Model selection in multivariate semiparametric regression

Zhuokai Li,[1] Hai Liu[2] and Wanzhu Tu[3]

Abstract

Variable selection in semiparametric mixed models for longitudinal data remains a challenge, especially in the presence of multiple correlated outcomes. In this paper, we propose a model selection procedure that simultaneously selects fixed and random effects using a maximum penalized likelihood method with the adaptive least absolute shrinkage and selection op-erator (LASSO) penalty. Through random effects selection, we determine the correlation structure among multiple outcomes and therefore address whether a joint model is necessary. Additionally, we include a bivariate nonparametric component, as approximated by tensor product splines, to accommodate the joint nonlinear effects of two independent variables. We use an adaptive group LASSO to determine whether the bivariate nonparametric com-ponent can be reduced to additive components. To implement the selection and estimation method, we develop a two-stage expectation-maximization (EM) procedure. The operating characteristics of the proposed method are assessed through simulation studies. Finally, the method is illustrated in a clinical study of blood pressure development in children.

Keywords

adaptive LASSO, adaptive group LASSO, EM algorithm, mixed effects, multivariate data

[1]Duke Clinical Research Institute, Durham, NC, USA
[2]Gilead Sciences, Inc., Foster City, CA, USA
[3]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA

Corresponding author:

Wanzhu Tu, Department of Biostatistics, Indiana University School of Medicine, 410 West 10th Street, Indianapolis, IN 46202, USA

Email: wtu1@iu.edu

This is the author's manuscript of the article published in final edited form as:

Li, Z., Liu, H., & Tu, W. (2017). A generalized semiparametric mixed model for analysis of multivariate health care utilization data. Statistical Methods in Medical Research, 26(6), 2909–2918. https://doi.org/10.1177/0962280215615159

1 Introduction

Longitudinally assessed multiple outcome data are abundant in clinical investigations. For example, systolic and diastolic blood pressure readings are always measured in pairs. Together, they quantify the arterial pressure that circulating blood exerts on the walls of blood vessels during a cardiac cycle. Although one could choose to analyze systolic and diastolic readings in separate models, simultaneous modeling of the two measures provides a more complete picture of the systemic circulation. The joint modeling approach not only improves estimation efficiency by borrowing information across the outcomes, but also provides an opportunity to compare the contributions of the same independent variable to different outcomes[1].

Among the available methods, multivariate semiparametric mixed effects models provide perhaps the most general analytical framework for such situations. Structurally, these models are extensions of the traditional mixed effects models[2], where the fixed effects characterize the influences of independent variables, and the random effects reflect the correlations among repeated measures within each outcome as well as the interdependence of multiple outcomes within each subject[3–5]. In addition, the inclusion of nonparametric functions has greatly enhanced the modeling flexibility for accommodating nonlinear effects of independent variables. Semiparametric mixed models have gained popularity in analyzing longitudinal data[6,7], and have been applied to multivariate settings[8,9]. These semiparametric modeling techniques have been successfully used to disseminate the concurrent nonlinear influences of biological regulators on blood pressure[10–12].

For such a general modeling framework, it is always desirable to have a systematic approach that helps to determine variable inclusion and functional forms through which key variables enter the model. For fixed effects, including unnecessary variables tends to reduce modeling efficiency and result in numerical instability while omitting relevant variables may lead to biased estimation[13]. For random effects, analysts often choose their inclusion based on subjective perceptions of the unknown correlation structure. While in generalized estimating equation (GEE) based models, correct specification of the correlation structure is less than essential[14], misspecification in mixed models could nonetheless affect the validity of variance estimation and statistical inference[15,16]. In a multivariate analysis, correct specification of the random effects takes on new importance because it determines how the multiple outcomes are related.

Variables can be selected using regularization methods including the least absolute shrinkage and selection operator (LASSO)[17], smoothly clipped absolute deviation (SCAD)[18], least angle regression (LARS)[19], elastic net[20], and the adaptive LASSO[21]. Further extensions of these methods have been proposed to select groups of variables, such as the group LASSO[22], the adaptive group LASSO[23,24], and the group bridge approach[25]. Through regularization, simultaneous selection of fixed and random effects has been discussed in linear mixed models[26,27] and genearalized linear mixed models[28]. More recently, variable selection and group selection methods have been applied to various nonparametric and semiparametric models. For example, Lin and Zhang[29] proposed the component selection and smoothing operator (COSSO) method for model selection and esti-mation in multivariate nonparametric regression. Huang et al.[30] used the adaptive group LASSO to select nonzero components in nonparametric additive models after obtaining an initial estima-tor with the group LASSO. Zhang et al.[31] proposed the linear and nonlinear discoverer (LAND) to determine the adequacy of linear effects of independent variables in partial linear models. Du et al.[32] discussed simultaneous selection of additive nonparametric and parametric components in semiparametric regression by applying double penalties. Variable selection has been further considered in semiparametric models for longitudinal data. Among the published work, Fan and Li[33] employed the SCAD penalty to select parametric covariate effects in a class of semiparametric models which did not require explicit specification of the correlation structure in longitudinal data. Ni et al.[34] proposed a double-penalized likelihood method for semiparametric mixed models, in which a shrinkage penalty was used in fixed effect selection and a roughness penalty was used for smooth function estimation. Other semiparametric models for longitudinal data in which variable selection has been investigated include partially linear varying coefficient models[35,36] and additive partial linear models[37,38].

The purpose of the current paper is to discuss model selection and structural discovery in multivariate semiparametric regression for longitudinal data. To the best of our knowledge, none of the existing variable selection methods are readily applicable to multivariate data in longitudinal settings. Multivariate mixed models usually have more complex random effects which accommodate both the within-outcome and cross-outcome correlations within each subject, so the challenge of model selection lies primarily in determining whether the outcomes are correlated through ran-

dom effects selection. This is the situation where selection methods are most relevant because of the increased model complexity. The research question is of greater importance because it helps justify or invalidate the joint modeling approach, an issue that has not been sufficiently addressed in the existing literature. Another question of interest discussed in this paper is whether two independent variables have joint nonlinear and interacting influences on the outcome variables, which we refer to as "structural discovery". This will also be addressed in the framework of model selection by performing group selection on bivariate nonparametric components.

Herein, we propose a model selection method for multivariate semiparametric mixed models. The method will (1) aid the simultaneous selection of the fixed and random effects, (2) determine the cross-outcome correlations to justify the joint modeling approach, and (3) discover the existence of joint nonlinear effects of two independent variables. Specifically, we present a two-stage model selection and estimation procedure. In Stage 1, we use the adaptive LASSO and adaptive group LASSO penalties to simultaneously select the fixed and random effects as well as the nonparametric components. In Stage 2, we obtain unbiased estimates of selected parameters by maximizing the observed likelihood function. The performance of the proposed method is demonstrated in simulation studies. We illustrate the method by analyzing data from a childhood blood pressure study.

2 Methods

2.1 Model Formulation

Suppose that there are m subjects in a longitudinal study and K outcomes are measured at each visit. For the ith subject, let Y_{ijk} be the kth outcome observed at the jth time point of repeated measurements, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, K$. We consider the following multivariate semiparametric mixed model:

$$Y_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T \boldsymbol{u}_{ik} + s_k(t_{1ij}, t_{2ij}) + \epsilon_{ijk}, \tag{1}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^T$ is a $p \times 1$ coefficient vector for the fixed effects $\boldsymbol{x}_{ij}$, $\boldsymbol{u}_{ik}$ is a $q \times 1$ vector of subject- and outcome-specific random effects for the corresponding covariates $\boldsymbol{z}_{ij}$ (which can

be a subset of $\boldsymbol{x}_{ij}$). We assume that the random effects $\boldsymbol{u}_{ik}$ follow a multivariate normal distribution $\mathrm{N}_q(\boldsymbol{0}, \boldsymbol{D}_{kk})$, the measurement errors ϵ_{ijk} independently follow a normal distribution $\mathrm{N}(0, \sigma_k^2)$, and $\boldsymbol{u}_{ik}$ and ϵ_{ijk} are independent. Let t_{1ij} and t_{2ij} be the continuous independent variables that potentially have nonlinear influences on the outcomes. Without loss of generality, we incorporate in the model a bivariate nonparametic smooth function s_k, which can be easily extended to multiple nonparametric components. The bivariate nonparametric function can also be expanded to three or more variables to allow for higher-order interactions. In practice, however, higher-order interactions are rarely explored in data analysis because they are generally hard to interpret. For this reason, we are focusing on the modeling and selection of bivariate nonparametric functions in this paper.

For smooth function s_k, we specify a tensor product basis[39] with marginal basis functions $\phi_{l_1}(t_{1ij})$, $l_1 = 1, \ldots, L_1$, $\psi_{l_2}(t_{2ij})$, $l_2 = 1, \ldots, L_2$, and all of their pairwise products. Examples of the marginal basis functions include truncated polynomials and B-splines. For regression splines, it is generally advisable to choose a modest number of knots depending on the specific research problem being studied so that the splines adequately represent the unknown true function while maintaining computational efficiency. Alternatively, penalized regression splines or smoothing splines may be used so the choice of number of knots is not as critical. The potential interactions between t_{1ij} and t_{2ij} are incorporated through the product terms in the tensor product basis, and therefore they can be selected while keeping the main effects intact. Assuming $\phi_1(t_{1ij}) = \psi_1(t_{2ij}) = 1$, we write $s_k(t_{1ij}, t_{2ij}) = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \alpha_{l_1,l_2,k} \phi_{l_1}(t_{1ij}) \psi_{l_2}(t_{2ij})$ where $\alpha_{l_1,l_2,k}$ are the coefficients associated with the corresponding basis functions. Using a vector form, we further write $s_k(t_{1ij}, t_{2ij}) = \boldsymbol{T}_{ij}^T \boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k$ is a vector of the coefficients $\alpha_{l_1,l_2,k}$ for $l_1 = 1, \ldots, L_1$ and $l_2 = 1, \ldots, L_2$, and $\boldsymbol{T}_{ij}$ is the corresponding vector of the tensor product basis functions.

For convenience, we rewrite model (1) into a matrix form. We define the response vector as $\boldsymbol{Y}_i = (Y_{i11}, \cdots, Y_{in_i1}, \cdots, Y_{i1K}, \cdots, Y_{in_iK})^T$, the fixed effect coefficient vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T)^T$, the subject-specific random effects as $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}^T, \cdots, \boldsymbol{u}_{iK}^T)^T$, the coefficient vector for the smooth functions as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_K^T)^T$, and the vector of measurement errors as $\boldsymbol{\epsilon}_i = (\epsilon_{i11}, \cdots, \epsilon_{in_i1}, \cdots, \epsilon_{i1K}, \cdots, \epsilon_{in_iK})^T$. Then model (1) can be rewritten as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{T}_i \boldsymbol{\alpha} + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\boldsymbol{u}_i$ follows a multivariate normal distribution $\mathrm{N}_{Kq}(\mathbf{0}, \boldsymbol{D})$. The covariance matrix $\boldsymbol{D}$ accommodates the within-subject correlations among the repeated measurements and across the multiple outcomes. Specifically, $\boldsymbol{D}$ can be written as

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} & \cdots & \boldsymbol{D}_{1K} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} & \cdots & \boldsymbol{D}_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{D}_{K1} & \boldsymbol{D}_{K2} & \cdots & \boldsymbol{D}_{KK} \end{pmatrix}, \tag{3}$$

where $\boldsymbol{D}_{jk}$ $(j, k = 1, \ldots, K)$ are $q \times q$ submatrices. The diagonal submatrices are the covariance matrices of the random effects within each of the outcomes, and the off-diagonal ones indicate the potential correlations across the outcomes. If the outcomes are not correlated, then $\boldsymbol{D}$ degenerates to a block diagonal matrix $\mathrm{diag}(\boldsymbol{D}_{11}, \cdots, \boldsymbol{D}_{KK})$. Additionally, the measurement errors are normally distributed, i.e., $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{Kn_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma_1^2 \boldsymbol{I}_{n_i}, \cdots, \sigma_K^2 \boldsymbol{I}_{n_i})$.

Cholesky decomposition of the covariance matrix $\boldsymbol{D}$ is a key step for the selection of random effects as it ensures that $\boldsymbol{D}$ remains positive semidefinite. Hence, $\boldsymbol{D}$ is decomposed as $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ where $\boldsymbol{\Gamma}$ is a $Kq \times Kq$ lower triangular matrix. Accordingly, the random effects $\boldsymbol{u}_i$ can be reparameterized as $\boldsymbol{u}_i = \boldsymbol{\Gamma}\boldsymbol{b}_i$ where $\boldsymbol{b}_i \sim \mathrm{N}_{Kq}(\mathbf{0}, \boldsymbol{I}_{Kq})$. Then model (2) becomes

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{b}_i + \boldsymbol{T}_i\boldsymbol{\alpha} + \boldsymbol{\epsilon}_i. \tag{4}$$

2.2 Penalized Likelihood

We employ a penalized regression method for simultaneous selection of fixed effects, random effects and joint nonlinear effects of two independent variables in the smooth functions. In particular, we aim to determine whether there are within-subject correlations across the outcomes by identifying the nonzero elements of $\boldsymbol{D}$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T, \boldsymbol{\sigma}^T)^T$ be a vector of all unknown parameters, where $\boldsymbol{\gamma}$ is a $\frac{Kq(Kq+1)}{2} \times 1$ vector of parameters of $\boldsymbol{\Gamma}$, and $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_K^2)$. We propose to maximize the following penalized (observed) log-likelihood function:

$$p\ell_o(\boldsymbol{\theta}) = \ell_o(\boldsymbol{\theta}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{5}$$

where $\ell_o(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \log f_o(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta})$ is the observed log-likelihood function, and $\eta_{\lambda_j}(\cdot)$ $(j = 1, \ldots, K+2)$ are nonnegetive and nondecreasing penalty functions for the fixed effects, random effects and smooth functions, with $\lambda_j > 0$ being the tuning parameters which control the amount of shrinkage.

A number of options for the penalty functions can be considered as discussed in Section 1. Here, we choose to use the adaptive LASSO penalty for its model selection consistency and easy implementation in practice[21]. For the fixed effects, the penalty function is defined as $\eta_{\lambda_1}(\boldsymbol{\beta}) = \lambda_1 \sum_{k=1}^{K}\sum_{l=1}^{p} |\widetilde{\beta}_{kl}|^{-1}|\beta_{kl}|$, where $\widetilde{\beta}_{kl}$ is the unpenalized maximum likelihood estimator (MLE). Note that it may not be necessary to penalize all of the fixed effect coefficients, for example, the intercept can be left out of the penalty function.

Selecting random effects in a multivariate model involves identifying important random effects for each outcome and determining the correlation structure across the outcomes. Similar to equation (3), we partition $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & & & \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} & & \\ \vdots & \vdots & \ddots & \\ \boldsymbol{\Gamma}_{K1} & \boldsymbol{\Gamma}_{K2} & \cdots & \boldsymbol{\Gamma}_{KK} \end{pmatrix},$$

where $\boldsymbol{\Gamma}_{jk}$ $(j = 1, \ldots, K,\ k = 1, \ldots, j)$ are $q \times q$ submatrices, and the diagonal submatrices $\boldsymbol{\Gamma}_{jj}$ are lower triangular. For a single outcome, the penalty is placed on the row vectors of $\boldsymbol{\Gamma}_{jj}$ in a grouped manner to ensure the positive semidefiniteness of $\boldsymbol{D}$[22,28]. If the elements of a certain row of $\boldsymbol{\Gamma}_{jj}$ are all shrunk to zero, the corresponding row and column vectors of $\boldsymbol{D}_{jj}$ will also be zero and thus the corresponding random effect will be removed from the model. In a multivariate setting, we propose to use the following penalty function: $\eta_{\lambda_2}(\boldsymbol{\gamma}) = \lambda_2 \sum_{j=1}^{K}\sum_{k=1}^{j}\sum_{l=1}^{q} \sqrt{c_{jkl}}||\widetilde{\gamma}_{jkl}||^{-1}||\gamma_{jkl}||$, where γ_{jkl} is the lth row of the submatrix $\boldsymbol{\Gamma}_{jk}$, $\widetilde{\gamma}_{jkl}$ is the unpenalized MLE, $||\widetilde{\gamma}_{jkl}|| = (\widetilde{\gamma}_{jkl}^T\widetilde{\gamma}_{jkl})^{1/2}$, and c_{jkl} is a normalizing constant to adjust for the varying sizes of γ_{jkl} (e.g., $c_{jkl} = dim(\gamma_{jkl})$). To account for the temporal correlations of the longitudinal measures, we do not penalize the random intercept for each outcome (i.e., γ_{jj1}, $j = 1, \ldots, K$), so the corresponding penalty terms are removed from $\eta_{\lambda_2}(\boldsymbol{\gamma})$. The proposed penalty function differs from the existing method in that instead of penalizing the row vectors of $\boldsymbol{\Gamma}$, we separate the penalization of the diagonal submatrices

$\boldsymbol{\Gamma}_{jj}$ and the off-diagonal submatrices $\boldsymbol{\Gamma}_{jk}$ $(j \neq k)$, so that the non-random elements in the within- and between-outcome variance components can be identified individually. Group penalties are imposed on $\boldsymbol{\Gamma}_{jj}$ for selecting the random effects within each outcome, whereas $\boldsymbol{\Gamma}_{jk}$ $(j \neq k)$ are penalized in order to determine the pairwise correlations between the outcomes.

To select the joint nonlinear effects in the smooth functions, we impose group penalties on the corresponding product terms in the tensor product basis. For s_k, the penalty function is defined as $\eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k) = \lambda_{2+k}||\widetilde{\alpha}_k^*||^{-1}||\alpha_k^*||$, where α_k^* is a $(L_1-1)(L_2-1)\times 1$ vector consisting of $\alpha_{l_1,l_2,k}$, $l_1 = 2,\ldots,L_1$, $l_2 = 2,\ldots,L_2$, and $\widetilde{\alpha}_k^*$ is the unpenalized MLE. Note that we allow different degrees of penalty for smooth functions s_k in penalized likelihood (5) through different tuning parameters λ_{2+k}, $k = 1, \ldots, K$. Depending on the research context, it may be appropriate to assume a common tuning parameter for smooth functions across multiple outcomes.

3 Computational Algorithm

The model selection and estimation procedure is implemented in two stages. In Stage 1, model selection is performed by maximizing the penalized likelihood function. Given a set of tuning parameters $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^{K+2}$, we use an EM algorithm to optimize (5) and obtain the maximum penalized likelihood estimator (MPLE) $\hat{\boldsymbol{\theta}}_\lambda$. The optimization procedure is carried out for different values of $\boldsymbol{\lambda}$. The optimal $\boldsymbol{\lambda}$ is selected based on a certain criterion, which will be discussed in detail in Section 3.2. The final form of the model will be determined according to the nonzero elements of $\hat{\boldsymbol{\theta}}_\lambda$. In Stage 2, we refit the model with selected fixed effects, random effects and smooth functions (with or without interactions) to obtain the MLE $\hat{\boldsymbol{\theta}}$.

3.1 EM algorithm

Consider $(\boldsymbol{Y}_i, \boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1,\ldots,m$, as the complete data and observed data, respectively. With the same penalty functions in (5), we write the penalized complete log-likelihood function as

$$p\ell_c(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{6}$$

where $\ell_c(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \log f_c(\boldsymbol{Y}_i, \boldsymbol{b}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta})$ is the complete log-likelihood function.

Denote the estimator of $\boldsymbol{\theta}$ at the sth iteration by $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)T}, \boldsymbol{\gamma}^{(s)T}, \boldsymbol{\alpha}^{(s)T}, \boldsymbol{\sigma}^{(s)T})^T$. In the E-step, for a given set of tuning parameters $\boldsymbol{\lambda}$, we calculate the expectation of the penalized complete log-likelihood (6) given the observed data and $\boldsymbol{\theta}^{(s)}$ as follows:

$$\begin{aligned} Q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) &= E[p\ell_c(\boldsymbol{\theta})|(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)_{i=1}^m, \boldsymbol{\theta}^{(s)}] \\ &= \frac{1}{m}\sum_{i=1}^m E[\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] \\ &\quad + \frac{1}{m}\sum_{i=1}^m E[\log f_b(\boldsymbol{b}_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] \\ &\quad - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^K \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \end{aligned} \tag{7}$$

where $f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) = \mathrm{N}_{Kn_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{b}_i + \boldsymbol{T}_i\boldsymbol{\alpha}, \boldsymbol{\Sigma}_i)$, and $f_b(\boldsymbol{b}_i) = \mathrm{N}_{Kq}(\boldsymbol{0}, \boldsymbol{I}_{Kq})$. Let $g_1(\boldsymbol{b}_i, \boldsymbol{\theta}) = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$, and $g_2(\boldsymbol{b}_i) = \log f_b(\boldsymbol{b}_i)$. The two terms of expectation in (7) can be written as

$$E[g_1(\boldsymbol{b}_i, \boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] = \int g_1(\boldsymbol{b}_i, \boldsymbol{\theta}) h(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) d\boldsymbol{b}_i, \tag{8}$$

and

$$E[g_2(\boldsymbol{b}_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] = \int g_2(\boldsymbol{b}_i) h(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) d\boldsymbol{b}_i, \tag{9}$$

where

$$\begin{aligned} h(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) &= \frac{f_c(\boldsymbol{Y}_i, \boldsymbol{b}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)})}{f_o(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)})} \\ &= \frac{f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) f_b(\boldsymbol{b}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)})}{\int f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) f_b(\boldsymbol{b}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}) d\boldsymbol{b}_i}. \end{aligned} \tag{10}$$

The q-dimensional integrals in (8), (9) and the denominator of (10) are usually intractable, so we use multivariate Gauss-Hermite quadrature rules[40] to for approximation. Denote the number of quadrature nodes for each dimension by n. Let $\boldsymbol{b}_d$ and w_d be the pre-specified quadrature nodes and weights respectively, $d = 1, \ldots, N_{\mathrm{GH}}$ where the total number of quadrature nodes is $N_{\mathrm{GH}} = n^q$. The first expectation term (8) can be approximated as

$$E[g_1(\boldsymbol{b}_i, \boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] \approx \sum_{d=1}^{N_{\mathrm{GH}}} w_d \exp(||\boldsymbol{b}_d||^2) g_1(\boldsymbol{b}_d, \boldsymbol{\theta}) h(\boldsymbol{b}_d|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}). \tag{11}$$

Since the second expectation term does not involve $\boldsymbol{\theta}$, it can be omitted in the M-step from the penalized Q-function (7), and thus we need to find $\boldsymbol{\theta}^{(s+1)}$ by maximizing

$$Q^*_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{12}$$

where $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{m}\sum_{i=1}^{m} E[g_1(\boldsymbol{b}_i, \boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}]$.

Considering that maximizing (12) with respect to $\boldsymbol{\theta}$ involves high-dimensional optimization, we propose the following expectation-conditional maximization (ECM)[41] algorithm which breaks down the M-step into several conditional maximization (CM) steps:

1. Given $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\sigma}^{(s)}$, find

$$\boldsymbol{\beta}^{(s+1)} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}|\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m\eta_{\lambda_1}(\boldsymbol{\beta}).$$

2. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\sigma}^{(s)}$, find

$$\boldsymbol{\gamma}^{(s+1)} = \arg\max_{\boldsymbol{\gamma}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m\eta_{\lambda_2}(\boldsymbol{\gamma}).$$

3. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\sigma}^{(s)}$, find

$$\boldsymbol{\alpha}^{(s+1)} = \arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^{(s)}|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m\sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k).$$

4. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\alpha}^{(s+1)}$, find

$$\boldsymbol{\sigma}^{(s+1)} = \arg\max_{\boldsymbol{\sigma}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\sigma}|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\sigma}^{(s)}).$$

5. Iterate the above steps until convergence to obtain the MPLE $\hat{\boldsymbol{\theta}}_\lambda$.

To start the optimization procedure, we fit the full model with all covariates and use the unpenalized parameter estimates as the initial values.

3.2 Tuning Parameter Selection

The performance of the proposed method depends on the appropriate selection of tuning parameters. Selection criteria that have been extensively used include cross validation (CV), generalized

cross-validation (GCV), and information criterion such as Akaike and the Bayesian information criteria (AIC and BIC)[42,43]. Here, we utilize the following form of BIC to select the optimal tuning parameters:

$$BIC_{\lambda} = -2\ell_o(\hat{\boldsymbol{\theta}}_{\lambda}) + \log(N)df_{\lambda},$$

where $\ell_o(\hat{\boldsymbol{\theta}}_{\lambda})$ is the value of the observed log-likelihood at the MPLE $\hat{\boldsymbol{\theta}}_{\lambda}$ obtained through the proposed EM algorithm for given $\boldsymbol{\lambda}$. In practice, $\ell_o(\hat{\boldsymbol{\theta}}_{\lambda})$ is approximated using the Gauss-Hermite quadrature rules described in Section 3.1. The sample size N in a multivariate setting is defined as $N = K\sum_{i=1}^{m} n_i$. The degrees of freedom df_{λ} is defined as the number of nonzero elements of $\hat{\boldsymbol{\theta}}_{\lambda}$. The proposed EM algorithm is repeated over a grid of tuning parameters, and the one that minimizes BIC_{λ} is considered optimal. In practice, the multidimensional grid search will be increasingly difficult when the number of outcomes becomes large. One can start with a crude search by using a relatively large interval between grid points to identify the neighborhoods where the optimal tuning parameters potentially reside, and then refine the search within those neighborhoods. The computational burden can also be alleviated by assuming a common tuning parameter for smooth functions across the outcomes in which case the total number of tuning parameter would drop to 3.

3.3 Implementation

The proposed computational algorithm is developed using R software. The M-steps in the EM algorithm are implemented using the `optim` function in the `stats` package[44]. The initial values of the parameters are obtained by fitting the full multivariate semiparametric model using the `gamm4` function in the `gamm4` package[45].

4 Simulation Study

To evaluate the performance of the proposed method, we conduct a simulation study in which we consider two settings. For each setting, we generate bivariate outcomes from the following model:

$$\begin{cases} Y_{ij1} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i1} + \bar{s}_1(t_{1ij}, t_{2ij}) + \epsilon_{ij1} \\ Y_{ij2} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i2} + \bar{s}_2(t_{1ij}, t_{2ij}) + \epsilon_{ij2}, \end{cases}$$

for $i = 1, \ldots, 200$ and $j = 1, \ldots, 5$.

In Setting 1, we set the fixed effect coefficients as $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15})^T = (1, 1, 3, 0, -1, 0)^T$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T = (1, 2, 0, -2, 0, 0)^T$. The corresponding covariates $\boldsymbol{x}_{ij} = (x_{ij0}, x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5})^T$ are generated independently from $\mathrm{N}(0, 1)$ except that the intercept $x_{ij0} = 1$. The subject-specific random effects are $(\boldsymbol{u}_{i1}^T, \boldsymbol{u}_{i2}^T)^T = (u_{i10}, u_{i11}, u_{i12}, u_{i20}, u_{i21}, u_{i22})^T \sim \mathrm{N}_6(\mathbf{0}, \boldsymbol{D})$ with

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1.25 & 0.75 & 0 & 0 & 0 \\ 0.5 & 0.75 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and the corresponding covariates $\boldsymbol{z}_{ij} = (z_{ij0}, z_{ij1}, z_{ij2})^T = (x_{ij0}, x_{ij1}, x_{ij2})^T$. Note that the outcomes are independent of each other since the 3×3 off-diagonal submatrices in $\boldsymbol{D}$ are $\mathbf{0}$. The smooth functions are given by $s_1(t_1, t_2) = t_1 + t_2$ and $s_2(t_1, t_2) = t_1 + t_2 + 2\exp(t_1)/(1.2 - t_2)$ with $t_1, t_2 \overset{iid}{\sim} \mathrm{Uniform}(0, 1)$, and $\bar{s}_1(t_{1ij}, t_{2ij})$ and $\bar{s}_2(t_{1ij}, t_{2ij})$ are the values of corresponding smooth functions centered over (t_{1ij}, t_{2ij}). The measurement errors are $\epsilon_{ij1} \sim \mathrm{N}(0, \sigma_1^2)$ and $\epsilon_{ij1} \sim \mathrm{N}(0, \sigma_2^2)$ with $\sigma_1 = 1$ and $\sigma_2 = 1.5$.

In Setting 2, the setup is the same as the previous setting except that the outcomes are correlated with the covariance matrix $\boldsymbol{D}$ given by

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.25 & 0.25 & 0 \\ 0.5 & 1.25 & 0.75 & 0.375 & 0.375 & 0 \\ 0.5 & 0.75 & 1.5 & 0.5 & 0.5 & 0 \\ 0.25 & 0.375 & 0.5 & 1.1875 & 0.6875 & 0 \\ 0.25 & 0.375 & 0.5 & 0.6875 & 0.6875 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

These two settings allow us to assess whether the proposed method can correctly determine the correlation structure between the outcomes.

The simulation is repeated 100 times for each setting using the proposed two-stage procedure. In Stage 1, we identify important fixed effects and random effects, and determine the presence

Table 1: Frequency of correct selection, incorrect inclusion and incorrect exclusion of model components in 100 simulation replications.

	Setting 1			Setting 2		
Model Component	Correct	Incorrect Inclusion	Incorrect Exclusion	Correct	Incorrect Inclusion	Incorrect Exclusion
Model	99	1	0	94	1	6
Fixed Effects	99	1	0	99	1	0
Random Effects	100	0	0	100	0	0
Correlation	100	0	–	94	–	6
s_1	100	0	–	100	0	–
s_2	100	–	0	100	–	0

of joint nonlinear effects in the bivariate functions. The between-outcome correlation structure is determined as part of the random effects selection. Then we estimate the selected effects and bivariate surfaces (or two additive univariate smooth functions) in Stage 2. For estimation, we use the same algorithm as described in Section 3.1, but remove the penalty terms from the likelihood function (6). In both settings, 4 quadrature nodes are used for each random effect; quadratic splines with 3 knots are used as the marginal basis for bivariate smooth functions.

Table 1 summarizes the selection results for the whole model and its components, including the fixed effects, the random effects, the interaction effects in the smooth functions, and the correlation between the outcomes. We report the numbers of times the model and individual components are correctly identified, as well as the frequencies of incorrect inclusion (an unimportant effect being selected) and incorrect exclusion (an important effect not being selected). Under both settings, the proposed algorithm achieves high rates of correct selection of the true model and its components. Specifically, it is able to identify the true fixed effects, random effects as well as the interactions with a correct selection rate $\geq 99\%$. The performance in terms of determining the between-outcome correlation is also satisfactory, although errors mostly occur when the outcomes are truly correlated.

Tables 2 and 3 show the estimated fixed effect coefficients and variance components from Stage 2.

The magnitude of bias in the estimation of non-zero parameters is generally small. Only

Table 2: Estimates of the fixed effect coefficients with empirical standard errors (SEs) averaged over 100 simulation replications.

		Setting 1		Setting 2	
Parameter	True Value	Estimate	SE	Estimate	SE
β_{10}	1	0.941	0.055	0.985	0.037
β_{11}	1	1.013	0.041	1.079	0.072
β_{12}	3	2.957	0.058	2.955	0.052
β_{13}	0	0	0	0	0
β_{14}	-1	-0.997	0.021	-0.993	0.023
β_{15}	0	0	0	0	0
β_{20}	1	1.062	0.056	0.904	0.059
β_{21}	2	2.077	0.051	1.923	0.064
β_{22}	0	-0.001	0.008	0.001	0.006
β_{23}	-2	-1.953	0.034	-1.908	0.213
β_{24}	0	0	0	0	0
β_{25}	0	0	0	0	0

one of the unimportant covariates was incorrectly included in the model. In addition, the mean squared errors (MSE) of the estimated smooth functions are: in Setting 1, $MSE(\hat{s}_1) = 0.010$, and $MSE(\hat{s}_2) = 0.347$; in Setting 2, $MSE(\hat{s}_1) = 0.039$, and $MSE(\hat{s}_2) = 0.382$. These results support the notion that the two-stage algorithm works well for both model selection and estimation.

Additional simulation studies are conducted to further evaluate the sensitivity of the proposed procedure to different choices of basis function and number of knots for the bivariate smooth functions. Data are simulated under Setting 2 with two correlated outcomes. We consider two choices of marginal basis functions, (1) quadratic splines with 4 knots, and (2) cubic splines with 3 knots. For each scenario we carry out 100 simulation replications. Model selection and estimation results are provided in the Supplement. In Supplementary Table 1, the structures of the bivariate smooth functions are all correctly identified using both types of marginal basis. Combined with the selection results for s_1 and s_2 in Table 1, it suggests that the selection procedure is robust to

Table 3: Estimates of the variance components with empirical standard errors (SEs) averaged over 100 simulation replications.

	Setting 1			Setting 2		
Parameter	True Value	Estimate	SE	True Value	Estimate	SE
D_{11}	1	0.923	0.058	1	0.864	0.137
D_{22}	1.25	1.306	0.084	1.25	1.364	0.098
D_{33}	1.5	1.771	0.165	1.5	1.776	0.127
D_{44}	1	1.044	0.062	1.1875	1.103	0.141
D_{55}	0.5	0.406	0.070	0.6875	0.742	0.074
D_{66}	0	0	0	0	0	0
σ_1	1	0.998	0.018	1	1.029	0.013
σ_2	1.5	1.595	0.047	1.5	1.553	0.040

the choice of basis functions and knots. The rates of correct selection for other model components are all $> 90\%$, and the overall selection performance for the model as a whole is reasonably well. Supplementary Tables 2 and 3 provide the estimation results averaged over the 100 replications. Bias is minimal in the estimated fixed effects and slightly larger for variance components in a magnitude comparable to that of the estimates in Tables 2 and 3. Using quadratic splines with 4 knots, the mean squared errors of the estimated smooth functions are $MSE(\hat{s}_1) = 0.029$ and $MSE(\hat{s}_2) = 0.329$; using cubic splines with 3 knots, $MSE(\hat{s}_1) = 0.027$ and $MSE(\hat{s}_2) = 0.253$. Compared to the estimation results using quadratic splines with 3 knots, there is a decrease in the mean squared error as the number of knots or the order of basis function increases.

The proposed method has a reasonably good computational efficiency. For example, the simulation study was performed on a Dell PowerEdge R930 server with a Linux operating system. The server has 4 Intel Xeon CPU E7-4850 with 8-core processors and 256 GB memory (shared by multiple users). Under Setting 2, it takes about an hour to perform model selection in Stage 1 for a given set of tuning parameters, and another hour for parameter estimation in Stage 2.

5 Data Application

This research is motivated by an ongoing study of blood pressure development in children. In this section, we illustrate the proposed model selection method by analyzing the data from this study. Detailed recruitment protocol of the study can be found in Pratt et al.[46] and the follow-up protocol in Tu et al.[47,48]. Briefly, study subjects were recruited from schools in Indianapolis, which were selected to provide a wide range of socioeconomic status. Enrolled subjects were followed twice a year for measurement of blood pressure, height, weight and upper arm circumference. For blood pressure, three readings were obtained at least two minutes apart, and the average of the last two was taken as the final measurement. Body mass index (BMI) was calculated based on height and weight as follows: BMI (kg/m^2) = weight/height^2. Overnight urine samples were also collected to determine urine volume and excretion rates of sodium and potassium.

A subset of the blood pressure data is used for this illustration. Of the 250 randomly selected subjects, 117 are males and 80 are blacks. The selected data include a total of 1776 follow-up visits, with a range of $1-18$ visits and an average of 7.1 visits per subject. The mean age at enrollment is 9.8 years (standard deviation = 2.7 years).

In this analysis, we consider systolic and diastolic blood pressure as paired outcomes, and use the proposed method to identify covariates associated with the outcomes. We start the model selection process with the following full model:

$$\begin{cases} Y_{ij1} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i1} + s_1(t_{1ij}, t_{2ij}) + \epsilon_{ij1} \\ Y_{ij2} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i2} + s_2(t_{1ij}, t_{2ij}) + \epsilon_{ij2}, \end{cases}$$

where Y_{ij1} and Y_{ij2} are the systolic and diastolic blood pressure respectively for the ith subject measured at the jth visit, $\boldsymbol{x}_{ij}$ is a vector of fixed effect covariates including an intercept, gender (male or female), race (black or other), birth weight (pound), mother's length of pregnancy (month), upper arm circumference (cm), urine volume (L), urinary sodium excretion rate (mmol/mg creatinine) and urine potassium excretion rate (mmol/mg creatinine), $\boldsymbol{z}_{ij}$ is a vector of random effect covariates including an intercept, birth weight and mother's length of pregnancy, $(\boldsymbol{u}_{i1}^T, \boldsymbol{u}_{i2}^T)^T \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$ are the subject-specific random effects, s_1 and s_2 are bivariate smooth functions of age (t_{1ij}) and BMI (t_{2ij}), and $\epsilon_{ij1} \sim \mathrm{N}(0, \sigma_1^2)$ and $\epsilon_{ij2} \sim \mathrm{N}(0, \sigma_2^2)$ are the independent measurement errors. We choose age and BMI as the bivariate nonparametric components because

a preliminary analysis (Figure 1) show that both tend to have nonlinear effects on blood pressure. We herein examine whether they interact with each other.

In Stage 1, the outcomes and the continuous covariates are standardized to ensure numerical stability before selection is performed. In Stage 2, the selected covariates are estimated in the original scale so that the coefficient estimates can be easily interpreted. Standard errors are calculated based on the observed Fisher information evaluated at the MLEs. The model selection and estimation results are summarized in Table 4. Zero estimates indicate that the corresponding covariates are not selected. For systolic blood pressure, the selected fixed effect covariates are gender, race and upper arm circumference; race and upper arm circumference are also selected for diastolic blood pressure. Based on the coefficient estimates, males have significantly higher systolic blood pressure than females. Comparing with other races, blacks tend to have higher systolic and diastolic blood pressure. Upper arm circumference, an indicator of obesity, is positively associated with both systolic and diastolic blood pressure.

As to the random effects, neither of the covariates, birth weight and mother's length of pregnancy, are selected for systolic or diastolic blood pressure. Table 4 provides the variance component estimates, i.e., square roots of the diagonal elements of $\boldsymbol{D}$, as well as σ_1 and σ_2. In addition, the systolic and diastolic blood pressure are highly correlated within each subject ($\rho = 0.78, \text{SE} = 0.074$), as suggested by the non-zero estimate of the off-diagonal element of $\boldsymbol{D}$.

The estimated bivariate smooth functions s_1 and s_2 are presented using the contour plots in Figure 2. Generally speaking, the systolic and diastolic blood pressure increase with both age and BMI. We also note that there are substantial interactions between the two. Specifically, the effect of BMI on blood pressure is stronger in older children over 12 years of age than in younger children. These observations lend support to the inclusion of bivariate smooth functions for depiction of the joint influences of age and BMI.

6 Discussion

The ability of determining the relevance of independent variables to outcomes of interest and incorporating variables in appropriate functional forms is of vital importance in scientific investigations. Model selection has long presented challenges to data analysts, who often struggled to

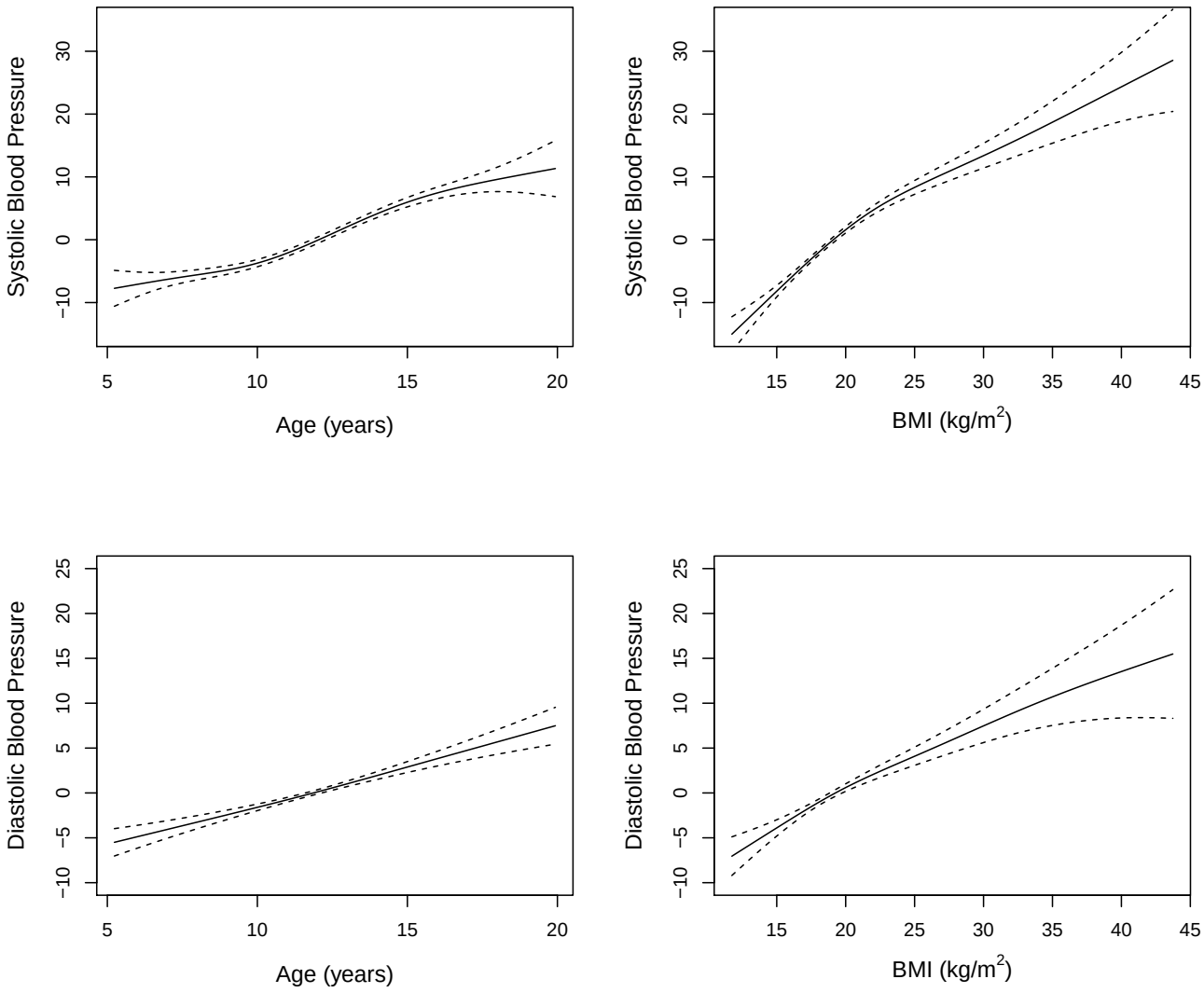

Figure 1: Marginal effects of age and BMI on systolic and diastolic blood pressure (subject to the centering constraint) (solid lines) with 95% confidence bands (dashed lines).

Table 4: Model selection and estimation results for the blood pressure data.

	Systolic Blood Pressure		Diastolic Blood Pressure	
Variable	Estimate	SE	Estimate	SE
Fixed effects				
Intercept	90.52	1.25	48.13	1.34
Male	2.69	0.45	0	–
Black	2.23	0.51	2.01	0.55
Birth weight	0	–	0	–
Mother's length of pregnency	0	–	0	–
Upper arm circumference	0.44	0.054	0.54	0.058
Urine volume	0	–	0	–
Urine sodium concentration	0	–	0	–
Urine potassium concentration	0	–	0	–
Variance components				
Intercept	6.09	0.46	5.99	0.49
Birth weight	0	–	0	–
Mother's length of pregnency	0	–	0	–
Error term	9.28	0.16	10.04	0.17

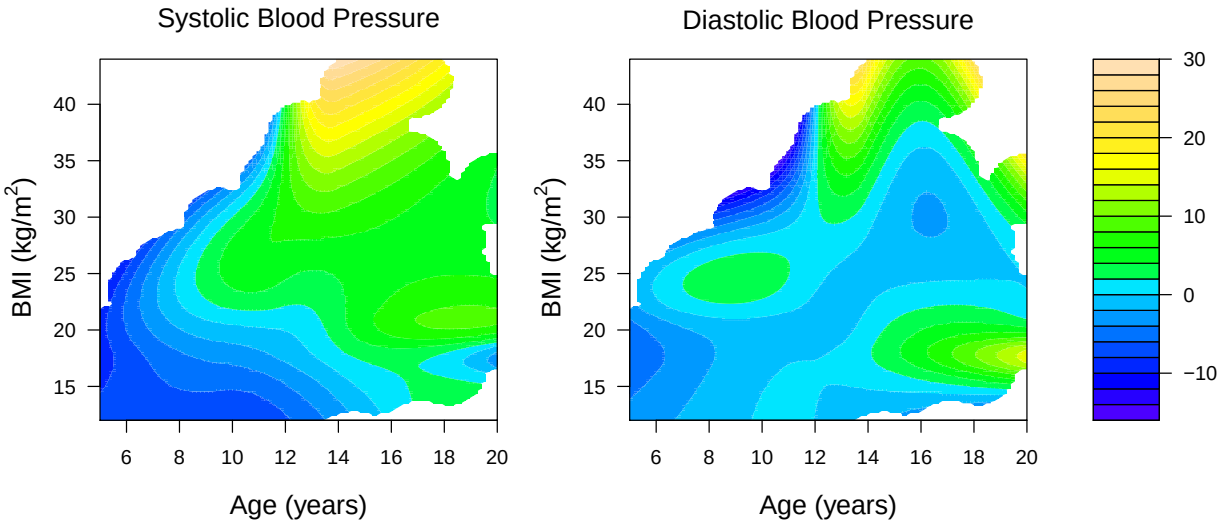

Figure 2: Contour plot of the estimated joint effects of age and BMI on systolic and diastolic blood pressure (subject to the centering constraint).

find appropriate selection methods and implementation algorithms. The situation has improved considerably in the last two decades since the publication of the LASSO by Tibshirani[17,49], which helps to lay the theoretical foundation of regularization methods. Applications of LASSO to various models have since alleviated barriers for model selection in most standard settings. This said, significant challenges remain for newly developed statistical models, such as the one discussed in the current paper.

In this paper, we consider model selection in multivariate semiparametric regression, a newer class of models that have been shown to be useful, yet for which selection methods have not been made available. To remedy, we present a two-stage model selection and estimation method for selecting fixed and random effects and for determining the presence of joint nonlinear effects in the form of bivariate smooth functions. To the best of our knowledge, the proposed selection method is the first for such models. In fact, there are relative few formal discussions of model selection in the context of multivariate models, a setting for which random effect selection plays an important role of determining unknown correlation structures. In other words, selection of random effects helps analysts to decide whether considering multiple outcomes in a joint structure is necessary. Along the same vein, selection of joint nonlinear effects depicted by bivariate surfaces is equally critical

because it facilitates the understanding of the concurrent influences of two independent variables. Our own scientific investigations have repeatedly demonstrated the scarcity of true linear effects in biological research and the danger of over-simplification with linear approximations. Through simulation studies, the proposed method has shown an excellent performance in addressing those issues.

This said, we are cognizant that post-selection inference is an important yet understudied topic for almost all data-driven model selection methods. The uncertainty involved in the model selection process needs to be accounted for in the subsequent inference such as confidence intervals and hypothesis testing. In linear regression, valid confidence intervals can be constructed by satisfying simultaneous coverage for all coefficients of the selected model or conditional coverage that conditions on the model being selected[50]. Further work is needed for making post-selection inference in the model setting discussed in this paper. Other future extensions can be considered to make the proposed method more widely applicable in situations where non-normal outcomes are of interest. We anticipate the extensions to be straightforward, although nontrivial. Notwithstanding these limitations, we believe the proposed method could be of use in a wide variety of investigations.

Funding

This research is partially funded by National Institutes of Health grants RO1 HL095086, P30 HS024384, and R01 AA025208.

Acknowledgement

The authors thank the Indiana University Centers for Aging Research and Biomedical Informatics for their support of our work.

References

1. Li Z, Liu H, Tu W. A Sexually Transmitted Infection Screening Algorithm Based on Semi-parametric Regression Models. *Stat Med.* 2015; 34(20): 2844–2857.

2. Laird NM, Ware JH. Random-effects Melection for Longitudinal Data. *Biometrics*. 1982; 38(4): 963–974.

3. Reinsel G. Multivariate Repeated-Measurement or Growth Curve Models with Multivariate Random-Effects Covariance Structure. *J Am Stat Assoc*. 1982; 77(377): 190–195.

4. Shah A, Laird N, Schoenfeld D. A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. *J Am Stat Assoc*. 1997; 92(438): 775–779.

5. Papageorgiou G, Hinde J. Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Stat Comput*. 2012; 22(1): 79–92.

6. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics*. 1990; 46: 673–687.

7. Ke C, Wang Y. Semiparametric nonlinear mixed-effects models and their applications. *J Am Stat Assoc*. 2001; 96(456): 1272–1281.

8. Coull BA, Staudenmayer J. Self-Modeling Regression for Multivariate Curve Data. *Stat Sinica*. 2004; 14: 695–711.

9. Liu H, Tu W. A Semiparametric Regression Model for Paired Longitudinal Outcomes with Application in Childhood Blood Pressure Development. *Ann Appl Stat*. 2012; 6(4): 1861–1882.

10. Tu W, Eckert GJ, Pratt JH, Jan Danser AH. Plasma levels of prorenin and renin in blacks and whites: their relative abundance and associations with plasma aldosterone concentration. *Am J Hypertens*. 2012; 25(9): 1030–1034.

11. Yu Z, Eckert GJ, Liu H, Pratt JH, Tu W. Adiposity has unique influence on the renin-aldosterone axis and blood pressure in black children. *J Pediatr*. 2013; 163(5): 1317–1322.

12. Tu W, Eckert GJ, Hannon TS, Liu H, Pratt LM, Wagner MA, et al. Racial differences in sensitivity of blood pressure to aldosterone. *Hypertension*. 2014; 63(6): 1212–1218.

13. Greene WH. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall, 2000.

14. Liang KY, Zeger SL. Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*. 1986; 73(1): 13–22.

15. Lange N, Laird NM. The Effect of Covariance Structures on Variance Estimation in Balance Growth-curve Models with Random Parameters. *J Am Stat Assoc*. 1989; 84(405): 241–247.

16. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*. 2011; 88(4): 973–985.

17. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J Roy Stat Soc B*. 1996; 58(1): 267–288.

18. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J Am Stat Assoc*. 2001; 96(456): 1348–1360.

19. Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Ann Stat*. 2004; 32(2): 407–499.

20. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J Roy Stat Soc B*. 2005; 67(2): 301–320.

21. Zou H. The Adaptive Lasso and Its Oracle Properties. *J Am Stat Assoc*. 2006; 101(476): 1418–1429.

22. Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables. *J Roy Stat Soc B*. 2006; 68(1): 49–67.

23. Wang H, Leng C. A Note on Adaptive Group Lasso. *Comput Stat Data An*. 2008; 52(12): 5277–5286.

24. Wei F, Huang J. Consistent Group Selection in High-Dimensional Linear Regression. *Bernoulli*. 2010; 16(4): 1369–1384.

25. Huang J, Ma S, Xie H, Zhang CH. A Group Bridge Approach for Variable Selection. *Biometrika*. 2009; 96(2): 339–355.

26. Bondell HD, Krishna A, Ghosh SK. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*. 2010; 66(4): 1069–1077.

27. Fan Y, Li R. Variable selection in linear mixed effects models. *Ann Stat*. 2012; 40(4): 2043–2068.

28. Ibrahim JG, Zhu H, Garcia RI, Guo R. Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*. 2011; 67(2): 495–503.

29. Lin Y, Zhang HH. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Ann Stat*. 2006; 34(5): 2272–2297.

30. Huang J, Horowitz JL, Wei F. Variable Selection in Nonparametric Additive Models. *Ann Stat*. 2010; 38(4): 2282–2313.

31. Zhang HH, Cheng G, Liu Y. Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models. *J Am Stat Assoc*. 2011; 106(495): 1099–1112.

32. Du P, Cheng G, Liang H. Semiparametric Regression Models with Additive Nonparametric Components and High Dimensional Parametric Components. *Comput Stat Data An*. 2012; 56(6): 2006–2017.

33. Fan J, Li R. New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *J Am Stat Assoc*. 2004; 99(467): 710–723.

34. Ni X, Zhang D, Zhang HH. Variable Selection in Semiparametric Mixed Models in Longitudinal Studies. *Biometrics*. 2010; 66(1): 79–88.

35. Zhao P, Xue L. Variable Selection in Semiparametric Regression Analysis for Longitudinal Data. *Ann Inst Stat Math*. 2012; 64(1): 213–231.

36. Wang K, Lu L. Simultaneous Structure Estimation and Variable Selection in Partial Linear Varying Coefficient Models for Longitudinal Data. *J Stat Comput Sim*. 2015; 85(7): 1459–1473.

37. Ma S, Song Q, Wang L. Simultaneous Variable Selection and Estimation in Semiparametric Modeling of Longitudinal/Clustered Data. *Bernoulli*. 2013; 19(1): 252–274.

38. Zeng X, Ma S, Qin Y, Li Y. Variable Selection in Strong Hierarchical Semiparametric Models for Longitudinal Data. *Stat Interface*. 2015; 8(3): 355–365.

39. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. New York, NY: Cambridge University Press, 2003.

40. Pinhero JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *J Comput Graph Stat*. 1995; 4(1): 12–35.

41. Meng X, Rubin DB. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*. 1993; 80(2): 267–278.

42. Akaike H. Information Theory and An Extension of the Maximum Likelihood Principle. In: Petrov BN, Csaki F, editors. Int Symp Inform Theory. Akademiai Kiado, Budapest, 1973. p. 267–281.

43. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978; 6(2): 461–464.

44. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2014. Available from: `http://www.R-project.org/`.

45. Wood SN, Scheipl F. gamm4: Generalized additive mixed models using mgcv and lme4, 2014. R package version 0.2-3. Available from: `http://CRAN.R-project.org/package=gamm4`.

46. Pratt JH, Jones JJ, Miller JZ, Wagner MA, Fineberg NS. Racial Differences in Aldosterone Excretion and Plasma Aldosterone Concentrations in Children. *New Engl J Med*. 1989; 321(17): 1152–1157.

47. Tu W, Eckert GJ, Saha C, Pratt JH. Synchronization of adolescent blood pressure and pubertal somatic growth. *J Clin Endocr Metab*. 2009; 94(12): 5019–5022.

48. Tu W, Eckert GJ, DiMeglio LA, Yu Z, Jung J, Pratt JH. Intensified Effect of Adiposity on Blood Pressure in Overweight and Obese Children. *Hypertension*. 2011; 58(5): 818–824.

49. Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med.* 1997; 16(4): 385–395.

50. Berk RB, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. *Ann Stat.* 2013; 41(2): 802–837.

Supplement to "Model selection in multivariate semiparametric regression"

Zhuokai Li, Hai Liu and Wanzhu Tu

Table 1: Frequency of correct selection, incorrect inclusion and incorrect exclusion of model components in 100 simulation replications under Setting 2 with different marginal basis functions for bivariate nonparametric components.

	Quadratic Splines with 4 knots			Cubic Splines with 3 knots		
Model Component	Correct	Incorrect Inclusion	Incorrect Exclusion	Correct	Incorrect Inclusion	Incorrect Exclusion
Model	84	7	9	88	5	7
Fixed Effects	94	6	0	97	3	0
Random Effects	99	1	0	98	2	0
Correlation	91	–	9	93	–	7
s_1	100	0	–	100	0	–
s_2	100	–	0	100	–	0

Table 2: Estimates of the fixed effect coefficients with empirical standard errors (SEs) averaged over 100 simulation replications under Setting 2 with different marginal basis functions for bivariate nonparametric components.

		Quadratic Splines with 4 knots		Cubic Splines with 3 knots	
Parameter	True Value	Estimate	SE	Estimate	SE
$\hat{\beta}_{10}$	1	1.004	0.097	0.998	0.103
$\hat{\beta}_{11}$	1	0.986	0.112	0.974	0.115
$\hat{\beta}_{12}$	3	3.007	0.125	2.999	0.128
$\hat{\beta}_{13}$	0	0	0	0	0
$\hat{\beta}_{14}$	-1	-0.996	0.046	-0.996	0.043
$\hat{\beta}_{15}$	0	0	0	0	0
$\hat{\beta}_{20}$	1	1.007	0.081	1.004	0.082
$\hat{\beta}_{21}$	2	2.010	0.091	2.009	0.098
$\hat{\beta}_{22}$	0	0	0	-0.001	0.007
$\hat{\beta}_{23}$	-2	-1.891	0.406	-1.904	0.375
$\hat{\beta}_{24}$	0	0.002	0.015	0.001	0.006
$\hat{\beta}_{25}$	0	-0.0004	0.025	0.002	0.018

Table 3: Estimates of the variance components with empirical standard errors (SEs) averaged over 100 simulation replications under Setting 2 with different marginal basis functions for bivariate nonparametric components.

		Quadratic Splines with 4 knots		Cubic Splines with 3 knots	
Parameter	True Value	Estimate	SE	Estimate	SE
$\hat{D}_{11}$	1	1.077	0.159	1.089	0.172
$\hat{D}_{22}$	1.25	1.380	0.235	1.404	0.224
$\hat{D}_{33}$	1.5	1.749	0.348	1.750	0.334
$\hat{D}_{44}$	1.1875	1.325	0.212	1.323	0.204
$\hat{D}_{55}$	0.6875	0.767	0.189	0.768	0.188
$\hat{D}_{66}$	0	0.001	0.014	0.003	0.018
$\hat{\sigma}_1$	1	1.020	0.032	1.019	0.030
$\hat{\sigma}_2$	1.5	1.871	0.167	1.857	0.136