML
0.0	(0.1,0.1)	39.19	38.10	52.80	42.70	40.99	52.29
0.2	(0.1,0.1)	44.69	41.77	56.45	50.17	46.10	57.75
0.0	(0.1,0.6)	51.03	52.98	51.50	124.13	198.06	260.56
0.2	(0.1,0.6)	59.31	60.67	56.76	135.57	226.30	396.12

Average Absolute Bias X 1000

		Outcome 1 SEMOR			Outcome 2 SEMOR		
ρ	$(\sigma_{e,1}, \sigma_{e,2})$	MM	KE	REML	MM	KE	REML
0.0	(0.1,0.1)	7.54	7.09	7.93	8.23	8.13	8.87
0.2	(0.1,0.1)	7.69	7.58	8.46	8.22	8.68	9.62
0.0	(0.1,0.6)	8.89	9.16	8.35	12.80	59.64	75.26
0.2	(0.1,0.6)	10.41	10.34	8.68	17.10	105.43	167.52

Table 2. Average RMSE and average absolute bias of the SEMOR-based mixed model (MM) estimates and two-stage estimates of Kneip and Engel (KE), as well as univariate penalized splines based on REML when the latent curve is $f(t) = 3 - 4(t - 0.5)^2$.

Average Root Mean Squared Error X 1000

		Outcome 1 SEMOR			Outcome 2 SEMOR		
ρ	$(\sigma_{e,1}, \sigma_{e,2})$	MM	KE	REML	MM	KE	REML
0.0	(0.1,0.1)	27.08	25.15	31.73	29.37	27.10	32.25
0.2	(0.1,0.1)	30.64	28.77	36.61	32.37	30.28	35.82
0.0	(0.1,0.6)	35.74	37.49	31.56	106.01	113.82	138.14
0.2	(0.1,0.6)	37.35	38.92	36.10	142.36	156.14	183.15

Average Absolute Bias X 1000

		Outcome 1 SEMOR			Outcome 2 SEMOR		
ρ	$(\sigma_{e,1}, \sigma_{e,2})$	MM	KE	REML	MM	KE	REML
0.0	(0.1,0.1)	2.66	5.50	6.32	5.43	6.32	6.33
0.2	(0.1,0.1)	4.44	5.59	6.78	6.16	8.56	8.21
0.0	(0.1,0.6)	2.75	7.23	5.68	14.38	15.14	24.58
0.2	(0.1,0.6)	5.76	6.07	6.13	10.23	30.72	41.41

6. Analysis of Respiratory Data

In the human respiratory study outlined in the Introduction, interest focuses on observations y_{scmj}, $j = 1, \ldots, T_{scm}$, from two outcomes ($m = 1, 2$) for eight subjects ($s = 1, \ldots, 8$), each observed under two conditions: filtered air exposure and ROFA exposures ($c = 1, 2$). A SEMOR model for this setting is

$$y_{scmj} = \lambda_{0scm} + \lambda_{1scm} f_s^c(t_{scmj}) + \varepsilon_{scmj}, \tag{5}$$

where $f_s^c(t_{scmj}) = \beta_{1s}^c t_{scmj} + \sum_{k=1}^K u_{sk}^c (t_{scmj} - \kappa_k)_+$, $u_{sk}^c \overset{\text{i.i.d.}}{\sim} N(0, \sigma_{u,sc}^2)$, $s = 1, \ldots, S$, $c = 1, 2$, and $\boldsymbol{\varepsilon}_{scm} = (\varepsilon_{scm1}, \ldots, \varepsilon_{scmT_{scm}})^\mathsf{T}$ are normally distributed with variance covariance matrix $\boldsymbol{\Sigma}_{scm} = \sigma_{\varepsilon,scm}^2 \mathbf{R}_{sc}$. Here, the (jj')th element of $\mathbf{R}_{sc}$ is $\rho_{sc}^{|t_j - t_{j'}|}$. As before, $\lambda_{1sc1} = 1.0$ for all s and c for identifiability. Thus, the TI outcome ($m = 1$) is the reference curve for each subject and condition. This model corresponds to fitting SEMOR model (3) to each subject separately. Figure 2 presents the fit of this model to the data shown in Figure 1. Here, the 95% pointwise confidence bands for each profile are obtained using (4) and applying the delta method to the outcome-specific fits $\widehat{\lambda}_{0,m} + [1 + (\widehat{\lambda}_{1,m} - 1) * I(m = 2)]\widehat{f}(t)$, where $I()$ is the indicator function. The quality of the model fit for the remaining seven subjects in the study is similarly good.

In this setting, primary scientific interest focuses on the differences between the curves under air and ROFA conditions for both the TI and TE outcomes, or $d_{s,m=1}(t) = [f_s^2(t) - f_s^1(t)]$ and $d_{s,m=2}(t) = [\lambda_{1s22} f_s^1(t) - \lambda_{1s12} f_s^0(t)]$ respectively. Figure 3 summarizes these difference curves, averaging them over the eight subjects in the study. We note that six of the eight subjects in each condition displayed the general pattern described by the average curve, supporting the fact that these averages represent the general trends and not an average of widely disparate curves for each condition. Also, differences based on univariate curves (not shown) have similar shapes, but the 95% pointwise confidence intervals are wider.

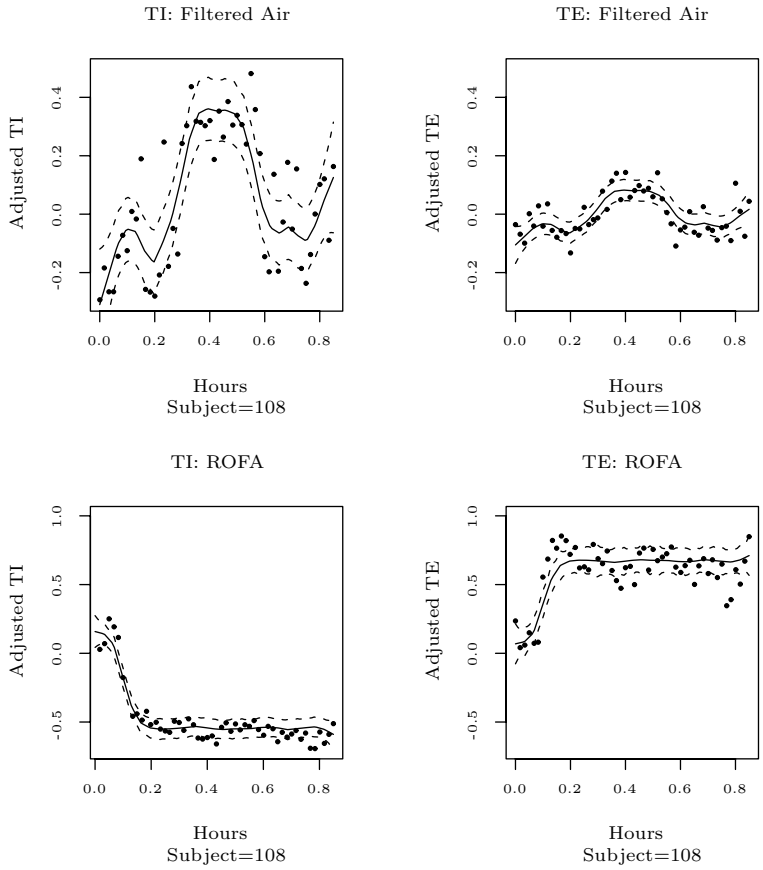

Figure 2. SEMOR fit to the data displayed in Figure 1.

Overall, the analysis suggests that, relative to control conditions, ROFA inhalation initially decreases one's time to inspiration up until approximately 30 minutes into the exposure, with this difference gradually disappearing toward the end of the exposure period. In contrast, results suggest that exposure initially

increases time to exhalation.

In addition to inference on each outcome, the subject-specific SEMOR model (5) yields inferences on the effect of exposure on the associations between the outcomes outcomes via the sign of λ_{1sc2}. Table 3 shows that four subjects exhibit negative correlation between outcomes (TI, TE) under ROFA ($c = 2$) conditions, whereas only one subject exhibits negative correlations under control conditions ($c = 1$).

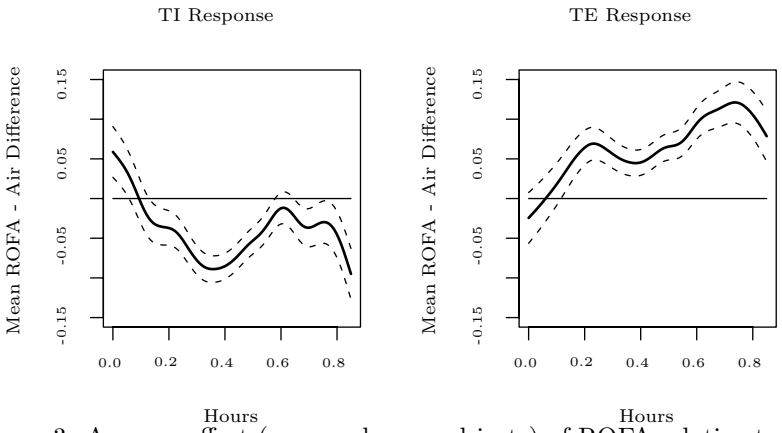

Figure 3. Average effect (averaged over subjects) of ROFA relative to clean air on log(time to inspiration) and log(time to exhalation).

Table 3. Subject-treatment specific "self-modeling" coefficients λ_{1sc2} obtained from the fit of SEMOR model (5) to the eight subjects in the human respiratory experiment.

Subject (s)	Air ($c = 1$) $\widehat{\lambda}_{1s12}$	ROFA ($c = 2$) $\widehat{\lambda}_{1s22}$
1	0.99	-0.20
2	0.85	1.02
3	0.28	-0.77
4	1.00	-0.38
5	0.11	-0.92
6	1.36	0.98
7	-0.95	1.34
8	2.14	1.81

7. Discussion

In this article we have proposed a penalized spline formulation of self-modeling regression for multivariate curve data. Wand (2003) showed that the mixed model representation of penalized splines allows one to easily incorporate complications

such as clustering, missing data, and measurement error into smoothing models. In this paper, we have shown that we can add multivariate models for multiple outcomes to this list, as this mixed model makes the connection between SEMOR regression and existing latent variable models for multiple outcomes immediate. Our approach differs from that of other penalized spline formulations of SEMOR (Altman (2001) and Altman and Villarreal (2001)) in that we use a fully likelihood-based framework for estimation and inference. This SEMOR framework allows one to easily estimate both the serial correlation within an outcome and the correlation among outcomes measured at a given timepoint.

We applied the models to explore the effects of residual oil fly ash inhalation on respiration in humans. Instead of using overall respiratory frequency as an endpoint, investigators were interested in the bivariate outcome (time to inspiration, time to exhalation). This multivariate analysis yielded greater insight into the mechanisms of air pollution inhalation, in that not only did it compare breathing patterns exhibited during the two exposure protocols, but it also yielded a formal mechanism for inference on the effect of exposure on the relationship between the two outcomes.

In the human respiratory example, we checked the fits of the SEMOR latent curve assumption by visually inspecting the difference in deviances from this model to that obtained from the univariate fits. The subject-specific SEMOR model fits extremely well for all subjects (see Figure 2), implying that more complicated SEMOR models are not necessary for this application. However, in other applications, one could extend the model to be more flexible in several ways. For instance, other SEMOR formulations use a linear time scale transformation (see Lindstrom (1995)). Alternatively, if interest focuses on $m > 2$ manifest curves, one could easily extend (3) to specify multiple latent curves. Alternatively, as in other hierarchical settings, information could be further pooled by placing constraints on parameters in the hierarchical formulation (5). For instance, a model that assumes that the associations between outcomes are homogeneous across subjects undergoing a particular exposure constrains $\lambda_{11cm} = \lambda_{12cm} = \ldots = \lambda_{1Scm}$ for all c and m. We fit this model to the respiratory data, but it fit the data poorly, as is suggested by the unrestricted estimates presented in Table 3.

Because (3) represents a generalization of existing latent variable models for multiple outcomes taken at a single time point, it inherits an important disadvantage associated with this class of models (Sammel, Lin and Ryan (1999) and Sammel and Ryan (2002)) that may make it unsuitable in some settings. In particular, the mixed-model representation of SEMOR implies that the loading parameters $\boldsymbol{\lambda}_1$ enter into the mean and covariance model for $\mathbf{Y}$, often leading to problems with robustness. As demonstrated by (6) in the appendix, this is

compounded by the fact that $\lambda_{1,m}$ also serves to adjust the smoothing parameter $\sigma^2_{\varepsilon,1}/\sigma^2_u$. In other applications, the more flexible formulation

$$y_{mj} = \beta_{0,m} + \beta_{1,m} t_{mj} + \delta_m \sum_{k=1}^{K} u_k (t_{mj} - \kappa_k)_+ + \varepsilon_{mj}$$

may be more robust. Again restrictions on some of the parameters would be necessary for identifiability. A disadvantage of this alternative formulation is interpretability and the lack of a global latent curve (Sammel, Lin and Ryan (1999)).

Another minor disadvantage of the SEMOR formulation is that, since the correlation among outcomes is captured solely through the nonlinear component of the model, the model does not cover all possible cases of correlated curves. For instance, the model does not cover the case in which observations taken on different outcomes are dependent, yet the profiles are linear (i.e., $\sigma^2_u = 0$). In such cases, however, if the latent curve assumption truly holds, then both profiles are likely to be flat and ordinary general linear models with correlated errors can be used to model the data. Thus, this does not appear to be a severe limitation of the model. If one curve is nonlinear and the other is constant, then the latent curve assumption itself is suspect and the SEMOR formulation is probably not the best model for the data. We note that we did not encounter any linear profiles in the human respiratory application.

An interesting problem for future research is to establish a theoretical framework for identifiability of mixed model representations of regression splines with data-driven smoothing parameter selection in the presence of autocorrelation. As noted in Section 2, this has been problematic in other smoothing contexts, but has not been formally investigated in this context. Our experience suggests that, using SAS PROC MIXED, REML iterations for model (2) with first-order serial correlation can converge to a weakly-identified local maximum of the likelihood if one does not provide the algorithm with good starting values of the variance components. However, once we start the algorithm at the variance components obtained from the model assuming independent errors, we have not observed any cases of non-identifiability. Others (Opsomer, Wang and Yang (2001), Currie and Durban (2002) and Durban and Currie (2003)) have reported similar empirical results for regression and smoothing splines in this regard.

Finally, the application of our model to the human respiratory data used subject-specific fixed effects. An alternative would be to extend the model to include subject-specific random effects. Such an extension would complicate model fitting though, and fully Bayesian approaches may be preferable. Overall, we find SEMOR models can be an effective strategy for modeling multivariate curve data.

Acknowledgements

We thank Matt Wand for helpful discussions during the initial phases of this research, and two referees for helpful comments. This research was partially supported by NIH grants ES 07142 and ES 10844.

Appendix

Let $\widetilde{\mathbf{X}}_m = [1 \quad \lambda_{1,m} t_{mj}]_{1 \le j \le T_m}$, $\boldsymbol{\beta}_m = [\lambda_{0,m} \quad \beta]^\mathsf{T}$, and $\widetilde{\mathbf{Z}}_m = \lambda_{1,m} Z_m$. From the definitions of $\mathbf{u}$ and $\boldsymbol{\varepsilon}_m$ in Section 3, it follows that the marginal distribution of $\mathbf{y}_m$ is

$$\mathbf{y}_m \sim \mathrm{MVN}_{T_m}\left(\widetilde{\mathbf{X}}_m \boldsymbol{\beta}_m, \sigma_u^2 \widetilde{\mathbf{Z}}_m \widetilde{\mathbf{Z}}_m^\mathsf{T} + \sigma_{e,m}^2 \mathbf{R}\right). \tag{6}$$

Let $T = \sum_{m=1}^{M} T_m$, $\widetilde{\mathbf{X}} = \mathrm{blockdiag} \widetilde{\mathbf{X}}_m$, $\widetilde{\mathbf{Z}} = [\widetilde{\mathbf{Z}}_1^\mathsf{T}, \ldots, \widetilde{\mathbf{Z}}_M^\mathsf{T}]^\mathsf{T}$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\mathsf{T}, \ldots, \boldsymbol{\beta}_M^\mathsf{T}]^\mathsf{T}$ and $\boldsymbol{\Sigma} = \mathrm{diag}\ \sigma_{e,m}^2 \otimes \mathbf{R}$. Using this notation, we can succinctly write the marginal distribution as $\mathbf{y} \sim MNV_T(\widetilde{\mathbf{X}}\boldsymbol{\beta}, \sigma_u^2 \widetilde{\mathbf{Z}} \widetilde{\mathbf{Z}}^\mathsf{T} + \boldsymbol{\Sigma})$, and the joint distribution of ($\mathbf{y}$, $\mathbf{u}$) as multivariate normal with

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\widetilde{\mathbf{X}}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix} \text{ and } \quad \mathrm{Cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}\sigma_u^2 \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\mathsf{T} + \boldsymbol{\Sigma} & \sigma_u^2 \widetilde{\mathbf{Z}}\\ \sigma_u^2 \widetilde{\mathbf{Z}}^\mathsf{T} & \sigma_u^2 \mathbf{I}_K\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y},\mathbf{y}} & \boldsymbol{\Sigma}_{\mathbf{y},\mathbf{u}}\\ \boldsymbol{\Sigma}_{\mathbf{y},\mathbf{u}}^\mathsf{T} & \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{u}}\end{pmatrix}.$$

Thus, using standard results, the conditional normal distribution of $\mathbf{u}|\mathbf{y}$ can be derived. Letting $\widetilde{\mathbf{C}} = [\widetilde{\mathbf{X}} \quad \widetilde{\mathbf{Z}}]$, the conditional mean can be simply expressed as the last K elements of

$$\left\{\widetilde{\mathbf{C}}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{C}} + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma_u^2 \mathbf{I}_K\end{pmatrix}\right\}^{-1} \widetilde{\mathbf{C}}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

(e.g., Robinson (1991)). A benefit of this form is that it illustrates that the SEMOR mixed model smoothes in a similar manner to the simpler mixed model presented in Section 2.

References

Altman, N. (2001). Inference for self-modeling regression with random effects. Technical Report. Department of Biometrics, Cornell University.

Altman, N. and Villarreal, J. C. (2001). Self-modeling regression with random effects using penalized splines. Department of Biometrics, Cornell University.

Altman, N. A. and Casella, G. (1995). Nonparametric empirical Bayes growth curve analysis. *J. Amer. Statist. Assoc.* **90**, 508-515.

Brumback, B. A., Ruppert, D. and Wand, M. P. (1999). Comment to "Variable selection and function estimation in additive nonparametric regression using a data-based prior". *J. Amer. Statist. Assoc.* **94**, 794-797.

Coull, B. A., Schwartz, J. and Wand, M. P. (2001). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics* **2**, 337-349.

Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statist. Modeling* **2**, 333-349.

Dockery, D. W., Pope, C. A., Xu, X. P., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G. and Speizer, F. E. (1993). An association between air-pollution and mortality in 6 United-States cities. *New England J. Medicine* **329**, 1753-1759.

Durban, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Comput. Statist.* **18**, 251-262.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11**, 89-121.

French, J., Kammann, E. E. and Wand, M. P. (2001). Comment on "Semiparametric nonlinear mixed-effects models and their applications." *J. Amer. Statist. Assoc.* **96**, 1285-1287.

Hansen, M. H. and Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statist. Sci.* **17**, 2-20.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-448.

Ke, C. L. and Wang, Y. D. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussion). *J. Amer. Statist. Assoc.* **96**, 1272-1298.

Kneip, A. and Engel, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23**, 551-570.

Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self modeling nonlinear regression. *Ann. Statist.* **11**, 82-112.

Ladd, W. M. and Lindstrom, M. J. (2000). Self-modeling for two-dimensional response curves. *Biometrics* **56**, 89-97.

Lawton, W. H. and Sylvestre, E. A. (1971). Self modeling curve resolution. *Technometrics* **13**, 617-633.

Lawton, W. H., Sylvestre, E. A. and Maggio, M. S. (1972). Self modeling nonlinear regression. *Technometrics* **14**, 513-532.

Lefkopoulou, M. and Ryan, L. M. (1993). Global tests for multiple binary outcomes. *Biometrics* **49**, 975-988.

Lin, X. H. and Carroll, R. J. (2001). Semiparametric regression for clustered data. *Biometrika* **88**, 1179-1185.

Lin, X. H. and Zhang, D. W. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Statist. Soc. Ser. B* **61**, 381-400.

Lindstrom, M. J. (1995). Self-modeling with random shift and scale parameters and a free-knot spline shape function. *Statist. Medicine* **14**, 2009-2021.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267-278.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York.

Opsomer, J., Wang, Y. D. and Yang, Y. H. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16**, 143-153.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials* **18**, 530-545.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statist. Sci.* **6**, 15-51.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11**, 735-757.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Chapman and Hall, New York.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I. and Zeger, S. L. (2000). Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *New England J. Medicine* **343**, 1742-1949.

Sammel, M. D., Lin, X. and Ryan, L. M. (1999). Multivariate linear mixed models for multiple outcomes. *Statist. Medicine* **18**, 2479-2492.

Sammel, M. D. and Ryan, L. M. (1996). Latent variable models with fixed effects. *Biometrics* **52**, 650-663.

Sammel, M. D. and Ryan, L. M. (2002). Effects of covariance misspecification in a latent variable model for multiple outcomes. *Statist. Sinica* **12**, 1207-1222.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **93**, 1403-1417.

Wand, M. P. (2003). Smoothing and mixed models. *Comput. Statist.* **18**, 223-249.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93**, 341-348.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics* **52**, 588-596.

Wang, Y., Guo, W. and Brown, M. B. (2000). Spline smoothing for bivariate data with applications to association between hormones. *Statist. Sinca* **10**, 377-397.

Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statist. Sci.* **16**, 275-294.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

Zhang, H. (1999). Analysis of infant growth curves using multivariate adaptive splines. *Biometrics* **55**, 452-459.

Zhang, D. W., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *J. Amer. Statist. Assoc.* **93**, 710-719.

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

E-mail: bcoull@hsph.harvard.edu

Department of Mathematics and Statistics, University of Massachusetts, Lederle Graduate Research Tower, Amherst, Massachusetts 01003, U.S.A.

E-mail: jstauden@math.umass.edu

(Received January 2003; accepted October 2003)