Published in final edited form as:

Stat Modelling. 2016 April ; 16(2): 91–113. doi:10.1177/1471082X15627004.

Using a latent variable model with non-constant factor loadings to examine PM_{2.5} constituents related to secondary inorganic aerosols

Zhenzhen Zhang¹, Marie S. O'Neill^{2,3}, and Brisa N. Sánchez¹

¹Department of Biostatistics, University of Michigan, Ann Arbor, USA

²Department of Environmental Health Sciences, University of Michigan, Ann Arbor, USA

³Department of Epidemiology, University of Michigan, Ann Arbor, USA

Abstract

Factor analysis is a commonly used method of modelling correlated multivariate exposure data. Typically, the measurement model is assumed to have constant factor loadings. However, from our preliminary analyses of the Environmental Protection Agency's (EPA's) PM_{2.5} fine speciation data, we have observed that the factor loadings for four constituents change considerably in stratified analyses. Since invariance of factor loadings is a prerequisite for valid comparison of the underlying latent variables, we propose a factor model that includes non-constant factor loadings that change over time and space using P-spline penalized with the generalized cross-validation (GCV) criterion. The model is implemented using the Expectation-Maximization (EM) algorithm and we select the multiple spline smoothing parameters by minimizing the GCV criterion with Newton's method during each iteration of the EM algorithm. The algorithm is applied to a one-factor model that includes four constituents. Through bootstrap confidence bands, we find that the factor loading for total nitrate changes across seasons and geographic regions.

Keywords

factor model; non-constant factor loading; P-spline; tensor product spline basis; generalized cross-validation; EM algorithm; PM_{2.5} constituents

1 Introduction

Fine particulate matter that is, air particles with diameter <2.5 micrometers (PM_{2.5}), is associated with adverse health effects (e.g., Dominici *et al.*, 2006; Schwartz *et al.*, 1996). Laden *et al.* (2000) suggested PM_{2.5} from different sources has different toxic effects, implying the importance of the chemical profile of the constituents of PM_{2.5}. 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 (Austin *et al.*, 2013) or latent variable models

(Gryparis *et al.*, 2007; Nikolov *et al.*, 2008). In particular, widely used source apportionment methods (Thurston *et al.*, 2011) use factor analytic 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 yield inconsistent summaries when applied to large geographic areas or long time periods. Such summaries depend on the correlation among constituents and correlations among constituents may vary with time and space (Austin *et al.*, 2013; Bell *et al.*, 2007). For instance, our preliminary analyses of PM_{2.5} constituent data issued by the Environmental Protection Agency (EPA) Air Quality System (AQS) for years 2008 and 2009 (see Section 5 for more details on the data) suggest that the correlation among constituents related to secondary inorganic aerosols (Salvador *et al.*, 2004; Thurston *et al.*, 2011) varies with season and geographic location (Figure 1). 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 preliminary factor analyses of this constituent data, stratified by time and region, yield factor loadings that differ across time and region for some constituents. In particular, the factor loading for total nitrate (Figure 2a) demonstrates large differences across regions and time. The estimated factor loadings tend to be larger in the winter months, implying a season effect. Furthermore, the estimated factor loadings are significantly higher in the midwest region, compared to other regions, for all periods except the first quarter of 2008. By contrast, the factor loading for ammonium ion exhibits fewer and less pronounced differences across either region or time (Figure 2b).

Differences in the factor model structure are not specifically modelled in current literature that uses factor analytic models applied to constituent data collected across large geographical areas. Although many models have been proposed to reduce data dimension and to predict outcomes based on many constituents simultaneously (Austin *et al.*, 2013; Choi *et al.*, 2009; Peng *et al.*, 2009), 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 modelling the association between constituents and health outcomes.

There is now a large literature on building models that analyze air pollution data from the perspective of time and space (e.g., Berrocal *et al.*, 2010; Daniels *et al.*, 2006; Huerta *et al.*, 2004; Shaddick and Wakefield, 2002; Sahu and Mardia, 2005, to name a few). However, we focus on models utilizing latent factors, given their connection to air pollution sources (Nikolov *et al.*, 2008, 2011) and because some of these models can be applied to understanding temporal and spatial patterns in the correlation of PM_{2.5} constituents specifically. For example, Gryparis *et al.* (2007) proposed semiparametric latent variable regression models for modelling multiple surrogates of a single pollution source. They used penalized splines to allow non-linear effects of covariates on the latent variable mean and used geoadditive models to account for the temporal and spatial correlations in the latent variable (Kammann and Wand, 2003). Lopes *et al.* (2008) used 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 was introduced through the factor loading matrix, that is, at a given time-point, monitors were treated as different variables. They modelled the temporal correlation among the latent factors using an autoregressive structure. Ippoliti *et al.* (2012) used a similar dynamic factor model but they coupled two sets of latent factors so that one latent factor could be predicted from the other.

More generally, models using latent variables 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 (see Sánchez *et al.*, 2005). In the last decade, much work has been done to relax model assumptions, including nonlinear relationships among latent variables (Arminger, 1998; Guo *et al.*, 2012; Lee *et al.*, 2003; Lu *et al.*, 2012; Song and Lu, 2010; Zhu and Lee, 1999), nonlinear links relating items to latent variables (Bauer, 2005; Gryparis *et al.*, 2007), as well as models with spatial structures on the latent variables (Fahrmeir and Raach, 2007; Hewson and Bailey, 2010; Liu *et al.*, 2005).

However, except for Zhang and Lee (2009) and Valentini *et al.* (2013), none of these existing advanced models specifically address the variation in the factor loadings exhibited among the constituent data (Figure 2a). Traditionally, the measurement model that links the observed items to the latent variables is assumed constant across the covariate space (e.g., across time and geographic location). A constant measurement model, also termed measurement model invariance (Horn and McArdle, 1992; Meredith, 1993), is needed for valid comparison of differences in latent variable 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 (Sánchez *et al.*, 2012; Tao *et al.*, 2015). Allowing the measurement model to vary across continuous covariates, however, is a more challenging problem. Zhang and Lee (2009) used a local maximum likelihood-based estimation procedure to model the factor loadings as varying with time. Valentini *et al.* (2013) gave the factor loadings a spatial structure and they modelled the factor loadings as conditionally independent multivariate Gaussian Markov Random Fields.

In this article, we propose a semiparametric factor analysis 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. We use penalized 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 (Hastie and Tibshirani, 1993). In Section 2, we propose the factor model that incorporates non-constant factor loadings, while Section 3, explains the estimation procedure. A key feature of our estimation algorithm is to include the generalized cross-validation (GCV) score within the EM iterations so that we can optimize the smoothing of the splines in a latent variable setting. In Section 4, we present a simulation study to assess the algorithm and estimation properties for certain model parameters and the estimated factor scores. We then analyze the constituent data set and draw inferences based on bootstrap approaches in Section 5. Section 6 ends with a discussion and future directions.

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. We use a latent variable model to summarize the information of $\mathbf{y}_i$. We 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.

We assume that the elements of $\mathbf{y}_i$ are linearly related to latent variables according to the measurement model,

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_i \eta_i + \boldsymbol{\varepsilon}_i, \quad (2.1)$$

where η_i is a latent variable, $\boldsymbol{\Lambda}_i = (\lambda_{1,i}, \dots, \lambda_{p,i})^T$ is a $P \times 1$ vector of factor loadings, $\boldsymbol{\mu}$ is a $P \times 1$ vector of scalars and $\boldsymbol{\varepsilon}_i$ is a $P \times 1$ vector of random errors. We let η_i be an i.i.d. random variable following the standard normal distribution. Since the mean of the latent variable is given, the observed variable means $\boldsymbol{\mu}$ are identifiable (Sánchez *et al.*, 2005). Furthermore, constraining the variance of the latent variable to 1 allows the identification of the factor loadings up to an additional sum to zero constraint (below). The error vectors $\boldsymbol{\varepsilon}_i$ are assumed to be independent of η_i , and to be independent and identically distributed (i.i.d.), and to follow a multivariate normal distribution, $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix. In our data application, the i.i.d. assumption is tenable since we first detrend the observed data. The normality assumption is also adequate since the log-transformed PM_{2.5} constituents are approximately normal (see Section 5 for more details).

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

In classical latent variable models, the factor loadings are estimated as constant parameters or sometimes allowed to differ across levels of a categorical variable (Horn and McArdle, 1992). However, in our proposed model, factor loadings are allowed to vary smoothly with time and/or location, resembling a varying coefficients model (Hastie and Tibshirani, 1993).

We 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 , $\lambda_{p,i}$ is specified in the form of an additive model,

$$\lambda_{p,i} = \lambda_{p,0} + f_{p,t}(t_i) + f_{p,uv}(u_i, v_i) + f_{p,tuv}(t_i, u_i, v_i), \quad (2.2)$$

where $f_{p,t}$, $f_{p,uv}$, $f_{p,tuv}$ are all centred functions (i.e., the sum of the function values across all the data points is zero), $\lambda_{p,0}$ 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. We will omit t_i , u_i , v_i in the following notations unless necessary.

We represent the non-constant factor loadings using penalized splines. Specifically, we construct univariate cubic B-spline basis functions for each of the covariates, t , u , v and use them to construct tensor product bases for the interaction terms $f_{p,uv}(u_i, v_i)$ and $f_{p,tuv}(t_i, u_i, v_i)$ (Wood, 2006, Section 4.1.8). Collecting the values of the factor loadings corresponding to the p th observed variable across all observations $i = 1, \dots, n$ into $\lambda_p = (\lambda_{p,1}, \dots, \lambda_{p,i}, \dots, \lambda_{p,n})^T$, we can express $\lambda_p = \lambda_p(\beta_p) = \mathbf{X}_p \beta_p$, where

$$\mathbf{X}_p = [1, \mathbf{X}_t, \mathbf{X}_{uv}, \mathbf{X}_{tuv}], \quad \beta_p^T = [\lambda_{p,0}, \beta_{p,t}^T, \beta_{p,uv}^T, \beta_{p,tuv}^T].$$

Note that we use $\Lambda_i = (\lambda_{1,i}, \dots, \lambda_{p,i}, \dots, \lambda_{P,i})^T$ to denote values of the P factor loadings for the i th observation and $\lambda_p = (\lambda_{p,1}, \dots, \lambda_{p,i}, \dots, \lambda_{p,n})^T$ 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 $i = 1, \dots, n$, and the columns corresponding to the components in equation (2.2) that have been suitably centered and constrained to remove redundancies that arise when creating the tensor product bases (see Wood, 2006, Section 1.8.1). Parameters β_p are the corresponding coefficients, which, except for the intercept $\lambda_{p,0}$, are penalized.

We use the squared first difference penalty (Eilers and Marx, 1996) to control the degrees of freedom of the smooth, which yields the desirable property of shrinking the smooth to the intercept $\lambda_{p,0}$. The penalty is constructed as a composite penalty that combines separate penalties for each of the component smooth in equation (2.2), that is, coefficients $\beta_{p,t}$, $\beta_{p,uv}$ and $\beta_{p,tuv}$ 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 Wood, 2006, Section 4.1.8 for a similar construction). In total, the penalty corresponding to each factor loading relies on six distinct smoothness 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 $\Omega_p(\beta_p, \omega_p) = \beta_p^T \mathbf{S}_p(\omega_p) \beta_p$, where $\mathbf{S}(\omega_p)$ is constructed as a block diagonal matrix whose diagonal blocks are 0, and $\omega_p \mathbf{S}_t$, $\omega_p \mathbf{S}_{u/v}$, $\omega_p \mathbf{S}_{v/tu}$, $\omega_p \mathbf{S}_{t/uv}$, $\omega_p \mathbf{S}_{u/tv}$ and $\omega_p \mathbf{S}_{v/tu}$ that correspond to the penalties of $\beta_{p,t}$, $\beta_{p,uv}$ and $\beta_{p,tuv}$. We define the overall penalty as

$\Omega(\beta, \omega) = \sum_{p=1}^P \beta_p^T \mathbf{S}_p(\omega_p) \beta_p$, where β is the vector of spline coefficients for all the factor loadings combined and ω denotes all the penalty parameters.

3 Estimation

We 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

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) - \frac{1}{2} \sum_{p=1}^P \Omega_p(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p), \quad (3.1)$$

where $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$, respectively, and $\boldsymbol{\theta}$ contains the unique parameters in $\boldsymbol{\Sigma}$ (i.e., residual variances σ_p^2 for each observed variable) in addition to $\{\mu_p, \lambda_{p,0}, \boldsymbol{\beta}_p | p = 1, \dots, P\}$. The E-step computes the expected log-likelihood using the conditional distribution of $\eta_i | \mathbf{y}_i$. The M-step estimates $\boldsymbol{\Sigma}$, and $\mu_p, \lambda_{p,0}, \boldsymbol{\beta}_p, p = 1, \dots, P$, by maximizing the expected log-likelihood. The smoothing parameters $\boldsymbol{\omega}_p$ are selected within each iteration of the EM.

3.1 EM algorithm

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

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

3.1.1 E-step—Even though we are to maximize the penalized likelihood as in equation (3.1), the penalty does not affect the E-step because the penalty does not involve the missing latent variable. 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. (We omit the superscript $(r-1)$ on $\boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ in the formulas below, but the calculations use parameter estimates from the $(r-1)$ th iteration).

$$\begin{aligned} \eta_i^{(r)} &= E(\eta_i | \mathbf{y}_i) \\ &= E(\eta_i) + \text{Cov}(\eta_i, \mathbf{y}_i) [\text{Cov}(\mathbf{y}_i)]^{-1} [\mathbf{y}_i - E(\mathbf{y}_i)] \\ &= \boldsymbol{\Lambda}_i^T (\boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^T + \boldsymbol{\Sigma})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}), \end{aligned}$$

$$\begin{aligned} (\eta_i^2)^{(r)} &= E(\eta_i^2 | \mathbf{y}_i) \\ &= \text{Cov}(\eta_i | \mathbf{y}_i) + [E(\eta_i | \mathbf{y}_i)]^2 \\ &= \text{Cov}(\eta_i) - \text{Cov}(\eta_i, \mathbf{y}_i) [\text{Cov}(\mathbf{y}_i)]^{-1} \text{Cov}(\mathbf{y}_i, \eta_i) + (\eta_i^{(r)})^2 \\ &= 1 - \boldsymbol{\Lambda}_i^T (\boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^T + \boldsymbol{\Sigma})^{-1} \boldsymbol{\Lambda}_i + (\eta_i^{(r)})^2. \end{aligned}$$

3.1.2 M-step—In the M-step, we estimate $\boldsymbol{\mu}, \boldsymbol{\Lambda}_p, \boldsymbol{\Sigma}$, from the first part of equation (3.1), including the penalty. Given the estimated latent variables 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). To see this, first we rearrange the log-likelihood. If we let

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

$$\begin{aligned} \sum_{i=1}^n \log L \left(\boldsymbol{\mu}, \boldsymbol{\Lambda}_i, \sum; \mathbf{y}_i | \boldsymbol{\eta}_i^{(r)} \right) - \frac{1}{2} \sum_{p=1}^P \Omega(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p) &= \sum_{i=1}^n \sum_{p=1}^P \log L \left(\mu_p, \lambda_{p,i}, \sigma_p^2; y_{p,i} | \eta_i^{(r)} \right) - \frac{1}{2} \sum_{p=1}^P \Omega(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p) \\ &= \sum_{p=1}^P \left\{ \log L \left(\mu_p, \boldsymbol{\lambda}_p(\boldsymbol{\beta}_p), \sigma_p^2; \mathbf{y}_{\cdot p} | \boldsymbol{\eta}^{(r)} \right) - \frac{1}{2} \Omega(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p) \right\}, \end{aligned} \quad (3.2)$$

where $\mathbf{y}_{\cdot p} = (y_{p,1}, \dots, y_{p,n})^T$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ are $n \times 1$ column vectors. We can do this rearrangement because $\boldsymbol{\Sigma}$ is a diagonal matrix. This allows us to maximize equation (3.2) separately for each p , that is, for each observed variable. If $\boldsymbol{\Sigma}$ is not diagonal, this step would instead be analogous to multivariate regression.

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

$$\begin{aligned} &\log L \left(\mu_p, \boldsymbol{\lambda}_p(\boldsymbol{\beta}_p), \sigma_p^2; \mathbf{y}_{\cdot p} | \boldsymbol{\eta}^{(r)} \right) - \frac{1}{2} \Omega(\boldsymbol{\beta}_p, \boldsymbol{\omega}_p) \\ &\propto -n \log(\sigma_p^2) - \sigma_p^{-2} \|\mathbf{y}_{\cdot p} - \mathbf{1}_n \mu_p - \boldsymbol{\eta}^{(r)} \circ \boldsymbol{\lambda}_p(\boldsymbol{\beta}_p)\|^2 - \boldsymbol{\beta}_p^T \mathbf{S}(\boldsymbol{\omega}_p) \boldsymbol{\beta}_p, \end{aligned} \quad (3.3)$$

where ‘ $\circ$ ’ is the entrywise Hadamard product and $\mathbf{S}(\boldsymbol{\omega}_p)$ is the previously described penalty matrix. Equation (3.3) is essentially a regression of $\mathbf{y}_{\cdot p}$ on $\tilde{\boldsymbol{\eta}}^{(r)}$ with $\boldsymbol{\beta}_p$ penalized, where

$$\tilde{\boldsymbol{\eta}}^{(r)} = \begin{bmatrix} \eta_1^{(r)} \mathbf{x}_1 \\ \vdots \\ \eta_n^{(r)} \mathbf{x}_n \end{bmatrix}, \quad (3.4)$$

$$\boldsymbol{\eta}^{(r)} \circ \boldsymbol{\lambda}_p(\boldsymbol{\beta}_p) = \boldsymbol{\eta}^{(r)} \circ (\mathbf{X}_p \boldsymbol{\beta}_p) = \begin{bmatrix} \eta_1^{(r)} \mathbf{x}_1 \boldsymbol{\beta}_p \\ \vdots \\ \eta_n^{(r)} \mathbf{x}_n \boldsymbol{\beta}_p \end{bmatrix} = \tilde{\boldsymbol{\eta}}^{(r)} \boldsymbol{\beta}_p$$

that is, we can write

replacing $\boldsymbol{\eta}^{(r)} \circ \boldsymbol{\lambda}_p(\boldsymbol{\beta}_p)$ with $\tilde{\boldsymbol{\eta}}^{(r)} \boldsymbol{\beta}_p$ in equation (3.3) and maximizing the penalized log-likelihood, we then obtain a Ridge regression solution for the coefficients

$$\begin{bmatrix} \mu_p^{(r)} \\ \boldsymbol{\beta}_p^{(r)} \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{1}_n^T \mathbf{1}_n & \mathbf{1}_n^T \tilde{\boldsymbol{\eta}}^{(r)} \\ (\tilde{\boldsymbol{\eta}}^{(r)})^T \mathbf{1}_n & (\tilde{\boldsymbol{\eta}}^{(r)})^T \tilde{\boldsymbol{\eta}}^{(r)} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{S}(\boldsymbol{\omega}_p) \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{1}_n^T \\ (\tilde{\boldsymbol{\eta}}^{(r)})^T \end{bmatrix} \mathbf{y}_{\cdot p}, \quad (3.5)$$

and

$$(\sigma_p^2)^{(r)} = n^{-1} \left[(\mathbf{y}_{\cdot p} - \mathbf{1}_n \mu_p^{(r)})^T (\mathbf{y}_{\cdot p} - \mathbf{1}_n \mu_p^{(r)}) - 2(\mathbf{y}_{\cdot p} - \mathbf{1}_n \mu_p^{(r)})^T \tilde{\boldsymbol{\eta}}^{(r)} \boldsymbol{\beta}_p^{(r)} + (\boldsymbol{\beta}_p^{(r)})^T (\tilde{\boldsymbol{\eta}}^T \tilde{\boldsymbol{\eta}})^{(r)} \boldsymbol{\beta}_p^{(r)} \right],$$

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

3.2 Generalized cross-validation

So far, we have treated the smoothing parameters $\boldsymbol{\omega}_p$ as given, but here we describe how to use the GCV score to select the smoothing parameters (Craven and Wahba, 1979; Golub *et al.*, 1979). GCV is commonly used in the selection of smoothing parameters for generalized additive models (GAM, see Wood, 2006, Section 4.5), 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 equation (3.6)). The smoothing parameters that give the smallest GCV score are selected. In the measurement model (see equation (2.1)), the latent variable, though actually $\tilde{\boldsymbol{\eta}}^{(r)}$ according to equation (3.4), is a natural counterpart of the covariate in a GAM model, so we make use of $\tilde{\boldsymbol{\eta}}^{(r)}$ and have

$$\text{GCV}_p = \frac{n \|\mathbf{y}_{\cdot p} - \mathbf{A}_p^{(r)} \mathbf{y}_{\cdot p}\|^2}{[n - \text{tr}(\mathbf{A}_p^{(r)})]^2}, \quad (3.6)$$

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 equation (3.5)) and $\text{tr}(\mathbf{A}_p)$ represents the effective number of parameters used to estimate $\boldsymbol{\lambda}_p$.

Since the latent variable is unknown and is itself to be computed from $\mathbf{y}_p$ and $\boldsymbol{\beta}^{(r)}$, use of $\|\mathbf{y}_{\cdot p} - \tilde{\boldsymbol{\eta}}^{(r)} \boldsymbol{\beta}^{(r)}\|^2$ is not valid if we 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, we 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 Wood, 2006, Section 4.6).

We 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. $\text{GCV}_p(\boldsymbol{\rho}_p)$ 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 $\text{GCV}_p(\boldsymbol{\rho}_p)$

can be approached by successively solving the quadratic minimum. Thus, the M-step is modified as follows. We first solve $\beta_p^{(r)}$ with $(\omega_p)^{(r-1)} = \exp((\rho_p)^{(r-1)})$ plugged in equation (3.5). Then we evaluate $\frac{\partial}{\partial \rho_p} \text{GCV}_p(\rho_p)$ and $\frac{\partial^2}{\partial (\rho_p)^2} \text{GCV}_p(\rho_p)$ at $(\rho_p)^{(r-1)}$, obtain the critical point of the quadratic approximation and re-evaluate the derivatives of $\text{GCV}_p(\rho_p)$ again at the critical point. The process is repeated until we get to the minimum of $\text{GCV}_p(\rho_p)$ and then $(\rho_p)^{(r-1)}$ is updated to $(\rho_p)^{(r)}$.

3.3 Simultaneous confidence band

We use bootstrap resampling to construct a 95% confidence interval for point estimates and 95% simultaneous confidence band for spline estimates (Ruppert *et al.*, 2003). We sample with replacement from the n observations to generate M new data sets of n observations each. In our data application, this resampling approach is valid since de-trending 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 ϵ_i of the actual data.

For single point estimates, such as $\lambda_{p,0}$, 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 $\hat{f}_{p,t}$, we first obtain C , the 95% quantile of

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

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 $\hat{f}_{p,t}(t) \pm C \sqrt{\text{Var}[\hat{f}_{p,t}^*(t) - \hat{f}_{p,t}(t)]}$. We obtain $\text{Var}[\hat{f}_{p,t}^*(t) - \hat{f}_{p,t}(t)]$ also from the bootstrap samples.

4. Simulation study

We conduct simulation studies to examine the advantages and disadvantages of this novel estimation method, as well as to verify that the estimation algorithm can recover model parameters. We first examine whether the algorithm can recover the true parameter values when the factor loadings vary with several covariates. For this purpose, we simulate 1000 data sets where P observed variables are associated with one latent factor and the factor loadings are smooth functions of three covariates in this form: $\lambda_{p,i} = \lambda_{p,0} + f_{p,uv}(u_i, v_i) + f_{p,t}(t_i)$. This model is similar to the one used for our data application. For each of the 1000 data sets, we generate data on an 11×11 grid (i.e., 121 (u_i, v_i)), and for each point on the grid we 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 our data example. The simulation shows that our algorithm can correctly estimate non-constant factor loadings that

change with multiple covariates, although some oversmoothing for $f_{p,uv}$ is observed (Figure A.1). Estimates for the other parameters are also unbiased (Table A.1).

Next, we further investigate how biases can enter various components of the model across different simulation scenarios where the factor loadings take on different functional forms. For this study, we let the factor loadings be functions of only one covariate. This simpler setting can enhance computational efficiency while still illustrating the impact of various estimating approaches under different scenarios. Specifically, we investigate how using non-constant factor loadings affects the estimation of model parameters and the factor scores. Examining estimation of factor scores gives insight as to how estimated source contributions in air pollution studies may be affected, for instance.

The simulations are set up with three different shapes for the true factor loadings. We let $\lambda_{p,i} = \lambda_{p,0} + f_p(t_i)$, where $t_i \in [0, 1]$, f_p is a centre 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 $p = 1, \dots, P$.

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 $\lambda_{p,i}$ deviates from $\lambda_{p,0}$. Figures 3a and 3b illustrate 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).

For each of the three shapes of factor loadings listed above, we hold some parameters constant, but we change the signal-to-noise ratio in the model as follows. For all the scenarios, we set $\mu_p = 1$ for all p and make $\lambda_{p,0}^2 + \sigma_p^2 = 8$ in all cases. Within each scenario, we let $\lambda_{p,0}$ take on the same value for all p , but we change the ratio, $\lambda_{p,0}^2 / (\lambda_{p,0}^2 + \sigma_p^2)$, 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, we 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, we generate 1000 simulated data sets, each with 1000 equally spaced data points within the range of t . For some scenarios, we vary the sample size (500, 5000) and the number of observed variables, $P = 6$ instead of 3. For each scenario, we estimate the factor loadings using three different approaches: (a) constant factor loadings, (b) non-constant

factor loadings using penalized splines and (c) non-constant factor loadings using unpenalized splines. For all the scenarios, we choose equally spaced knots that give 10 spline coefficients (including $\lambda_{p,0}$).

We report the simulation results primarily focusing on scenarios where (a) each simulated data set has 1000 equally spaced data points within the range of t_b , (b) $\lambda_{p,0}^2/(\lambda_{p,0}^2 + \sigma_p^2) = 0.8$ and (c) $\kappa = 1.25$ for the second and third types of $f_p(t_i)$. The results for these primary scenarios are provided in Tables 1 and 2.

We look at the results of our simulation study from two aspects. First, we compare the bias and standard error of $\hat{\lambda}_{p,0}, \hat{\sigma}_p^2$ among different estimation approaches and simulation scenarios (Table 1). This comparison primarily gives insight into possible loss of efficiency in constant factor loadings when they are instead estimated as non-constants, as well as gives insight regarding bias in residual variances, which ultimately affect how much weight a given observed variable is given when estimating the latent factor. When the factor loadings are constant, the bias and efficiency of $\hat{\lambda}_{p,0}, \hat{\sigma}_p^2$ are similar across estimation approaches, although $\hat{\sigma}_p^2$ is slightly underestimated when we use unpenalized splines since without penalization, the splines overfit the data. However, when the factor loadings are non-constant and estimated as constant, $\hat{\lambda}_{p,0}, \hat{\sigma}_p^2$ are both biased. The bias becomes trivial when we correctly estimate the factor loadings as non-constant. We also find that the penalty has little effect on the standard deviations of $\hat{\lambda}_{p,0}, \hat{\sigma}_p^2$ (compare ‘penalized’ with ‘un-penalized’ in Table 1), possibly because these two parameters measure the overall signal-to-noise ratio and thus they are less affected by the smoothness of $\hat{f}_p(t)$ as long as the main shape of the pattern is captured. However, the penalty reduces the mean integrated squared error of $\hat{f}_p(t)$ (see Table A.1), since the penalized splines are smoother.

Next, we compare how estimation approaches affect the factor score (Table 2). We define factor score as $\hat{\eta}_i = E(\eta_i | \mathbf{y}_i, \hat{\theta})$, where $\hat{\theta}$ 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 shrunk towards the centre compared with

η_i ($|E[\hat{\eta}_i | \eta_i]| = |\mathbf{\Lambda}_i^T (\mathbf{\Lambda}_i \mathbf{\Lambda}_i^T + \sum_{\epsilon}^{-1} \mathbf{\Lambda}_i \eta_i)| < |\eta_i|$). Thus, $\hat{\eta}_i$ is potentially more biased for values of η_i that are farther away from the centre 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 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), we also examine bias and MSE by percentiles of η_i . Since the distribution of the factor score is symmetric, only three percentile categories are listed in Table 2. The upper three categories have the same bias in negative sign (thus the overall zero bias) and the same MSE. To gain further insight regarding estimation of the factor score at different covariate values, we also draw scatterplots of the estimated factor score versus the true 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 we examine bias by percentiles of the

distribution of the true latent variable and when we examine the factor scores at different covariate values t (Table 2 and Figure 4a–4b). We find that when the factor loadings are constant, the bias and MSE are similar if we estimate factor loadings either as constant or as non-constant using penalized splines. If we estimate the factor loadings using unpenalized splines, the MSE is unsurprisingly larger. However, when the factor loadings are non-constant, we obtain the smallest bias and MSE when we estimate the factor loadings using penalized splines. Conversely, the bias and MSE are 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 4a–4b). 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 scatterplot of factor scores versus true latent variable in Figure 4b marked according to the values of t).

In scenarios where the variance in the observed variables is smaller (i.e., $\lambda_{p,0}^2/(\lambda_{p,0}^2 + \sigma_p^2)$ and κ are smaller), we 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 $\hat{\sigma}_p^2$ and increased MSE among all the parameter estimates along with the increase of MSE for the factor score.

5 Data application

We applied our model to four constituents from the US PM_{2.5} speciation data issued by EPA for years 2008 and 2009. The data set we used included 108 monitors stationed in part of the midwest, south and northeast (see Figure 5). Thirty-six monitors had measurements in every three days when they were functioning, but the majority had measurements only in every six days. Thus, we took every other observation from the thirty-six monitors that had measurements in every three days so that data from all monitors were equally spaced.

Many of the constituents in the EPA data were zero inflated and would require special treatment beyond the scope of this article. Thus, we selected constituents that, after taking the logarithm, had an approximate normal distribution. Based on the exploratory factor analysis (see Section 1 and Figure 1), we decided to look at four constituents, sulphur, ammonium ion, total nitrate, sulphate, that are associated with secondary aerosols (Nikolov *et al.*, 2008) and fit into a one-factor structure.

Since the concentration of constituents change across time and space, the mean of each observed variable (μ in equation (2.1)) would need to be modelled as a smooth function. Although this is in principle possible with our modelling approach, we opted instead to detrend the data and work with residuals, that is, first we regressed each constituent on time and spatial coordinates using GAM and then we used the residuals from these GAMs as the observed variables. Note that this approach not only allowed us to treat μ as constant but also effectively removed spatial and temporal correlation that enabled us to conduct inference using bootstrap.

The factor loadings were initially modelled according to equation (2.2), but the interaction term $f_{p,tuv}$ was insignificant based on the bootstrap confidence band, and thus we decided to keep only f_t and f_{uv} . We used cubic B-splines as the basis functions; we chose the number of equally spaced knots so that β_t and β_{uv} had 10 and 25 parameters, respectively. Altogether, β had 34 parameters. Each component function then had its own set of smoothing parameters, a total of three smoothing parameters.

Table 3 summarizes the estimates of all the parameters in the model except the spline coefficients. (We do not list μ in the table because they are zero for the detrended data.) Figure 5 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 engulf zero entirely (not shown). From Figure 5a, we can see that the factor loading in the midwest is significantly higher in the midwest region than in the south. Figure 5b 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 can be mostly explained by the fact that formation of ammonium nitrate particles is favoured in warmer environments (Salvador *et al.*, 2004). 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 6).

6 Discussion

We propose a latent variable model where the factor loadings are allowed to vary across continuous covariate values. This modelling approach extends classical factor analysis models and can help shed light into how the relationship between measured variables and underlying constructs may be modified by covariates, and thus can help improve estimation of underlying latent variables. This model is motivated by analysis of PM_{2.5} data that showed a noticeable difference for some of the factor loadings across time and region. Variations of factor analysis 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 temporal and spatially varying correlations among constituents can help improve summaries of PM_{2.5} constituents collected from a wide geographical area and time span. Our model allows us 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.

We use the penalized regression spline to model factor loadings as functions of multiple covariates and we 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 variable model, we 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 variable as the predictor in a single M-step. We 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 (available as supplementary material), thus enhancing applicability of our method.

Alternatively, we can penalize the spline by treating the spline coefficients as random variables, similar to the likelihood approach used for smoothing parameter selection in additive models (Ruppert *et al.*, 2003; Wood, 2006). Under this approach, the variance of the random coefficients acts as the smoothing parameter and it can be estimated using restricted maximum likelihood (REML). Reiss and Ogden (2009) discusses the advantage of REML over GCV based on a theoretical comparison. However, this approach would be more computationally intensive for a latent factor model because the likelihood does not have a closed form.

In this article, we use bootstrap to draw inference, which can be computationally intensive. Future work can include developing methods to test the significance of different component functions of the non-constant factor loading more efficiently. In our model, we assume that the random errors in equation (2.1) are uncorrelated. In our data application, we de-trended the data before applying our modelling strategy to remove spatial and temporal autocorrelation using GAM. Conditional residuals from the model had negligible correlation, suggesting that an independent error structure may be sufficient for the de-trended data that we have analyzed. Incorporating alternative correlation structures is a desirable next step to avoid the two-step approach we have used. Alternatively, as in the work of Gryparis *et al.* (2007), the mean trend of the latent variable could also be used to capture the temporal and spatial correlation; this extension could also be used to readily predict constituents in space and time, by first predicting the latent variable at a given time and location.

In our current data analysis, we have modelled every constituent with a penalized regression spline and then assessed 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 (Kneip and Gasser, 1988; Jiang *et al.*, 2013). The extension of our model to more than one latent variable would be fairly straightforward after proper identifiability constraints are applied. The model can also easily accommodate data missing at random, since it is likelihood based. However, a potentially more challenging problem is the case when observed variables are not measured simultaneously at all locations, that is, missing by design.

The model and algorithm implemented in this article provide a platform for developing more comprehensive measurement models that can allow us to apply flexible dimension reduction strategies. The measurement model for the factor analysis that we have proposed can incorporate known information (such as well-understood sources of air pollution) while learning from the data by allowing for non-constant factor loadings.

Acknowledgments

The authors acknowledge support from: NIH grants R01ES016932; R01ES017022; P01ES022844; P20ES018171; UM NIEHS Core Center P30 ES017885 and EPA grants RD834800 and RD83543601. The contents of this study are solely the responsibility of the authors and do not necessarily represent the official views of the The National Institute of Environmental Health Sciences (NIEHS) or the National Institute of Health or EPA.

References

- Arminger G. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis—Hastings algorithm. *Psychometrika*. 1998; 63:271–300.
- Austin E, Coull BA, Zanobetti A, Koutrakis P. A framework to spatially cluster air pollution monitoring sites in US based on the PM_{2.5} composition. *Environmental International*. 2013; 59:244–54.
- Bauer DJ. The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*. 2005; 10:305–16. [PubMed: 16221030]
- Bell ML, Dominici F, Ebisu K, Zeger SL, Samet JM. Spatial and temporal variation in PM_{2.5} chemical composition in the United States for health effects studies. *Environmental Health Perspectives*. 2007; 115:989–95. [PubMed: 17637911]
- Berrocal VJ, Gelfand AE, Holland DM. A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics*. 2010; 4:1942–75. [PubMed: 21853015]
- Choi J, Fuentes M, Reich BJ. Spatial-temporal association between fine particulate matter and daily mortality. *Computational Statistics and Data Analysis*. 2009; 53:2989–3000. [PubMed: 19652691]
- Craven P, Wahba G. Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*. 1979; 31:377–403.
- Daniels MJ, Zhou ZG, Zou H. Conditionally specified space-time models for multivariate processes. *Journal of Computational and Graphical Statistics*. 2006; 15:157–77.
- Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*. 2006; 295:1127–34. [PubMed: 16522832]
- Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science*. 1996; 11:89–102.
- Fahrmeir L, Raach A. A Bayesian semi-parametric latent variable model for mixed responses. *Psychometrika*. 2007; 72:327–46.
- Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good Ridge parameter. *Technometrics*. 1979; 21:215–23.
- Gryparis A, Coull BA, Schwartz J, Suh HH. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society Series*. 2007; 56:183–209.
- Guo R, Zhu HT, Chow SM, Ibrahim JG. Bayesian lasso for semiparametric structural equation models. *Biometrics*. 2012; 68:567–77. [PubMed: 22376150]
- Hastie T, Tibshirani R. Varying-coefficient models. *Journal of the Royal Statistical Society Series B*. 1993; 55:757–96.
- Hewson PJ, Bailey TC. Modelling multivariate disease rates with a latent structure mixture model. *Statistical Modelling*. 2010; 10:241–64.
- Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*. 1992; 18:117–44. [PubMed: 1459160]
- Huerta G, Sanso B, Stroud JR. A spatiotemporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society: Series C*. 2004; 53:231–48.
- Ippoliti L, Valentini P, Gamerman D. Space–time modelling of coupled spatiotemporal environmental variables. *Journal of the Royal Statistical Society: Series C*. 2012; 61:175–200.
- Jiang Q, Wang HS, Xia YC, Jiang GH. On a principal varying coefficient model. *Journal of the American Statistical Association*. 2013; 108:228–36.

- Kamman EE, Wand MP. Geoadaptive models. *Journal of the Royal Statistical Society: Series C*. 2003; 52:1–18.
- Kneip A, Gasser T. Convergence and consistency results for self-modeling nonlinear-regression. *Annals of Statistics*. 1988; 16:82–112.
- Laden F, Neas LM, Dockery DW, Schwartz J. Association of fine particulate matter from different sources with daily mortality in six US cities. *Environmental Health Perspectives*. 2000; 108:941–47. [PubMed: 11049813]
- Lee SY, Song XY, Lee JCK. Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics*. 2003; 28:111–34.
- Liu X, Wall MM, Hodges JS. Generalized spatial structural equation models. *Bio-statistics*. 2005; 6:539–57.
- Lopes HF, Salazar E, Gamerman D. Spatial dynamic factor analysis. *Bayesian Analysis*. 2008; 3:759–92.
- Lu B, Song XY, Li XD. Bayesian analysis of multi-group nonlinear structural equation models with application to behavioral finance. *Quantitative Finance*. 2012; 12:477–88.
- Meredith W. Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*. 1993; 58:525–43.
- Nikolov MC, Coull BA, Catalano PJ, Diaz E, Godleski JJ. 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*. 2008; 57:357–78.
- Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. *Environmetrics*. 2011; 22:165–78.
- Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, Dominici F. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*. 2009; 117:957–63. [PubMed: 19590690]
- Reiss PT, Ogden RT. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B*. 2009; 71:505–23.
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression* (Cambridge series in statistical and probabilistic mathematics). Cambridge, New York: Cambridge University Press; 2003.
- Sahu SK, Mardia KV. A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society: Series C*. 2005; 54:223–44.
- Salvador P, Artinano B, Alonso DG, Querol X, Alastuey A. Identification and characterisation of sources of PM₁₀ in Madrid (Spain) by statistical methods. *Atmospheric Environment*. 2004; 38:435–47.
- Sánchez BN, Budtz-Jørgensen E, Ryan LM, Hu H. Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association*. 2005; 100:1443–55.
- Sánchez BN, Kang S, Mukherjee B. A latent variable approach to study gene—environment interactions in the presence of multiple correlated exposures. *Biometrics*. 2012; 68:466–76. [PubMed: 21955029]
- Schwartz J, Dockery DW, Neas LM. Is daily mortality associated specifically with fine particles? *Journal of the Air and Waste Management Association*. 1996; 46:927–39. [PubMed: 8875828]
- Shaddick G, Wakefield J. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C*. 2002; 51:351–72.
- Song XY, Lu ZH. Semiparametric latent variable models with Bayesian P-splines. *Journal of Computational and Graphical Statistics*. 2010; 19:590–608.
- Tao Y, Sánchez BN, Mukherjee B. Latent variable models for gene—environment interactions in longitudinal studies with multiple correlated exposures. *Statistics in Medicine*. 2015; 34:1227–241. [PubMed: 25545894]
- Thurston GD, Ito K, Lall R. A source apportionment of U.S. fine particulate matter air pollution. *Atmospheric Environment*. 2011; 45:3924–36. [PubMed: 24634604]

- Valentini P, Ippoliti L, Fontanella L. Modeling US housing prices by spatial dynamic structural equation models. *Annals of Applied Statistics*. 2013; 7:763–98.
- Wood, SN. *Generalized Additive Models: An Introduction with R* (Texts in statistical science). 2006.
- Zhang WY, Lee SY. Nonlinear dynamical structural equation models. *Quantitative Finance*. 2009; 9:305–14.
- Zhu HT, Lee SY. Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*. 1999; 52:225–42.

APPENDIX

Table A.1

Bias and standard error (in parentheses) of $\hat{\mu}_p$, $\hat{\lambda}_{0,p}$, $\hat{\sigma}_p^2$ for simulations where the factor loadings change with three covariates.

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\lambda}_{1,0}$	$\hat{\lambda}_{2,0}$	$\hat{\lambda}_{3,0}$	$\hat{\lambda}_{4,0}$	$\hat{\sigma}_1^2$
0.000 (0.023)	0.001 (0.023)	0.001 (0.024)	0.000 (0.025)	−0.007 (0.025)	−0.009 (0.024)	−0.001 (0.025)	0.002 (0.025)	0.003 (0.066)

Table A.2

Mean integrated squared error (MISE) of $\hat{f}_p(t)$ for different simulation scenarios and estimating approaches.

True factor loadings	Estimating approach	$\hat{f}_1(t)$	$\hat{f}_2(t)$	$\hat{f}_3(t)$
Constant	Penalized	3.82	4.25	3.80
	Unpenalized	45.28	45.31	44.63
Non-constant cyclic	Penalized	37.48	32.61	33.12
	Unpenalized	51.36	41.96	42.12
Non-constant non-cyclic	Penalized	30.01	39.94	37.78
	Unpenalized	45.84	50.23	52.18

Figure A. 1.

$\hat{f}_{1,UV}$, $\hat{f}_{1,t}$ and their 95% confidence bands, (a), (b) respectively show $f_{1,UV}$, $\hat{f}_{1,UV}$ as contour plots. (c) plots $f_{1,t}$ (circles), $\hat{f}_{1,t}$ (solid line) and its confidence band (shaded area). Results for the other $\hat{f}_{p,UV}$, $\hat{f}_{p,t}$ are similar.

Figure 1.

Array of heat maps shows how the Pearson correlation among the four PM_{2.5} constituents (log-transformed) change across different strata defined by region (row) and time period (column). The column headers denotes the year (Yr1: 2008, Yr2: 2009) and the quarter (Qtr) within that year. The row headers indicate the three US regions (also see Section 5 and Figure 5). The strength of correlation is represented with different shades of green. We can see that the differences in correlations are primarily between total nitrate and the other constituents.

Figure 2.

(a) and (b) 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 and dotted lines, respectively. These results suggest that there is temporal and spatial difference for the factor loading of total nitrate (a). In comparison, ammonium ion (b) has a rather constant factor loading as the confidence intervals (represented as error bars) of the different estimates mostly overlap.

Figure 3.

(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 $\kappa = 1$.

Figure 4.

(a)–(b). 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 marker symbol distinguishes η_i by tertiles of t_p , where circle, diamond and triangle refer to the lower, middle and upper tertile, respectively. The solid line is where $(\eta_i, \hat{\eta}_i)$ falls if $\hat{\eta}_i = \eta_i$.

Figure 5.

(a) is the contour plot of $\hat{f}_{p,uv}$ for the factor loading of total nitrate. The bigger scattered dots are the monitor locations and the smaller arrayed 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 1 January of 2008 and $t=2$ is 31 December of 2009).

Figure 6.

Factor scores from two models: (a) all factor loadings estimated as constant parameters (y -axis), (b) all factor loadings estimated as non-constant (x -axis). Factor scores representing observations from different seasons and regions are plotted as follows: warmer regions in cold months (circle), colder regions in cold months (diamond), warmer regions in warm months (square) and colder regions in warm months (triangle); warm months refer to May, June, July, August; cold months refer to November, December, January, February; warmer regions refer to Mississippi, Georgia, Alabama and colder regions refer to the midwest except Missouri.

Table 1

Bias and standard deviation (in parentheses) of $\hat{\lambda}_{1,0}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$ for different simulation scenarios and estimating approaches.

True factor loadings	Estimating approach	$\hat{\lambda}_{1,0}$	$\hat{\lambda}_{2,0}$	$\hat{\lambda}_{3,0}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
Constant	Constant	-0.005 (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)

Bias and mean squared error (MSE) of the factor score for different simulation scenarios and estimating approaches. IBIAS and IMSE show the bias and MSE integrated over the entire distribution of the latent variable η ; the remaining columns show the bias and MSE integrated within ranges of the distribution of η , for example, the ' $< 10\%$ ' column shows the bias at the lower tail of the distribution of η . The bias and MSE were symmetric, hence only columns for the lower half of the distribution are shown. As the values show, bias and MSE are larger at the tails of the distribution of η .

Table 2

True factor loadings	Estimating approach	Bias			MSE				
		IBIAS	< 10%	10% ~ 25%	25% ~ 50%	IMSE	< 10%	10% ~ 25%	25% ~ 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

Table 3Parameter estimates for four PM_{2.5} constituent data.

Constituent		
$\lambda_{p,0}$	Constant factor loadings	Non-constant factor loadings
Sulphur	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)
Sulphate	0.976 (0.951, 1.000)	0.955 (0.925, 1.107)
σ_p^2	Constant factor loadings	Non-constant factor loadings
Sulphur	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)
Sulphate	0.047 (0.032, 0.064)	0.053 (0.033, 0.075)