USING P-SPLINES TO ESTIMATE NONLINEAR COVARIATE EFFECTS IN LATENT FACTOR MODELS

by

Zhenzhen Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2015

Doctoral Committee:

Associate Professor Brisa Sánchez, Chair
Professor Thomas Braun
Associate Professor Marie O'Neill
Assistant Professor Min Zhang

ACKNOWLEDGEMENTS

I first want to thank my advisor, Dr. Brisa Sánchez, for creating the ideas for my thesis projects and also for editing all the writings. I also want to thank my committee members for meeting with me regularly and providing insightful feedbacks. Particularly, I want to thank Dr. Marie O'Neill for helping me obtain the $PM_{2.5}$ constituent data and the Children's Center at SPH for providing the ELEMENT data. PhD time can be a special period for everyone who undertakes this path. The wilderness moments are sometimes the most memorable. But more memorable are the smiles that guide me in peace. You are the streams, the breezes, the fragrances to my soul. Your little acts of kindness are starlets that light up the dark sky. I am grateful to count all your names (but I have too little time now to do it in alphabetical order), Caroline Cheng, Katie Huang, Min A Jhun, KK Leung, Barbara Scanlon, Yancy Lo, Clement Ma, Lisa Henn, Hsiaomin Huang, Karen Peterson, Caroline Kittle, Steve Gray, Lisa Ruby, Linda Jones, Nada Radakovich, Jincheng Shen, Yumeng Li, Alice Chao, Yibin Jiang, Xingyuan Fan, Branden Stolz, George Dentel, Cari Carson, Leanne Kang, Linda White. Lastly, I want to thank my grandmother for your altruistic love. I am glad God has preserved you till this day and He will unite us in eternity. I also want to thank my aunt who supported me when I first started and my mother who often prays for me. This thesis is too small to reflect God's glory, even though I have been generous with compliments for myself. Still, a lily of the field He knows and adorns. *Soli Deo Gloria!*

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

Latent factor models are commonly used for multivariate data analysis. These models assume that the correlation among observed variables comes from a few shared unobserved latent factors. Latent factors and observed variables are linked through a set of regression models termed "measurement models". In the measurement model, observed variables are regressed on the latent factors, and the regression coefficients are termed factor loadings. Typically the number of latent factors is much smaller than the number of observed variables, so latent factor models can also serve as a data reduction step. After the latent factors are estimated (i.e., factor scores are derived), they can be used as either the outcomes or predictors for further analyses. In this thesis I develop novel approaches to estimate and test nonlinear associations between latent factors and observed variables (i.e., non-constant factor loadings), as well as estimate and test the nonlinear relationship between an observed continuous predictor and multiple observed outcomes that measure a latent factor.

In Chapter II and III I develop strategies to estimate non-constant factor loadings. The models developed in these two chapters are aimed at taking into account the varying correlation among multivariate exposure data (see Figure 2.1 for an example). In the context of latent factor models, the varying correlation suggests the

factor loadings should vary with certain covariates and should be estimated as non-constant. In the literature several papers start to use splines in latent factor models (or more generally, structural equation models). These papers develop methods that can estimate a nonlinear mean trend of the latent factor [1, 2, 3, 4, 5], or a nonlinear relationship between outcome latent factor and predictor latent factor [1, 4, 5, 6], or both. But few have looked at the possibility of using splines to estimate non-constant factor loadings [7]. A key contribution of the first two chapters is the development of methods to estimate non-constant factor loadings.

We let the non-constant factor loadings be composed of two parts, a constant part and a functional part. We estimate the functional part using penalized regression splines (P-spline). Splines have been widely used as an approach to estimating non-parametric functions [8, 9, 10, 11, 12, 13, 14]. There are several different ways of implementing the splines [15]. In this thesis I use P-splines [16, 17] because of their simplicity and flexibility. Other commonly used methods include cubic regression splines [11, 18, 19] and thin plate splines [9, 11, 20]. P-splines are based on B-spline bases [21]. They are smoothed through a difference penalty on adjacent spline coefficients [16]. The order of the difference penalty determines the order of polynomial the splines shrink toward [22].

The functional part of the non-constant factor loadings can depend on one or more covariates. In Chapter II, the functional part is estimated as the sum of a univariate function of time and a bivariate function of spatial coordinates. I use the tensor product [17, 23, 24, 25, 26, 27] of B-spline bases to construct the splines for bivariate functions. In Chapter III, the non-constant factor loadings are univariate functions of age only. As for estimating the smoothing parameters of P-splines, I use generalized cross-validation method (GCV) in Chapter II, and maximum likelihood

method in Chapter III. These techniques have been widely used in generalized additive models [15], but in the context of latent factor models, these techniques are not directly applicable. So a large part of these two chapters is to develop new estimating strategies.

In Chapter II, I use the EM algorithm[28] for estimation, treating the latent factor as missing. Given the latent factor, I estimate the non-constant factor loadings using Ridge regression under given smoothing parameters. A highlight of my algorithm is that it optimizes a GCV-type criterion [11, 12, 29, 30, 31] during each EM iteration so that I can estimate the smoothing parameters based on prediction error using the estimated latent factor. I also use multiple smoothing parameters to penalize the wiggliness of splines that are associated with each covariate (time and spatial coordinates) and Newton's method [15, 32] is used for the optimization. In Section 2.4 and 2.5, I also show how correctly estimating the non-constant factor loadings can help with estimating the latent factor.

In Chapter III, I still use EM algorithm for the estimation, but I estimate the smoothing parameter using maximum likelihood. A key contribution of this approach is that tests of measurement invariance can be developed [33, 34]. Measurement invariance is important because in many studies, especially in psychometric analyses, researchers want to make sure the observed variables measure the same latent factor for all members of the population under study [35, 36]. A non-constant factor loading would suggest the measurement invariance assumption is violated. Usually measurement invariance is tested by stratifying the data into different groups and testing if the factor loadings are significantly different across groups [33, 37, 38, 39]. This method is less efficient if there is reason to suspect the factor loading changes with a continuous variable because some information will be lost when a continuous

variable is arbitrarily divided into categories.

To enable testing in Chapter III, I treat the spline coefficients as random [40, 41, 42, 43] and the smoothing parameter as equal to the variance of the spline coefficients after proper transformation [22, 24, 44]. Testing measurement invariance is thus the same as testing if the variance of the spline coefficients is zero. I develop a likelihood ratio test (LRT) for this hypothesis. However, since both the latent factors and the spline coefficients are unobserved, the marginal likelihood does not have a closed form. Therefore, I use a Monte-Carlo E-step [45] based on the Gibbs sampler. Through simulation (Section 3.4) I show that estimating the factor loading as a continuous function improves the power for testing measurement invariance. I also show that in practice, a pointwise confidence interval can be a simple alternative method for testing because it can be a byproduct of the Monte-Carlo EM algorithm and its power can be as good as the LRT.

In Chapter IV, I use P-splines to estimate an overall nonlinear exposure effect for multiple outcomes. The exposure is only a single predictor, but I have multivariate outcome data. I stay in the framework of a latent factor model which takes into account the correlation among multivariate outcomes. In this chapter my focus is not on the factor loading, but on estimating a nonlinear exposure effect on the mean of the latent factor. Compared with the literature that also adds a nonlinear mean trend for the latent factor using splines in a Bayesian framework, my work in Chapter IV uses restricted maximum likelihood (REML) and linear mixed model to estimate the P-spline. The model is also developed from the perspective of estimating and testing an overall nonlinear exposure effect. Several papers [46, 47, 48] have developed models that aim to improve efficiency through estimating an overall linear exposure effect. Here I extend that idea to nonlinear exposure effects.

In summary, in Chapter II and III I use confirmatory factor analysis to only model multivariate exposure data. My objective is to improve the estimation of the latent factor by introducing non-constant factor loading into the model. The estimated latent factors can be used as predictors for health outcomes in future analyses. In Chapter IV I develop approaches to test for nonlinear effect of observed predictors on a latent factor. Taken together, Chapters II through IV significantly advance the modeling approaches for multivariate data within the framework of latent factor models.

CHAPTER II

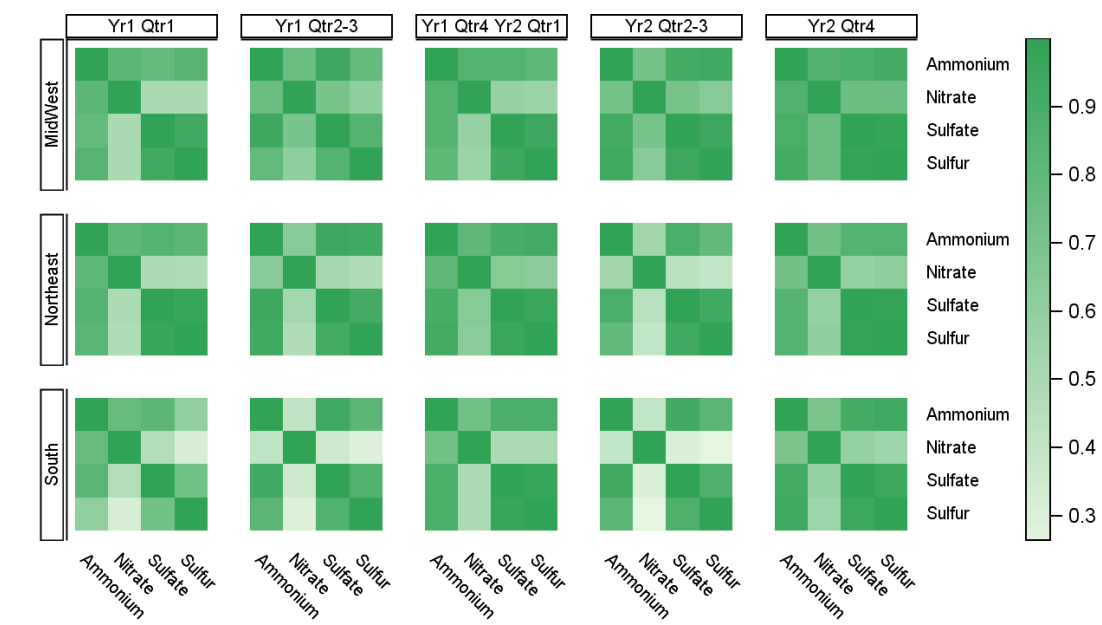
Using a Latent Factor Model with Non-Constant Factor Loadings to Examine the Varying Correlation among Four $PM_{2.5}$ Constituents

2.1 Introduction

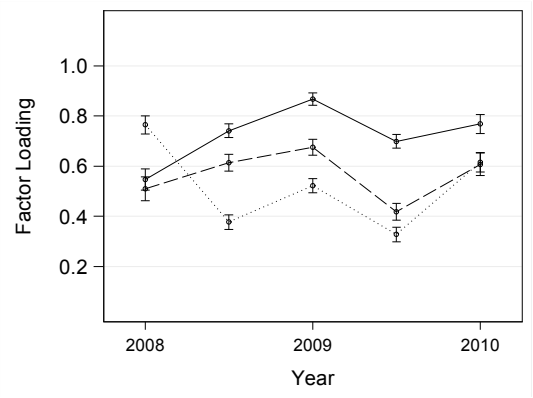
Fine particulate matter $PM_{2.5}$, air particles with aerodynamic diameter of 2.5 micrometers or less, is associated with adverse health effects [49, 50]. The study by Laden et al. [51] suggests $PM_{2.5}$ from different sources has different toxic effects, implying the importance of the chemical profile of the constituents of $PM_{2.5}$. Although the chemical constituents of $PM_{2.5}$ that determine its toxicity are not yet well understood, many different studies have associated several of its constituents, for example, nitrates, elemental carbon, organic carbon, with mortality and various other health risks [52, 53, 54, 55, 56]. Because the constituents are typically correlated, studies that analyze the relationship between constituents and health outcomes often use methods that summarize the correlated constituents, for example, using clustering [57] or latent factor models [2, 58]. In particular, widely used source apportionment methods [59] use latent factor models to summarize constituent data in terms of their relationship to air pollution sources, such as power plant or vehicle exhaust, that can be potentially regulated. Since the air pollution sources are not directly observed, source contributions are treated as latent factors and estimated from the data.

However, methods to summarize $PM_{2.5}$ constituents may insufficiently capture spatial and temporal differences in the correlations among constituents when applied to large geographic areas or long time periods. Summaries depend on the correlation among constituents, and correlations among constituents may vary with time and space [57, 60]. For instance, my preliminary analyses of $PM_{2.5}$ constituent data issued by the EPA for the years 2008 and 2009 (see Section 2.5 for more details on the data) suggests that the correlation among constituents related to secondary inorganic aerosols [59, 61] varies with season and geographic location (Figure 2.1a). Nitrate exhibits the greatest differences in terms of its correlation with other constituents, both across regions and across time within region. As a result, it is not surprising that in preliminary factor analyses of this constituent data, stratified by time and region, factor loadings differ across time and region for some constituents (Figure 2.1b and 2.1c). However, these differences in the factor model structure are not specifically modeled in current literature that uses latent factor models applied to constituent data collected across large geographical areas [59]. Although many models have been proposed to reduce data dimension and to predict outcomes based on many constituents simultaneously [54, 57, 62], a more thorough understanding of how the correlations among $PM_{2.5}$ constituents change with time and space may allow us to gain further insights on the validity and interpretation of such dimension-reduction approaches and thus potentially improve modeling of the association between constituents and health outcomes.

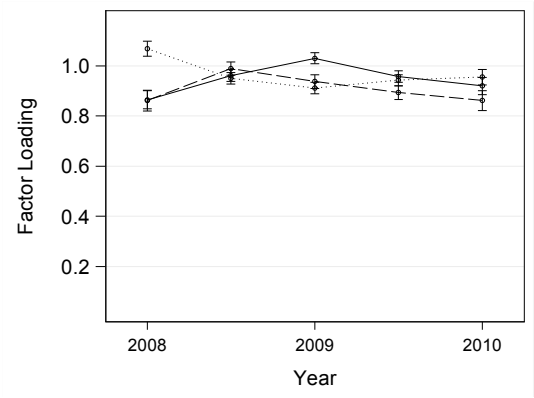
In the literature multiple papers have described models that analyze air pollution data from the perspective of time and space [63, 64, 65, 66]. However, I focus on models utilizing latent factors, given their connection to air pollution sources [58, 67] and because some of these models can be applied to understanding temporal and



(a) Correlation among constituents



(b) Total nitrate



(c) Ammonium ion

Figure 2.1: Preliminary analyses of four $PM_{2.5}$ constituents stratified by time period and region. (a) shows the correlation matrix of the four constituents for strata defined by region (row) and time period (column). Data for the four constituents from 2008 to 2009 are downloaded from the EPA Air Quality System (AQS) website. The header for time denotes the year (Yr) and quarter (Qtr) within that year. The strength of correlation is represented with different shades of green. We can see the differences in correlations are primarily between total nitrate and the other constituents. (b) and (c) compare the factor loadings for a particular constituent from stratified confirmatory factor analyses. The y-axis shows the size of the factor loading. Each dot on the plot is the factor loading from analysis of one stratum. The Midwest, Northeast, and South regions are represented by solid, dashed, dotted lines, respectively. These results suggest there is temporal and spatial difference for the factor loading of total nitrate (b). In comparison, ammonium ion (c) has a rather constant factor loading as the error bars of the different estimates mostly overlap.

spatial patterns in the correlation of $PM_{2.5}$ constituents specifically. For example, Gryparis et al. [2] propose semiparametric latent factor models for modeling multiple surrogates of a single pollution source. They use a penalized spline formulation to allow nonlinear effects of covariates on the mean trend of latent factor and they also use geoadditive models [68] to account for the temporal and spatial correlations in the latent factor. Lopes et al. [69] use a spatial dynamic factor model to incorporate temporal and spatial correlations among observations of a single pollutant from different monitors over a certain time period. In that model, the spatial correlation is introduced through the factor loading matrix. That is, at a given time point, monitors are treated as different variables. The temporal correlation is modeled by an autoregressive structure among the latent factors. Ippoliti et al. [70] use a similar dynamic factor model but they couple two sets of latent factors so that one latent factor could be predicted from the other.

More generally, models using latent factors have gained much popularity in research that relates environmental exposures to health outcomes because they can reduce the dimension of the data by summarizing correlated exposure measurements [71]. In the last decade, much work has been done to relax model assumptions, including nonlinear relationships among latent factors [4, 72, 73, 74, 75, 76], nonlinear links relating observed variables to latent factors [2, 77], as well as models with spatial structures on the latent factors [1, 78, 79].

However, except for the model proposed by Zhang et al. [7], none of these existing advanced models specifically address the variation in the factor loadings exhibited in the constituent data (Figure 2.1b). Traditionally, the measurement model that links the observed variables to the latent factors is assumed constant across the covariate space (e.g., across time and geographic location). A constant measurement model,

also termed measurement model invariance [33, 80], is needed for valid comparison of differences in the latent factor means across covariate levels or groups. However, it has been shown that allowing the measurement model to vary across values of categorical covariates may be desirable when estimating exposure-health outcome associations, potentially because the flexible measurement model enables better estimation of the latent factor itself [81, 82]. Allowing the measurement model to vary across continuous covariates, however, is a more challenging problem.

In this paper I propose a semiparametric latent factor model with non-constant factor loadings that change with multiple covariates, in particular time and space, as motivated by the EPA's $PM_{2.5}$ constituents data. I use P-splines to estimate factor loadings that change with multiple covariates. This construction for the factor loadings bears a resemblance to the classic varying coefficients model [83]. In Section 2.2 I propose the latent factor model that incorporates non-constant factor loadings, while Section 2.3 explains the estimation procedure. A key feature of my estimation algorithm is to include the GCV score within the EM iterations so that I can optimize the smoothing parameters for P-splines in the framework of a latent factor model. In Section 2.4 I present a simulation study to assess the algorithm, and estimation properties for certain model parameters and the estimated factor scores. I then analyze the constituent data set and draw inferences based on bootstrap approaches in Section 2.5. Section 2.6 ends with a discussion and future directions.

2.2 Model

Let $\mathbf{y}_i$ be a vector of P observed variables measured at time t_i and location (u_i, v_i) denoted by two coordinates, typically longitude and latitude. I use a latent factor model to summarize the information of $\mathbf{y}_i$. I introduce non-constant factor

loadings in order to incorporate temporal and spatial differences in the marginal correlations among observed variables and to enable assessment of the invariance of the measurement model across time and space.

I assume the elements of $\mathbf{y}_i$ are linearly related to a latent factor according to the measurement model,

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\lambda}_{\cdot i}\eta_i + \boldsymbol{\epsilon}_i, \tag{2.1}$$

where η_i is a latent factor, $\boldsymbol{\lambda}_{\cdot i} = (\lambda_{1i}, \ldots, \lambda_{pi}, \ldots, \lambda_{Pi})^T$ is a $P \times 1$ vector of factor loadings, $\boldsymbol{\mu}$ is a $P \times 1$ vector of scalars, and $\boldsymbol{\epsilon}_i$ is a $P \times 1$ vector of random errors. I let η_i be an i.i.d. random variable following the standard normal distribution. Since the mean of the latent factor is given, the observed variable means $\boldsymbol{\mu}$ are identifiable [71]. Further, constraining the variance of the latent factor to equal one allows the identification of the factor loadings up to an additional sum to zero constraint (below). The errors $\boldsymbol{\epsilon}_i$ are assumed to be independent of η_i, and to be i.i.d. and follow a multivariate normal distribution, $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_\epsilon)$, where $\boldsymbol{\Sigma}_\epsilon = \text{diag}(\sigma^2_{\epsilon,1}, \ldots, \sigma^2_{\epsilon,P})$. Although a completely unstructured $\boldsymbol{\Sigma}_\epsilon$ would lead to identifiability problems, a diagonal $\boldsymbol{\Sigma}_\epsilon$ is not strictly required but simplifies estimation (Section 3). In my data application the i.i.d. assumption is tenable since I first detrend the observed data (see Section 5 for more details).

2.2.1 Non-constant factor loadings that vary with time and space

In classical latent factor models, the factor loadings are estimated as constant parameters, or sometimes allowed to differ across levels of a categorical variable [80]. However, in my proposed model, factor loadings are allowed to vary smoothly with time and/or location, resembling a varying coefficients model [83].

I model the non-constant factor loadings as a smooth function of (t_i, u_i, v_i). That

is, the value of the p-th factor loading for sample i, λ_{pi}, is specified in the form of an additive model,

$$\lambda_{pi} = \lambda_{p0} + f_{p,t}(t_i) + f_{p,uv}(u_i, v_i) + f_{p,tuv}(t_i, u_i, v_i), \tag{2.2}$$

where $f_{p,t}$, $f_{p,uv}$, $f_{p,tuv}$ are all centered functions (i.e., the sum of the function values across all the data points is zero), and λ_{p0} represents the constant part of the factor loading. The functions $f_{p,t}$, $f_{p,uv}$ respectively capture the main effects of time and location on factor loadings, while $f_{p,tuv}$ tells us whether the location effect changes over time, or whether there is interaction between time and location. I will omit t_i, u_i, v_i in the following notation unless necessary.

I represent the non-constant factor loadings using P-splines. Specifically, I construct univariate cubic B-spline basis functions for each covariate, t, u, v, and use them to construct tensor product bases for the interaction terms $f_{p,uv}(u_i, v_i)$ and $f_{p,tuv}(t_i, u_i, v_i)$ (see Section 4.1.8 of the book by Wood [15]). Collecting the values of the factor loadings corresponding to the p-th observed variable across all observations $i = 1, \ldots, n$ into $\boldsymbol{\lambda}_{p\cdot} = (\lambda_{p1}, ..., \lambda_{pi}, ..., \lambda_{pn})^T$, I can express $\boldsymbol{\lambda}_{p\cdot} = \mathbf{X}_p \boldsymbol{\beta}_p$ where

$$\mathbf{X}_p = [1, \mathbf{X}_t, \mathbf{X}_{uv}, \mathbf{X}_{tuv}], \ \boldsymbol{\beta}_p^T = \left[\lambda_{p0}, \boldsymbol{\beta}_{p,t}^T, \boldsymbol{\beta}_{p,uv}^T, \boldsymbol{\beta}_{p,tuv}^T\right]. \tag{2.3}$$

Note that I use $\boldsymbol{\lambda}_{\cdot i} = (\lambda_{1i}, \ldots, \lambda_{pi}, \ldots, \lambda_{Pi})^T$ to denote values of the P factor loadings for the i-th observation, and $\boldsymbol{\lambda}_{p\cdot} = (\lambda_{p1}, ..., \lambda_{pi}, ..., \lambda_{pn})^T$ as the factor loadings for the p-th observed variable. The design matrix $\mathbf{X}_p$ contains the values of the basis functions evaluated for each observation, with rows corresponding to observations $i = 1, ..., n$, and the columns corresponding to the components in (2.2) which have been suitably centered and constrained to remove redundancies that arise when creating the tensor product bases (see Section 1.8.1 of the book by [15]). Parameters $\boldsymbol{\beta}_p$ are the corresponding coefficients, which, except for the intercept λ_{p0}, are penalized.

I use the squared first difference penalty [16] to control the degrees of freedom of the smooth, which yields the desirable property of shrinking the smooth to the intercept λ_{p0}. The penalty is constructed as a composite penalty that combines separate penalties for each of the components of the smooth in (2.2), that is, the coefficients $\boldsymbol{\beta}_{p,t}, \boldsymbol{\beta}_{p,uv}$, and $\boldsymbol{\beta}_{p,tuv}$ are penalized separately. Furthermore, the interaction terms use tensor product bases and have separate penalties for each covariate conditional on the other covariates in the interaction (see Section 4.1.8 of the book by [15] for a similar construction). In total, the penalty corresponding to each factor loading relies on six distinct smoothing parameters. These penalties are constructed separately for each of the P factor loadings in the model. The penalty for the spline coefficients of the p-th factor loading is $\Omega_p(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p) = \boldsymbol{\beta}_p^T \mathbf{S}_p(\boldsymbol{\omega}_p)\boldsymbol{\beta}_p$, where $\boldsymbol{\omega}_p$ denote the six smoothing parameters, $\mathbf{S}(\boldsymbol{\omega}_p)$ is constructed as a block diagonal matrix whose diagonal blocks are 0, and $\omega_{p,t}\mathbf{S}_t$, $\omega_{p,u|v}\mathbf{S}_{u|v}+\omega_{p,v|u}\mathbf{S}_{v|u}$, $\omega_{p,t|uv}\mathbf{S}_{t|uv}+\omega_{p,u|tv}\mathbf{S}_{u|tv}+\omega_{p,v|tu}\mathbf{S}_{v|tu}$ that correspond to the penalties of $\boldsymbol{\beta}_{p,t}, \boldsymbol{\beta}_{p,uv}$, and $\boldsymbol{\beta}_{p,tuv}$. I define the overall penalty as $\Omega(\boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{p=1}^{P} \boldsymbol{\beta}_p^T \mathbf{S}_p(\boldsymbol{\omega}_p)\boldsymbol{\beta}_p$ where $\boldsymbol{\beta}$ is the vector of spline coefficients for all the factor loadings combined and $\boldsymbol{\omega}$ denotes all the penalty parameters.

2.3 Estimation

I use the EM algorithm to facilitate the estimation, where the complete data structure includes the observed data $\mathbf{y}_i$ augmented with unobserved variables η_i. The penalized log-likelihood of the augmented data is

$$\log \mathcal{L}(\boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\eta}) - \frac{1}{2}\sum_{p=1}^{P} \Omega_p(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p), \tag{2.4}$$

where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, and $\boldsymbol{\eta} = (\eta_1, \ldots \eta_n)$, respectively, and $\boldsymbol{\theta} = \{\mu_p, \lambda_{p0}, \sigma^2_{\epsilon,p}, \boldsymbol{\beta}_p | \; p = 1, \ldots, P\}$. The E-step computes the expected log-likelihood using the conditional distribution of $\eta_i|\mathbf{y}_i$. The M-step estimates $\boldsymbol{\theta}$ by maximizing the expected

log-likelihood. The smoothing parameters $\boldsymbol{\omega}_p$ are selected within each iteration of the EM.

2.3.1 EM algorithm

Using the assumption of a normal distribution and omitting the proportionality constant, the complete data log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\eta}) \propto -\sum_{i=1}^{n} \left\{ \log |\boldsymbol{\Sigma}_\epsilon| + (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\lambda}_{\cdot i}\eta_i)^T \boldsymbol{\Sigma}_\epsilon^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\lambda}_{\cdot i}\eta_i) + \eta_i^2 \right\}.$$

E-step. Even though I am to maximize the penalized likelihood in (2.4), the penalty does not affect the E-step because the penalty does not involve the missing latent factor. The expected values of η_i and η_i^2 are needed in the E-step. Using the conditional normal distribution, these expectations at the (r)-th iteration are listed as follows. (I omit the superscripts "$(r-1)$" in the formulas below, but the calculations use parameter estimates from the $(r-1)$-th iteration).

$$\begin{aligned}
\eta_i^{(r)} &= \mathrm{E}(\eta_i \,|\, \mathbf{y}_i) \\
&= \mathrm{E}(\eta_i) + \mathrm{Cov}(\eta_i, \mathbf{y}_i)[\mathrm{Cov}(\mathbf{y}_i)]^{-1}[\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i)] \\
&= \boldsymbol{\lambda}_{\cdot i}^T (\boldsymbol{\lambda}_{\cdot i}\boldsymbol{\lambda}_{\cdot i}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}),
\end{aligned}$$

$$\begin{aligned}
(\eta_i^2)^{(r)} &= \mathrm{E}(\eta_i^2 \,|\, \mathbf{y}_i) \\
&= \mathrm{Cov}(\eta_i \,|\, \mathbf{y}_i) + [\mathrm{E}(\eta_i \,|\, \mathbf{y}_i)]^2 \\
&= \mathrm{Cov}(\eta_i) - \mathrm{Cov}(\eta_i, \mathbf{y}_i)[\mathrm{Cov}(\mathbf{y}_i)]^{-1}\mathrm{Cov}(\mathbf{y}_i, \eta_i) + (\eta_i^{(r)})^2 \\
&= 1 - \boldsymbol{\lambda}_{\cdot i}^T (\boldsymbol{\lambda}_{\cdot i}\boldsymbol{\lambda}_{\cdot i}^T + \boldsymbol{\Sigma}_\epsilon)^{-1} \boldsymbol{\lambda}_{\cdot i} + (\eta_i^{(r)})^2.
\end{aligned}$$

M-step. The M-step estimates $\boldsymbol{\theta}$ from the first part of (2.4), including the penalty. Given the estimated latent factor at the (r)-th iteration, this step can be thought of as fitting P varying coefficient models, one for each of the P observed variables

(e.g., for each pollutant). In order to see this, first I rearrange the log-likelihood as below, which is possible under the assumption that $\mathbf{\Sigma}_\epsilon$ is a diagonal matrix.

Let y_{pi} denote each of the P observed variables in $\mathbf{y}_i$, then I can write the objective penalized log-likelihood as

$$\begin{aligned} & \textstyle\sum_{i=1}^{n}\sum_{p=1}^{P}\log L\left(\mu_p,\lambda_{pi},\sigma^2_{\epsilon,p};\, y_{pi}\,|\,\eta_i^{(r)}\right)-\tfrac{1}{2}\sum_{p=1}^{P}\Omega(\boldsymbol{\beta}_p,\boldsymbol{\omega}_p) \\ = & \textstyle\sum_{p=1}^{P}\left\{\log L\left(\mu_p,\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p),\sigma^2_{\epsilon,p};\,\mathbf{y}_{p.}\,|\,\boldsymbol{\eta}^{(r)}\right)-\tfrac{1}{2}\Omega(\boldsymbol{\beta}_p,\boldsymbol{\omega}_p)\right\}. \end{aligned} \tag{2.5}$$

where $\mathbf{y}_{p.}=(y_{p1},\ldots,y_{pn})^T$ and $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_n)^T$ are $n\times 1$ column vectors. This allows us to maximize (2.5) separately for each p, that is, for each observed variable. If $\mathbf{\Sigma}_\epsilon$ is not diagonal, this step would instead be analogous to multivariate regression.

For each observed variable $\mathbf{y}_{p.}$, the penalized log-likelihood given $\boldsymbol{\eta}^{(r)}$ is

$$\begin{aligned} & \log L\left(\mu_p,\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p),\sigma^2_{\epsilon,p};\,\mathbf{y}_{p.}\,|\,\boldsymbol{\eta}^{(r)}\right)-\tfrac{1}{2}\Omega(\boldsymbol{\beta}_p,\boldsymbol{\omega}_p) \\ \propto\ & -n\log(\sigma^2_{\epsilon,p})-\sigma^{-2}_{\epsilon,p}\left[\mathbf{y}_{p.}-\mathbf{1}_n\mu_p-\boldsymbol{\eta}^{(r)}\circ\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p)\right]^T\left[\mathbf{y}_{p.}-\mathbf{1}_n\mu_p-\boldsymbol{\eta}^{(r)}\circ\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p)\right] \\ & -\boldsymbol{\beta}_p^T\mathbf{S}(\boldsymbol{\omega}_p)\boldsymbol{\beta}_p, \end{aligned} \tag{2.6}$$

where "$\circ$" is the entry-wise Hadamard product and $\mathbf{S}(\boldsymbol{\omega}_p)$ is the previously described penalty matrix. Equation (2.6) is essentially a regression of $\mathbf{y}_{p.}$ on $\tilde{\boldsymbol{\eta}}^{(r)}$ with $\boldsymbol{\beta}_p$ penalized, where

$$\tilde{\boldsymbol{\eta}}^{(r)}=\begin{bmatrix}\eta_1^{(r)}\mathbf{x}_1\\ \vdots\\ \eta_n^{(r)}\mathbf{x}_n\end{bmatrix}. \tag{2.7}$$

That is, I can write $\boldsymbol{\eta}^{(r)}\circ\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p)=\boldsymbol{\eta}^{(r)}\circ(\mathbf{X}_p\boldsymbol{\beta}_p)=\begin{bmatrix}\eta_1^{(r)}\mathbf{x}_1\boldsymbol{\beta}_p\\ \vdots\\ \eta_n^{(r)}\mathbf{x}_n\boldsymbol{\beta}_p\end{bmatrix}=\tilde{\boldsymbol{\eta}}^{(r)}\boldsymbol{\beta}_p$. Hence, replacing $\boldsymbol{\eta}^{(r)}\circ\boldsymbol{\lambda}_{p.}(\boldsymbol{\beta}_p)$ with $\tilde{\boldsymbol{\eta}}^{(r)}\boldsymbol{\beta}_p$ in (2.6) and maximizing the penalized log-likelihood,

I then obtain a Ridge regression solution for the coefficients

$$\begin{bmatrix} \mu_p^{(r)} \\ \boldsymbol{\beta}_p^{(r)} \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{1}_n^T \mathbf{1}_n & \mathbf{1}_n^T \tilde{\boldsymbol{\eta}}^{(r)} \\ (\tilde{\boldsymbol{\eta}}^T)^{(r)} \mathbf{1}_n & (\tilde{\boldsymbol{\eta}}^T \tilde{\boldsymbol{\eta}})^{(r)} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{S}(\boldsymbol{\omega}_p) \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{1}_n^T \\ (\tilde{\boldsymbol{\eta}}^T)^{(r)} \end{bmatrix} \mathbf{y}_{p.}, \tag{2.8}$$

and

$$\begin{aligned} (\sigma^2_{\epsilon,p})^{(r)} &= n^{-1} \Big[(\mathbf{y}_{p.} - \mathbf{1}_n \mu_p^{(r)})^T (\mathbf{y}_{p.} - \mathbf{1}_n \mu_p^{(r)}) - 2 (\mathbf{y}_{p.} - \mathbf{1}_n \mu_p^{(r)})^T \tilde{\boldsymbol{\eta}}^{(r)} \boldsymbol{\beta}_p^{(r)} \\ &\quad + (\boldsymbol{\beta}_p^T)^{(r)} (\tilde{\boldsymbol{\eta}}^T \tilde{\boldsymbol{\eta}})^{(r)} \boldsymbol{\beta}_p^{(r)} \Big], \end{aligned}$$

where $(\tilde{\boldsymbol{\eta}}^T \tilde{\boldsymbol{\eta}})^{(r)} = \sum_i^n \mathbf{x}_i^T (\eta_i^2)^{(r)} \mathbf{x}_i = [\, \mathbf{x}_1^T (\eta_1^2)^{(r)}, \ldots, \mathbf{x}_n^T (\eta_n^2)^{(r)}] \, \mathbf{X}_p$.

2.3.2 Generalized cross validation

So far I have treated the smoothing parameters $\boldsymbol{\omega}_p$ as given, but here I describe how to use the GCV score to select the smoothing parameters [29, 84]. GCV is commonly used in the selection of smoothing parameters for Generalized Additive Models (GAM, see Section 4.5 of the book by [15]), where the response variable is to be predicted from known covariates using a spline function. GCV penalizes the residual sum of squares with the effective number of parameters (see (2.9) below). The smoothing parameters that give the smallest GCV score are selected. In the measurement model (2.1) the latent factor, though actually $\tilde{\boldsymbol{\eta}}^{(r)}$ according to (2.7), is a natural counterpart of the covariate in a GAM model, so I make use of $\tilde{\boldsymbol{\eta}}^{(r)}$ and have

$$\mathrm{GCV}_p = \frac{n \| \mathbf{y}_{p.} - \mathbf{A}_p^{(r)} \mathbf{y}_{p.} \|^2}{[\, n - \mathrm{tr}(\mathbf{A}_p^{(r)})]^2}, \tag{2.9}$$

where

$$\mathbf{A}_p^{(r)} = \begin{bmatrix} \mathbf{1}_n^T \\ (\tilde{\boldsymbol{\eta}}^T)^{(r)} \end{bmatrix}^T \left\{ \begin{bmatrix} \mathbf{1}_n^T \mathbf{1}_n & \mathbf{1}_n^T \tilde{\boldsymbol{\eta}}^{(r)} \\ (\tilde{\boldsymbol{\eta}}^T)^{(r)} \mathbf{1}_n & (\tilde{\boldsymbol{\eta}}^T \tilde{\boldsymbol{\eta}})^{(r)} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{S}(\boldsymbol{\omega}_p) \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{1}_n^T \\ (\tilde{\boldsymbol{\eta}}^T)^{(r)} \end{bmatrix}$$

is the influence matrix (compare with (2.8)) and $\mathrm{tr}(\mathbf{A}_p)$ represents the effective number of parameters used to estimate $\boldsymbol{\lambda}_{p.}$.

Since the latent factor is unknown and is itself to be computed from $\mathbf{y}_p$ and $\boldsymbol{\beta}^{(r)}$, use of $\|\mathbf{y}_{.p} - \tilde{\boldsymbol{\eta}}^{(r)}\boldsymbol{\beta}^{(r)}\|^2$ is not valid if I evaluate GCV_p after the EM algorithm converges. This is because $\tilde{\boldsymbol{\eta}}^{(r)}$ from the last iteration varies with the smoothing parameters. Instead, I evaluate GCV_p within each iteration of the EM algorithm where $\tilde{\boldsymbol{\eta}}^{(r)}$ is held as the fixed covariate in the regression-based M-step, and update the smoothing parameters with $(\boldsymbol{\omega}_p)^{(r)}$ that gives the smallest GCV_p for each M-step. This approach of selecting smoothing parameters for each EM iteration is similar to the performance iteration method used in GAM models (see Section 4.6 of the book by Wood [15]).

I minimize GCV_p using a numerical approach based on Newton's method. Here GCV_p is regarded as a function of $\boldsymbol{\rho}_p = \log(\boldsymbol{\omega}_p)$ because $\boldsymbol{\omega}_p$ has to be constrained as positive. $\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ is approximated around the current estimate of $(\boldsymbol{\rho}_p)^{(r)}$ with a quadratic function based on the first and second derivatives of GCV_p with regard to $\boldsymbol{\rho}_p$. Then the minimum of $\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ can be approached by successively solving the quadratic minimum. Thus, the M-step is modified as follows. I first solve $\boldsymbol{\beta}_p^{(r)}$ with $(\boldsymbol{\omega}_p)^{(r-1)} = \exp\left((\boldsymbol{\rho}_p)^{(r-1)}\right)$ plugged in (2.8). Then I evaluate $\frac{\partial}{\partial \boldsymbol{\rho}_p}\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ and $\frac{\partial^2}{\partial (\rho_p^\beta)^2}\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ at $(\boldsymbol{\rho}_p)^{(r-1)}$, obtain the critical point of the quadratic approximation and reevaluate the derivatives of $\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ again at the critical point. The process is repeated until I get to the minimum of $\mathrm{GCV}_p(\boldsymbol{\rho}_p)$ and then $(\boldsymbol{\rho}_p)^{(r-1)}$ is updated to $(\boldsymbol{\rho}_p)^{(r)}$.

2.3.3 Simultaneous confidence band

I use bootstrap resampling to construct a 95% confidence interval for point estimates and 95% simultaneous confidence band for spline estimates [40]. I sample with

replacement from the n observations to generate M new data sets of n observations each. In my data application, this resampling approach is valid since detrending the data removes spatial and temporal correlation. The same EM algorithm is repeated on these data sets to get bootstrap estimates. This bootstrap strategy is based on the assumption that there is little temporal and spatial correlation for the residual error $\boldsymbol{\epsilon}_i$ of the actual data.

For single point estimates, such as λ_{p0}, the 2.5% and 97.5% quantile of the bootstrap samples is taken as the bounds of the 95% confidence interval. For spline estimates, like $\hat{f}_{p,t}$, I first obtain C, the 95% quantile of

$$\sup_{t\in\mathbf{t}} \left| \frac{\hat{f}^*_{p,t}(t) - \hat{f}_{p,t}(t)}{\sqrt{\text{Var}[\hat{f}^*_{p,t}(t) - \hat{f}_{p,t}(t)]}} \right|, \tag{2.10}$$

where $\hat{f}^*_{p,t}$ is the bootstrap estimate of $f_{p,t}$ and $\mathbf{t}$ is the support of t. Then the 95% simultaneous confidence band is constructed as $\hat{f}_{p,t}(t) \pm C\sqrt{\text{Var}[\hat{f}^*_{p,t}(t) - \hat{f}_{p,t}(t)]}$. I obtain $\text{Var}[\hat{f}^*_{p,t}(t) - \hat{f}_{p,t}(t)]$ also from the bootstrap samples.

2.4 Simulation

I conduct simulation studies to examine the advantages and disadvantages of this novel estimation method, as well as verify that the estimation algorithm can recover model parameters. I first test the algorithm through simulations where I let the non-constant factor loadings change with three covariates in this form, $\lambda_{pi} = \lambda_{p0} + f_{p,uv}(u_i, v_i) + f_{p,t}(t_i)$. This model is similar to the one estimated for my motivating example. I generate data on an 11 by 11 grid (i.e., 121 (u_i, v_i)), and for each point on the grid I have 101 time points. Each simulated data set has four observed variables and a sample size of $11 \times 11 \times 101 = 12,221$, which is a similar size to the constituent data set I use. The simulation shows that my algorithm can correctly

estimate non-constant factor loadings that change with multiple covariates, although some oversmoothing for $f_{p,uv}$ is observed. Estimates for the other parameters are also unbiased.

I further investigate how using non-constant factor loadings affects the estimation of model parameters and the factor scores through different simulation scenarios. Examining estimation of factor scores gives insight as to how estimated source contributions in air pollution studies may be influenced under various scenarios, for instance. In this second set of simulations the model again includes one latent factor associated with P observed variables, and I let the non-constant factor loadings vary with one covariate, $t_i \in [0, 1]$. I let λ_{pi} denote the factor loading for the p-th observed variable evaluated at t_i.

The simulations are set up with three different shapes for the true factor loadings. I let $\lambda_{pi} = \lambda_{p0} + f_p(t_i)$, where f_p is a centered function (i.e., $\sum_{i=1}^{n} f_p(t_i) = 0$), and the three types of $f_p(t_i)$ are as follows, for $p = 1, \ldots, P$.

Constant: $f_p(t_i) = 0$,

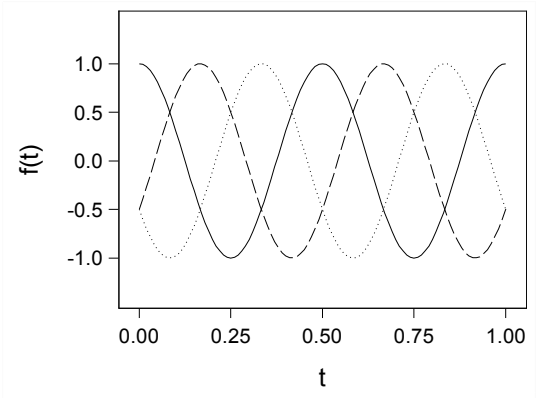
Non-constant, cyclic: $f_p(t_i) = \kappa \left[c_p + \cos\left(4\pi t^*_{p,i}\right)\right]$, where $t^*_{p,i} = t_i - \frac{p-1}{2P}$,

Non-constant, non-cyclic:

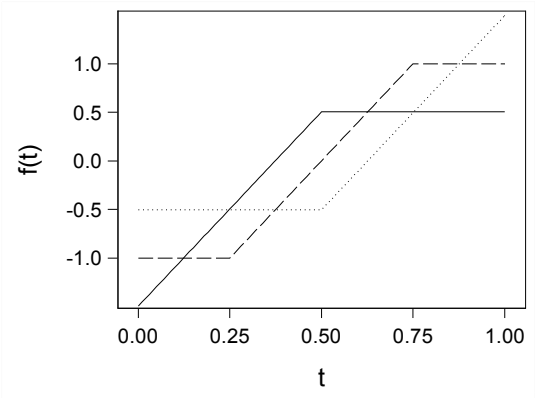
$$f_p(t_i) = \kappa \left[c_p + 4t^*_{p,i} I\left(-\frac{1}{4} \leq t^*_{p,i} \leq \frac{1}{4}\right) + I\left(t^*_{p,i} > \frac{1}{4}\right) - I\left(t^*_{p,i} < -\frac{1}{4}\right)\right],$$

where $t^*_{p,i} = t_i - \frac{3}{4} + \frac{p-1}{2(P-1)}$.

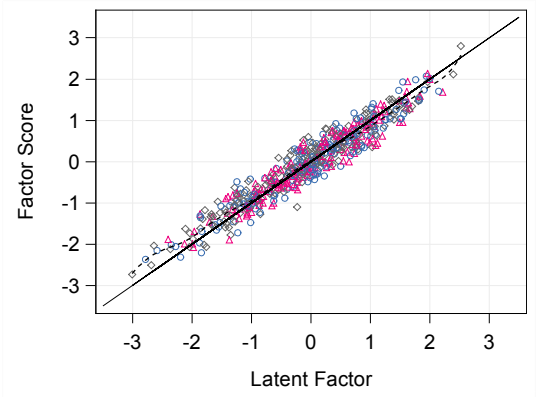
In the formulas, c_p is the constant that ensures $\sum_{i=1}^{n} f_p(t_i) = 0$ and κ is an amplitude parameter that changes the magnitude by which λ_{pi} deviates from λ_{p0}. Figure 2.2 (a), (b) illustrates the shape of $f_p(t_i)$ for the non-constant factor loadings. In the context of air pollution studies, cyclic factor loadings would resemble seasonal patterns, whereas the non-constant non-cyclic pattern would resemble spatial gradients (e.g., north to south).



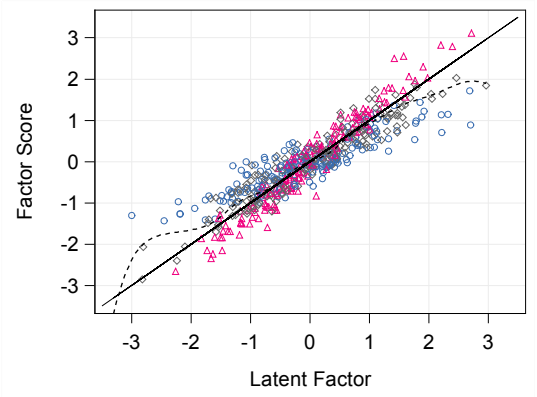
(a) Non-constant, cyclic



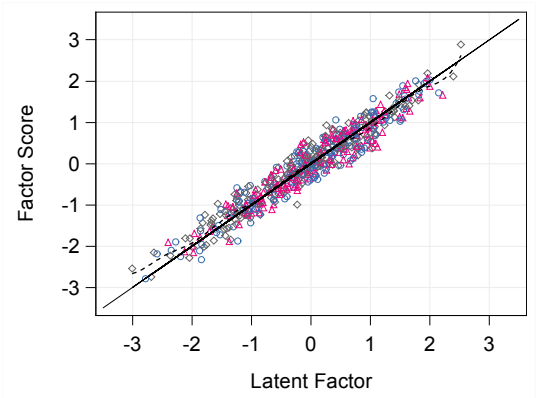
(b) Non-constant, non-cyclic



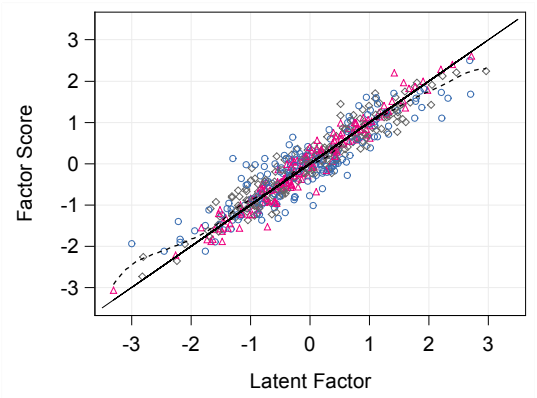
(c) Cyclic pattern, estimated as constant



(d) Non-cyclic pattern, estimated as constant



(e) Cyclic pattern, estimated as non-constant



(f) Non-cyclic pattern, estimated as non-constant

Figure 2.2: Bias in factor scores when non-constant factor loadings are estimated as constant. (a)-(b): Shape of $f_p(t_i)$ for the simulation study. Each plot shows three curves differentiated by line styles when a scenario has three observed variables and the amplitude parameter $\kappa = 1$. (c)-(f): Simulation results showing plots of $\hat{\eta}_i$ against η_i when factor loadings are estimated as constant or non-constant, when true factor loadings are non constant. The color scheme and marker symbol distinguish η_i by tertiles of t_i, where blue circle, gray diamond, pink triangle refer to the lower, middle, upper tertile, respectively. The solid line is where $(\eta_i, \hat{\eta}_i)$ falls if $\hat{\eta}_i = \eta_i$.

For each of the three shapes of factor loadings listed above, I hold some parameters constant, but I change the signal-to-noise ratio in the model as follows. For all the scenarios, I set $\mu_p = 1$ for all p, and make $\lambda_{p0}^2 + \sigma_p^2 = 8$ in all cases. Within each scenario I let λ_{p0} take on the same value for all p, but I change the ratio, $\lambda_{p0}^2/(\lambda_{p0}^2 + \sigma_{\epsilon,p}^2)$, among the different scenarios. This ratio, which was set at 0.3, 0.5, and 0.8, quantifies the percentage of variance for the p-th observed variable explained by the latent factor when the factor loading is constant. For the non-constant factor loadings, I also use different values of κ (0.75, 1.25), which also influence the percentage of variance explained by the latent factor.

For each scenario, I generate 1000 simulated data sets, each with 1000 equally spaced data points within the range of t. For some scenarios I vary the sample size (500, 5000) and the number of observed variables, P=6 instead of 3. For each dataset I estimate the factor loadings using three different approaches: (1) constant factor loadings, (2) non-constant factor loadings using P-splines (penalized), (3) non-constant factor loadings using unpenalized splines (all smoothing parameters equal to zero). In all cases, I choose equally spaced knots that give 10 spline coefficients (including λ_{p0}).

In Tables 2.1 and 2.2 I report the simulation results primarily focusing on scenarios where (1) each simulated data set has 1000 equally spaced data points within the range of t_i, (2) $\lambda_{p0}^2/(\lambda_{p0}^2 + \sigma_{\epsilon,p}^2) = 0.8$, (3) $\kappa = 1.25$ for the second and third types of $f_p(t_i)$. Comments on other scenarios are given further below.

I evaluate the results of my simulation study from two vantage points. First, I compare the bias and standard error of $\hat{\lambda}_{p0}$, $\hat{\sigma}_{\epsilon,p}^2$ among different estimation approaches and simulation scenarios (Table 2.1). This comparison gives insight into possible loss of efficiency in constant factor loadings when they are instead esti-

Table 2.1: Bias and standard error of $\hat{\lambda}_{p0}$, $\hat{\sigma}^2_{\epsilon,p}$ for different simulation scenarios and estimating approaches

True Factor Loadings	Estimating Approach	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\sigma}^2_{\epsilon,1}$	$\hat{\sigma}^2_{\epsilon,2}$	$\hat{\sigma}^2_{\epsilon,3}$
Constant	Constant	-0.005 [a] (0.072)	-0.003 (0.071)	-0.003 (0.073)	-0.002 (0.114)	0.000 (0.109)	-0.004 (0.115)
	Penalized	-0.007 (0.072)	-0.005 (0.072)	-0.004 (0.073)	-0.008 (0.114)	-0.007 (0.110)	-0.010 (0.115)
	Unpenalized	-0.011 (0.072)	-0.009 (0.071)	-0.008 (0.073)	-0.018 (0.114)	-0.017 (0.109)	-0.020 (0.115)
Non-Constant Cyclic	Constant	-0.082 (0.082)	-0.078 (0.086)	-0.082 (0.084)	1.175 (0.199)	1.152 (0.197)	1.165 (0.202)
	Penalized	-0.018 (0.070)	-0.018 (0.072)	-0.018 (0.073)	-0.011 (0.106)	-0.021 (0.101)	-0.014 (0.104)
	Unpenalized	-0.008 (0.071)	-0.007 (0.073)	-0.007 (0.073)	-0.013 (0.106)	-0.022 (0.101)	-0.015 (0.103)
Non-Constant Non-Cyclic	Constant	0.077 (0.077)	0.198 (0.086)	0.075 (0.082)	0.249 (0.129)	-0.015 (0.121)	0.261 (0.121)
	Penalized	-0.058 (0.073)	-0.069 (0.070)	-0.063 (0.073)	-0.021 (0.112)	-0.014 (0.105)	-0.007 (0.113)
	Unpenalized	-0.006 (0.073)	-0.009 (0.074)	-0.007 (0.073)	-0.028 (0.112)	-0.017 (0.105)	-0.014 (0.112)

[a]The bias is the top value and the standard error is the bottom value in parentheses.

mated as non-constants, as well as give insight regarding bias in residual variances, which ultimately affect how much weight a particular observed variable is given when estimating the latent factor. When the factor loadings are constant, the bias and efficiency of $\hat{\lambda}_{p0}$, $\hat{\sigma}^2_{\epsilon,p}$ are similar across estimation approaches, although $\hat{\sigma}^2_{\epsilon,p}$ is slightly underestimated when I use unpenalized splines since without penalization the splines overfit the data. However, when the factor loadings are non-constant and estimated as constant, $\hat{\lambda}_{p0}$, $\hat{\sigma}^2_{\epsilon,p}$ are both biased. The bias becomes trivial when I correctly estimate the factor loadings as non-constant.

Next, I compare how estimation approaches affect the factor score (Table 2.2). I define the factor score as $\hat{\eta}_i = \mathrm{E}_{\hat{\boldsymbol{\theta}}}(\eta_i|\mathbf{y}_i)$, where $\hat{\boldsymbol{\theta}}$ represents all the parameter estimates. In other words, the factor score $\hat{\eta}_i$ is the same as $\hat{\eta}_i^{(r)}$ from the E-step, but evaluated at convergence. Note that $\hat{\eta}_i$ is shrunken toward the center compared with η_i ($|\mathrm{E}(\hat{\eta}_i|\eta_i)| = |\boldsymbol{\lambda}_{.i}^T(\boldsymbol{\lambda}_{.i}\boldsymbol{\lambda}_{.i}^T + \boldsymbol{\Sigma}_\epsilon)^{-1}\boldsymbol{\lambda}_{.i}\eta_i| < |\eta_i|$). Thus, $\hat{\eta}_i$ is potentially more biased for values of η_i that are farther away from the center of its distribution. Therefore, in addition to examining the overall bias, $\frac{1}{n}\sum_{i=1}^n(\hat{\eta}_i - \eta_i)$ and overall mean squared error (MSE), $\frac{1}{n}\sum_{i=1}^n(\hat{\eta}_i - \eta_i)^2$ of the factor score (η_i is the true value of the latent factor from a given simulated data set), I also examine bias and MSE by percentiles

Table 2.2: Bias and MSE of the factor score for different simulation scenarios and estimating approaches

True Factor Loadings	Estimating Approach	Bias [a]				MSE			
		Overall	$< 10\%$	$10\% \sim 25\%$	$25\% \sim 50\%$	Overall	$< 10\%$	$10\% \sim 25\%$	$25\% \sim 50\%$
Constant	Constant	0.001	0.134	0.073	0.026	0.079	0.092	0.078	0.073
	Penalized	0.001	0.133	0.072	0.026	0.079	0.092	0.078	0.074
	Unpenalized	0.001	0.142	0.068	0.024	0.082	0.101	0.080	0.075
Non-Constant Cyclic	Constant	0.001	0.182	0.100	0.035	0.080	0.105	0.079	0.070
	Penalized	0.001	0.117	0.058	0.020	0.072	0.086	0.071	0.068
	Unpenalized	0.001	0.126	0.061	0.020	0.073	0.089	0.072	0.068
Non-Constant Non-Cyclic	Constant	-0.002	0.205	0.111	0.037	0.164	0.381	0.158	0.080
	Penalized	-0.002	0.143	0.074	0.024	0.103	0.137	0.101	0.089
	Unpenalized	-0.002	0.176	0.087	0.028	0.103	0.137	0.101	0.090

[a]Since the distribution of the factor score is symmetric, only three percentile categories are listed in this table. The upper three categories have the same bias in negative sign (thus the overall zero bias) and the same MSE.

of η_i. To gain further insight regarding estimation of the factor score at different covariate values, I also draw scatterplots of the estimated factor score vs. the true latent factor, with points on the scatterplot receiving different symbols according to tertiles of the values of the covariate t.

The overall average bias and MSE for the factor scores are similar across estimation approaches; however, differences emerge when I examine bias by percentiles of the distribution of the true latent factor and when I examine the factor scores at different covariate values of t (Table 2 and Figure 2.2 (c)-(f). I find that when the factor loadings are constant, the bias and MSE are similar if I estimate factor loadings either as constant or as non-constant using penalized splines. If I estimate the factor loadings using unpenalized splines, the MSE is unsurprisingly larger. However, when the factor loadings are non-constant, I get the smallest bias and MSE when I estimate the factor loadings using P-splines. Conversely, the bias and MSE is largest when the factor loadings are estimated as constant, and this is most notable for scenarios with the non-cyclic factor loading. In this case, $\hat{\eta}_i$ has a higher MSE for values of η_i at the two ends of its distribution (Figure 2.2 (c)-(f)). This larger MSE at the tails of the distribution is due to different directions in the bias of the factor score, which depends on covariate values (see scatter plot of factor scores vs. true

latent factor in Figure 2.2 (d) marked according to the values of t).

In scenarios where the variance in the observed variables is smaller (i.e., $\lambda_{p0}^2/(\lambda_{p0}^2 + \sigma_{\epsilon,p}^2)$ and κ are smaller), I find that the advantage of correctly estimating the non-constant factor loadings using penalized splines is less obvious. Similarly, when the noise in the data gets larger but the true factor loadings are constant, incorrectly estimating them as non-constant results in slightly increased bias in $\hat{\sigma}_{\epsilon,p}^2$ and increased MSE among all the parameter estimates along with the increase of MSE for the factor score.

2.5 Data Application

I apply my model to the US $PM_{2.5}$ speciation data issued by EPA for years 2008 and 2009. The data set I use includes 108 monitors stationed in parts of the Midwest, South and Northeast (see Figure 2.3). Thirty-six monitors had measurements every three days when they were functioning, but the majority only every six days. Thus, I only take every other observation from the thirty-six monitors that had measurements every three days so that data from all monitors are equally spaced.

Many of the chemical constituents in the EPA data are zero-inflated and would require special treatment beyond the scope of this thesis. Thus I select constituents that, after taking the logarithm, have an approximate normal distribution. Based on the exploratory factor analysis (see Section 2.1 and Figure 2.1), I decide to look at four constituents (sulfur, ammonium ion, total nitrate, sulfate) that are associated with secondary aerosols [58] and fit into a one factor structure.

Since the concentrations of constituents change across time and space, the mean of each observed variable ($\boldsymbol{\mu}$ in (2.1)) would need to be modeled as a smooth function. Although this is in principle possible with my modeling approach, I opt instead to

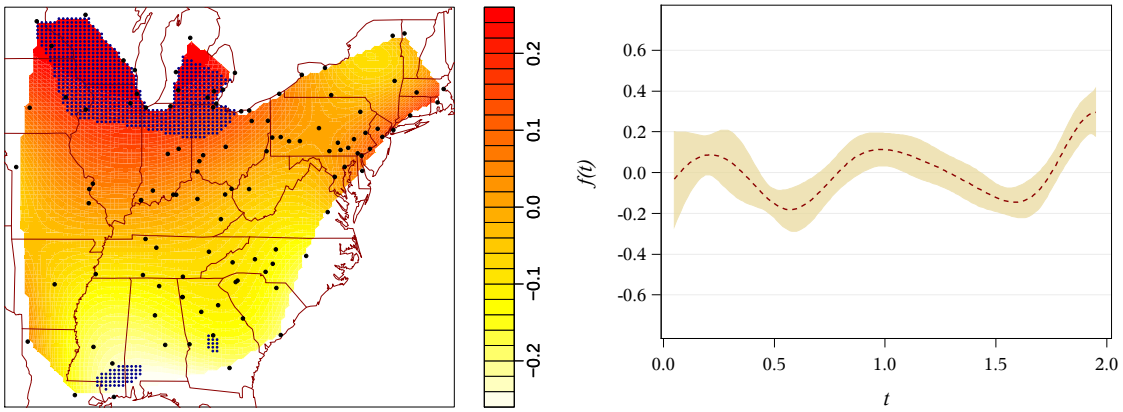
Table 2.3: $\hat{\lambda}_{p0}$, $\hat{\sigma}^2_{\epsilon,p}$ from the data application

Constituent		
$\hat{\lambda}_{p0}$	Factor loadings constant	Non-constant factor loadings
Sulfur	0.937 (0.890, 0.995)	0.944 (0.894, 1.110)
Ammonium Ion	0.929 (0.883, 0.982)	0.883 (0.841, 1.042)
Total Nitrate	0.606 (0.581, 0.632)	0.550 (0.526, 0.627)
Sulfate	0.976 (0.951, 1.000)	0.955 (0.925, 1.107)
$\hat{\sigma}^2_{\epsilon,p}$		
Sulfur	0.122 (0.068, 0.186)	0.106 (0.048, 0.153)
Ammonium Ion	0.136 (0.109, 0.163)	0.111 (0.089, 0.136)
Total Nitrate	0.633 (0.605, 0.658)	0.578 (0.544, 0.610)
Sulfate	0.047 (0.032, 0.064)	0.053 (0.033, 0.075)

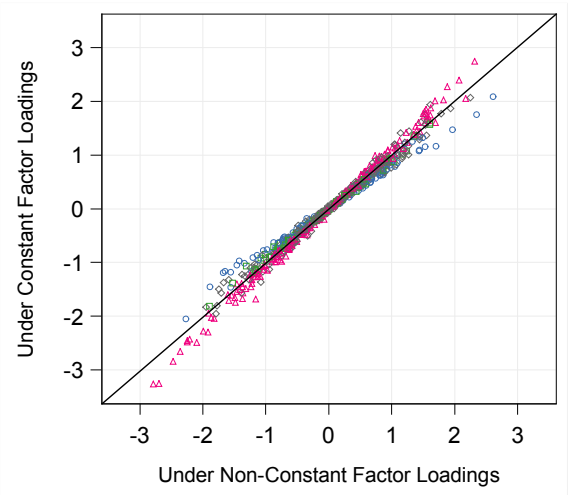
detrend the data and work with residuals. That is, first I regress each constituent on time and spatial coordinates using GAM, and then I use the residuals from these GAMs as the observed variables. Note that this approach not only allows us to treat $\boldsymbol{\mu}$ as constant, but also effectively removes spatial and temporal correlation enabling inference using bootstrap as described in Section 2.3.3.

The factor loadings were initially modeled according to (2.2), but the interaction term $f_{p,tuv}$ was insignificant based on the bootstrap confidence band, and thus I decide to keep only $f_{p,t}$ and $f_{p,uv}$. I use cubic B-splines as the basis functions; I choose the number of equally spaced knots so that $\boldsymbol{\beta}_{p,t}$ and $\boldsymbol{\beta}_{p,uv}$ have 10 and 25 parameters, respectively. Altogether, $\boldsymbol{\beta}_p$ have 34 parameters. Each component function then has its own set of smoothing parameters, a total of three smoothing parameters.

Table 2.3 summarizes the estimates of all the parameters in the model except the spline coefficients. (I do not list $\boldsymbol{\mu}$ in the table because they are zero for the detrended data.) Figure 2.3 shows the time and spatial components of the non-constant factor loading for total nitrate. For the other constituents, the non-constant part is not deemed significant since the confidence bands enclose zero entirely (not shown). From Figure 2.3 (a) I find that the factor loading for total nitrate is significantly higher in



(a) (b)



(c)

Figure 2.3: $\hat{f}_{p,uv}$, $\hat{f}_{p,t}$ from the data application for total nitrate. (a) is the contour plot of $\hat{f}_{p,uv}$ for the factor loading of total nitrate; black dots are the monitor locations and blue dots represent areas that are significantly different from zero according to the 95% confidence band constructed using bootstrap. (b) is $\hat{f}_{p,t}$, and 95% confidence band, for the factor loading of total nitrate ($t = 0$ is January 1 of 2008 and $t = 2$ is December 31 of 2009). (c) shows the factor score $\hat{\eta}_i$ for the $PM_{2.5}$ constituent data from the model with all constant factor loadings against $\hat{\eta}_i$ from the model with all non-constant factor loadings according to season and region: warmer regions in cold months (blue circle), colder regions in cold months (gray diamond), warmer regions in warm months (brown square), and colder regions in warm months (pink triangle); warm months refer to May, June, July and August, cold months refer to November, December, January and February, warmer regions refer to Mississippi, Georgia and Alabama, and colder regions refer to the Midwest except Missouri.

the Midwest than in the South. Figure 2.3 (b) shows that the factor loading for total nitrate fluctuates significantly over the two years in a periodic fashion, indicating a strong seasonal pattern. These patterns are likely due to the fact that formation of ammonium nitrate particles is favored in warmer environments [61]. These non-constant factor loadings have an appreciable impact on the estimated factor scores. In comparison to the factor scores from the model with non-constant factor loadings (which are essentially unbiased as shown in simulations), the factor scores from the model with constant factor loadings have systematic bias depending on season and region (Figure 2.3 (c)).

2.6 Discussion

In this chapter I propose a latent factor model where the factor loadings are allowed to vary across continuous covariate values. This modeling approach extends classical latent factor models and can help shed light on how the relationship between observed variables and underlying latent factors may be modified by covariates, and thus can help improve estimation of the underlying latent factors. This model is motivated by analysis of the EPA $PM_{2.5}$ constituent data that shows a noticeable difference for some of the factor loadings across time and region. Variations of latent factor models have long been used to summarize $PM_{2.5}$ constituents and gain insight into the sources the particles arose from, a process termed source apportionment. However, such data reduction methods have typically been applied to smaller geographical areas. Factor structures that can better capture the temporally and spatially varying correlations among constituents can help improve summaries of $PM_{2.5}$ constituents collected from a wide geographical area and a long time span. My model allows me to capture time-and-space-varying correlations among constituents

through the temporal and spatial pattern of factor loadings while maintaining a connection to interpretable sources. This ability of the model may help shed light on differing associations of the pollution mixture with health outcomes measured over several years; for example, in time series studies of mortality or hospital admissions.

I use P-splines to estimate factor loadings as functions of multiple covariates and I control the smoothness related to each covariate with its own smoothing parameter. In order to optimize the smoothing parameters in the context of a latent factor model, I propose the use of GCV within each iteration of the EM algorithm. This bypasses the issue of lack of predictors for the marginal distribution of the model by using the latent factor as the predictor in a single M-step. I also incorporate Newton's method into the EM iterations to facilitate the optimization of multiple smoothing parameters at the same time. The algorithm is successfully implemented in SAS, thus enhancing applicability of my method.

In this paper I use bootstrap to draw inference, which can be computationally intensive. Future work can include developing methods to more efficiently test the significance of different component functions of the non-constant factor loading. In my model I assume the random errors in (2.1) are uncorrelated. In my data application I detrend the data prior to applying my modeling strategy to remove spatial and temporal autocorrelation using GAM. Conditional residuals from the model have negligible correlation, suggesting an independent error structure may be sufficient for the detrended data I am analyzing. Incorporating alternative correlation structures is a desirable next step to avoid the two step approach I have used. Alternatively, as in the model proposed by Gryparis et al. [2], the mean trend of the latent factor could also be used to capture the temporal and spatial correlation. One could also explore extending the model for the latent factor to include spatio-temporal trends

as well as other covariates [67]. In my current data analysis I model every constituent with a P-spline and then assess whether it is invariant across time and space. Estimating each spline with its own set of smoothing parameters is demanding, although feasible, computationally. It may also be less efficient to build up a model with this much complexity when there are a considerable number of observed items. However, the non-constant factor loadings for many of the constituents may potentially change in similar patterns, due to their common source. Therefore, one potential extension to further improve estimation efficiency is to model multiple non-constant factor loadings with fewer number of curves that capture the essential shape [85, 86].

The model and algorithm implemented in this paper provides a platform for developing more comprehensive measurement models that can be used to to apply flexible dimension reduction strategies. The measurement model I propose can incorporate known information (such as well understood sources of air pollution) while learning from the data by allowing for non-constant factor loadings.

CHAPTER III

Testing Measurement Invariance in a Latent Factor Model Using P-Splines

3.1 Introduction

Latent factor models are typically used to summarize multivariate data for the purpose of deriving or relating factor scores to other covariates [87]. These models relate observed variables that reflect the underlying latent factors through a system of regression equations termed as the measurement model. However, in order to maintain a consistent and valid interpretation of the latent factors, certain measurement invariance conditions need to be satisfied [33]. Measurement invariance means the parameters in the measurement model are the same regardless of how the data is grouped in terms of covariates. Otherwise, measurement bias is introduced because the effect of variables other than the latent factor on the observed variables is not accounted for [35]. If bias is introduced, then the factor scores will be biased estimates of the underlying latent factor. The two most critical invariance conditions are invariance of the intercept and factor loadings in the measurement model. Simulation studies have shown that it is harder to detect bias coming from non-constant factor loadings [37, 38, 39, 88].

Assessing the measurement invariance assumption is an established step for scale development and validation in social and behavioral studies, where measurement

invariance is traditionally studied through multi-group factor analysis [87]. Multi-group factor analysis divides the data into groups according to covariates, for example, age, gender. Then a different measurement model is fitted for each group, and differences in parameters across groups are tested. The main disadvantage of this approach is its less satisfactory bias detection rate, especially for the factor loadings when the measurement bias is related to a continuous covariate, due to arbitrary group membership assignment that results in less accuracy and efficiency of parameter estimates [39, 88].

Novel models that allow factor loadings to vary as continuous functions of some covariates offer alternative strategies for addressing the issue of measurement invariance [89, 90, 7]. Zhang et al. [7] uses bootstrap simulations to construct confidence intervals for non-constant factor loadings and examine whether a factor loading varies across covariate values. Barendse et al. [88] compares restricted factor analysis (RFA) with multi-group models with respect to their ability to detect measurement bias; they compare the fit of models with and without a non-constant factor loading using LRT. Their simulation study shows RFA models have better bias detection rates than multi-group models. However, the Type I error for testing one of the factor loading as being non-constant is higher than nominal when measurement bias for the other observed variables is not taken into account in the model.

In this chapter I propose a modeling strategy that can be used to estimate factor loadings that vary smoothly across covariate values and to test if the factor loadings deviate significantly from a constant. I use P-splines to model the factor loadings as varying coefficients [83], where the spline coefficients are treated as random. In this way the smoothing process is incorporated into the likelihood [40, 91]. I then test whether a factor loading is constant using LRT of whether the variance of the spline

coefficients differs from zero.

However, because the value of the parameter of interest is at the boundary of the parameter space, the asymptotic distribution of this test statistic is not easy to derive. This test for non-zero variance components related to observed covariates in the mixed model framework has been extensively studied [92, 93, 94, 95, 96, 97]. Except for the paper by Stoel et al. [98], I am not aware of this type of testing problem within the framework of latent factor models. Following the method proposed by Greven et al. [96], I use a parametric bootstrap method to approximate the null distribution of the LRT statistic and compare the power of LRT in my model to the multi-group model. I investigate how the performance of the test for one factor loading changes with the estimation of other factor loadings. Even though the finite sample null distribution proposed in [94] is a faster method, it is less straightforward for my model because of the unobserved latent factor.

This work is motivated by research conducted as part of the Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) project. The ELEMENT project collected an array of lead (Pb) exposure biomarkers and several health outcomes in three pregnancy cohorts in Mexico City. In particular, latent factor models have been used to summarize lead exposure biomarkers collected from the mother and or infant near the time of birth [81]. In this model the latent factor represents the underlying but unobserved true exposure. However, the factor loadings may vary by age of the mother, because the biological processes involved in lead metabolism of lead in the body vary by age. If the factor loadings do vary by age, the estimated factors would be a biased measure of the latent exposure.

In Section 3.2 I present the model and the hypothesis test of interest. Section 3.3 details the estimation method and some technical aspects of the nested-Monte-Carlo

EM (MCEM) algorithm in the context of my proposed method [99]. I also talk about computing the likelihood using Monte-Carlo integration and the construction of confidence intervals for the non-constant factor loadings. In Section 3.4 I carry out simulation studies to investigate Type I error and power of testing using LRT and confidence intervals. I also evaluate the use of parametric bootstrap in obtaining the critical value for LRT. In Section 3.5 I apply the proposed methods to data from the ELEMENT project. Section 3.6 ends with a discussion and future directions.

3.2 Model

I model the factor loadings as non-constant functions of covariate z_i so that the measurement model can differ across covariate values. For exposition, suppose there is only one latent factor. Let y_{pi} be the p-th observed variable measured on subject i and $\eta_i \overset{\text{iid}}{\sim} N(0,1)$ be the latent factor underlying the P observed variables. The observed variables are related to the latent factor as

$$y_{pi} = \lambda_{pi}\eta_i + \epsilon_{pi}, \quad p = 1, \ldots, P, \quad i = 1, \ldots, n. \tag{3.1}$$

The factor loading λ_{pi} for the p-th observed variable includes a subscript for subjects, i, because it depends on the covariate value z_i for the ith subject, as shown below. I assume the residual error $\epsilon_{pi} \overset{\text{iid}}{\sim} N(0, \mathbf{\Sigma}_\epsilon)$ is independent of η_i. Without loss of generality, I omit the mean for y_{pi} as y_{pi} can easily be centered to have zero mean; centering observed variables y_{pi} is common data pre-processing practice in factor analysis [87]. In this modeling strategy, further detrending to remove covariate effects on the mean on y_{pi} (e.g., see Section 3.5) ensures that non-constant values of λ_{pi} are due to differences in the correlation among observed values, above and beyond any potential differences in the average relationship between y_{pi} and z_i.

I use P-splines to estimate λ_{pi}: $\lambda_{pi} = \lambda_{p0} + f_p(z_i) = \lambda_{p0} + \mathbf{x}_i\boldsymbol{\beta}_p$, where $f_p(z_i)$ has

mean zero across the covariate values (i.e., $\sum_{i=1}^{n} f_p(z_i) = 0$). Mean zero for $f_p(z_i)$ is needed for identifiability, and also helps interpretation since, when $f_p(z_i)$ is zero, the model simplifies to standard factor analysis. In other words, $f_p(z_i)$ represents the non-constant component of the factor loading. I use cubic B-spline basis functions to model $f_p(z_i)$, and let $\mathbf{x}_i$ be the spline basis based on z_i; i.e., $\mathbf{x}_i$ contains the value of each basis function evaluated at z_i [15]. The spline coefficients, which differ across the P observed variables, are $\boldsymbol{\beta}_p$.

I rewrite (3.1) as a regression model. First, I collect data from all subjects as: $\mathbf{y}_{p\cdot} = [y_{p1}, \ldots, y_{pn}]^T$, $\boldsymbol{\epsilon}_{p\cdot} = [\epsilon_{p1}, \ldots, \epsilon_{pn}]^T$, $\mathbf{X}_\eta = [\mathbf{x}_1^T\eta_1, \ldots, \mathbf{x}_n^T\eta_n]^T$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^T$. Then (3.1) becomes

$$\mathbf{y}_{p\cdot} = \boldsymbol{\eta}\lambda_{p0} + \mathbf{X}_\eta\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_{p\cdot}. \tag{3.2}$$

I aim to test if, for a particular observed variable $\mathbf{y}_{p\cdot}$, its factor loading is constant, i.e.,

$$H_0 : \lambda_{pi} = \lambda_{p0} \quad \text{for all } i \quad \text{vs.} \quad H_a : \lambda_{pi} \neq \lambda_{p0} \quad \text{for some } i. \tag{3.3}$$

This is equivalent to testing $H_0 : \boldsymbol{\beta}_p = 0 \quad \text{vs.} \quad H_a : \boldsymbol{\beta}_p \neq 0$.

In Section 3.3, I penalize the spline by setting $\boldsymbol{\beta}_p$ as random variables; specifically, $\boldsymbol{\beta}_p \sim N(\mathbf{0}, \sigma_{p,b}^2\mathbf{I}_K)$. This approach is related to representing a regression spline model in a mixed model setting [40]. In this way, $\sigma_{b,p}^2$ takes the role of the smoothing parameter for $f_p(z_i)$: a smaller $\sigma_{b,p}^2$ leads to smaller $\boldsymbol{\beta}_p$ and smoother splines, and when $\sigma_{b,p}^2 \to 0$, $f_p \to 0$, which makes $\lambda_{pi} \to \lambda_{p0}$. Note that in this model the covariance structure of $\boldsymbol{\beta}_p$ corresponds to setting the penalty matrix $\mathbf{S}_p$ as $\mathbf{I}_K$ in the quadratic penalty $\boldsymbol{\beta}_p^T\mathbf{S}_p\boldsymbol{\beta}_p$ of a penalized regression spline model. Even though the identity penalty matrix is mostly used when a truncated polynomial basis is used [40], I use the first difference penalty matrix $\mathbf{S}_p$ [16], and transform it to an identity

by modifying $\mathbf{X}_\eta$ according to the spectral decomposition of $\mathbf{S}_p$ [24]. The test in (3.3) then becomes

$$H_0 : \sigma^2_{b,p} = 0 \quad \text{vs.} \quad H_a : \sigma^2_{b,p} > 0. \tag{3.4}$$

3.3 Estimation

3.3.1 Likelihood

I use the EM algorithm for estimation, and treat η_i, $\boldsymbol{\beta}_p$ as the augmented missing data. The log-likelihood given the augmented data is $\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\beta}_p) \propto$

$$-\frac{1}{2}\sum_{p=1}^{P}\left\{ n\log(\sigma^2_{\epsilon,p}) + \frac{\|\mathbf{y}_{p\cdot} - \boldsymbol{\eta}\lambda_{p0} - \mathbf{X}_\eta\boldsymbol{\beta}_p\|^2}{\sigma^2_{\epsilon,p}} + K\log(\sigma^2_{b,p}) + \frac{\|\boldsymbol{\beta}_p\|^2}{\sigma^2_{b,p}}\right\}, \tag{3.5}$$

where $\boldsymbol{\theta}$ represents all the parameters. In writing (3.5) I assume all factor loadings are non-constant. However, if the factor loading for the p-th observed variable is instead assumed as constant a priori, I only need to remove $\sigma^2_{b,p}$ from the likelihood and set $\boldsymbol{\beta}_p = 0$.

3.3.2 E-Step

The E-step evaluates $Q\left(\boldsymbol{\theta}\left|\hat{\boldsymbol{\theta}}^{(r-1)}\right.\right) = \mathrm{E}\left\{\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\eta},\boldsymbol{\beta}_p)\left|\mathbf{y}, \hat{\boldsymbol{\theta}}^{(r-1)}\right.\right\}$ at the r-th iteration, where $\hat{\boldsymbol{\theta}}^{(r-1)}$ represents the parameter estimates from the previous iteration. Because $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ are the missing data the conditional expectation of all terms involving these quantities needs to be calculated. This essentially means computing $\boldsymbol{\eta}^{(r)}$, $(\boldsymbol{\eta}^T\boldsymbol{\eta})^{(r)}$, $(\mathbf{X}_\eta\boldsymbol{\beta}_p)^{(r)}$, $(\boldsymbol{\eta}^T\mathbf{X}_\eta\boldsymbol{\beta}_p)^{(r)}$, $(\boldsymbol{\beta}_p^T\mathbf{X}_\eta^T\mathbf{X}_\eta\boldsymbol{\beta}_p)^{(r)}$, and $(\boldsymbol{\beta}_p^T\boldsymbol{\beta}_p)^{(r)}$, where, for example, $\boldsymbol{\eta}^{(r)} = \mathrm{E}\left(\boldsymbol{\eta}\left|\mathbf{y};\hat{\boldsymbol{\theta}}^{(r-1)}\right.\right)$.

The conditional distribution of $(\boldsymbol{\eta}, \boldsymbol{\beta}_p|\mathbf{y};\boldsymbol{\theta})$ has no analytical form, thus I use Monte-Carlo EM (MCEM) for estimation. I draw samples from the conditional distribution $(\boldsymbol{\eta}, \boldsymbol{\beta}_p|\mathbf{y};\boldsymbol{\theta}^{(r-1)})$ through Gibbs sampler to alternatively sample from the two conditional distributions:

$$\eta_i|(\mathbf{y}_{.i}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_P) \sim N(\boldsymbol{\Lambda}_i^T \mathbf{B}_i \mathbf{y}_{.i}, 1 - \boldsymbol{\Lambda}_i^T \mathbf{B}_i \boldsymbol{\Lambda}_i),$$

where for simplicity of notation I have collected terms as $\mathbf{y}_{.i} = [y_{1i}, \ldots, y_{Pi}]^T$, $\boldsymbol{\Lambda}_i = [\lambda_{1i}, \ldots, \lambda_{Pi}]^T$, $\mathbf{B}_i = (\boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^T + \boldsymbol{\Sigma}_\epsilon)^{-1}$, and

$$\boldsymbol{\beta}_p|(\mathbf{y}_{p.}, \boldsymbol{\eta}) \sim N\left(\hat{\boldsymbol{\beta}}_p(\boldsymbol{\eta}), \sigma_{b,p}^2(\mathbf{I}_K - \mathbf{D}_p \mathbf{X}_\eta^T \mathbf{X}_\eta)\right),$$

where $\hat{\boldsymbol{\beta}}_p(\boldsymbol{\eta}) = \mathbf{D}_p \mathbf{X}_\eta^T (\mathbf{y}_{p.} - \lambda_{p0}\boldsymbol{\eta})$, $\mathbf{D}_p = (\mathbf{X}_\eta^T \mathbf{X}_\eta + \rho_p \mathbf{I}_K)^{-1}$, $\rho_p = \sigma_{\epsilon,p}^2 / \sigma_{b,p}^2$. If I assume the factor loadings for some observed variables are constant, then I replace λ_{pi} with λ_{p0} for all the constant factor loadings in the first conditional distribution.

After sampling $(\boldsymbol{\eta}^{(m)}, \boldsymbol{\beta}_p^{(m)}), m = 1, \ldots, M^{(r)}$, using the Gibbs sampler, I can evaluate $Q\left(\boldsymbol{\theta} \middle| \hat{\boldsymbol{\theta}}^{(r-1)}\right)$ by setting the expectation of all the statistics as an average of those calculated using each sample. For example,

$$(\boldsymbol{\eta}^T \mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r)} \approx \frac{1}{M^{(r)}} \sum_{m=1}^{M^{(r)}} (\boldsymbol{\eta}^{(m)})^T \mathbf{X}_\eta^{(m)} \boldsymbol{\beta}_p^{(m)}.$$

3.3.3 M-Step

The M-step then updates the parameter estimates as

$$\hat{\boldsymbol{\theta}}^{(r)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta} \middle| \hat{\boldsymbol{\theta}}^{(r-1)}\right).$$

The computation is similar to obtaining estimates from a linear regression model:

$$\begin{aligned}
\lambda_{p0}^{(r)} &= \left\{(\boldsymbol{\eta}^T \boldsymbol{\eta})^{(r)}\right\}^{-1} (\boldsymbol{\eta}^T)^{(r)} \left\{\mathbf{y}_{p.} - (\mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r)}\right\}, \\
(\sigma_{\epsilon,p}^2)^{(r)} &= \frac{1}{n}\Big\{\mathbf{y}_{p.}^T \mathbf{y}_{p.} - 2\lambda_{p0}^{(r)} \mathbf{y}_{p.}^T \boldsymbol{\eta}^{(r)} + (\lambda_{p0}^{(r)})^2 (\boldsymbol{\eta}^T \boldsymbol{\eta})^{(r)} \\
&\quad -2\mathbf{y}_{p.}^T (\mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r)} + 2\lambda_{p0}^{(r)} (\boldsymbol{\eta}^T \mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r)} + (\boldsymbol{\beta}_p^T \mathbf{X}_\eta^T \mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r)}\Big\}, \\
(\sigma_{b,p}^2)^{(r)} &= \frac{1}{K}(\boldsymbol{\beta}_p^T \boldsymbol{\beta}_p)^{(r)}.
\end{aligned}$$

Since the Monte-Carlo E-step is computationally intensive, I implement a nested MCEM to more efficiently use the $(\boldsymbol{\eta}^{(m)}, \boldsymbol{\beta}_p^{(m)})$ samples from the E-step. Specifically, because the conditional distribution of $(\boldsymbol{\beta}_p|\mathbf{y}, \boldsymbol{\eta})$ has an analytical form, within each iteration I nest a few additional EM cycles to get better estimates of $\boldsymbol{\beta}_p$ by treating $\boldsymbol{\eta}^{(m)}$ as known variables. In this way, the whole algorithm converges faster and thus saving the computational cost of the E-step [99]. (see Appendix A)

3.3.4 Likelihood Ratio Test (LRT)

Since I use random spline coefficients to incorporate the smoothing of splines into the likelihood, I can use LRT to test my hypothesis of interest, that is, whether the p-th variable has a constant factor loading. Note that I am only testing for the factor loading of one variable, that is, $\sigma^2_{b,p}$ for a particular p according to the test in (3.4). I let other variables not being tested have either constant or non-constant factor loadings under both the null and alternative hypotheses. The LRT test statistic is constructed as $2\left\{\log \mathcal{L}(\hat{\boldsymbol{\theta}}_a|\mathbf{y}) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\mathbf{y})\right\}$, where $\hat{\boldsymbol{\theta}}_a$ represents the parameter estimates from the model where $\sigma^2_{b,p}$ is estimated and $\hat{\boldsymbol{\theta}}_0$ represents the parameter estimates from the model where $\sigma^2_{b,p} = 0$.

The asymptotic theory about the LRT statistic says that if the parameter in the null hypothesis is on the boundary of the parameter space, then the asymptotic distribution for the null LRT statistic does not follow a simple chi-square distribution. For instance, the seminal paper by Stram et al. [93] says that in a linear mixed model setting, the asymptotic null distribution of LRT statistic for the test of a single variance component, similar to (3.4) follows $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$.

However, Crainiceanu et al. [94] argues that using $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ gives a conservative test, because the asymptotic assumption is rarely reached in real cases and also is not directly applicable to penalized spline models. They propose an approach to

find the exact distribution when there is only one variance component in the model. Greven et al. [96] generalizes the technique proposed by Crainiceanu et al. [94] to models where only one variance component is being tested. But this method cannot be immediately applied to my model because it would be based on the eigenvalues of $\mathbf{X}_\eta$ in (3.2), but in my model $\boldsymbol{\eta}$ is unobserved. They also suggest using parametric bootstrap to derive the 95-th percentile of the LRT statistic under the null hypothesis. This approach can be implemented either using full-scale bootstrap or through the use of a parametric family of chi-square mixture distributions proposed in the paper by Greven et al. [96]. This mixture distribution takes the form of $p\chi_0^2 + (1-p)a\chi_1^2$, where p is the probability of the statistic being zero, and a is a scaling factor that provides more flexibility for the distribution.

For a given data set, deriving the null distribution using this chi-square mixture distributions is straightforward. First, M^* parametric bootstrap data sets are generated using the parameters from the null model. LRT statistics are computed for each of the M^* data sets and are used to estimate p and a using the method of moments as described by the paper by Greven et al. [96]. Since the 95-th percentile of the null LRT is derived from a fitted parametric family of distributions, a fewer number of bootstrap samples M^* is needed to obtain equally precise estimates of the tail quantiles compared to a full parametric bootstrap. After estimating $\hat{p}$ and $\hat{a}$, I set the 95-th percentile of the null LRT statistic as $\hat{a}F^{-1}_{\chi_1^2}(\frac{0.95-\hat{p}}{1-\hat{p}})$, where $F_{\chi_1^2}$ represents the cumulative density function of a χ_1^2 distribution. P-values can also be derived using $(1-\hat{p})F^{-1}_{\chi_1^2}\left[\frac{1}{\hat{a}}\{\text{LRT statistic}\}\right]$. I use this distribution for hypothesis testing. As a comparison, I also examine the use of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ since it would lead to substantial computational savings.

3.3.5 Monte-Carlo Integration

In order to form the LRT statistic, I need to calculate the likelihood. However, since the likelihood given the observed data is analytically intractable, I use Monte-Carlo integration with importance sampling to simulate the likelihood. The numerical integration is achieved by averaging the integrand evaluated at each sampled $\boldsymbol{\beta}^{(s)}$. In the following equations, I let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_P^T]^T$, where g represents the probability density function (pdf) associated with the model, h represents the pdf of the importance sampling distribution. The likelihood is given by:

$$
\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{y}) &= \int g_{\mathbf{y}|\boldsymbol{\beta},\hat{\boldsymbol{\theta}}}\, g_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}}\, d\boldsymbol{\beta} \\
&= \int g_{\mathbf{y}|\boldsymbol{\beta},\hat{\boldsymbol{\theta}}} \cdot (g_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}}/h_{\boldsymbol{\beta}}) \cdot h_{\boldsymbol{\beta}}\, d\boldsymbol{\beta} \\
&\approx \tfrac{1}{M_s} \textstyle\sum_{s=1}^{M_s} \left\{ g_{\mathbf{y}|\boldsymbol{\beta}^{(s)},\hat{\boldsymbol{\theta}}} \cdot (g_{\boldsymbol{\beta}^{(s)}|\hat{\boldsymbol{\theta}}}/h_{\boldsymbol{\beta}^{(s)}}) \right\},
\end{aligned}
\tag{3.6}
$$

where I can obtain $g_{\mathbf{y}|\boldsymbol{\beta},\hat{\boldsymbol{\theta}}} = \prod_{i=1}^{n} g_{\mathbf{y}_{\cdot,i}|\boldsymbol{\beta},\hat{\boldsymbol{\theta}}}$ and $g_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}} = \prod_{p=1}^{P} g_{\boldsymbol{\beta}_p|\hat{\boldsymbol{\theta}}}$ based on

$$
\mathbf{y}_{\cdot,i}|\boldsymbol{\beta},\hat{\boldsymbol{\theta}} \sim N\left(\mathbf{0}, \hat{\boldsymbol{\Lambda}}_i(\boldsymbol{\beta})\hat{\boldsymbol{\Lambda}}_i(\boldsymbol{\beta})^T + \hat{\boldsymbol{\Sigma}}_\epsilon\right), \quad \text{and} \quad \boldsymbol{\beta}_p|\hat{\boldsymbol{\theta}} \sim N(\mathbf{0}, \hat{\sigma}_{b,p}^2 \mathbf{I}_K).
$$

The optimal choice of $h_{\boldsymbol{\beta}}$ is the unknown $g_{\boldsymbol{\beta}|\mathbf{y},\hat{\boldsymbol{\theta}}}$, because the summand in (3.6) would become $g_{\mathbf{y}}$ and is thus unrelated to the sampling. Instead, I take the mean and covariance matrix of the $\boldsymbol{\beta}_p$ samples drawn from the distribution of $\boldsymbol{\beta}_p|\mathbf{y}$ at E-step and construct a multivariate normal distribution to draw samples of $\boldsymbol{\beta}_p^{(s)}$.

3.3.6 Confidence Intervals

Since the construction of LRT statistic and the ensuing parametric bootstrap is computationally intensive, I also examine the use of confidence intervals for testing purposes. Confidence intervals can be directly derived using samples obtained at the E-step. Then I can construct pointwise confidence intervals or simultaneous

confidence bands, each with advantages and disadvantages. For each $\boldsymbol{\beta}_p^{(m)}$ drawn using the Gibbs sampler (see Section 3.3.2) at convergence of the EM algorithm, I evaluate f_p at z_j (z_j can be different from the covariate values z_i in the data) using $f_p^{(m)}(z_j) = \mathbf{x}_j \boldsymbol{\beta}_p^{(m)}$, $m = 1, 2, \ldots, M^{(\infty)}$ (∞ is used to denote convergence).

The 95% pointwise confidence intervals for $f_p(z_j)$ are constructed using the 2.5-th and 97.5-th percentiles of the $M^{(\infty)}$ $f_p^{(m)}(z_j)$ samples. I then test whether $f_p(z_j) = 0$ by looking at whether the pointwise confidence intervals covers zero or not. The null hypothesis in (3.3) is rejected if the pointwise confidence intervals for any $f_p(z_j)$ fails to cover zero (that is, $f_p(z_j) = \lambda_{pj} - \lambda_{p0} \neq 0$ for some j, provided the z_j being examined are dense enough).

However, since the pointwise confidence intervals are typically too narrow for function estimates, this testing procedure has the potential to be liberal. It may also have a low coverage of the true f_p because 95% coverage for each $f_p(z_i)$ does not guarantee simultaneous 95% coverage. Therefore, I also construct the simultaneous confidence bands according to [40]. I first obtain C, the 95-th percentile of

$$C^{(m)} = \sup_i \left| \frac{f_p^{(m)}(z_i) - \hat{f}_p(z_i)}{\sqrt{\text{Var}\left\{ f_p^{(m)}(z_i) - \hat{f}_p(z_i) \right\}}} \right|,$$

where $\hat{f}_p(z_i) = \frac{1}{M^{(\infty)}} \sum_{i=1}^{M^{(\infty)}} f_p^{(m)}(z_i)$. Then the 95% simultaneous confidence band is constructed as $\hat{f}_{p,t}(t) \pm C \sqrt{\text{Var}\left\{ f_p^{(m)}(z_i) - \hat{f}_p(z_i) \right\}}$. I obtain $\text{Var}\left\{ f_p^{(m)}(z_i) - \hat{f}_p(z_i) \right\}$ using the same set of $f_p^{(m)}(z_i)$. Again, testing proceeds by rejecting the null hypothesis if the simultaneous confidence band for any $f_p(z_j)$ fails to cover zero.

3.4 Simulation

My simulation studies have two objectives. First, I examine whether using the approximations $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ and $p\chi_0^2 + (1-p)a\chi_1^2$ adequately control Type I error,

compared to a full parametric bootstrap. Second, I address the Type I error and power of the LRT (including using $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ as the null distribution), as well as using tests based on confidence intervals described in Section 3.3.6.

For all the data-generating scenarios (further described below), I let each of 10000 simulated data sets have three observed variables, $\mathbf{y}_{p.}, p = 1, 2, 3$, and 2000 data points with values of z_i equally spaced within the range of $[0, 1]$. I set $\lambda^2_{p0} + \sigma^2_{\epsilon,p}$, the total variance of the observed variables, as 8 and let $\lambda^2_{p0}/(\lambda^2_{p0} + \sigma^2_{\epsilon,p}) = 0.5$ so that I have a medium signal-to-noise ratio, which is also in the middle of the range of the signal-to-noise ratio in my data example.

I vary the shapes of f_1, f_2 for different simulation scenarios, organized in two groups. In the first group of scenarios (Figure 3.1 and Table 3.1 rows 1, 2, 3), I set the first factor loading as non-constant, that is, $f_1 \neq 0$, but set $f_2 = 0$, $f_3 = 0$. I let f_1 take on two different shapes,

(1) $f_1(z_i) = \kappa\{-0.1\cos(6\pi z_i)\}$,

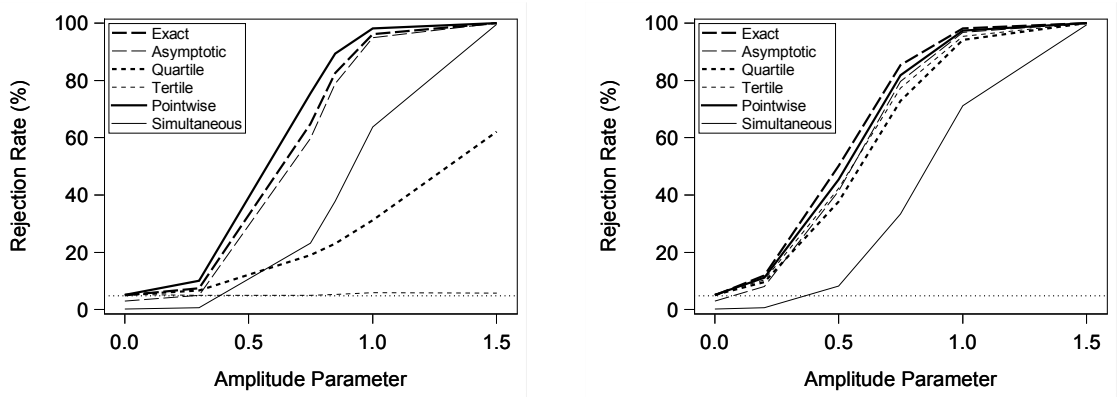
(2) $f_1(z_i) = \kappa\{c + (1.6z_i - 0.8z_i^2) - 0.1\cos(6\pi z_i)\}$.

The first shape has a cyclic pattern, while the second shape adds a monotone trend. In the formulas, c is the constant that ensures $\sum_{i=1}^{n} f_1(z_i) = 0$ and κ is an amplitude parameter that changes the magnitude of f_1. In the second group of scenarios (Table 3.1 rows 4, 5, 6) I let f_1 be the same as in the first group, but I let $f_2(z_i) = 0.6(z_i - 0.5)$ (f_3 still remains zero).

For each scenario in both groups, I estimate two random coefficient models: one where only f_1 is estimated (left of Table 3.1) and one where both f_1 and f_2 are estimated (right of Table 3.1). This allows me to assess any potential bias in tests when the fitted model is correct or misspecified, as well as potential loss in power when the models used are more flexible than necessary. As a comparison, I also fit

multi-group models that allow the factor loadings to be different across groups of data points. This is a commonly used approach in psychological and social studies when the research question is whether the underlying latent factor represents the observed variables in the same way for different sub-groups of the population. Since the factor loading within each group is still assumed constant, this approach can overlook important differences in the factor loadings and the estimation also depends on how the group is assigned [88]. In this study I use tertiles and quartiles of z_i to group the data, and I jointly model the different groups. The non-constant factor loadings take the following form (if I use tertiles of z_i), $\lambda_{pi} = \lambda_{p0} + \alpha_1 I(z_i <= Q_1) + \alpha_2 I(Q_1 < z_i <= Q_2) + \alpha_3 I(z_i > Q_2)$, where Q_1, Q_2 represent the first and second tertiles of z_i, respectively. Since I am interested in the factor loadings, which are the main sources of measurement bias, all groups are constrained to share the same residual variances, $\sigma^2_{\epsilon,p}$. I use LRT for testing in the multi-group models. The LRT statistic asymptotically follows a chi-square distribution with degrees of freedom equal to 2 and 3 for tertile and quartile models, respectively.

I first focus on Type I error ($\kappa = 0$). Using LRT, I find that $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ gives about a 2.8 - 3.0% rejection rate (first row of Table 3.1). The simulation suggests that more than half the time the spline estimate shrinks to zero, which makes $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ a conservative null distribution. Since a conservative null distribution lowers the power of the test, in practice a better approximation for the null distribution is needed. Options for such approximation can be a full-scale parametric bootstrap or using a mixture $p\chi^2_0 + (1-p)a\chi^2_1$ (see Section 3.3.4). I evaluate in this same scenario the performance of $p\chi^2_0 + (1-p)a\chi^2_1$, compared to the full-scale bootstrap, following the same procedure in [96]. The results show that this approximation can perform as well as the full-scale bootstrap approach (see Appendix B of the

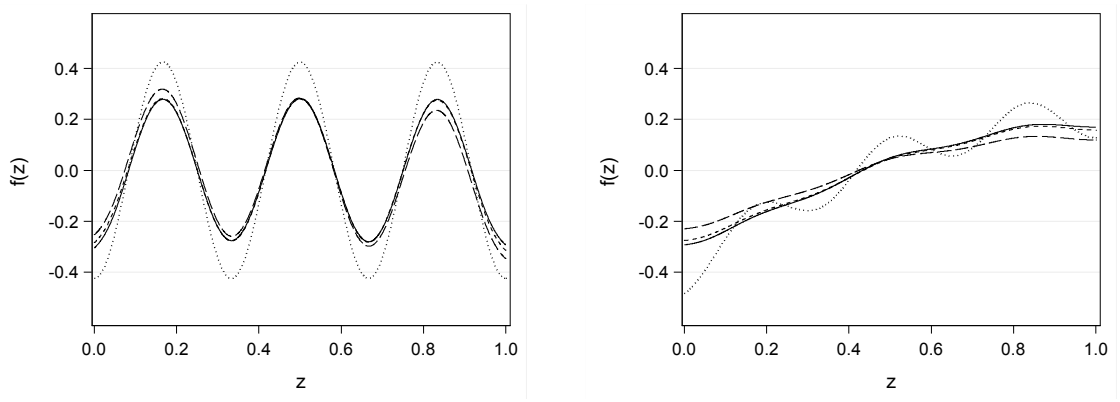


(a) Cyclic Shape (b) Monotone Trend Added

Figure 3.1: Power curve of LRT for testing H_0: $f_1 = 0$ vs. $f_1 \neq 0$, for two shapes of f_1 in models where only f_1 is non-zero (first group of scenarios) and being estimated. The curves are based on simulation scenarios from the first group where $\kappa = 0, 0.2, 0.5, 0.75, 1, 1.5$ for the cyclic shape and $\kappa = 0, 0.3, 0.75, 0.85, 1, 1.5$ for the monotone trend added. For random coefficient models, we use either the 95-th percentile of the LRT statistic from the scenario where $f_1 = 0$ (referred to as 'Exact') or the 95-th percentile of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (referred to as 'Asymptotic') as the critical value. We also plot the rejection rate when we use pointwise confidence intervals (referred to as 'Pointwise') or simultaneous confidence band (referred to as 'Simultaneous') for testing. For multi-group models, we plot the rejection rate when tertiles or quartiles of z_i are used to group the data. The horizontal dotted line near the bottom indicates the 5% rejection rate.

Supplementary Materials), even with bootstrap sample size M^* as low as 500. I will use $p\chi_0^2 + (1-p)a\chi_1^2$ to derive the 95-th percentile of the true null distribution for the data analysis in Section 3.5. Nevertheless, to conserve computation cost, in the following scenarios where $\kappa \neq 0$, I use the 95-th percentile of the simulated LRT statistic from the null scenario, which is similar to the full-scale bootstrap, as the critical value.

I also examine the rejection rate of the pointwise confidence intervals and simultaneous confidence band described in Section 3.3.6. For each of the 10000 simulations, I construct two types of confidence intervals using $M^{(\infty)} = 1000$ Monte-Carlo samples for $\boldsymbol{\beta}_p^{(m)}$. I examine the confidence intervals at 500 different values of z_j equally spaced within the range of $[0, 1]$ to determine whether they fail to cover zero (rejection rate) or whether they cover the true f_p (coverage rate). I find that the test based on pointwise confidence intervals is conservative: its rejection rate is 5.2% (also



(a) Cyclic Shape (b) Monotone Trend Added

Figure 3.2: Comparison of $\hat{f}_1$ from different scenarios in Table 3.1. The dotted line is the true f_1. The solid line is $\hat{f}_1$ from the scenario where $f_2 = 0$ and f_2 is not estimated. The long dashed line is $\hat{f}_1$ from the scenario where $f_2 \neq 0$ but f_2 is incorrectly fixed at zero during estimation. The short dashed line is $\hat{f}_1$ from the scenario where $f_2 \neq 0$ and f_2 is estimated.

when $\lambda_{p0}^2/(\lambda_{p0}^2 + \sigma_{\epsilon,p}^2) = 0.3$ or 0.8). However, the simultaneous confidence band only rejects 0.2%. Pointwise confidence intervals eliminate the need for a wider simultaneous confidence band when testing is the main purpose. The pointwise confidence intervals perform well because my model gives a smooth estimate of $\hat{f}_1$ and this makes the coverage of zero among neighboring $f_1(z_j)$ highly associated. In order to make sure my results are not influenced by the sample size $M^{(\infty)}$ of $\boldsymbol{\beta}_p^{(m)}$ and z_j, I have also let $M^{(\infty)} = 10000$ when constructing the confidence intervals and have also used 1000 z_j (vs. 500) values when assessing the coverage, and the results do not change. Tests from multi-group analyses preserve their nominal rate since the distributions are known.

Next I use increasing values of κ so that f_1 is non-zero, and the rejection rate reflects the power of the test. The larger κ is, the more f_1 deviates from zero, and thus the power should increase. Figure 3.1 shows a comparison of the different testing approaches when data are generated with $f_1 \neq 0$, $f_2 = 0$ and the correct model is estimated. Since $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ is conservative for LRT, it also lowers the power of the test

Table 3.1: Rejection rate for testing H_0: $f_1 = 0$ vs. $f_1 \neq 0$ using the random coefficient model and the multi-group model. For f_1, null, cyclic, and monotone refer to the three shapes f_1 takes. The amplitude parameter for the cyclic and monotone shapes are 0.85 and 0.75, respectively. For f_2, null, non refer to whether the true f_2 is set as constant or not. For random coefficient models in both groups, we use either the 95-th percentile of the LRT statistic from the scenario where f_1 and f_2 are both null (referred to as 'Exact') or the 95-th percentile of $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ as the critical value for the other scenarios in the group. For multi-group models, we examine the rejection rate when tertiles or quantiles of z_i are used to group the data.

f_1	f_2	Only f_1 is estimated				f_1 and f_2 are both estimated			
		Random Coefficient		Multi-Group		Random Coefficient		Multi-Group	
		Exact	$\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$	Quartile	Tertile	Exact	$\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$	Quartile	Tertile
Null	Null	5.0	3.0	5.1	5.3	5.0	2.8	5.0	5.0
Cyclic	Null	82.5	78.7	23.2	5.3	82.6	78.8	23.0	5.3
Monotone	Null	85.3	79.7	72.9	77.3	85.2	79.0	71.2	75.6
Null	Non	8.0	5.1	7.0	7.7	5.0	3.0	5.0	5.0
Cyclic	Non	83.5	80.1	26.3	7.9	82.6	78.7	23.1	5.3
Monotone	Non	67.3	59.2	52.2	58.0	81.0	74.1	71.0	75.4

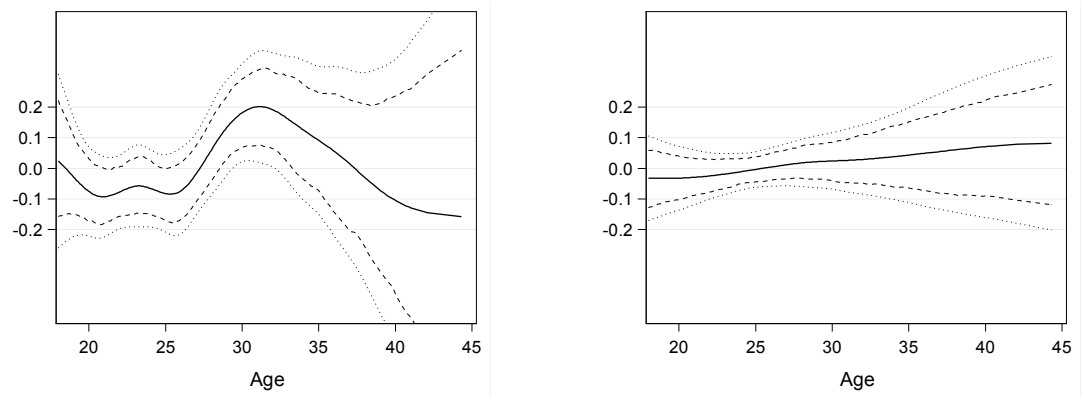
compared with using the 95-th percentile of the true null distribution. In comparison, the power of using pointwise confidence intervals for testing is better than the true power of LRT when f_1 has the cyclic shape and almost as good as the true power when f_1 has the monotonic trend (Figure 3.1), while the simultaneous confidence band has very low power. Thus the computationally efficient pointwise confidence intervals can be a preferable approach to estimating whether a factor loading is constant, although I can obtain a P-value using LRT. However, if I want to use confidence intervals to know the plausible range for the true curve, then the simultaneous confidence band has much better coverage than the pointwise confidence intervals (see Appendix C of the Supplementary Materials). The multi-group models have lower power than my model because they cannot capture the shape of f_1, especially when f_1 has the cyclic pattern. For example, in scenarios where f_1 is cyclic, I cannot detect a non-zero f_1 when I use tertiles because f_1 has a three-fold repeated pattern. When I unnecessarily estimate two non-constant factor loadings, I do not lose power (compare in Table 3.1 the two groups of scenarios when f_2 is null). This is because my model gives a stable estimate of f_1 whether f_2 is also estimated or not. The rejection rate of using confidence intervals when f_2 is estimated and the true f_2 is null is similar to when

f_2 is null and not estimated.

In the second group of simulation scenarios I investigate how estimating or ignoring f_2 in the model affects the estimation and testing of f_1. If I do not estimate f_2, $\hat{f}_1$ is biased (Figure 3.2). This is because the factor loading is associated with the scale of the observed variable, and the relative scale between two observed variables is related to the relative magnitude of their factor loadings. If I assume $f_2(z_i) = 0$ for all z_i, then, with the parameter settings I have, $\hat{\lambda}_{2i}$ will be positively biased for $z_i < 0.5$ and negatively biased for $z_i > 0.5$. This will in turn influence $\hat{\lambda}_{1i}$ in the same direction because the relative scale between $y_{1,i}$ and $y_{2,i}$ does not change. Therefore, $\hat{f}_1(z_i)$ will be positively biased for $z_i < 0.5$ and negatively biased for $z_i > 0.5$, which inflates Type I error and increases power when f_1 is cyclic because the estimated amplitude of $\hat{f}_1$ is positively biased. However, power is lower when f_1 has a monotonic trend because the amplitude is instead attenuated. The multi-group approach suffers similar consequences when the model is misspecified. When I estimate both f_1 and f_2, the bias in $\hat{f}_1$ nearly disappears. However, since a small amount of bias remains if f_1 has a monotone trend (Figure 3.2), the power is still smaller than that from the scenario where f_2 is actually zero (81.0% compared to 85.2% in Table 3.1).

3.5 Data Application

I use data collected from 880 mother-child pairs from the ELEMENT project. Mothers were between 18-44 years old at recruitment (mean=25.8, SD=5.0) The ELEMENT project recruited pregnant women in Mexico City between 1994 and 2003 to investigate the long-term consequences of lead exposure on child development. The project took prenatal and postnatal measurements from mothers and also followed the children longitudinally [100, 101]. The four observed variables I am



(a) Cord Blood Lead

(b) Tibia Bone Lead

Figure 3.3: The non-constant component for the factor loadings of cord blood lead and tibia bone lead. The solid line shows $\hat{f}_p(z_i)$ from Model A. The dashed line and the dotted line show the the 95% pointwise confidence intervals and the simultaneous confidence band, respectively.

interested in are cord blood lead, blood lead one month after delivery, patella bone lead, and tibia bone lead of the mothers. These lead biomarkers can be conceptualized as manifestations of latent lead exposure during pregnancy [81]; hence, I use a one factor model to summarize them. I am interested in looking at whether and how the factor loadings for each observed variable differ with maternal age. Maternal age has the potential to modify the factor loadings because the biological processes involved in lead metabolism in the body (e.g. bone resorption rate [101]) depend on maternal age. Because of this, the correlation between blood and bone lead measures may depend on maternal age as well. In turn, the factor loadings, which rely on correlations among observed variables, would also vary with maternal age. As described in Section 3.2, I first detrend the lead biomarkers to remove any potential covariate effects on the mean. To detrend the data (i.e., center lead biomarkers at zero), for each biomarker, I use an additive model [15] with indicators for participant's cohort membership and a smooth term of maternal age (recall three pregnancy cohorts are included in the ELEMENT project). I take the residuals from these models as the input to my latent factor model.

Theoretically some factor loadings can be shrunk exactly to a constant in my model when the variance of the random spline coefficients, $\hat{\sigma}^2_{b,p}$, is zero. However, the estimation is built on the normal distribution of $\boldsymbol{\beta}_p$, so even when $\hat{\sigma}^2_{b,p}$ approaches zero as the algorithm converges, it will not be exactly zero since a distribution to sample $\boldsymbol{\beta}_p$ from is needed. Thus, in an initial exploratory analyses where all four factor loadings were estimated as non-constant, I examined convergence of $\sigma^2_{b,p}$ for all four variables. I decide to set the factor loading for blood lead after delivery and patella bone lead factor loadings as constant and proceed to model the factor loadings of cord blood lead and tibia bone lead as varying with maternal age.

Next I conduct analyses in two models to illustrate the testing approaches previously described. In Model A the factor loadings of both cord blood lead and tibia bone lead are modeled as non-constant, while in Model B only one of the two observed variables, either cord blood lead (Model B1) or tibia bone lead (Model B2), has a non-constant factor loading. In both models, and for both biomarkers, I find that the factor loadings deviate from a constant, although the deviation for cord blood lead's factor loading is much stronger (Figure 3.3).

The patterns of the two factor loadings are similar between Model A and Model B. However, as in the simulations, if a truly non-constant factor loading is estimated as constant, the estimation for the other non-constant factor loadings can be affected. In this data analysis I find that the amplitudes of $\hat{f}_p$ for both cord blood lead and tibia bone lead are attenuated when the other factor loading is estimated as constant; that is, Model B1 and Model B2 compared with Model A (figure not shown). This attenuation can also be seen numerically by examining $\hat{\sigma}^2_{p,b}$, the variance of the random spline coefficients (Table 3.2). A bigger $\hat{\sigma}^2_{b,p}$ is related to a more pronounced deviation of $\hat{f}_p$ from zero. I find that $\hat{\sigma}^2_{b,p}$ from Model B is smaller than Model A.

Table 3.2: Parameter estimates for the model fitted under the alternative hypothesis, and LRT p-values for the ELEMENT study data. The P-values are obtained using the 50:50 mixture: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$; the parametric boostrap distribution (PBoot): $\hat{p}\chi_0^2 + (1-\hat{p})\hat{a}\chi_1^2$; and the multi-group analysis with maternal age in tertiles [MG(Tert)] or in quartiles [MG(Quart)]. $M^* = 1000$ parametric bootstrap samples are used to fit $\hat{p}\chi_0^2 + (1-\hat{p})\hat{a}\chi_1^2$. The parametric bootstrap samples are generated from parameters obtained by fitting the model under the null hypothesis, i.e., the factor loading being tested is estimated as constant, but the other factor loadings are constant (Models B1 and B2) or not (Model A).

Model or Observed Variable	Estimates			P-values			
	$\hat{\lambda}_{p,0}$	$\hat{\sigma}^2_{p,\epsilon}$	$\hat{\sigma}^2_{p,b}$	50:50 Mix	PBoot	MG(Tert)	MG(Quart)
Model A							
Cord Blood Lead	0.68	0.52	0.049	0.005	0.0035	0.0001	0.027
Tibia Bone Lead	0.32	0.89	0.0023	0.29	0.18	0.42	0.19
Blood Lead After Delivery	0.82	0.32	-	-	-	-	-
Patella Bone Lead	0.41	0.83	-	-	-	-	-
Model B1							
Cord Blood Lead	0.68	0.52	0.044	0.006	0.004	0.00011	0.029
Tibia Bone Lead	0.32	0.9	-	-	-	-	-
Blood Lead After Delivery	0.83	0.31	-	-	-	-	-
Patella Bone Lead	0.4	0.84	-	-	-	-	-
Model B2							
Cord Blood Lead	0.68	0.53	-	-	-	-	-
Tibia Bone Lead	0.32	0.89	0.0012	0.37	0.23	0.48	0.2
Blood Lead After Delivery	0.83	0.31	-	-	-	-	-
Patella Bone Lead	0.4	0.84	-	-	-	-	-

The pointwise confidence intervals in both Model A and B suggest that the factor loading of cord blood lead is significantly non-constant, while there is no significant evidence that the factor loading for tibia bone lead is non-constant (Figure 3.3). This conclusion agrees with LRT using critical values from either $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ or $p\chi_0^2 + (1-p)a\chi_1^2$ (Table 3.2). In both Model A and B, $\hat{p} \approx 0.7$ and $\hat{a} \approx 0.9$. The P-value based on $\hat{p}\chi_0^2 + (1-\hat{p})\hat{a}\chi_1^2$ is smaller than the P-value based on $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, as expected. The P-values obtained through the multiple group analyses depend on whether tertiles or quartiles of the maternal age distribution are used to group the observations. In particular, the P-value obtained when using tertiles is smaller for cord blood lead, which makes sense because the pattern of the factor loading is U-shaped, and thus age tertiles capture the difference of middle tertile compared to the other two more accurately. In contrast, the P-value for the tibia lead factor loading, which follows a linear pattern, is smaller when using quartiles than tertiles.

In these data, the factor loading of cord blood lead differs depending on maternal age. The non-linear pattern in the non-constant component of the factor loading for cord blood implies that the factor loading among the youngest and oldest mothers is lower compared to mothers in the center of the age distribution. This implies that the use of a latent factor model with constant factor loadings may introduce measurement bias into summaries created from it. In other words, the factor scores created from a model assuming constant factor loadings may not correctly rank overall exposure levels among study participants, and the bias in rankings would be related to maternal age. The impact of this bias on estimated associations between this latent factor is out of the scope of this chapter and will be investigated in future work.

3.6 Discussion

In this paper I propose a latent factor model that estimates non-constant factor loadings using random spline coefficients. This type of model is important because it allows the measurement model to differ across values of another covariate, and thus enables assessment of measurement bias. Compared with multi-group models that also allow factor loadings to differ across covariate values, my model is not subject to power loss due to poor choices of group assignment for categorizing continuous covariates. It also provides me with more flexibility in capturing the shape of the non-constant pattern. I apply my methods to data collected as part of the ELEMENT project, where a one factor model is used to summarize four lead exposure biomarkers, and I show that there are significant differences in the measurement model across maternal age.

I implement the model using the EM algorithm. Because the random spline

coefficients are unobserved in addition to the latent factor, the algorithm calls for more involved EM approaches. I use Monte-Carlo sampling method to facilitate the E-step. More advanced data augmentation techniques can be used to make the algorithm more efficient [102].

I am able to carry out LRT by using Monte-Carlo integration to calculate the likelihood and parametric bootstrap to approximate the null distribution of the LRT statistic. As stated in the paper by Crainiceanu et al. [94], the classical $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ for testing one variance component is conservative. But a more general alternative $p\chi_0^2+(1-p)a\chi_1^2$ proposed in [96] can approximate well the null distribution of the LRT statistic in my model. The method of moments used in the paper by Greven et al. [96] also proves to be a robust method for avoiding the numerical imprecision when the variance component is shrunk to zero. My model also allows me to easily construct confidence intervals for the non-constant component of the factor loading using the posterior distribution of the spline coefficients. I find that the pointwise confidence interval is as powerful as using the LRT. This can be the preferred method in practice because it avoids the computational burden associated with LRT, although it does not yield a p-value.

In summary, I propose novel approaches to estimate factor loadings that vary smoothly as functions of covariates and to test for measurement invariance. Future studies should investigate the impact of measurement bias when examining associations between covariates and the latent factor, or when the latent factor is a predictor of other outcomes.

CHAPTER IV

Estimating an Overall Nonlinear Exposure Effect for Multiple Continuous Outcomes in a Latent Factor Model

4.1 Introduction

Most epidemiological studies collect measurements on multiple outcomes in order to assess the potential hazard of a particular exposure. Such a study design is based either on the nature of the outcome, for example, a psychometric test battery, or on concerns that the exposure effect is typically small and hard to detect from an observational study using a single outcome. Regressions of multiple outcomes on the exposure may provide higher power to detect the association between exposure and outcomes. Since the multivariate outcomes are usually related to the same trait and thus correlated, modeling the correlation among them may help avoid loss of power when correcting for multiple testing [103]. This chapter is motivated by a study examining the effect of lead (Pb) exposure in the ELEMENT study, measured as the concentration of lead in children's blood, on children's behavioral development assessed with the Behavior Assessment System for Children (BASC) rating scales. The BASC battery includes 14 subscores, each capturing a specific aspect of development. In addition to modeling the effect of exposure on multiple outcomes, one of the main interests in this chapter is to assess the non-linear effect of lead on behavioral development. Previous studies have shown some evidence for

the low-dose effect of lead which reaches a ceiling when blood lead concentration is high [104].

A number of papers have developed models that incorporate multivariate outcomes to assess the association with a given exposure. Generally these models account for the correlation among the outcomes and estimate a linear exposure effect on different outcomes in a single model, thereby allowing a joint test for the overall exposure effect [103]. Some models directly estimate an overall exposure effect and use various scaling approaches to obtain outcome-specific exposure effects. Jia et al. [105] scale the whole linear predictor including covariate effects and estimate the scaling parameters. In the paper by Sammel et al. [46], the approach is to model the correlation among outcomes through a latent factor model and scales the overall exposure effect with the factor loadings of each outcome. By contrast, Lin et al. [47] first model the covariance among outcomes using random effects and use the residual variance for scaling the exposure effects. In the paper by Roy et al. [48], however, the scaling is based on the total standard deviation of the outcomes, including their covariance. In comparison, several papers model outcome-specific exposure effects by first assuming a common overall effect (termed 'domain effect'), and model outcome-specific effects as random deviations from the domain effect [106, 107]. In general, the aim of these available approaches is to reduce the dimension of the model and to improve estimation efficiency.

I use a latent factor model to summarize the correlation among the BASC subscores, and I estimate the exposure effect on the latent factors. Under this model construction, the outcome-specific exposure effect is the factor-specific exposure effect scaled by the factor loadings. My model is similar to that developed by Sammel et al. [46] but it also borrows the idea of 'domain effect' from the paper by Thurston

et al. [106]. However, a key innovation is that I also develop an approach to estimate an overall nonlinear lead exposure effect (Section 4.2, 4.3), and develop an approach to test the deviation from a linear exposure effect (Section 4.4). Through simulation studies (4.5) I show the relative advantages and disadvantages of modeling an overall nonlinear exposure effect using my proposed approach compared to two other modeling approaches. I use the proposed model, and alternative models, to estimate the effect of lead exposure on the developmental domains captured by the BASC battery. The discussion in Section 4.7 ends with my conclusions and potential directions for future work.

4.2 Model

I consider three models that estimate the nonlinear exposure effect of x_i on multiple continuous outcomes $y_{p,i}$. I let $i = 1, \ldots, n$ denote different subjects and $p = 1, \ldots, P$ denote different outcomes. These three models can all be viewed as special cases of a model with the following general form:

$$y_{p,i} = Z_i \boldsymbol{\alpha}_p + f_p(x_i) + \zeta_{p,i}, \tag{4.1}$$

where Z_i is a $1 \times M$ vector of adjusting covariates (for ease of notation intercept is also included here as a column of 1) and $\boldsymbol{\alpha}_p$ is the vector of related coefficients, x_i is a scalar measure of the exposure, $f_p(x_i)$ is an unspecified function that captures the nonlinear relationship between $y_{p,i}$ and x_i, $\zeta_{p,i}$ is a residual error; a model on the covariance matrix of $\zeta = \zeta_{1,i}, \ldots, \zeta_{P,i}$ may be used to account for the correlation among the P outcomes measured for one subject.

It is of interest to test whether $f_p(x_i)$ deviates from a simple linear exposure effect; under the null hypothesis I have

$$H_0 : f_p(x_i) = x_i \beta_p. \tag{4.2}$$

Under the alternative hypothesis,

$$H_a : f_p(x_i) = x_i\beta_p + \mathbf{X}_i\boldsymbol{\gamma}_p,$$

where X_i is a $1 \times K$ vector of spline bases, $\boldsymbol{\gamma}_p$ is a $K \times 1$ vector of spline coefficients. Here, $X_i\boldsymbol{\gamma}_p$ captures the deviation of the fitted exposure effect from being a linear effect. The term $X_i\boldsymbol{\gamma}_p$ is parameterized as cubic P-splines, with the coefficients penalized such that the spline shrinks toward a first-degree polynomial, that is, a linear term [16, 22]. The penalty is implemented using a restricted maximum likelihood (REML) approach, where the coefficients $\boldsymbol{\gamma}_p$ are treated as random. Therefore, the smoothing parameter becomes the variance of the random coefficients (see Section 4.9 in the book by Ruppert et al. [40], for example).

Model A considers separate regressions for each outcome. This approach is the most straightforward, but does not capitalize on the potential correlation among outcomes to gain efficiency:

$$y_{p,i} = Z_i\boldsymbol{\alpha}_p + x_i\beta_p + X_i\boldsymbol{\gamma}_p + \epsilon_{p,i}, \tag{4.3}$$

where $\epsilon_{p,i} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\epsilon,p})$. Since I model each outcome separately, I disregard the correlation among outcomes. Using matrix form, I then write (4.3) as

$$\mathbf{y}_p = \mathbf{Z}\boldsymbol{\alpha}_p + \mathbf{x}\beta_p + \mathbf{X}\boldsymbol{\gamma}_p + \boldsymbol{\epsilon}_p, \tag{4.4}$$

where $\mathbf{y}_p = [y_{p,1}, \ldots, y_{p,n}]^T$, $\mathbf{Z} = [Z_1^T, \ldots, Z_n^T]^T$, $\mathbf{x} = [x_1, \ldots, x_n]^T$, $\mathbf{X} = [X_1^T, \ldots, X_n^T]^T$, $\boldsymbol{\epsilon}_p = [\epsilon_{p,1}, \ldots, \epsilon_{p,n}]^T$. This approach requires estimation of P smoothing parameters, and independently tests the P exposure effects.

Model B considers modeling the correlation among outcomes and jointly estimating a separate exposure effect for each outcome. Many potential correlation structures can be chosen for the outcomes. Given that in the case of the BASC

the multiple scores are nested into four domains, I consider using a factor analytic structure for the residual errors. This factor analytic structure is also supported by initial exploratory analysis. I first modify (4.3) by adding the latent factor η_i to impose the factor analytic structure,

$$y_{p,i} = Z_i\boldsymbol{\alpha}_p + x_i\beta_p + X_i\boldsymbol{\gamma}_p + \lambda_p\eta_i + \epsilon_{p,i}, \tag{4.5}$$

where $\eta_i \overset{\text{iid}}{\sim} N(0,1)$, λ_p is the factor loading. The variance of η_i is fixed at 1 to ensure identifiability. If I compare (4.5) with (4.1), I can see that $\zeta_{p,i} = \lambda_p\eta_i + \epsilon_{p,i}$ and is correlated for different p within same subject i. Using Kronecker product, I rewrite this model in matrix form as follows,

$$\mathbf{Y} = (\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\alpha} + (\mathbf{I}_P \otimes \mathbf{x})\boldsymbol{\beta} + (\mathbf{I}_P \otimes \mathbf{X})\boldsymbol{\gamma} + (\boldsymbol{\lambda} \otimes \boldsymbol{\eta}) + \boldsymbol{\epsilon}, \tag{4.6}$$

where $\mathbf{Y} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_P^T]^T$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_P^T]^T$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]^T$, $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_P^T]^T$, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_P]^T$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^T$. Compared to Model A, Model B can take into account the correlation among multiple outcomes to gain efficiency, although it still estimates P exposure effects. However, Model B also has the advantage of allowing for a joint test of whether exposure affects any of the outcomes (e.g., in the case of a linear effect only, this model enables testing of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_P = 0$ vs. H_a : at least one differs from 0).

Model C assumes an overall exposure effect; the model takes the form

$$\mathbf{Y} = (\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\alpha} + (\boldsymbol{\lambda} \otimes \mathbf{x})\tilde{\beta} + (\boldsymbol{\lambda} \otimes \mathbf{X})\tilde{\boldsymbol{\gamma}} + (\boldsymbol{\lambda} \otimes \boldsymbol{\eta}) + \boldsymbol{\epsilon}. \tag{4.7}$$

Compared with (4.6), I only estimate an overall exposure effect. The linear effect is captured by $\tilde{\beta}$, and the deviation from linearity is captured by $\tilde{\boldsymbol{\gamma}}$. The overall exposure effect is scaled with the factor loadings, $\boldsymbol{\lambda}$, so that the magnitude of the exposure effect for each outcome is different.

This model can be more appealing than Model B (4.6) when the nonlinear exposure effect for each outcome is similar and when it is plausible to think the exposure affects the multivariate outcomes through certain underlying traits which in turn can be well represented by the latent factor. If I switch the order of Kronecker product and matrix multiplication in (4.7), e.g., $(\boldsymbol{\lambda}\otimes\mathbf{X})\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\lambda}\otimes\mathbf{X})(1\otimes\tilde{\boldsymbol{\gamma}}) = (\boldsymbol{\lambda}\times 1)\otimes(\mathbf{X}\tilde{\boldsymbol{\gamma}}) = \boldsymbol{\lambda}\otimes(\mathbf{X}\tilde{\boldsymbol{\gamma}})$, then I can rewrite (4.7) as

$$\mathbf{Y} = (\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\alpha} + \boldsymbol{\lambda}\otimes(\mathbf{x}\tilde{\beta} + \mathbf{X}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\eta}) + \boldsymbol{\epsilon}. \tag{4.8}$$

In (4.8) it is apparent that the individual exposure effect comes from an overall effect on the mean of the latent factor. Model C is advantageous because it enables testing for a single exposure effect, and a single test of deviation from linearity.

4.3 Estimation

As I stated in Section 4.2, I use P-splines to estimate the nonlinear exposure effects and I estimate the P-splines using REML. For Model A and B, I model the spline coefficients as $\boldsymbol{\gamma}_p \overset{\text{iid}}{\sim} N(0, \sigma^2_{\gamma,p}\mathbf{I}_K)$ (after proper matrix transformation, see Appendix and also these papers [40, 16, 22]). The computation can be directly carried out using standard software; I use PROC MIXED in SAS.

Since Model C involves scaling the exposure effect using factor loadings, I develop an iterative procedure similar to that used by Lin et al. [47]. Given the factor loadings, I estimate the nonlinear exposure effect and the other parameters using PROC MIXED. Then I use Newton's method to update the factor loadings.

Specifically, I run PROC MIXED for the following model,

$$\tilde{\mathbf{Y}} = (\boldsymbol{\Lambda}^{-1}\otimes\mathbf{I}_n)\mathbf{Y} = (\mathbf{I}_P\otimes\mathbf{Z})\tilde{\boldsymbol{\alpha}} + (\mathbf{1}_P\otimes\mathbf{x})\tilde{\beta} + (\mathbf{1}_P\otimes\mathbf{X})\tilde{\boldsymbol{\gamma}} + (\mathbf{1}_P\otimes\boldsymbol{\eta}) + \tilde{\boldsymbol{\epsilon}}, \tag{4.9}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_P)$ so that $\boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda} = \mathbf{1}_P$, $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\Lambda}^{-1}\otimes\mathbf{I}_M)\boldsymbol{\alpha}$, $\tilde{\boldsymbol{\epsilon}} = (\boldsymbol{\Lambda}^{-1}\otimes\mathbf{I}_n)\boldsymbol{\epsilon}$, $\tilde{\boldsymbol{\gamma}} \sim N(0, \tilde{\sigma}^2_\gamma\mathbf{I}_K)$. In (4.9), the outcomes are 'standardized' by the given factor

loadings. In this step of the fitting procedure, a factor analytic correlation structure is being used for the multivariate outcome, but the factor loadings are fixed at 1. (I use 'standardize' to refer to multiplication by $\mathbf{\Lambda}^{-1}$ and 'scale' refers to multiplication by $\mathbf{\Lambda}$.)

I can justify this iterative approach by examining the likelihood for the original model (see 4.7). For notational purposes, I let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$, $\tilde{\boldsymbol{\mu}} = \mathrm{E}(\tilde{\mathbf{Y}}) = (\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_n)\boldsymbol{\mu}$, $\mathbf{V} = \mathrm{Cov}(\mathbf{Y})$, $\tilde{\mathbf{V}} = \mathrm{Cov}(\tilde{\mathbf{Y}}) = (\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_n)\mathbf{V}(\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_n)$. I replace $\boldsymbol{\mu}$, $\mathbf{V}$ in the following REML criterion with $\mathbf{Y} = (\mathbf{\Lambda} \otimes \mathbf{I}_n)\tilde{\mathbf{Y}}$, $\boldsymbol{\mu} = (\mathbf{\Lambda} \otimes \mathbf{I}_n)\tilde{\boldsymbol{\mu}}$, $\mathbf{V} = (\mathbf{\Lambda} \otimes \mathbf{I}_n)\tilde{\mathbf{V}}(\mathbf{\Lambda} \otimes \mathbf{I}_n)$, and get the following identity,

$$\begin{aligned} & \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right] \\ = & \frac{|\mathbf{\Lambda}|^{-n}}{\sqrt{(2\pi)^n |\tilde{\mathbf{V}}|}} \exp\left[-\frac{1}{2}(\tilde{\mathbf{Y}}-\tilde{\boldsymbol{\mu}})^T (\tilde{\mathbf{V}})^{-1}(\tilde{\mathbf{Y}}-\tilde{\boldsymbol{\mu}})\right]. \end{aligned}$$

Therefore, given $\mathbf{\Lambda}$, the likelihood is maximized by solving the MLE for Model (4.9).

If I define the REML criterion for the original model (4.7) as

$$\mathcal{L}_R = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \iint \exp\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right] d\tilde{\boldsymbol{\alpha}} d\tilde{\beta},$$

then I can also use REML to estimate the variance components in model (4.9), following the same reasoning as in the MLE case. The definition of the REML criterion also follows that of a linear mixed model, where the likelihood is integrated over the fixed effects. But for Model C I integrate over the standardized fixed effects.

I carry out the iterative algorithm as follows.

(1) Given $\hat{\boldsymbol{\lambda}}^{(r-1)}$, I standardize $\mathbf{Y}$ and get $\tilde{\mathbf{Y}}^{(r)}$.

(2) I run model (4.9) in PROC MIXED using $\tilde{\mathbf{Y}}^{(r)}$ as the outcome and estimate $\tilde{\boldsymbol{\mu}}$, $\tilde{\mathbf{V}}$, denoted as $\widehat{\tilde{\boldsymbol{\mu}}}^{(r)}$, $\widehat{\tilde{\mathbf{V}}}^{(r)}$.

(3) Then I update $\hat{\boldsymbol{\lambda}}$ as

$$\hat{\boldsymbol{\lambda}}^{(r)} = \hat{\boldsymbol{\lambda}}^{(r-1)} - \mathbf{U}^{-1}\mathbf{u},$$

where the p-th entry of the $P \times 1$ vector $\mathbf{u}$ is $\frac{\partial}{\partial \lambda_p} \log \mathcal{L}_R$, and the (p, p')-th entry of the $P \times P$ matrix $\mathbf{U}$ is $\frac{\partial^2}{\partial \lambda_p \partial \lambda_{p'}} \log \mathcal{L}_R$. $\mathbf{u}$, $\mathbf{U}$ can be computed as follows.

Let $\mathbf{r}_p^{(r)} = \Delta_p \left(\widehat{\tilde{\mathbf{V}}}^{(r)} \right)^{-1} \left(\tilde{\mathbf{y}}_p^{(r)} - \tilde{\boldsymbol{\mu}}_p^{(r)} \right)$, $\mathbf{W}_{p,p'}^{(r)} = \Delta_p \left(\widehat{\tilde{\mathbf{V}}}^{(r)} \right)^{-1} \Delta_p$, where $\Delta_p = \mathbf{I}_n \otimes \delta_p$ and the (p, p')-th entry of the $P \times P$ matrix δ_p is 1 and otherwise all zero, then

$$\frac{\partial}{\partial \lambda_p} \log \mathcal{L}_R = \frac{1}{\lambda_p^{(r)}} \left[(\tilde{\mathbf{y}}_p^{(r)})^T \mathbf{r}_p^{(r)} - n \right]$$

$$\frac{\partial^2}{\partial \lambda_p \partial \lambda_{p'}} \log \mathcal{L}_R = \frac{1}{\lambda_p^{(r)} \lambda_{p'}^{(r)}} \left\{ I(p' = p) \left[2(\tilde{\mathbf{y}}_p^{(r)})^T \mathbf{r}_p^{(r)} - n \right] + (\tilde{\mathbf{y}}_p^{(r)})^T \mathbf{W}_{p,p'}^{(r)} \tilde{\mathbf{y}}_p^{(r)} \right\}$$

4.4 Hypothesis Testing

For Model A, I test (4.2) for each outcome as testing

$$H_0 : \sigma_{\gamma,p}^2 = 0 \quad \text{vs.} \quad H_a : \sigma_{\gamma,p}^2 > 0. \tag{4.10}$$

For Model B, I perform a joint test of whether any exposure effect deviates from linearity. The test is equivalent to

$$H_0 : \sigma_{\gamma,p}^2 = 0 \text{ for all } p \quad \text{vs.} \quad H_a : \sigma_{\gamma,p}^2 > 0 \text{ for some } p. \tag{4.11}$$

For Model C, the test is equivalent to

$$H_0 : \tilde{\sigma}_\gamma^2 = 0 \quad \text{vs.} \quad H_a : \tilde{\sigma}_\gamma^2 > 0. \tag{4.12}$$

It is not straightforward to derive the null distribution of the LRT statistic because the parameter under the null is on the boundary of the parameter space. As stated also in Section 3.1, Stram et al. [93] derived an asymptotic distribution for testing variance components in a linear mixed model. Crainiceanu et al. [94], however, argued against the use of the asymptotic distribution, especially for cases like using

REML for the estimation of penalized splines. Since then, some methods [108, 109] have been developed for testing one variance component, but those methods cannot be directly applied to my test of interest. The test in (4.10) involves multiple variance components, so only the questionable asymptotic distribution from the paper by Stram et al. [93] can be used for this testing purpose. Even though the test in (4.12) only examines one variance component, the existing methods would be applied to the model in (4.9), where I do not take into account of the degrees of freedom lost through estimating $\boldsymbol{\lambda}$.

Therefore, I use parametric bootstrap to obtain the null distribution of the LRT statistic and use the 95-th percentile of the simulated LRT statistic as the cutoff value for testing. However, for Model A, I test each outcome in a separate regression and reject H_0 of (4.10) if the LRT test statistic is larger than the cutoff in any of the P regressions. Since this approach is subject to multiple testing, I use Bonferroni adjustment. Note that for Model A, I can also derive the exact null distribution of the LRT statistic by using the approach developed in these two papers [94, 97].

In the data application of Section 4.6, I also first use LRT to test whether the linear exposure effect is significant while not assuming a nonlinear exposure effect in the model (Table 4.5). For Model A, I test for each p,

$$H_0 : \beta_p = 0 \quad \text{vs.} \quad H_a : \beta_p \neq 0. \tag{4.13}$$

For Model B, I jointly test whether there is a linear exposure effect for any outcome, which is

$$H_0 : \beta_p = 0 \text{ for all } p \quad \text{vs.} \quad H_a : \beta_p \neq 0 \text{ for some } p. \tag{4.14}$$

In Model B I can also carry out the test in (4.13) and obtain a P-value for the linear exposure effect of each outcome, but the estimates and P-value will be the same as

in Model A (see Appendix D).

For Model C, I test the overall linear exposure effect,

$$H_0 : \tilde{\beta} = 0 \quad \text{vs.} \quad H_a : \tilde{\beta} \neq 0. \tag{4.15}$$

I also use LRT to compare Model B and C so that I can assess whether the more parsimonious Model C can adequately represent each individual exposure effect. For this purpose, I test if the exposure effects scaled by the corresponding factor loadings are the same across all P outcomes:

$$\begin{gathered} H_0 : \beta_p/\beta_{p'} = \lambda_p/\lambda_{p'} \text{ for all } (p, p') \text{ pairs} \\ \text{vs.} \\ H_a : \beta_p/\beta_{p'} \neq \lambda_p/\lambda_{p'} \text{ for some } (p, p') \text{ pair .} \end{gathered} \tag{4.16}$$

Since I am testing fixed effects and not estimating random coefficients, I use maximum likelihood instead of REML for these tests, and I also assume the LRT statistic under the null follows the asymptotic chi-square distribution. The degree of freedom for the tests in (4.13) and (4.16) is one. It is P for the test in (4.14) and $P - 1$ for the test in (4.16).

4.5 Simulation

This simulation study compares how well the three different models in Section (4.2) can detect and estimate the nonlinear exposure effect. This section includes two simulations. I let Model C be correct in the first one and incorrect in the second one.

4.5.1 Simulation 1

Data Generation

First, I generate data where the true exposure effect agrees with Model C. For convenience I omit covariates in this simulation study. I let $P = 3$, and generate data

Table 4.1: Prediction error and power of LRT for all three models from Simulation 1

MSE of the Predicted Values $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$	Model	$\kappa=0$	$\kappa=0.5$	$\kappa=0.75$
0.3	A	0.74	1.31	1.68
	B	0.72	1.33	1.96
	C	0.59	1.09	1.18
0.5	A	0.44	0.94	1.07
	B	0.42	1.05	1.73
	C	0.37	0.77	0.79
0.8	A	0.28	0.65	0.65
	B	0.24	0.91	1.75
	C	0.26	0.58	0.58
Rejection Rate of LRT $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$	Model	$\kappa=0$	$\kappa=0.5$	$\kappa=0.75$
0.3	A	4.8	20.5	50.8
	B	-	9.1	21.9
	C	-	36.2	67.8
0.5	A	4.6	33.0	70.9
	B	-	7.6	19.3
	C	-	41.8	80.5
0.8	A	3.7	43.0	83.0
	B	-	5.9	8.7
	C	-	52.4	89.5

from the model:

$$y_{p,i} = \mu_p + \lambda_p \left[0.25x_i + \kappa \sin(\pi x_i) + \eta_i\right] + \epsilon_{p,i}, \tag{4.17}$$

where $\mu_p = 2$, $\lambda_p = 3$, $\epsilon_{p,i} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$ for all three outcomes. I change $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$ so that the signal-to-noise ratio varies across scenarios. I also change the size of the nonlinear exposure effect by letting κ take on three different values: 0, 0.5, 0.75. For all the simulations I generate data for a fixed set of 100 x_i randomly chosen from $N(0, 0.5^2)$. The number of simulations is 10,000 for the null scenario ($\kappa = 0$) and 1,000 for the other two.

Results regarding mean squared error

To assess estimation, I first look at the prediction error of $\hat{y}_{p,i} = \hat{\mu}_p + \hat{f}_p(x_i)$, where $f_p(x_i)$ is the nonlinear exposure effect in (4.1). In Table (4.1) the prediction error

is given as the mean squared error (MSE), $\frac{1}{nP}\sum_{p=1}^{P}\sum_{i=1}^{n}(\hat{y}_{p,i}-y_{p,i})^2$. I use the average MSE across all outcomes because they have the same magnitude in my simulation scenarios. My results show that Model C has the smallest MSE because it uses information from all the outcomes to estimate an overall effect rather than a separate exposure effect for each outcome. The difference between the MSE of Model A and Model C gets larger as $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$ increases, showing the loss of precision in Model A due to estimating a similar nonlinear exposure effect separately. However, even though Model B jointly estimates the separate exposure effects in a framework that takes into account the correlation among outcomes, it has larger MSE than Model A. This is because when I use REML to estimate P-splines, the information about the splines lies in the correlation among outcomes conditioned on the fixed effects (the linear exposure effect), and the correlation structure implied by Model B differs greatly from the simulation scenario where the overall nonlinear exposure effect also contributes to the correlation among outcomes (see detailed explanation below).

Results regarding power

Since one of my objectives is to test the nonlinear exposure effect using LRT, I also examine the power of the LRT in this simulation study. In order to make this simulation study computationally feasible, I obtain the null distribution of the LRT statistic using the generated data sets only, instead of carrying out a parametric bootstrap for each data set, similar to the approach proposed by Greven et al. [96]. The results I obtain this way will be similar to the full parametric bootstrap, but with less noise since this approach is the same as carrying out a parametric bootstrap based on the true parameters.

I find that Model C has higher power than Model A, but Model B has very

Table 4.2: Prediction error and power of LRT for all three models from Simulation 2

MSE of the Predicted Values $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$	Model	$\hat{\mathbf{y}}_1$	$\hat{\mathbf{y}}_2$	$\hat{\mathbf{y}}_3$
0.5	A	1.10	1.13	0.43
	B	1.26	1.26	0.56
	C	1.17	1.16	1.19
Rejection Rate of LRT $\lambda_p^2/(\lambda_p^2+\sigma_{\epsilon,p}^2)$				
0.5	A	59.2		
	B	61.9		
	C	56.4		

low power compared with Model A and C. The reason why Model B performs worse than the other two is because the data generated from this simulation scenario follows Model C. This means the nonlinear exposure effects for different outcomes are similar, or "correlated". Since I use P-splines and REML to estimate the nonlinear exposure effects, the linear component of the exposure effect is treated as a fixed effect and the coefficients that capture the nonlinear component are treated as random effects. So, essentially the estimation of the nonlinear component is determined by the correlation structure of the $\mathbf{Y}$ conditioned on the linear fixed effect. In this simulation scenario, the correlations among outcomes conditional on the linear fixed effect come not only from the shared latent factor, but also from the similar nonlinear exposure effects. In Model C, $\text{Cov}(\mathbf{Y}) = (\tilde{\sigma}_\gamma^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^T) \otimes (\mathbf{X}\mathbf{X}^T) + (\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \otimes \mathbf{I}_n + \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_n$, where $\boldsymbol{\Sigma}_\epsilon = \text{diag}(\sigma_{\epsilon,1}^2, \ldots, \sigma_{\epsilon,P}^2)$, so the correlation among different outcomes (indicated by the left-hand side of the Kronecker product) comes from both the random coefficients and the latent factor. However, in Model B, $\text{Cov}(\mathbf{Y}) = \boldsymbol{\Sigma}_\gamma \otimes (\mathbf{X}\mathbf{X}^T) + (\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \otimes \mathbf{I}_n + \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_n$, where $\boldsymbol{\Sigma}_\gamma = \text{diag}(\sigma_{\gamma,1}^2, \ldots, \sigma_{\gamma,P}^2)$. This means the correlation among different outcomes only comes from the latent factor and the nonlinear exposure effects are independent from each other. Therefore, this correlation structure does not fit the simulation scenario and the extra correlation induced by the overall nonlinear expo-

sure effect will be estimated by the latent factor. As a result, the factor loadings will be inflated and the diagonal of $(\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \otimes \mathbf{I}_n$ will also be inflated. This in turn will reduce the diagonal of $\boldsymbol{\Sigma}_\gamma \otimes (\mathbf{X}\mathbf{X}^T)$, which means the variance of the random coefficients (the diagonal of $\boldsymbol{\Sigma}_\gamma$) will be underestimated. In other words, the information of the nonlinear exposure effect is taken away by the latent factor in Model B. One way to improve Model B might be to let $\boldsymbol{\Sigma}_\gamma$ be unstructured, which would imply the P different splines are correlated with each other. But it is not straightforward to implement and may bring about additional issues, for example, how to ensure $\text{Cov}(\mathbf{Y})$ is positive definite.

4.5.2 Simulation 2

In practice, even though the data does not necessarily follow Model C as in Simulation 1, some correlation among outcomes will be related to the nonlinear exposure effect if the effects are similar across outcomes. So, Model C has a more general appeal than Model B. At the same time, however, Model C is built on a stronger assumption about how the overall exposure effect relates to the individual exposure effect for each outcome. Therefore, if the exposure effect for a particular outcome differs greatly from the scaled overall exposure effect, Model C may not be the best model. In order to illustrate this, I conduct a second simulation where I remove the exposure effect in (4.17) for the third outcome variable, $\mathbf{y}_3$ when generating the data. As a result, the scale for this outcome variable is not λ_3, but zero. I then repeat the same simulation study under $\kappa = 0.75$, and the results are listed in Table (4.2).

I find that Model C cannot properly estimate $\mathbf{y}_3$, compared with the other two models, since the MSE for $\hat{\mathbf{y}}_3$ is largest in Model C. However, the MSE for $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are not much worse than the other two models, even though they become larger compared to Table (4.1). As for the power of LRT in detecting any nonlinear

exposure effect, all three models are similar. Model B performs much better than in the scenarios of Table (4.1). This is because the relative scale of the correlation that is induced by the overall exposure effect now differs greatly from the relative scale of the factor loadings. Therefore, in this scenario, the latent factor cannot cover the correlation related to the overall exposure effect. As a result, I can get a better estimate of $\mathbf{\Sigma}_\gamma$ in Model B.

4.6 Data Application

I examine the potential nonlinear relationship between lead (Pb) exposure and children's behavior. Behavior is measured with the Behavior Assessment System for Children, Second Edition (BASC-2) on 516 children from The Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) project. The children's concurrent blood lead levels are taken as their lead exposure measurements (see Table 4.3 for a summary of the descriptive statistics for all the covariates and the outcomes).

BASC-2 scales consist of 14 subscores divided into four categories: internalizing problems, externalizing problems, behavioral symptoms index, adaptive skills. I also let the outcomes that are under the same category be related to the same latent factor. Then for each category I run the three models described in Section 4.2. For outcomes from the first three categories, except anxiety and attention problems, I use the logarithmic scale because the distribution under the original scale is highly skewed. I select covariates to adjust for based on prior knowledge [104]; all analyses are adjusted for child's age, gender, birth weight, and mother's age at delivery, education, socioeconomic level. Since the blood lead level is highly skewed, I use the logarithmic scale. Using the nonlinearity tests detailed in Section 4.4, I want to test whether the lead exposure effect significantly deviates from log-linearity. (Below

Table 4.3: Participant Characteristics

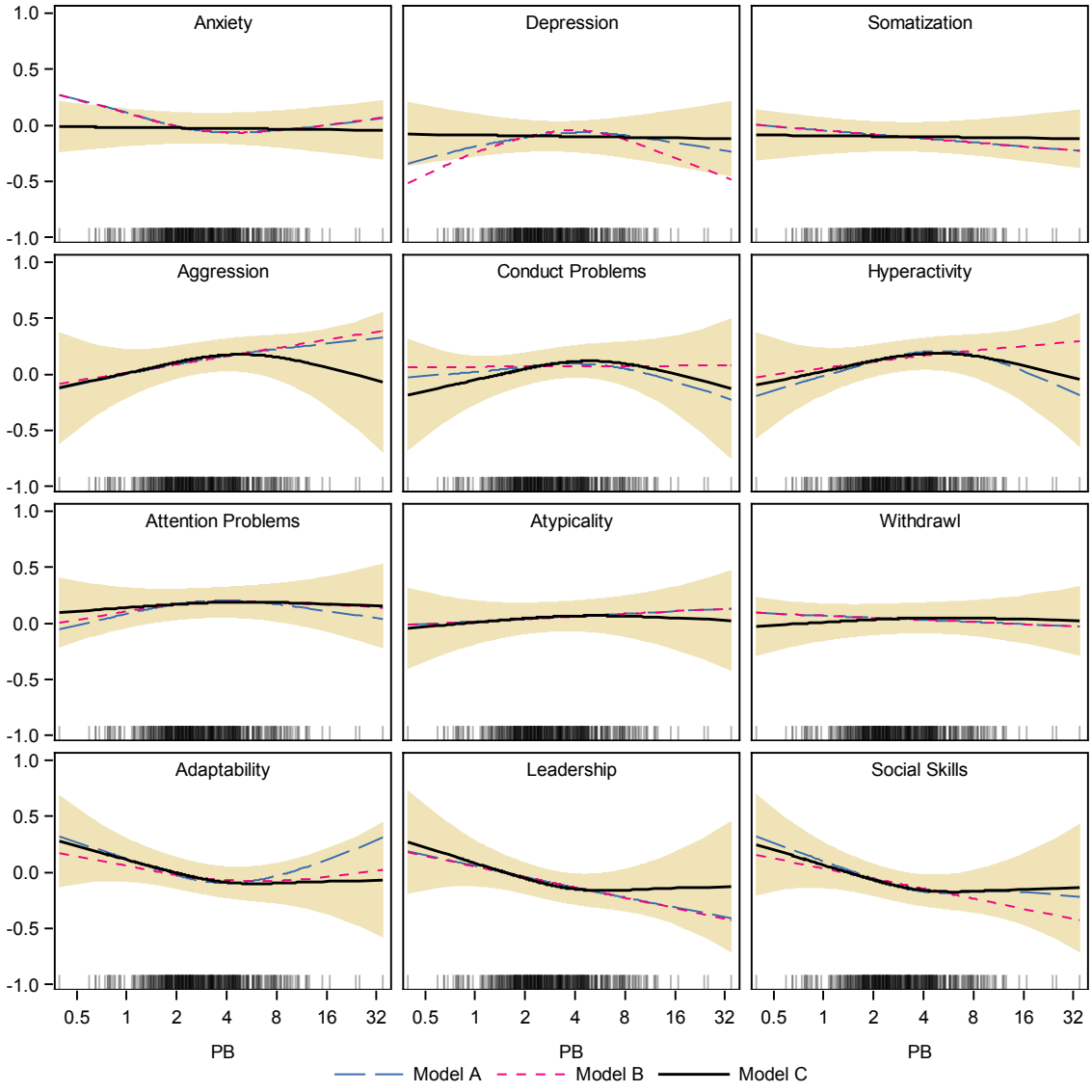
Characteristics	Mean(SD)	Range
Behavioral Assessment		
Internalizing Problems		
Anxiey	53.3 (10.0)	29 – 89
Depression	51.5 (10.6)	37 – 103
Somatization	49.8 (10.3)	36 – 91
Externalizing Problems		
Aggression	47.8 (8.6)	36 – 86
Conduct Problems	50.5 (9.8)	34 – 97
Hyperactivity	49.3 (9.7)	31 – 87
Behavioral Symptoms Index		
Attention Problems	51.7 (10.7)	33 – 81
Atypicality	50.9 (9.1)	41 – 97
Withdrawl	49.2 (10.5)	34 – 92
Adaptive Skills		
Adaptability	49.4 (9.8)	19 – 69
Activities of Daily Living	46.9 (10.7)	19 – 70
Functional Communication	47.9 (10.2)	10 – 68
Leadership	48.7 (10.7)	22 – 73
Social Skills	46.2 (10.6)	20 – 70
Participant's Characteristics		
Blood Lead Level, μg/dL	3.4 (2.9)	0.4 – 34.8
Age (Years)	10.8 (2.6)	7 – 15.5
Sex (Female%)	50.4	
Birth Weight (kg)	3.1 (0.4)	1.2 – 4.5
Maternal Characteristics		
Maternal Age at Delivery (Years)	26.3 (5.1)	18 – 44
Maternal Education (Years)	10.7 (2.9)	1 – 20
Mothers' Marital Status (%)	77.3	
Smoking During Pregnancy (%)	3.7	
Paternal Education (Years)	10.4 (3.5)	0 – 22
SES Level Index	8.7 (3.3)	1.0 – 18.5

'linear' lead exposure effect refers to log-linear effect, and 'nonlinear' means the exposure effect is non log-linear.)

Fig 4.1 and Table 4.4 show the results. Based on the exposure effects from Model A, I find the lead exposure effects are nonlinear for several outcomes. In general, the nonlinear patterns indicate the exposure effect reaches a plateau at around 5 μg/dL of blood lead level. The directions of the estimated nonlinear exposure effects are the same for test items from the category of externalizing problems, behavioral symptoms index, adaptive skills. But for internalizing problems, the directions are opposite for anxiety and depression.

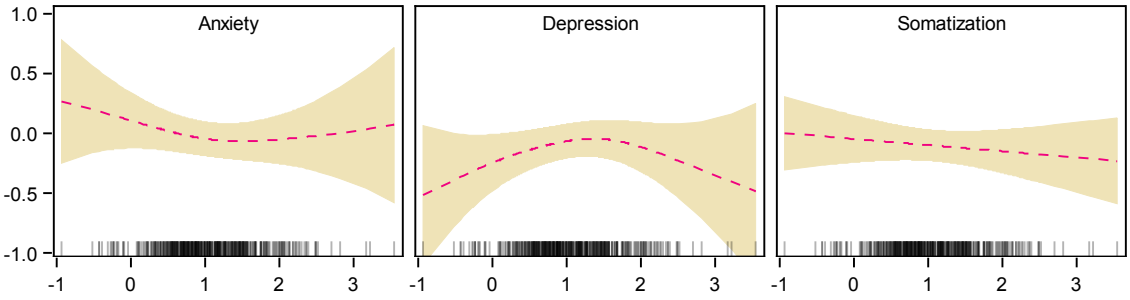
This suggests Model C, which estimates an overall exposure effect, would not be appropriate for the category of internalizing problems because the positive correlation among the test items make the factor loadings all positive. However, the relative scale of the nonlinear exposure effects for anxiety and depression imply the signs of the factor loadings should be opposite. Therefore, in this case Model B would be a better choice than Model C because the patterns of nonlinear exposure effects are very different across test items. Model A does not detect any deviation from linearity due to its lower power (Table 4.4), but the LRT that jointly tests all items (P-value = 0.046) indicates there is a nonlinear exposure effect for at least one of the test item. This in turn suggests that a model assuming linear exposure effects (Table 4.5) would not be appropriate (alse see Figure 4.2).

Figure 4.1 also shows the estimated nonlinear exposure effects for Model C, and Table 4.4 lists the P-values for the LRT testing lack of linearity. Since for the category of internalizing problems, the effects of anxiety and depression have opposite patterns, Model C averages the effects and gives an essentially null association for this category. For the other three categories, the nonlinear exposure effects have more



(a)

Figure 4.1: Nonlinear exposure effects of concurrent blood lead for each of the BASC-2 outcomes. The blood lead level is transformed to the logarithmic scale, and all the outcomes are standardized with their respective mean and standard deviation. Other model settings are the same as Table 4.5. Each row represents outcomes from different categories. Two outcomes from the fourth category that have similar exposure effects to adaptability, leadership, social skills are eliminated from the graph for a uniform layout. The confidence intervals are constructed using the covariance matrix of the fixed and random effects from Model C.



(a)

Figure 4.2: Nonlinear exposure effect for test items of internalizing problems under Model B. The plot is similar to that of Figure 4.1, except that the confidence intervals are constructed using the covariance matrix of the fixed and random effects from Model B.

Table 4.4: P-values for the LRT of whether the exposure effects deviate from linearity. I carry out the tests as described in Section 4.4.

Outcome	Model A	Model B	Model C
Internalizing Problems		0.046	1.00
Anxiey	0.15		
Depression	0.17		
Somatization	1.00		
Externalizing Problems		0.61	0.13
Aggression	0.28		
Conduct Problems	0.17		
Hyperactivity	0.09		
Behavioral Symptoms Index		0.50	0.29
Attention Problems	0.16		
Atypicality	1.00		
Withdrawl	1.00		
Adaptive Skills		0.59	0.15
Adaptability	0.06		
Activities of Daily Living	0.23		
Functional Communication	0.31		
Leadership	0.32		
Social Skills	0.16		

Table 4.5: Estimates and P-values of the linear lead exposure effects on BASC-2 scales. All the estimates are obtained from models where I only include the linear component of the exposure effect. The other model settings are the same as in Figure 4.1. The P-values are obtained using LRT described in Section 4.4. The italic numbers under Model B (the original numbers are the same as Model A, for reasons stated in Appendix D) are estimates standardized by the factor loadings. They are comparable to the italic numbers under Model C, where $\tilde{\beta}$ in (4.9) is also standardized by the factor loadings. 'B vs. C' under Model C refers to the P-values for LRT testing whether Model B fits significantly better than Model C (4.16)

	Model A		Model B		Model C		
Outcome	Estimate	P-value	Estimate	P-value	Estimate	P-value	B vs. C
Internalizing Problems				0.46	*-0.015*	0.87	0.28
Anxiey	-0.064	0.38	*-0.118*				
Depression	0.041	0.57	*0.058*				
Somatization	-0.052	0.46	*-0.096*				
Externalizing Problems				0.23	*0.072*	0.36	0.17
Aggression	0.105	0.15	*0.132*				
Conduct Problems	0.003	0.96	*0.004*				
Hyperactivity	0.074	0.29	*0.098*				
Behavioral Symptoms Index				0.87	*0.033*	0.71	0.75
Attention Problems	0.033	0.65	*0.055*				
Atypicality	0.033	0.64	*0.048*				
Withdrawl	-0.027	0.71	*-0.054*				
Adaptive Skills				0.41	*-0.138*	0.078	0.75
Adaptability	-0.055	0.44	*-0.080*				
Activities of Daily Living	-0.065	0.37	*-0.090*				
Functional Communication	-0.102	0.16	*-0.130*				
Leadership	-0.135	0.052	*-0.172*				
Social Skills	-0.129	0.070	*-0.169*				

similar patterns. Thus, as exemplified in the simulation study, it is more appropriate and powerful to use Model C for testing lack of linearity. The P-values of LRT under Model C are smaller than Model B and also smaller than Model A in general. Nonetheless, they are still all non-significant and suggest that linear models would be sufficient to describe the associations for these three categories in the current data.

Table 4.5 shows the estimates from linear models. It also includes the P-values of LRT for testing whether Model B (H_a) fits significantly better than Model C (H_0) (see (4.16) in Section 4.4). For all the three categories where the linear models are of interest, LRT suggests Model B does not give a significantly better fit than Model C. So I base my conclusions on the more powerful Model C. There is a marginally significant lead exposure effect for the category of adaptive skills (P-value = 0.078).

I find that higher lead exposure is overall associated with lower scores for test items related to adaptive skills. This overall association is likely driven by the association between lead exposure and the two test items for leadership ad social skills (see the P-values for individual items under Model A and B).

4.7 Discussion

In this chapter I develop a model to estimate and test nonlinear exposure effects for multiple continuous outcomes where the outcomes are believed to measure the same trait and are thus correlated. I use a single latent factor to account for their correlation and jointly estimate the nonlinear exposure effect. The idea can also be extended to multiple latent factors. I was motivated by a study that examined the effects of lead exposure on the psychometric test battery Conners' Rating Scales-Revised (CRS-R). I find that lead exposure has a similar ceiling effect across the different subscores in BASC-2, particularly for tests within the categories of externalizing problems and adaptive skills. I estimate an overall nonlinear exposure effect and scale it by the respective factor loadings to obtain outcome-specific effects. The overall exposure effect can be interpreted as an effect on the mean of the latent factor.

I use P-splines and REML within the framework of latent factor models to estimate the nonlinear exposure effect. In the literature of latent factor models a few papers [1, 2, 3, 4, 5] have also included nonlinear mean trend estimated by splines, but those approaches are not well suited for testing due to the extra complexity. I extend the estimating strategy developed in these two papers [47, 110], which not only estimates the linear exposure effect, but also can be used to estimate and test the nonlinear exposure effect. The likelihood-based estimating approach can allow me to carry out a likelihood ratio test when the objective of the analysis also includes

testing whether the exposure effect is linear. This testing method consists of testing a variance component, which has long been a difficult problem. Even though I only need to test one variance component for the overall exposure effect, I cannot apply existing methods [111, 109] directly on (4.9) because the factor loadings are also estimated. I use parametric bootstrap for testing. Future work can include studying whether existing approaches can provide an adequate answer to the hypothesis testing question.

Besides directly estimating the nonlinear exposure effect on the shared trait that underlies different outcomes, my model has the advantage of pooling information from different outcomes for the estimation. Compared with models that estimate individual exposure effects, this approach reduces prediction error and also increases testing power. However, this strategy of using an overall nonlinear exposure effect is also based on stronger assumptions. Our simulation study has shown that if an individual exposure effect differs greatly from the scaled overall exposure effect, then its estimation will be compromised. Future work can investigate strategies to effectively select between models that estimate the overall exposure effect and the individual exposure effects.

CHAPTER V

Conclusion

Latent factor models are useful for summarizing information among multiple outcomes. In this thesis I apply semiparametric methods to latent factor models in order to estimate and test non-constant factor loadings and nonlinear effects of observed predictors on latent variable means. Compared with existing semiparametric latent factor models in the literature that use Bayesian methods (e.g., [5]), my approaches stay in the frequentist framework. In particular, Chapter II and III develop a new class of models that estimate and test non-constant factor loadings using semiparametric methods.

I use P-splines to estimate semiparametric factor loadings and semiparametric covariate effects on factor means, so that deviations from the classic constant factor loadings and a linear covariate effects can be estimated in the latent factor framework. P-splines use a difference penalty on adjacent spline coefficients that is easy to implement. As for estimating the smoothing parameters for the splines, I adapt GCV and maximum likelihood methods commonly used in GAM to latent factor models because latent factors are unobserved. In Chapter II I optimize a GCV-type criterion during each EM iteration. In Chapter III I use a Monte-Carlo E-step based on Gibbs sampler to get around the computational issue that comes forth when both the la-

tent factor and its factor loadings are unobserved random variables. In Chapter IV I propose an iterative algorithm and within each iteration I fit a typical linear mixed model using outcomes standardized by the factor loadings. My simulation studies have shown that these novel estimation strategies can achieve desirable smoothness for spline estimates.

The semiparametric latent factor models I develop in this thesis can help researchers detect potential non-constant factor loadings and nonlinear covariate effects when they use latent factor models to summarize certain variables from the data. In Chapter II I show that correctly modeling factor loadings as non-constant can improve estimation of the factor score. I use a latent factor to specify the correlation structure among four highly correlated $PM_{2.5}$ constituents. The semiparametric estimation strategy allows me to identify the factor loading for total nitrate as non-constant. This not only reveals how the association between total nitrate and the other three constituents change with season and geographical region, but it can also reduce the measurement error of the factor score. For the four constituents I investigate, the factor score can be interpreted as an index for the level of secondary inorganic aerosols. Therefore, my method can potentially help future analyses if the factor score is used to study the association between health outcomes and this pollution source.

In Chapter III I develop methods for testing whether the factor loading is non-constant. My results show that the parametric bootstrap method from [96] for linear mixed models can also control the Type I error of LRT for my latent factor model. Given some of the smoothing properties of the MCEM algorithm under the null, I can also use the pointwise confidence interval for testing purposes. I use a latent factor model with non-constant factor loadings to estimate the underlying lead exposure

represented by four types of lead measurements on mothers from the ELEMENT study. Using the testing approaches I develop, I find that among older mothers the factor loading for cord blood lead is significantly higher. This may indicate for older mothers the latent factor measures more of the true prenatal lead exposure for children. Also, it implies that the other lead measurements (e.g., blood lead measured one month after delivery) can serve as better surrogate measures for cord blood lead among older mothers. In this chapter, I have also compared the power of LRT between the multi-group factor analysis with my semiparametric latent factor model. My model has better estimation properties and has higher testing power.

In Chapter IV I estimate a nonlinear mean trend for the latent factor, with a particular emphasis on estimating and testing deviations from linearity for the effects of a covariate on a latent factor. My algorithm is based on standard linear mixed model and is successfully implemented using PROC MIXED in SAS. This algorithm can also be easily extended to models that have multiple independent latent factors within the framework of confirmatory factor analysis. Even though similar models have been developed using Bayesian approaches, I make the connection between my semiparametric latent factor model to a class of linear mixed models that estimate an overall exposure effect for multiple outcomes. I apply the model to studying the potential effect of lead exposure on children's behaviors as measured by the psychometric battery BASC-2. My main interest is to estimate an overall nonlinear exposure effect on the latent factor that represents the underlying trait of certain behavioral outcomes. My results show that the negative impact of lead exhibits a ceiling effect on items that measure externalizing problems and adaptive skills. This also agrees with previous studies that use a segmented regression approach to investigate the relationship between blood lead and ADHD-like behavior [104]. My

approach, however, does not require a priori ideas about the pattern of exposure effects and is more data-driven.

Future extensions of my thesis work can include relaxing certain model assumptions and further studying the potential advantage of my proposed modeling and testing strategies. In Chapter II, I impose the strong assumption that observations are independent from each other. Even though I can remove certain parts of the spatiotemporal correlation by detrending the data, extensions of the model could allow the latent factor or residuals to follow a spatiotemporal process. Bayesian methods can be used to implement the more complicated model and investigate its effect on the estimation and inference about the non-constant factor loadings. As for the application of the models from Chapters II and III, the next step is to include health outcomes and study the impact of measurement bias on estimating the association between latent factor and outcomes. In Chapter IV I have shown the advantage of estimating an overall exposure effect, both in terms of its efficiency and interpretation. Another advantage of this approach over directly estimating individual exposure effects in a joint model is that the testing of nonlinearity can be potentially simpler, because I only need to test one variance component. Although the estimation of the overall exposure effect is involved with factor loadings, existing methods of testing one variance component may provide a good approximation and it would be of interest to study this strategy through simulation. Also, I have observed a tradeoff between the two models when I need to decide whether to estimate an overall exposure effect or estimate each individual effect. So, it would be useful to develop model selection approaches that can help make such a decision based on the actual pattern of the exposure effects.

In conclusion, classical latent factor models are popular tools for analyzing mul-

tivariate data but they are also limited by certain model assumptions. In this thesis I have developed three semiparametric latent factor models motivated by the idea of relaxing the assumption of constant factor loading and constant/linear mean trend. The estimating and testing approaches from this work add to existing methods that broaden the scope of latent factor models and can also serve as a basis for future methodology development in modeling and especially more efficiently testing nonlinear components in latent factor models.

APPENDICES

APPENDIX A

Nested MCEM Algorithm in Chapter III

Since the conditional distribution of $(\boldsymbol{\beta}_p|\mathbf{y},\boldsymbol{\eta})$ has an analytical form, I nest a few additional EM cycles within each iteration to get better estimates of $\boldsymbol{\beta}_p$. In this way, the whole algorithm converges faster and thus saving the computational cost of the E-step.

Let $Q_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r')},\boldsymbol{\theta}^{(r)}) = \mathrm{E}\left\{\mathrm{E}\left(\log\mathcal{L}(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta},\boldsymbol{\beta}_p)|\mathbf{y},\boldsymbol{\eta},\boldsymbol{\theta}^{(r')}\right)\middle|\,\mathbf{y},\boldsymbol{\theta}^{(r)}\right\}$, then the nested EM algorithm repeats the following cycle T times, for $t=1,\ldots,T$, after sampling $\boldsymbol{\eta}^{(m)}, m=1,\ldots,M^{(\infty)}$ from $(\boldsymbol{\eta},\boldsymbol{\beta}_p|\mathbf{y},\boldsymbol{\theta}^{(r)})$ in the preceding Monte Carlo E-step.

E-step: Compute $Q_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r+\frac{t-1}{T})},\boldsymbol{\theta}^{(r)})$, which means computing the inner expectation under each $\boldsymbol{\eta}^{(m)}$ and then average across all m. The inner expectation essentially involves computing

$$(\mathbf{X}_\eta\boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} = \mathrm{E}\left(\mathbf{X}_\eta\boldsymbol{\beta}_p|\mathbf{y}_{.i},\boldsymbol{\eta}^{(m)},\boldsymbol{\theta}^{(r+\frac{t-1}{T})}\right),$$
$$(\boldsymbol{\beta}_p^T\boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} = \mathrm{E}\left(\boldsymbol{\beta}_p^T\boldsymbol{\beta}_p|\mathbf{y}_{.i},\boldsymbol{\eta}^{(m)},\boldsymbol{\theta}^{(r+\frac{t-1}{T})}\right).$$
$$(\boldsymbol{\beta}_p^T\mathbf{X}_\eta^T\mathbf{X}_\eta\boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} = \mathrm{E}\left(\boldsymbol{\beta}_p^T\mathbf{X}_\eta^T\mathbf{X}_\eta\boldsymbol{\beta}_p|\mathbf{y}_{.i},\boldsymbol{\eta}^{(m)},\boldsymbol{\theta}^{(r+\frac{t-1}{T})}\right),$$

Recall $\boldsymbol{\beta}_p|(\mathbf{y}_{p.},\boldsymbol{\eta}) \sim N\left(\hat{\boldsymbol{\beta}}_p(\boldsymbol{\eta}),\sigma_{b,p}^2(\mathbf{I}_K-\mathbf{D}_p\mathbf{X}_\eta^T\mathbf{X}_\eta)\right)$, and let $\hat{\boldsymbol{\beta}}_p^{(m)} = \hat{\boldsymbol{\beta}}_p(\boldsymbol{\eta}^{(m)})$ and $\mathbf{X}_\eta^{(m)}$, $\mathbf{D}_p^{(m)}$ be $\mathbf{X}_\eta$, $\mathbf{D}_p$ with $\boldsymbol{\eta}^{(m)}$ plugged in. Then setting the parameters equal to $\boldsymbol{\theta}^{(r+\frac{t-1}{T})}$ we obtain

$$(\mathbf{X}_\eta\boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} = \mathbf{X}_\eta^{(m)}\hat{\boldsymbol{\beta}}_p^{(m)},$$
$$(\boldsymbol{\beta}_p^T\boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} = \mathrm{tr}\left\{\hat{\boldsymbol{\beta}}_p^{(m)}(\hat{\boldsymbol{\beta}}_p^T)^{(m)} + \sigma_{b,p}^2\left(\mathbf{I}_K-\mathbf{D}_p^{(m)}(\mathbf{X}_\eta^T)^{(m)}\mathbf{X}_\eta^{(m)}\right)\right\}.$$

$$
\begin{aligned}
& (\boldsymbol{\beta}_p^T \mathbf{X}_\eta^T \mathbf{X}_\eta \boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})} \\
= \quad & \mathrm{tr}\left[(\mathbf{X}_\eta^T)^{(m)} \mathbf{X}_\eta^{(m)} \left\{ \hat{\boldsymbol{\beta}}_p^{(m)} (\hat{\boldsymbol{\beta}}_p^T)^{(m)} + \sigma_{b,p}^2 \left(\mathbf{I}_K - \mathbf{D}_p^{(m)} (\mathbf{X}_\eta^T)^{(m)} \mathbf{X}_\eta^{(m)} \right) \right\} \right]
\end{aligned}
$$

,

M-step: Set $\boldsymbol{\theta}^{(r+\frac{t}{T})} = \arg\max_{\boldsymbol{\theta}} Q_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r+\frac{t-1}{T})}, \boldsymbol{\theta}^{(r)})$, which basically uses the same M-step formulas as in the original MCEM. But the statistics involved are averaged across all $\boldsymbol{\eta}^{(m)}$. For example,

$$
(\mathbf{X}_\eta \boldsymbol{\beta}_p)^{(r+\frac{t-1}{T})} = \frac{1}{M^{(r)}} \sum_{m=1}^{M^{(r)}} (\mathbf{X}_\eta \boldsymbol{\beta}_p)_{(m)}^{(r+\frac{t-1}{T})}.
$$

APPENDIX B

Comparison of $p\chi_0^2 + (1-p)a\chi_1^2$ to the Full-Scale Bootstrap in Chapter III

Suppose I generate M simulations and compute the LRT statistics. Then I randomly pick M^* LRT statistics from the M simulations and estimate $\hat{p}$ and $\hat{a}$ using the method of moments proposed in [96]. I use the 95-th percentile of $p\chi_0^2 + (1-p)a\chi_1^2$ as the critical value for deciding whether to reject the null hypothesis among the rest $M - M^*$ simulations. This process is repeated 1000 times where each time I draw a different sample of size M^* from the same M simulations. I then obtain the average rejection rate and its standard error. For the full-scale bootstrap approach, I use the 95-th percentile of the M^* LRT statistics as the critical value each time. Although the M^* simulations I use are not strictly parametric bootstrap samples because the simulations use the true parameter values instead of estimates from the model, this approach still allows us to compare $p\chi_0^2 + (1-p)a\chi_1^2$ to the full-scale bootstrap method while saving computational cost.

In the table below I list the rejection rate for testing H_0: $f_1 = 0$ vs. $f_1 \neq 0$

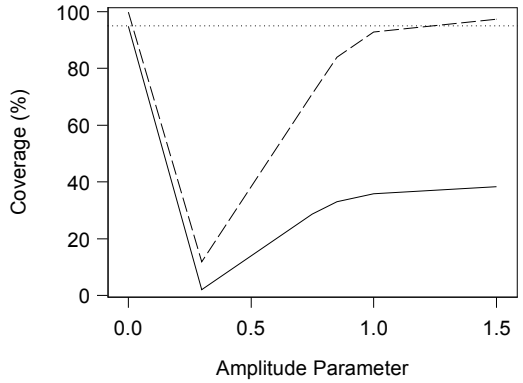
M^*	$p\chi_0^2 + (1-p)a\chi_1^2$	Full-Scale Bootstrap
100	5.58 (1.95)	5.33 (2.20)
200	5.32 (1.36)	5.16 (1.58)
500	5.15 (0.83)	5.05 (1.01)
1000	5.07 (0.60)	5.00 (0.72)
2000	5.08 (0.46)	5.02 (0.55)

between using $p\chi_0^2 + (1-p)a\chi_1^2$ and the full-scale bootstrap under the scenario where the truth is $f_p = 0$ for all p and only f_1 is estimated. The value on the left is the average rejection rate, while the value in parentheses is the standard error of the rejection rate. I get similar results from the scenario where I also estimate f_2.

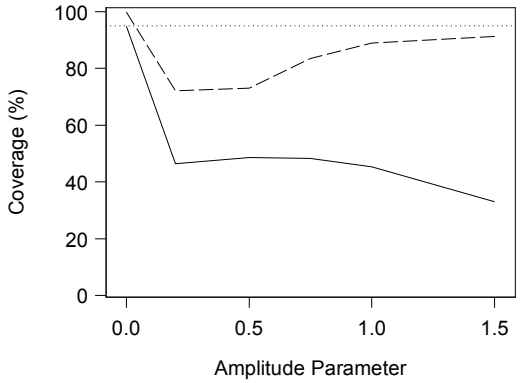
APPENDIX C

Coverage of the Pointwise Confidence Interval and Simultaneous Confidence Band in Chapter III

In the graph below I show the coverages of the pointwise confidence interval and simultaneous confidence band for two shapes of f_1 in models where only f_1 is non-zero and being estimated. The curves are based on simulation scenarios from the first group where $\kappa = 0, 0.2, 0.5, 0.75, 1, 1.5$ for the cyclic shape and $\kappa = 0, 0.3, 0.75, 0.85, 1, 1.5$ for the monotone trend added ($f_2 = 0,\ f_3 = 0$). Curves for the pointwise confidence interval are shown with solid lines, while curves for the simultaneous confidence interval are shown with dashed lines. The horizontal dotted line near the top indicates 95% coverage. The coverage is lower than nominal for the simultaneous confidence band mainly because the over-smoothing of the splines. This effect is especially notable when f_1 has a cyclic shape and κ is close to zero.



(a) Cyclic Shape



(b) Monotone Trend Added

APPENDIX D

Proof Showing Why Model A and B Have the Same Coefficient Estimates in Table 4.4

Since I assume multivariate normal distribution for the observations, the likelihood can be written in the following form (omitting the part unrelated to the regression coefficients),

$$(\mathbf{Y} - \mathcal{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathcal{X}\boldsymbol{\beta}),$$

where $\mathbf{Y}$ is the column vector of all the observations as used in the text, $\mathcal{X}$ is the design matrix, $\boldsymbol{\beta}$ is the column vector of all regression coefficients. Model A and Model B differ in $\boldsymbol{\Sigma}^{-1}$, the covariance structure of $\mathbf{Y}$.

The data and model structure I use satisfy the following sufficient conditions for Model A and B to have the same coefficient estimates:

(1) The data has a balanced two-level structure, the outcome level and the subject level. There are P outcomes and each of them has measurements for all the n subjects. The correlations among the P outcomes are the same for all subjects, but there is no correlation among the subjects. This implies the correlation structure of $\mathbf{Y}$ can be written as $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{Y}) = \boldsymbol{\Sigma}_P \otimes \mathbf{I}_n$. I use $\boldsymbol{\Sigma}_P$ to denote the correlation among the P outcomes for a given subject. (The order of the Kronecker product implies the observations in $\mathbf{Y}$ are arranged in a subject-major order.)

(2) Each outcome has its own individual mean structure, but they all have the same

form. This means the design matrix can be written as $\mathcal{X} = \mathbf{I}_p \otimes \mathbf{X}_n$, where I use $\mathbf{X}_n$ to refer to the design matrix for one outcome.

These two sufficient conditions do not put any constraint on the actual form of $\boldsymbol{\Sigma}$. Essentially Model A and B have the same coefficient estimates because these two conditions allow the data to be grouped into i.i.d. units, and each unit follows a multivariate normal distribution that has an individual mean structure for each variable but they all have the same form. If this is the case, then the estimation of the mean and the covariance are independent.

Below I prove $\hat{\boldsymbol{\beta}}$ is not related to $\boldsymbol{\Sigma}$. The solution of $\hat{\boldsymbol{\beta}}$ is the weighted least squares, $(\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathcal{X})^{-1}\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$. Then I plug in $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_P \otimes \mathbf{I}_n$, $\mathcal{X} = \mathbf{I}_p \otimes \mathbf{X}_n$ and I can show that $\hat{\boldsymbol{\beta}} = \left[\mathbf{I}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\right]\mathbf{Y}$. The proof relies on the switch of Kronecker product and matrix multiplication in the following manner,

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}),$$

if $\mathbf{AC}$ and $\mathbf{BD}$ are valid.

First, I have $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{I}_n$. Then, I can get

$$\begin{aligned}
\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathcal{X} &= (\mathbf{I}_p \otimes \mathbf{X}_n)^T(\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{I}_n)(\mathbf{I}_p \otimes \mathbf{X}_n) \\
&= (\mathbf{I}_p \otimes \mathbf{X}_n^T)(\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{I}_n)(\mathbf{I}_p \otimes \mathbf{X}_n) \\
&= (\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{X}_n^T)(\mathbf{I}_p \otimes \mathbf{X}_n) \\
&= \boldsymbol{\Sigma}_P^{-1} \otimes (\mathbf{X}_n^T\mathbf{X}_n)
\end{aligned}$$

and also

$$\begin{aligned}
\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y} &= (\mathbf{I}_p \otimes \mathbf{X}_n)^T(\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= (\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{X}_n^T)\mathbf{Y}.
\end{aligned}$$

Plugging these back into $\hat{\boldsymbol{\beta}} = (\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathcal{X})^{-1}\mathcal{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, I then have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left[\boldsymbol{\Sigma}_P^{-1} \otimes (\mathbf{X}_n^T\mathbf{X}_n)\right]^{-1} (\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{X}_n^T)\mathbf{Y} \\
&= \left[\boldsymbol{\Sigma}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\right] (\boldsymbol{\Sigma}_P^{-1} \otimes \mathbf{X}_n^T)\mathbf{Y} \\
&= \left[\mathbf{I}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\right] \mathbf{Y}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \left[\mathbf{I}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\right] \mathrm{Cov}(\mathbf{Y}) \left[\mathbf{I}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\right]^T \\
&= \left[\mathbf{I}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\right] (\boldsymbol{\Sigma}_P \otimes \mathbf{I}_n) \left[\mathbf{I}_P \otimes \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\right] \\
&= \boldsymbol{\Sigma}_P \otimes (\mathbf{X}_n^T\mathbf{X}_n)^{-1}.
\end{aligned}$$

The diagonal of $\hat{\boldsymbol{\Sigma}}_P$ is the same (the variance of residuals for each outcome) when $\boldsymbol{\Sigma}_P$ assumes either the unstructured, diagonal, or factor-analytic structure. So the standard errors of the coefficient estimates are also the same under Model A and B.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] L. Fahrmeir and A. Raach. A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika*, 72(3):327–346, 2007.

[2] A. Gryparis, B. A. Coull, J. Schwartz, and H. H. Suh. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 56:183–209, 2007.

[3] N. Bliznyuk, C. J. Paciorek, J. Schwartz, and B. A. Coull. Nonlinear predictive latent process models for integrating spatio-temporal exposure data from multiple sources. *The Annals of Applied Statistics*, 8(3):1538–1560, 2014.

[4] X. Y. Song and Z. H. Lu. Semiparametric latent variable models with Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 19(3):590–608, 2010.

[5] X. Y. Song, Z. H. Lu, J. H. Cai, and E. H. S. Ip. A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika*, 78(4):624–647, 2013.

[6] X. Y. Song, Z. H. Lu, and X. N. Feng. Latent variable models with nonparametric interaction effects of latent variables. *Statistics in Medicine*, 33(10):1723–1737, 2014.

[7] W. Y. Zhang and S. Y. Lee. Nonlinear dynamical structural equation models. *Quantitative Finance*, 9(3):305–314, 2009.

[8] C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

[9] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models : a roughness penalty approach*. Monographs on statistics and applied probability. Chapman & Hall, London ; New York, 1st edition, 1994.

[10] R. L. Eubank. *Nonparametric regression and spline smoothing*. Statistics, textbooks and monographs. Marcel Dekker, New York, 2nd edition, 1999.

[11] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1990.

[12] T. Hastie and R. Tibshirani. *Generalized additive models*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York, 1st edition, 1990.

[13] J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 53(1):233–243, 1991.

[14] M. G. Shi, R. E. Weiss, and J. M. G. Taylor. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 45(2):151–163, 1996.

[15] S. N. Wood. *Generalized additive models : an introduction with R*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[16] P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102, 1996.

[17] S. Lang and A. Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

[18] C. Gu. *Smoothing spline ANOVA models*. Springer series in statistics. Springer, New York, NY, 2nd edition, 2013.

[19] P. Lancaster and K. Šalkauskas. *Curve and surface fitting : an introduction*. Academic Press, London ; Orlando, 1986.

[20] S. N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 65:95–114, 2003.

[21] C. De Boor. *A practical guide to splines*. Applied mathematical sciences. Springer, New York, 2nd edition, 2000.

[22] L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14(3):731–761, 2004.

[23] S. N. Wood. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036, 2006.

[24] S. N. Wood, F. Scheipl, and J. J. Faraway. Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23(3):341–360, 2013.

[25] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25(4):1371–1425, 1997.

[26] B. D. Marx and P. H. C. Eilers. Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22, 2005.

[27] Z. H. Chen. Fitting multivariate regression-functions by interaction spline models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 55(2):473–491, 1993.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 39(1):1–38, 1977.

[29] P. Craven and G. Wahba. Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1979.

[30] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13(4):1378–1402, 1985.

[31] C. Gu and D. Xiang. Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*, 10(3):581–591, 2001.

[32] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *Siam Journal on Scientific and Statistical Computing*, 12(2):383–398, 1991.

[33] W. Meredith. Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993.

[34] W. Meredith and J. A. Teresi. An essay on measurement and factorial invariance. *Medical Care*, 44(11):S69–S77, 2006.

[35] G. J. Mellenbergh. Item bias and item response theory. *International Journal of Educational Research*, 13(2):127–143, 1989.

[36] K. G. Jöreskog. Simultaneous factor analysis in several populations. *Psychometrika*, 36(4):409–426, 1971.

[37] N. Schmitt and G. Kuljanin. Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18(4):210–222, 2008.

[38] R. J. Vandenberg and C. E. Lance. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1):4–70, 2000.

[39] M. Barendse, F. Oort, and G. Garst. Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *Asta-Advances in Statistical Analysis*, 94(2):117–127, 2010.

[40] D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge ; New York, 2003.

[41] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[42] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 73:3–36, 2011.

[43] P. T. Reiss and R. T. Ogden. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 71:505–523, 2009.

[44] S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

[45] G. C. G. Wei and M. A. Tanner. A Monte-Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[46] M. Sammel, X. H. Lin, and L. Ryan. Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*, 18(17-18):2479–2492, 1999.

[47] X. H. Lin, L. Ryan, M. Sammel, D. W. Zhang, C. Padungtod, and X. P. Xu. A scaled linear mixed model for multiple outcomes. *Biometrics*, 56(2):593–601, 2000.

[48] J. Roy, X. H. Lin, and L. M. Ryan. Scaled marginal models for multiple continuous outcomes. *Biostatistics*, 4(3):371–383, 2003.

[49] J. Schwartz, D. W. Dockery, and L. M. Neas. Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association*, 46(10):927–939, 1996.

[50] F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, and J. M. Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10):1127–34, 2006.

[51] F. Laden, L. M. Neas, D. W. Dockery, and J. Schwartz. Association of fine particulate matter from different sources with daily mortality in six US cities. *Environmental Health Perspectives*, 108(10):941–947, 2000.

[52] R. T. Burnett, J. Brook, T. Dann, C. Delocla, O. Philips, S. Cakmak, R. Vincent, M. S. Goldberg, and D. Krewski. Association between particulate- and gas-phase components of urban air pollution and daily mortality in eight Canadian cities. *Inhalation Toxicology*, 12:15–39, 2000.

[53] B. Ostro, W. Feng, R. Broadwin, S. Green, and M. Lipsett. The effects of components of fine particulate air pollution on mortality in California: Results from CALFINE. *Environmental Health Perspectives*, 115(1):13–19, 2007.

[54] R. D. Peng, M. L. Bell, A. S. Geyh, A. McDermott, S. L. Zeger, J. M. Samet, and F. Dominici. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*, 117(6):957–63, 2009.

[55] H. H. Suh, A. Zanobetti, J. Schwartz, and B. A. Coull. Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in Atlanta, Georgia. *Environmental Health Perspectives*, 119(10):1421–8, 2011.

[56] J. R. Krall, G. B. Anderson, F. Dominici, M. L. Bell, and R. D. Peng. Short-term exposure to particulate matter constituents and mortality in a national study of U.S. urban communities. *Environmental Health Perspectives*, 121(10):1148–53, 2013.

[57] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis. A framework to spatially cluster air pollution monitoring sites in US based on the $PM_{2.5}$ composition. *Environmental International*, 59:244–54, 2013.

[58] M. C. Nikolov, B. A. Coull, P. J. Catalano, E. Diaz, and J. J. Godleski. Statistical methods to evaluate health effects associated with major sources of air pollution: a case-study of breathing patterns during exposure to concentrated Boston air particles. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 57:357–378, 2008.

[59] G. D. Thurston, K. Ito, and R. Lall. A source apportionment of U.S. fine particulate matter air pollution. *Atmospheric Environment*, 45(24):3924–3936, 2011.

[60] M. L. Bell, F. Dominici, K. Ebisu, S. L. Zeger, and J. M. Samet. Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. *Environmental Health Perspectives*, 115(7):989–995, 2007.

[61] P. Salvador, B. Artinano, D. G. Alonso, X. Querol, and A. Alastuey. Identification and characterisation of sources of PM_{10} in Madrid (Spain) by statistical methods. *Atmospheric Environment*, 38(3):435–447, 2004.

[62] J. Choi, M. Fuentes, and B. J. Reich. Spatial-temporal association between fine particulate matter and daily mortality. *Computational Statistics & Data Analysis*, 53(8):2989–3000, 2009.

[63] V. J. Berrocal, A. E. Gelfand, and D. M. Holland. A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics*, 4(4):1942–1975, 2010.

[64] M. J. Daniels, Z. G. Zhou, and H. Zou. Conditionally specified space-time models for multivariate processes. *Journal of Computational and Graphical Statistics*, 15(1):157–177, 2006.

[65] G. Huerta, B. Sanso, and J. R. Stroud. A spatiotemporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 53:231–248, 2004.

[66] G. Shaddick and J. Wakefield. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 51:351–372, 2002.

[67] M. C. Nikolov, B. A. Coull, P. J. Catalano, and J. J. Godleski. Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. *Environmetrics*, 22(2):165–178, 2011.

[68] E. E. Kammann and M. P. Wand. Geoadditive models. *Journal of the Royal Statistical Society. Series C-Applied Statistics*, 52(1):1–18, 2003.

[69] H. F. Lopes, E. Salazar, and D. Gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008.

[70] L. Ippoliti, P. Valentini, and D. Gamerman. Spacetime modelling of coupled spatiotemporal environmental variables. *Journal of the Royal Statistical Society: Series C-Applied Statistics*, 61(2):175–200, 2012.

[71] B. N. Sánchez, E. Budtz-Jørgensen, L. M. Ryan, and H. Hu. Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association*, 100(472):1443–1455, 2005.

[72] G. Arminger. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63(3):271–300, 1998.

[73] R. Guo, H. T. Zhu, S. M. Chow, and J. G. Ibrahim. Bayesian lasso for semiparametric structural equation models. *Biometrics*, 68(2):567–577, 2012.

[74] S. Y. Lee, X. Y. Song, and J. C. K. Lee. Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics*, 28(2):111–134, 2003.

[75] B. Lu, X. Y. Song, and X. D. Li. Bayesian analysis of multi-group nonlinear structural equation models with application to behavioral finance. *Quantitative Finance*, 12(3):477–488, 2012.

[76] H. T. Zhu and S. Y. Lee. Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical & Statistical Psychology*, 52:225–242, 1999.

[77] D. J. Bauer. The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, 10(3):305–316, 2005.

[78] P. J. Hewson and T. C. Bailey. Modelling multivariate disease rates with a latent structure mixture model. *Statistical Modelling*, 10(3):241–264, 2010.

[79] X. Liu, M. M. Wall, and J. S. Hodges. Generalized spatial structural equation models. *Biostatistics*, 6(4):539–557, 2005.

[80] J. L. Horn and J. J. McArdle. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3-4):117–144, 1992.

[81] B. N. Sánchez, S. Kang, and B. Mukherjee. A latent variable approach to study gene-environment interactions in the presence of multiple correlated exposures. *Biometrics*, 68(2):466–476, 2012.

[82] Y. Tao, B. N. Sánchez, and B. Mukherjee. Latent variable models for gene-environment interactions in longitudinal studies with multiple correlated exposures. *Statistics in Medicine*, 2014. ahead of print.

[83] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 55(4):757–796, 1993.

[84] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good Ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[85] A. Kneip and T. Gasser. Convergence and consistency results for self-modeling nonlinear-regression. *Annals of Statistics*, 16(1):82–112, 1988.

[86] Q. Jiang, H. S. Wang, Y. C. Xia, and G. H. Jiang. On a principal varying coefficient model. *Journal of the American Statistical Association*, 108(501):228–236, 2013.

[87] T. A. Brown. *Confirmatory factor analysis for applied research.* Methodology in the social sciences. Guilford Press, New York, 2006.

[88] M. T. Barendse, F. J. Oort, C. S. Werner, R. Ligtvoet, and K. Schermelleh-Engel. Measurement bias detection through factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(4):561–579, 2012.

[89] A. Klein and H. Moosbrugger. Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(4):457–474, 2000.

[90] F. J. Oort. Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: a Multidisciplinary Journal*, 5(2):107–124, 1998.

[91] B. A. Brumback, D. Ruppert, and M. P. Wand. Variable selection and function estimation in additive nonparametric regression using a data-based prior - comment. *Journal of the American Statistical Association*, 94(447):794–797, 1999.

[92] S. G. Self and K. Y. Liang. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[93] D. O. Stram and J. W. Lee. Variance-components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1177, 1994.

[94] C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 66:165–185, 2004.

[95] C. Crainiceanu, D. Ruppert, G. Claeskens, and M. P. Wand. Exact likelihood ratio tests for penalised splines. *Biometrika*, 92(1):91–103, 2005.

[96] S. Greven, C. M. Crainiceanu, H. Kuchenhoff, and A. Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4):870–891, 2008.

[97] O. E. Lee and T. M. Braun. Permutation tests for random effects in linear mixed models. *Biometrics*, 68(2):486–493, 2012.

[98] R. D. Stoel, F. G. Garre, C. Dolan, and G. Van den Wittenboer. On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, 11(4):439–455, 2006.

[99] D. A. van Dyk. Nesting EM algorithms for computational efficiency. *Statistica Sinica*, 10(1):203–225, 2000.

[100] T. Gonzalez-Cossio, K. E. Peterson, L. H. Sanin, E. Fishbein, E. Palazuelos, A. Aro, M. Hernández-Avila, and H. Hu. Decrease in birth weight in relation to maternal bone-lead burden. *Pediatrics*, 100(5):856–862, 1997.

[101] M. M. Téllez-Rojo, M. Hernández-Avila, H. Lamadrid-Figueroa, D. Smith, L. Hernández-Cadena, A. Mercado, A. Aro, J. Schwartz, and H. Hu. Impact of bone lead and bone resorption on plasma and whole blood lead levels during pregnancy. *American Journal of Epidemiology*, 160(7):668–678, 2004.

[102] D. A. van Dyk and X. L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[103] F. B. Yoon, G. M. Fitzmaurice, S. R. Lipsitz, N. J. Horton, N. M. Laird, and S. L. T. Normand. Alternative methods for testing treatment effects on the basis of multiple outcomes: Simulation and case study. *Statistics in Medicine*, 30(16):1917–1932, 2011.

[104] S. Huang, H. Hu, B.N. Sánchez, K.E. Peterson, A.S. Ettinger, H. Lamadrid-Figueroa, L. Schnaas, A. Mercado-García, R.O. Wright, N. Basu, D.E. Cantonwine, M. Hernández-Avila, and M.M. Téllez-Rojo. Association of low concurrent blood lead with hyperactivity/impulsivity, but not inattentiveness. 2015. Submitted:.

[105] J. Jia and R. E. Weiss. Common predictor effects for multivariate longitudinal data. *Statistics in Medicine*, 28(13):1793–1804, 2009.

[106] S. W. Thurston, D. Ruppert, and P. W. Davidson. Bayesian models for multiple outcomes nested in domains. *Biometrics*, 65(4):1078–1086, 2009.

[107] B. A. Coull, J. P. Hobert, L. M. Ryan, and L. B. Holmes. Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association*, 96(456):1194–1204, 2001.

[108] F. Scheipl, S. Greven, and H. Kuchenhoff. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52(7):3283–3299, 2008.

[109] Y. J. Wang and H. H. Chen. On testing an unspecified function through a linear mixed effects model with multiple variance components. *Biometrics*, 68(4):1113–1125, 2012.

[110] M. D. Sammel and L. M. Ryan. Latent variable models with fixed effects. *Biometrics*, 52(2):650–663, 1996.

[111] C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *Journal of Multivariate Analysis*, 91(1):35–52, 2004.