A novel principal component analysis for spatially misaligned multivariate air pollution data

Roman A. Jandarov,
University of Cincinnati, USA

and Lianne A. Sheppard, Paul D. Sampson and Adam A. Szpiro
University of Washington, Seattle, USA

[Received February 2015. Revised February 2016]

Summary. We propose novel methods for predictive (sparse) principal component analysis with spatially misaligned data. These methods identify principal component loading vectors that explain as much variability in the observed data as possible, while also ensuring that the corresponding principal component scores can be predicted accurately by means of spatial statistics at locations where air pollution measurements are not available. This will make it possible to identify important mixtures of air pollutants and to quantify their health effects in cohort studies, where currently available methods cannot be used. We demonstrate the utility of predictive (sparse) principal component analysis in simulated data and apply the approach to annual averages of particulate matter speciation data from national Environmental Protection Agency regulatory monitors.

Keywords: Air pollution; Dimension reduction; Principal component analysis; Spatial misalignment; Universal kriging

1. Introduction

One of the first well-documented air pollution events was in Belgium in 1930. In the Meuse Valley incident a thick fog lasted for 5 days, which led to hundreds of people suffering from respiratory symptoms and 60 deaths during the following days (Nemery et al., 2001). Another major event was the famous London fog in 1952 which led to an estimated excess death toll of over 4000 (Logan, 1953). More recently, the number of extra deaths after this event was re-estimated to be nearly 12000 (Bell and Davis, 2001).

Since these events, extensive research has been conducted on the health effects of ambient air pollution exposure. A growing body of literature presents evidence of the adverse effects of long-term air pollution exposure on various health end points (Samet et al., 2000; Pope et al., 2002; Brook et al., 2007). However, many of these studies focus on a single pollutant. For example, during recent years, much attention has been devoted to studying the role of particulate matter in air pollution, specifically particulate matter less than 2.5 mm in aerodynamic diameter (PM$_{2.5}$) and its effects on health (Miller et al., 2007; Pope and Dockery, 2006).

It is known that ambient air pollution is a complex mixture that contains multiple pollutants. The health studies that focus on understanding effects of a single pollutant ignore the fact that typically people are simultaneously exposed to two or more pollutants, and these can
combine either to amplify or to alleviate their overall effects on the health end points. The
associations between the health end points and the single proxy pollutants that are found in
these studies are, therefore, more likely to be the effect of the mixture (that may change by space
and time), rather than the effect of the individual pollutant (Crouse et al., 2010). For example,
health effects of PM$_{2.5}$ could change depending on characteristics of the particles, including
shape, solubility, acidity or chemical composition of the pollutant (Vedal et al., 2012). Hence,
in contrast with studies based on a single-pollutant approach, investigating health effects of
multipollutant exposure can be useful in epidemiological studies for two main reasons:

(a) it may provide a better understanding of potential combined effects of individual pollu-
tants and interactions between the pollutants and
(b) it may result in easier interpretation of these effects to help to inform air quality manage-
ment policy.

Analysis of health effects of long-term average, spatially varying multipollutant air pollution
exposure by using cohort data has two main challenges:

(a) dimensionality of the multipollutant data and
(b) the spatial misalignment of the monitoring data and cohort study participants.

The first challenge is fairly common in any analysis of high dimensional data when there are
many independent variables of interest. In the context of understanding health effects of ex-
posure to multiple pollutants, dimensionality can present difficulties because these pollutants
could potentially be highly correlated. Subsequently, this means that estimating and interpreting
parameters of interest in health models to reveal the association between the health end point
and the predictors could be non-trivial.

The second challenge arises because multipollutant exposures are not always available at all
subject locations in a cohort study. This challenge is obviously not unique to multipollutant
exposure studies. For example, in epidemiological studies of health effects of a single pollutant,
the common framework to deal with this issue is to use prediction models to assign exposures
to study participants on the basis of the monitoring data that are available at different locations.
To investigate the effects of multipollutant exposure, this framework requires building multiple
prediction models and assigning exposures to each pollutant at all study participant locations.
This means that, if one of these prediction models produces inaccurate predictions, the results
of the health analysis may become unreliable.

Although there are methods to resolve the issues with dimensionality and spatial misalign-
ment in various contexts, these methods do not deal with these challenges simultaneously. To
understand health effects of multipollutant exposures and health end points, these challenges
need to be solved together. In this paper, we propose an approach for dimension reduction that
can be applied to multipollutant spatial data that resolves these issues in a computationally
fast unified approach. Our approach seeks to find sparse principal component (PC) scores that
explain a large proportion of the variance in the data while also ensuring that mixtures that are
derived from these components are predictable at health end point locations. Predictions of the
lower dimensional component scores can then be effectively used in a health effect analysis. We
show that the new approach is preferable to the sequential two-step approach of dimension re-
duction followed by spatial prediction, which may result in PC scores that are difficult to predict
at unmeasured locations. We apply our approach to national multipollutant air pollution data
from the US Environmental Protection Agency.

Two widely used models for predicting exposures at subject locations are based on land use
regression with geographic information system (GIS) covariates (Brauer et al., 2003; Hoek et al.,
Principal Component Analysis for Air Pollution Data

2008) and universal kriging (UK) (Jerrett et al., 2005; Künzli et al., 2005). In addition to using GIS covariates to construct the mean structure of the model as in land use regression, UK also allows for spatial dependence by modelling correlations between residuals at different locations (Kim et al., 2009; Mercer et al., 2011; Sampson et al., 2013). Since this usually results in more accurate predictions of pollutants, we use UK throughout the paper.

The rest of the paper is organized as follows. Section 2 describes the available monitoring multipollutant data and geographic covariates. Section 3 describes in detail the traditional and predictive principal component analysis (PCA) approaches. Here, in Section 3.2, we also briefly give an overview of UK that is used in predicting pollutants and PC scores at new locations. In Section 4, we present results from our simulation analysis to demonstrate key differences between different methods. In Section 5, we describe the application of PCA approaches to the monitoring data. Finally, in Section 6, we summarize our methods and results and discuss our statistical approach and conclusions.

The programs that were used to analyse the data can be obtained from http://wileyonlinelibrary.com/journal/rss-datasets

2. Data

2.1. Monitoring data

Air pollution data are collected nationally by a network of monitors called the air quality system (AQS). Most of these data are accessible at the Environmental Protection Agency’s (EPA) Web site http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm. Whereas measurements of PM$_{2.5}$ are available at almost all AQS monitor locations, detailed data on other pollutants and PM$_{2.5}$ components are collected at only two subnetworks of the AQS network. One of the subnetworks of the AQS with multipollutant data is called the chemical speciation network (Environmental Protection Agency, 2009). Chemical speciation network monitoring locations are generally selected to assess population multipollutant exposures and therefore are mostly located in urban areas. Another subnetwork of the AQS with multipollutant monitoring data is called the ‘Interagency monitoring of protected visual environments’ (IMPROVE) network (Eldred et al., 1988). IMPROVE monitoring locations are located in national parks and some rural areas. Data collected by the IMPROVE network are available at the ‘Visibility information exchange web system’, at http://views.cira.colostate.edu/web/. IMPROVE monitoring locations are selected to assess pollution effects in sparsely populated environments. Therefore, most of these monitors do not represent exposure levels for (the majority of) subjects living in populous areas.

The AQS monitors are located throughout the USA. However, as the availability of funding and monitoring aims evolve, some new monitors become available or/and some old ones are discontinued, and monitoring schedules are changed. For example, the monitors at a given location that collected PM$_{2.5}$ data from 2000 to 2002 may not have collected PM$_{2.5}$ data from 2003 to 2005. Furthermore, an individual monitor that collected PM$_{2.5}$ data every day in 2002 may have collected data every third day in 2003. Another monitor that collects PM$_{2.5}$ data every day during the winter may collect data only every third day in the summer. Since these features complicate the analysis of the data, we apply the methods that are described in this paper to annual averages data for 2010. In 2010, the data from the AQS monitors were at a daily resolution. These data are converted to annual averages by using completeness rules, making sure that each monitor had at least 10 data points per quarter and a maximum of 45 days between measurements. Subsequently, before the analysis, the annual averages were square root transformed.
Initially, we had 7375 monitors with incomplete data for 28 pollutants: PM$_{2.5}$, PM$_{10}$, nitrogen oxides NO$_2$ and NO$_x$, sulphur dioxide (SO$_2$), ozone (O$_3$), elemental carbon (C), organic C, sulphate (SO$_4$), nitrate (NO$_3$), aluminium (Al), arsenic (As), bromine (Br), cadmium (Cd), calcium (Ca), chromium (Cr), copper (Cu), cobalt (Co), iron (Fe), potassium (K), manganese (Mn), sodium (Na), sulphur (S), silicon (Si), selenium (Se), nickel (Ni), vanadium (V) and zinc (Zn). Since we needed complete multipollutant data for our analysis, our first goal was to obtain a list of pollutants with a large number of monitors without missing measurements for all pollutants in the list. After cleaning the original data, our final list included 19 pollutants: PM$_{2.5}$, elemental C, organic C, Al, As, Br, Ca, Cr, Cu, Fe, K, Mn, Na, Ni, S, Si, V and Zn. For these pollutants, the number of monitors with full data was equal to 284. The pollutants that were removed from the analysis were PM$_{10}$, NO$_2$, NO$_x$, SO$_2$, O$_3$, SO$_4$, NO$_3$, Cd and Co. We note that adding any of the pollutants PM$_{10}$, NO$_2$, NO$_x$, SO$_2$ or O$_3$ to our final list reduced the number of monitors from 284 to a number between 44 and 55, whereas adding any of SO$_4$, NO$_3$, Cd or Co reduced the number of monitors to 123. We also note that S is a good surrogate for SO$_4$, and Cd and Co are not particularly good tracers of major sources believed to be scientifically important, and NO$_3$ has sampling artefact issues associated with its volatility.

The measurements for the pollutants from the final list were from the chemical speciation (130 monitors) and IMPROVE (154 monitors) networks. The data from these networks for these pollutants were reported in units of milligrams per cubic metre. Additionally, we note that in 2010 the chemical speciation and IMPROVE networks collected data by using similar methods. For example, for PM$_{2.5}$, both networks used Federal reference methods to collect data. IMPROVE data and AQS Federal reference methods data collected concurrently at the same locations were sufficiently comparable to be considered equivalent. Table 1 shows the list of retained pollutants with related summary statistics. Here, on the basis of $R^2$s (defined as

| Pollutant | Mean (mg m$^{-3}$) | Standard deviation (mg m$^{-3}$) | $R^2$ | MSE |
|-----------|------------------|-------------------------------|-------|-----|
| PM$_{2.5}$ | 7.351           | 3.814                         | 0.933 | 0.00156 |
| Elemental C | 0.417           | 0.336                         | 0.866 | 0.00299 |
| Organic C | 1.533           | 0.875                         | 0.877 | 0.00729 |
| Al | 0.050           | 0.036                         | 0.582 | 0.01484 |
| As | 0.001           | 0.001                         | 0.817 | 0.00057 |
| Br | 0.002           | 0.002                         | 0.730 | 0.00450 |
| Ca | 0.052           | 0.046                         | 0.513 | 0.00001 |
| Cr | 0.001           | 0.003                         | 0.497 | 0.00013 |
| Cu | 0.003           | 0.004                         | 0.555 | 0.00689 |
| Fe | 0.065           | 0.068                         | 0.545 | 0.00039 |
| K | 0.051           | 0.027                         | 0.668 | 0.00009 |
| Mn | 0.002           | 0.006                         | 0.269 | 0.00243 |
| Na | 0.091           | 0.114                         | 0.531 | 0.01473 |
| S | 0.566           | 0.352                         | 0.965 | 0.00004 |
| Si | 0.135           | 0.092                         | 0.577 | 0.00477 |
| Se | 0.001           | 0.002                         | 0.525 | 0.00086 |
| Ni | 0.001           | 0.001                         | 0.649 | 0.00283 |
| V | 0.001           | 0.001                         | 0.783 | 0.00282 |
| Zn | 0.008           | 0.011                         | 0.717 | 0.00010 |
\[ R^2 = \max\{0, 1 - \Sigma(x_i - \hat{x}_i)^2 / \Sigma(x_i - \bar{x}_i)^2\} \], where \(x_i\) is the observed and \(\hat{x}_i\) is the predicted values of the pollutants) and mean-squared errors (defined as \(\text{MSE} = \Sigma(x_i - \hat{x}_i)^2 / n\)) obtained by using cross-validation and UK (reviewed in detail in Section 3.2), we see that, whereas some pollutants are highly predictable (e.g. PM$_{2.5}$ or S have high \(R^2\)), some pollutants cannot be predicted well (e.g. Mn). We note here that predictions that are used in calculations of \(R^2\)s and MSEs are obtained from single-pollutant UK models to show descriptively how individually predictable the pollutants are.

2.2. Geographic covariates

The description below closely follows Bergen et al. (2013). For all monitor and subject locations, we obtained approximately 600 GIS covariates. These data were from various external sources, such as TeleAtlas, the US Census Bureau and the US Geological Survey. The list of GIS covariates includes distances to A1, A2 and A3 roads (census feature class codes), land use within a given buffer size, population density within a given buffer and the normalized difference vegetation index, which measures the level of vegetation in a monitor’s vicinity. Here, census feature class code A1 roads are limited access highways; A2 and A3 roads are other major roads such as county and state highways without limited access (Mercer et al., 2011). For the normalized difference vegetation index first, a series of 23 16-day composite satellite images from the year 2006 were obtained. Then, the index was converted by the University of Maryland from the \(-1\) to \(1\) scale to the 0–255-pixel brightness scale. On this scale, water has a value of approximately 50 and areas with dense vegetation have values around 200. For each location of interest, for each image, all pixels with a centroid within a certain distance of the location were averaged (radii included 250 m, 500 m, 1 km and 5 km). For each buffer size, five summary numbers were calculated from the series of 23 averages for each location: the 25th, median and 75th percentile of the entire year’s series, the median of the expected ‘high vegetation’ season, defined as April 1st–September 30th, and the median of the expected ‘low vegetation’ season, defined as the rest of the year.

Before further analysis, all GIS covariates are preprocessed to remove uninformative and unreliable variables with very small variability or outliers: we eliminated variables with more than 85% identical values, and those with the most extreme standardized outliers more than seven. We then log-transformed and truncated all distance variables at 10 km and computed additional ‘compiled’ distance variables such as the minimum distance to major roads and the minimum distance to any ports. All selected covariates are then mean centred and scaled by their respective standard deviations.

Additionally, we note that, because of a large number of GIS covariates (some of which are possibly correlated), using all GIS variables directly in land use regression or UK models is not practical. A commonly used solution is to build a model by first reducing the dimensionality of the GIS covariates into a smaller set of variables. These extracted variables are then used to construct the mean structure in the land use regression and UK models. Various dimension reduction and variable selection algorithms have been proposed to deal with dimensionality of GIS covariates. These algorithms include exhaustive search, stepwise selection and shrinkage by the lasso (see Tibshirani (1996) and Mercer et al. (2011)). Since most of these variable selection methods may be computationally expensive, an alternative dimension reduction approach is to apply PCA to the GIS data. Similarly to analogous approaches based on partial least squares (Sampson et al., 2011; Abdi, 2003), PCA allows us to reduce the number of original GIS covariates by using linear transformations into a smaller set of variables that explain most of the variability in the GIS data (see Section 3 for more on PCA). The PCA scores that are obtained
from GIS covariates can then be used to construct the mean structure in a land use regression or UK model instead of using all individual GIS covariates. Therefore, instead of using all GIS covariates, we use PCA scores obtained from GIS data only. To avoid confusion, throughout the paper, we refer to these lower dimensional variables simply as GIS covariates. We also note that all methods in this paper can be applied with the full set of preprocessed GIS covariates as well.

3. Methods

3.1. Review of sparse principal component analysis for exposure dimension reduction

We begin by reviewing dimension reduction methods (specifically PCA and sparse PCA) and their application to cohort studies when the exposure data are not spatially misaligned. We introduce notation along the way.

Suppose that, for each subject $i = 1, \ldots, n$ in a cohort study, we observe a scalar health endpoint $y_i$ and a $p$-dimensional exposure vector $x_i$. In linear regression, the most straightforward approach to assessing associations between $y_i$ and $x_i$ is to estimate the parameter vector $\beta$ in

$$Y = \beta_0 + X \beta + \epsilon,$$

possibly also including interactions. Here, $Y = (y_1, \ldots, y_n)^T$ is the vector of end points and $X$ is the $n \times p$ exposure matrix with rows $x_i$. In settings when the number of monitors or subjects is limited compared with the number of exposure covariates and/or when there is a complex dependence between the covariates, it can be beneficial to reduce the dimensionality of the exposures.

A dimension reduction method identifies low dimensional structure in the exposure data and given this structure provides a mapping from a full exposure vector to the corresponding low dimensional representation. PCA is a widely used technique that can extract a small number of important variables from high dimensional data (Jolliffe, 1986). In PCA, the low dimensional structure is a $p \times k$ matrix $V$ comprised of $k$ representative $p$-dimensional exposure vectors $\{v_i\}_{i \leq k}$, and the low dimensional representation is an $n \times k$ matrix $U$ comprised of $n$ subject-specific $k$-dimensional exposure vectors such that $X \sim UV^T$ for $k < p$ and $X = UV^T$ for $k = p$. Using this representation, we can estimate $\tilde{\beta}$ in

$$Y = \tilde{\beta}_0 + U \tilde{\beta} + \tilde{\epsilon}.$$ 

For $k \ll p$, the regression coefficients in $\tilde{\beta}$ are generally easier to estimate and (may be easier to) interpret than those in $\beta$. We clarify that, in the original regression problem, the exposure variables could be highly correlated. This means that the components of $\beta$ corresponding to each exposure variable are difficult to interpret as the amount of change in the end point variable $y$ per unit increase in the exposure variable when all the other exposures are kept constant. In the exposure data, some pollutants always occur and vary together, making it unrealistic to assume that pollutants can be kept constant when we change the values of other pollutants. In contrast, the coefficients $\tilde{\beta}$ are lower dimensional than $\beta$ and they correspond to mixtures that are usually almost uncorrelated. This implies that the components of $\tilde{\beta}$ can be interpreted as the change in the end point per 1-unit increase in the mixture with other principal mixtures fixed.

We focus on unsupervised dimension reduction methods that are not informed by the association between $Y$ and $X$. An alternative is supervised dimension reduction, which may sometimes identify representations that are more strongly indicative of the relationship between $Y$ and $X$. 

However, we consider it promising to work with unsupervised methods, as the resulting low dimensional representations have the potential of being important for the multiple health end points one faces in many epidemiology studies.

If we apply PCA to $X$, we find orthonormal direction or loading vectors $v_1, \ldots, v_p$ and corresponding PC scores $u_1 = Xv_1, \ldots, u_p = Xv_p$. Here, dimension reduction using PCA is then accomplished by using the first $k < p$ PC scores to represent the data, with $k$ chosen such that these few scores and their corresponding loadings explain most of the variation in the data $X$.

In traditional PCA, all entries in the loading vectors are non-zero, which can make it difficult to interpret individual PC scores. Several sparse PCA methods have been proposed, with the goal of modifying PCA so that the loading vectors contain many 0s to improve interpretability. We briefly review an approach that was proposed by Shen and Huang (2008), which exploits the connection between PCA and low rank approximation of matrices. For $k < p$, let us define

$$X^{(k)} = \sum_{l=1}^k u_l v_l^T,$$

where $u_1$ and $v_1$ are the first $k$ PC score and loading vectors. It can be shown that $X^{(k)}$ is the closest rank $k$ approximation to $X$ under the Frobenius norm. For two matrices $X$ and $\tilde{X}$, the Frobenius norm is defined as $\|X - \tilde{X}\|_F^2 = \text{tr}((X - \tilde{X})(X - \tilde{X})^T)$. This implies that we can conduct PCA by solving a series of optimization problems. Suppose, for example, that our goal is to find the best rank 1 approximation to the matrix $X$. Since any $n \times p$ rank 1 matrix can be written as $\tilde{u}\tilde{v}^T$ with $n$-dimensional unit vector $\tilde{u}$ and $p$-dimensional vector $\tilde{v}$, we can find the best rank 1 approximation to $X$ by solving the optimization problem

$$\min_{\tilde{u}, \tilde{v}} \|X - \tilde{u}\tilde{v}^T\|$$

with respect to $\tilde{u}$ and $\tilde{v}$, under the constraint $\|\tilde{u}\| = 1$. We then set $v_1 = \tilde{v}/\|\tilde{v}\|$ and $u_1 = Xv_1$ to obtain the first PC loading and score vectors. Note that we constrain $\|\tilde{u}\| = 1$ in the optimization, as this turns out to be helpful in the development of sparse PCA (Shen and Huang, 2008). We find subsequent pairs $(u_l, v_l)$, $l > 1$, by solving corresponding one-rank approximation problems with residual matrices obtained at the previous step.

As described in Shen and Huang (2008), an interpretable approach to enforce sparsity in the loadings $v$ can be developed via low rank approximations. To explain the algorithm, we focus on defining and calculating the first sparse PC score and its loading vector. We obtain subsequent PC scores and their loadings by simply considering the residual matrices of the sequential matrix approximations. The main idea is to add a regularization penalty on the loadings via equation (2). In other words, to find the first sparse PC loadings and scores, we solve the minimization problem

$$\min_{\tilde{u}, \tilde{v}} \|X - \tilde{u}\tilde{v}^T\| + P_\lambda(\tilde{v}),$$

with respect to $\tilde{u}$ and $\tilde{v}$, under the constraint $\|\tilde{u}\| = 1$. Here, $P_\lambda(y) = \Sigma \lambda |y_i|$ is an $L_1$ (lasso) penalty function with a penalty parameter $\lambda \geq 0$. This problem can easily be solved by a fast alternating algorithm; see Shen and Huang (2008) for details. Similarly to above, after $\tilde{u}$ and $\tilde{v}$ have been obtained, the first PC score is defined by $u_1 = Xv_1$ (where the loading vector is again $v_1 = \tilde{v}/\|\tilde{v}\|$). After iterating this procedure for the first $k$ PCs, we define our reduced dimensional exposure vector at monitor locations as $U = (u_1, \ldots, u_k)$. In what follows, we refer to this approach as traditional sparse PCA.

We note that the penalty parameter $\lambda \geq 0$ controls the degree of sparsity in the loadings. We shall discuss various options for selecting the penalty parameter in later sections, when we
describe our new predictive sparse PCA algorithm. If \( \lambda = 0 \) this algorithm recovers the standard PC loadings and scores.

### 3.2. Review of universal kriging for predicting principal component scores

We can apply (sparse) PCA as described above to reduce the dimensionality of spatially misaligned exposure monitoring data. However, our ultimate goal is to predict the PC scores at subject locations where exposure data are not available. A straightforward way to solve this problem is to apply (sparse) PCA without any modification to the monitoring data and then to use spatial statistics to predict the PC scores at subject locations. We review UK in this context.

First, let \( \mathbf{u} = \{ u(s) \} \) be a PC score (say, the first PC score), where \( s \in E \), and \( E \) is a set of geographical co-ordinates. If \( s \) is a monitor location, \( u(s) \) can be calculated. Our goal is to predict \( u(s^*) \) at subject locations \( s^* \). We model \( u(s) \) as

\[
  u(s) = \mathbf{Z}(s)^T \alpha + w(s), \quad \text{for } s \in E, \tag{4}
\]

where \( \mathbf{Z}(s) \) is a set of covariates that are associated with each location \( s \) and \( \alpha \) is a vector of coefficients. Traditionally, in UK, we use only the GIS covariates. Spatial dependence is then incorporated by modelling \( \{ w(s) : s \in E \} \) as a zero-mean stationary Gaussian process: if \( w = (w(s_1), \ldots, w(s_n))^T \) and \( \xi \) are the parameters of the model, then

\[
  w|\xi \sim N\{ \mathbf{0}, \Sigma(\xi) \}, \tag{5}
\]

where \( \Sigma(\xi) \) is a symmetric and positive definite covariance matrix of \( w \). One commonly used covariance function is the ‘exponential’ covariance function with parameters \( \xi = (\psi, \kappa, \phi) \) (with \( \psi, \kappa, \phi > 0 \)), which has the form \( \Sigma(\xi)_{ij} = \psi I(i = j) + \kappa \exp(-||s_i - s_j||/\phi) \), where \( I \) is the indicator function and \( ||s_i - s_j|| \) is the Euclidean distance between locations \( s_i, s_j \in E \). This model is interpreted in the following manner: \( \psi \) (which is often called the nugget effect) is the non-spatial error associated with each location, and \( \kappa \) (which is called the sill) and \( \phi \) (which is called the range) are parameters that define the scale and range of the spatial dependence respectively.

We note that, throughout the paper, we use the ‘exponential’ covariance function to model the covariance structure of the process. We also note that, although there are more flexible covariance functions (e.g. the Matérn function) that can be used to model the covariance matrix, we chose to use an exponential covariance function as we have found that it works well with air pollution data from EPA monitors (see Sampson et al. (2013) and Bergen et al. (2013)). Moreover, we believe that the results and conclusions in this paper should be robust with regard to the form of the covariance function.

If \( \mathbf{u} \) is the vector of observed PC scores modelled as a Gaussian process, our goal is to obtain predictions of the score \( \mathbf{u}^* \) at new locations in \( E \). Since \( \mathbf{u} \) is a Gaussian process, the joint distribution of \( \mathbf{u} \) and \( \mathbf{u}^* \) given \( \xi \) and \( \alpha \), the covariance and mean parameters respectively, can be written as

\[
  \left( \begin{array}{c} \mathbf{u} \\ \mathbf{u}^* \end{array} \right) | \xi, \alpha \sim N\left\{ \left( \begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right), \left( \begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right) \right\}, \tag{6}
\]

where \( \mu_1 \) and \( \mu_2 \) are the means of \( \mathbf{u} \) and \( \mathbf{u}^* \) respectively and \( \Sigma_{11}, \Sigma_{12}, \Sigma_{21} \) and \( \Sigma_{22} \) are block partitions of the covariance matrix \( \Sigma(\xi) \), given \( \xi \) and \( \alpha \). \( \mu_1 \) and \( \mu_2 \) are functions of \( \alpha \) and \( \Sigma_{11}, \Sigma_{12}, \Sigma_{21} \) and \( \Sigma_{22} \) are functions of \( \xi \). Consequently, using multivariate normal theory (Anderson, 2003), \( \mathbf{u}^* | \mathbf{u}, \alpha, \xi \) is normally distributed with the mean and covariance

\[
  E(\mathbf{u}^* | \mathbf{u}, \alpha, \xi) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{u} - \mu_1),
\]

\[
  \text{var}(\mathbf{u}^* | \mathbf{u}, \alpha, \xi) = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}. \tag{7}
\]
To obtain predictions at new locations, we first obtain estimates of \((\xi, \alpha)\) and \((\hat{\xi}, \hat{\alpha})\), via maximization of the multivariate normal likelihood function \(\pi(u|\xi, \alpha)\). Then, we use the formula for \(E(u^*|u, \alpha, \xi)\) in equation (7) as predictions of \(u^*\).

3.3. Predictive sparse principal component analysis

The two-stage procedure that was outlined in Sections 3.1 and 3.2 can be described as (sparse) PCA followed by spatial prediction. The main drawback to this procedure is that there is no mechanism in the first stage to ensure that we find scores that can be accurately predicted by a spatial model in the second stage. Suppose that there is a column of \(X\) that accounts for a large portion of the exposure variability but does not have much spatial structure. With complete data, we would weight this column highly in \(V\), but with misaligned data this would result in one or more columns of the score matrix \(U\) being poorly predicted at subject locations. This happens because traditional PCA does not know that our multipollutant data are spatial data. Therefore, although the traditional (sparse) PCA does give more weight to pollutants with higher variability, it does not distinguish overall variability from spatial variability. It can be observed that some pollutants could be highly variable, while also being well informed by geographical covariates or splines. These pollutants can be predicted well by a geostatistical model. At the same time, other pollutants could be highly variable and spatially unpredictable. This implies that traditional (sparse) PCA may not necessarily result in scores that have better or worse predictability compared with individual predictability of the pollutants. In this section, we extend the sparse PCA approach that was described in Shen and Huang (2008) to resolve this issue by adding a constraint to equation (3) to force the PCs to be predictable from the available data. This new algorithm is termed predictive (sparse) PCA. We develop an alternating minimization algorithm to solve the corresponding optimization problem. In contrast with traditional (sparse) PCA, in predictive (sparse) PCA, we inform the algorithm that the data are spatial by including the GIS variables and spline terms. Thus, in the new PCA, we are reducing (or eliminating in some cases) the contribution of the pollutants that are highly variable but spatially unpredictable.

Let \(\tilde{Z}(s)\) be a set of geographic covariates and thin plate spline basis functions calculated at each location \(s\). We shall use \(\tilde{Z}(s)\) to guide our selection of PC loadings to ensure that they are spatially predictable. As we noted in Section 3.2, in UK, \(Z(s)\) does not include splines. We include thin plate spline basis functions in \(Z(s)\) here so that our predictive (sparse) PCA algorithm incorporates both geographic covariates and spatial smoothing in the prediction. It is known that there is a link between interpolation using splines and kriging (see Matheron (1981), Hutchinson and Gessler (1994) and Dubrule (1984)). We can think of spline interpolation here as low rank kriging with fixed covariance and with a fixed degree of polynomial trend to approximate the spatial variability. In other words, the spline basis functions stand in for the kriging component in identifying loading vectors with predictable scores. When it comes time to predict scores optimally at subject locations we revert to UK with only GIS covariates in the mean model. In our approach, we use rank 10 thin plate regression splines. This is comparable with using 10 knots in conventional knot-based spline interpolation (see Wood (2003)). Although our choice of a rank 10 spline model is somewhat arbitrary, we remark that the spline model is intermediate in the sense that it is used to select PC loadings but not in the final prediction of scores at new locations. Our simulation and data analysis results in Sections 4 and 5 demonstrate that rank 10 splines are effective, if not necessarily optimal. Other methods for selecting the rank of thin plate splines have been proposed, including hypothesis testing based on generalized cross-validation, Mallows’s \(C_p\) criteria or maximization of the likelihood function (see Wood (2003)), but we do not explore these in the present setting.
Let $\tilde{Z}$ be the $(n \times m)$-dimensional matrix of the available covariates and spline terms calculated at each monitor location. To find the first PCs of the multipollutant data, we propose to optimize

$$\|X - (\tilde{Z}\alpha/\|\tilde{Z}\alpha\|)\tilde{v}^T + P_\lambda(\tilde{v})\|_F$$

with respect to $\tilde{\alpha}$ and $\tilde{v}$. Here, we use the $L_1$ penalty function $P_\lambda(\cdot)$ again. Note that this optimization problem is similar to that in equation (3). The difference is that $\tilde{\alpha}$ is now constrained to be equal to $\tilde{Z}\alpha/\|\tilde{Z}\alpha\|$ to force better predictability of the PC scores. The reason to normalize $\tilde{Z}\alpha$ is twofold. First, this increases identifiability of the parameters $\tilde{\alpha}$ and $\tilde{v}$. Second, it allows us to construct a fast algorithm to find a solution (see lemmas 1 and 2 below). After optimal $\hat{\alpha}$ and $\hat{v}$ for the first PC have been obtained, since $(\tilde{Z}\alpha/\|\tilde{Z}\alpha\|)\hat{v}^T$ explains some part of $X$, we define the residual matrix by

$$\tilde{X} = X - (\tilde{Z}\alpha/\|\tilde{Z}\alpha\|)\tilde{v}^T.$$  

This matrix is then used to find the second PC score and its loadings. The subsequent pairs of component scores and loading vectors can be found by considering analogously defined residual matrices by using the $\tilde{\alpha}$ and $\tilde{v}$ from the previous step.

To solve the optimization problems in equation (8) with respect to $\tilde{\alpha}$ and $\tilde{v}$, following ideas in Shen and Huang (2008), we use an iterative algorithm that is described below. We first consider the following lemmas.

**Lemma 1.** For a fixed $\tilde{v}$, the $\hat{\alpha}$ that minimizes equation (8) has the form

$$\hat{\alpha} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{W},$$

where $\tilde{W} = XX^T\tilde{v}/\tilde{v}^T\tilde{v}$.

**Proof.** When $\tilde{v}$ is fixed, it is easy to see that minimizing equation (8) is equivalent to minimizing

$$\|W - \tilde{Z}\alpha/\|\tilde{Z}\alpha\|^2_F,$$

which is equivalent to optimizing

$$\|W - \tilde{Z}\alpha\|^2_F \quad \text{subject to } \|\tilde{Z}\alpha\|^2 = 1.$$  

Hence, applying the method of Lagrange multipliers to the latter, we obtain

$$\hat{\alpha} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{W}.$$  

**Lemma 2.** For a fixed $\tilde{\alpha}$, the optimal $\hat{v}$ that minimizes equation (8) is obtained at

$$\hat{v} = h_\lambda(X^T\tilde{Z}\tilde{\alpha}/\|\tilde{Z}\tilde{\alpha}\|),$$

where $h_\lambda(y) = \text{sgn}(|y| - \lambda)_+.$

**Proof.** First, let $w = \tilde{Z}\alpha/\|\tilde{Z}\alpha\|$. Then, following Shen and Huang (2008), equation (8) can be minimized componentwise for each $j$ by minimizing

$$\sum_i x_{ij}^2 - 2(X^T w)_{ij}\tilde{v}_j + \tilde{v}_j^2 + P_\lambda(\tilde{v}).$$

Since $P_\lambda(\tilde{v}) = 2\lambda \sum_j |\tilde{v}_j|$, the proof of lemma 2 then follows from simple calculations.

We now propose the following algorithm to solve our optimization problem.

(a) We first choose initial values for $\tilde{\alpha}$ and $\tilde{v}$. 
(b) For a fixed $\tilde{v}$, we update $\tilde{\alpha}$ by using lemma 1.
(c) For a fixed $\tilde{\alpha}$, we update $\tilde{v}$ by using lemma 2.
(d) We repeat steps (b) and (c) until convergence.

Suppose that we are interested in the first $k$ PCs. Denote the parameters that are estimated by solving equation (8) $k$ times by $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_k$ and $\tilde{v}_1, \ldots, \tilde{v}_k$. Denote the corresponding residual matrices by $\tilde{X}_1, \ldots, \tilde{X}_k$, where $\tilde{X}_1 = X$, $\tilde{X}_2 = \tilde{X}_1 - (\tilde{Z}\tilde{\alpha}_1/\|\tilde{Z}\tilde{\alpha}_1\|)\tilde{v}_1^T$, etc. For $l = 1, \ldots, k$ we can now define the $l$th PC loadings by $v_l = \|\tilde{v}_l\|$. Then, we define PC scores as follows:

$$u_1 = X \frac{\tilde{v}_1}{\|\tilde{v}_1\|} = Xv_1,$$
$$u_2 = X \frac{\tilde{v}_2}{\|\tilde{v}_2\|} = Xv_2, \ldots$$

As in Section 3.2 we use UK to predict the PC scores at subject locations without exposure data, but we now expect prediction accuracy to be improved since the loadings were selected to ensure predictability of the scores. Here, using the GIS covariates in predictive PCA and in UK may seem to raise suspicion with overfitting. To clarify that this is not so, we note that the GIS covariates are not used to define the PC scores in equation (10). We use the GIS covariates only to guide the estimation of the loadings in the predictive (sparse) PCA. Additionally, to ensure that we are not overfitting the data, we validate our approach by using out-of-sample tenfold cross-validation.

3.4. Selection of penalty parameters
We describe two methods to select penalty parameters in sparse PCA. Recall that, in both traditional and predictive approaches, the penalty parameters $\lambda$ control the sparsity of the loading vectors for each PC. Since PCA algorithms are sequential, it is easy to see that different $\lambda$ can be used for different PCs.

The first approach is to maximize the prediction accuracy of the obtained PC scores by using tenfold cross-validated $R^2$ for each PC. The second approach is to maximize prediction accuracy of the original pollutants by minimizing the out-of-sample Frobenius distance $\|X - \hat{U}v^T\|_F$ with respect to $\lambda$ for each PC. Here, $\hat{U}$ is defined as the matrix of predictions of the PC scores at monitor locations obtained via tenfold cross-validation using UK. We use the term ‘out of sample’ because these predictions are derived from applying PCA (predictive or traditional) followed by UK to each of the 10 training sets first and predicting the scores at the left-out test locations. Therefore, minimizing this distance is equivalent to approximating $X$ via predictions of PC scores, which requires that the PC scores be predictable and that the loading vectors are not highly correlated.

4. Simulation studies
4.1. Simulated exposure analysis
In this section, we apply the methods that were discussed in Section 3 to simulated data. Our goal is to understand the main differences between the traditional and predictive versions of PCA and sparse PCA. In this and the following sections, we focus on the first three PCs. We use the proportion of the variance in the original data explained by the PC scores as a criterion to select the number of components in the analysis. With three PC scores, we can explain about 70–80% of the original variance in the simulated and real data sets in Section 5. The threshold values of 70–80% are often used as a rule of thumb in the PCA literature (see Miller (1998) and Kim
and Mueller (1978)). In our paper, three is also sufficiently small to be useful in presenting the material without creating excessively large tables. We note that using the amount of explained variance in the data with lower dimensional scores is not the only way to select the number of components in the analysis. For example, other commonly used methods include approaches based on the Kaiser criterion (Kaiser, 1960) and scree tests (Cattell, 1966). Various statistical and non-statistical methods and tests to find optimal numbers of components are discussed in detail in Peres-Neto et al. (2005), Bryant and Yarnold (1995) and Gorsuch (1983).

We first randomly choose 400 locations from the list of 970 monitor locations. At these locations, we have access to all GIS covariates. Using the notation from Section 3.3, let \( \tilde{Z} \) be the matrix of the GIS covariates and thin plate splines obtained by using the co-ordinates of the selected locations. Using \( \tilde{Z} \), we generate 19-dimensional multipollutant data \( X \), by simulating its columns \( X[, j] \), using linear models of the form

\[
X[, j] = \tilde{Z}\gamma_j + \epsilon_j,
\]

where \( j = 1, \ldots, 19 \), with \( \gamma_j \) fixed coefficient vectors and independently and identically distributed noise vectors \( \epsilon_j \) distributed according to a multivariate normal distribution with mean 0 and covariance matrix \( \sigma_j I \).

We choose the simulation parameters \( \gamma_j \) and \( \sigma_j \) so that some pollutants depend only on GIS covariates; some pollutants depend only on thin plate splines and some pollutants are just noise. In addition, we also generate some pollutants by using a combination of GIS covariates and splines. As can be seen from the model above, all of this is easily achieved by making some of the \( \gamma_j \)s close to 0. Finally, we also introduce correlation between pollutants by forcing some of the \( \gamma_j \)s to be close to each other for different pollutants to make the simulated data reflect some of the correlation structure that is observed in the actual monitoring data.

Additionally, we consider two different scenarios. Under the first scenario, we choose the parameters in the data-generating model so that most of the simulated pollutants are predictable by using UK. Under the second scenario, we make sure that most of the pollutants cannot be predicted well by using a spatial model. We note that this can be achieved by varying the values of the coefficients \( \gamma_j \) and the variances of the error terms \( \sigma_j \): increasing the values of \( \gamma_j \) corresponding to the GIS covariates increases the predictability of the pollutants by using the geographical features of the locations, which in turn increases the predictability by using UK since the GIS variables are a part of the kriging model; increasing the values of \( \gamma_j \) corresponding to the thin plate splines results in smoother pollutant surfaces, which increases predictability by using UK since kriging captures spatial variability; and, finally, adding random noise with larger variances \( \sigma_j \) decreases the predictability by making it difficult to detect the true signal.

At each simulation, we have 400 locations with multipollutant data. In assessing the performance of different models, we split the simulated data into two parts with 300 and 100 locations. We apply the methods to data from 300 locations. Then, we predict the pollutants and scores at the remaining 100 location that were not used in building the model. We compare these predictions with known values of the pollutants and scores. We note here that the known scores are defined by projecting the observed values of pollutants at 100 locations to fixed loadings obtained from applying PCA to data from 300 monitors.

Tables 2 and 3 summarize results from the simulation study under different scenarios. In Tables 2 and 3, simulated examples are run 100 times and all summaries are based on averages. The ordering of PC scores is from the highest predictable to the lowest predictable score in terms of \( R^2 \). In other words, at each simulation, the best predictable PC score is called the first score, the second best predictable score is called the second score etc. Here, we define the absolute correlation as an average of three absolute correlations between all three scores, and sparseness
Table 2. Summary: simulation results for data with pollutants with high predictability†

| Method                        | Predictability of scores ($R^2$) | Absolute correlation (average) | Sparseness (%) |
|-------------------------------|----------------------------------|--------------------------------|----------------|
|                               | PC 1    | PC 2    | PC 3    |                               |                |
| Without penalty               |         |         |         |                               |                |
| Traditional PCA               | 0.96    | 0.87    | 0.70    | 0.04                           | 0.00           |
| Predictive PCA                | 0.96    | 0.88    | 0.71    | 0.05                           | 0.00           |
| Maximized scores, with penalty|         |         |         |                               |                |
| Traditional PCA               | 0.97    | 0.91    | 0.84    | 0.38                           | 35.22          |
| Predictive PCA                | 0.97    | 0.92    | 0.84    | 0.43                           | 35.53          |
| Maximized pollutants, with penalty|    |         |         |                               |                |
| Traditional PCA               | 0.96    | 0.87    | 0.66    | 0.11                           | 19.96          |
| Predictive PCA                | 0.96    | 0.87    | 0.74    | 0.18                           | 20.04          |

†$R^2$ is obtained by using UK. Absolute correlation is defined as the average of three absolute correlations between the scores. Sparseness is the fraction of zero-valued loadings. Here, ‘maximized scores’ means that the penalty was selected to maximize the predictability of the PC scores, and ‘maximized pollutants’ means that the penalty was selected to maximize the predictability of the pollutants.

Table 3. Summary: simulation results for data with pollutants with low predictability†

| Method                        | Predictability of scores ($R^2$) | Absolute correlation (average) | Sparseness (%) |
|-------------------------------|----------------------------------|--------------------------------|----------------|
|                               | PC 1    | PC 2    | PC 3    |                               |                |
| Without penalty               |         |         |         |                               |                |
| Traditional PCA               | 0.76    | 0.60    | 0.27    | 0.02                           | 0.00           |
| Predictive PCA                | 0.85    | 0.76    | 0.56    | 0.08                           | 0.00           |
| Maximized scores, with penalty|         |         |         |                               |                |
| Traditional PCA               | 0.93    | 0.79    | 0.28    | 0.31                           | 52.47          |
| Predictive PCA                | 0.95    | 0.90    | 0.83    | 0.42                           | 69.96          |
| Maximized pollutants, with penalty|    |         |         |                               |                |
| Traditional PCA               | 0.79    | 0.66    | 0.28    | 0.08                           | 19.14          |
| Predictive PCA                | 0.88    | 0.80    | 0.61    | 0.13                           | 25.69          |

†$R^2$ is obtained by using UK. Absolute correlation is defined as the average of three absolute correlations between the scores. Sparseness is the fraction of zero-valued loadings. Here, ‘maximized scores’ means that the penalty was selected to maximize the predictability of the PC scores, and ‘maximized pollutants’ means that the penalty was selected to maximize the predictability of the pollutants.

is the fraction of loadings equal to 0. Here, since correlations with different signs can result in an average close to 0, looking at the average of absolute correlations is a more informative summary of dependence in scores.

Table 2 presents the first simulation scenario where all pollutants are highly predictable by using a spatial model. We see that traditional and predictive PCA perform similarly when there is no penalty and when the sparseness penalty parameters are chosen to maximize predictability of
the scores. However, when the penalty parameters are chosen to maximize predictability of the pollutants, we see that the predictive method increases the $R^2$ for the third PC score compared with traditional sparse PCA.

Table 3 presents the second simulation scenario where only some pollutants are highly predictable by using a spatial model. We expect to see more of a benefit from predictive (sparse) PCA in this setting, and indeed the advantage is more pronounced regardless of how the penalty is chosen. With no penalty, predictive PCA increases predictability of the PC scores without increasing the correlations between the scores. When the penalty is chosen to maximize predictability of scores, the average correlation between the scores tends to be undesirably high. When the penalty is selected to maximize predictability of pollutants, both the traditional and the predictive sparse PCA methods result in relatively small correlations between its scores, but prediction accuracy is better for predictive sparse PCA.

4.2. Simulated health analysis

In this section, we provide a simulated example analysis of the health effects of exposure to multiple pollutants by using traditional and predictive sparse PCA. We follow the idea laid out in Section 4.1 and assign $N_1 = 300$ locations as monitor locations and $N_2 = 7075$ locations for our cohort subjects. We select these locations from the full list of 7375 available EPA monitoring sites for convenience so that we do not have to recalculate the geographical covariates. We note that, in Section 4.1, we used the EPA AQS sites without missing data, but this does not affect the results of this analysis. We also note that we intentionally make the number of monitor locations small (300), compared with the number of subject locations in our simulated cohort data (7075), as this is usually so in real cohort studies where sizes of cohorts are much larger than the number of the monitoring sites. We generate the true exposure surface measurements at $N_1 + N_2$ locations by using the methods that were described in Section 4.1. For brevity, however, we consider only the second scenario where we made most of the pollutants spatially unpredictable. From Section 4.1, we know that the predictive PCA results in PC scores that are more predictable than the scores from traditional PCA. Our primary goal here is to see how this affects the estimated health effects in the health analysis. Under the assumption that our simulated subjects live at the selected $N_2$ addresses, for each subject, we generate five subject-specific covariates. These are $A$, age, $R$, race, $I$, income, $E$, education, and $S$, smoking status. We assume that age is a uniform continuous variable between 30 and 80, and race, income, education and smoking status are all categorical variables with four categories with realistic proportions. We assume that we are interested in two health end points, $Y_1$ and $Y_2$. For example, the first end point $Y_1$ could be a cardio-vascular health end point; the second end point $Y_2$ could be a respiratory end point. In the true health model, we use the following models to generate the health end points:

\begin{align}
Y_1 &= 0.35 P_1 + 0.5 A + 2 R + 0.5 I - 1 E + 1 S + \text{error}, \\
Y_2 &= 0.6 P_5 + 0.6 P_8 + 0.4 A + 0.2 S + \text{error},
\end{align}

where we assume that both error terms are independent and normally distributed with standard deviation equal to 10. Here, $P_1$, $P_5$ and $P_8$ are the simulated pollutants.

Finally, assuming that the true pollutant measurements are available only at the monitor locations, we want to study the effects of multipollutant exposure on $Y_1$ and $Y_2$ by using traditional and predictive (sparse) PCA. When we again apply PCA approaches to the simulated monitoring data, we look at only the first three PCs. As noted before, with three components we can explain
approximately 70% of the original variability in the data. There are various strategies that could be used to select the number of components to consider in the analysis. Selecting the number of components by using these approaches would be worthwhile but is beyond the scope of this paper. For each component, we choose the penalty parameters to maximize predictability of the pollutants. We see from the previous section that this yields scores that are more predictable and less intercorrelated than those that result from the other methods to choose the penalty parameters in both traditional and predictive PCA. Once predictions of the PC scores have been obtained, we fit health effects regression models with these as the main predictors. We also include all the subject-specific covariates. In fitting these models, we do not expect to recover the exact true parameters and variables that are used to generate the data (equations (11) and (12)). In contrast with variable selection methods, both the predictive and the traditional PCA are unsupervised methods to extract mixtures of pollutants that occur together. When we study the effect of these mixtures on our health end points by using predictions of the scores, we want to determine whether the scores are significantly associated with the end points and whether these mixtures contain the true predictors that we used in generating the data.

Results from one run of this analysis are given in Tables 4 and 5. First, Table 4 shows the individual prediction $R^2$ for each pollutant and the loadings from the two PCA approaches. These loadings can be used to interpret the mixtures that are derived from PCA. Here, the prediction $R^2$s for the PC scores from traditional sparse PCA were 0.11, 0.86 and 0.77; the prediction $R^2$s for PC scores from predictive sparse PCA were 0.89, 0.74 and 0.81. We note that, when we calculate the $R^2$s, we use the predicted and the true values of the scores. These results are consistent with the results that were shown in Section 4.1. We also note that pollutants 1 and 5 show up with fairly large loadings in the predictive PCs whereas the contributions of pollutants 1 and 5 to the PC scores are small in the traditional PCA. This is because pollutants 1 and 5 are relatively spatially predictable compared with the other pollutants (see Table 4).

### Table 4. Loadings of PC scores

| Pollutant | $R^2$ | Loadings for traditional sparse PCA | Loadings for predictive sparse PCA |
|-----------|-------|-----------------------------------|-----------------------------------|
|           |       | PC 1  | PC 2  | PC 3  | PC 1  | PC 2  | PC 3  |
| 1         | 0.97  | 0.14  | 0.18  | 0     | 0.25  | 0.85  | 0     |
| 2         | 0.19  | 0.5   | 0     | 0     | 0     | 0.15  | 0     |
| 3         | 0.07  | 0     | 0     | 0     | 0     | 0     | 0.09  |
| 4         | 0.19  | 0.5   | 0     | 0     | 0     | 0.16  | 0     |
| 5         | 0.41  | 0.08  | 0     | 0     | 0     | 0.35  | 0     |
| 6         | 0.41  | 0     | 0     | 0.5   | 0     | 0     | 0.44  |
| 7         | 0.94  | 0     | −0.15 | 0.59  | −0.16 | 0     | 0.7   |
| 8         | 0.06  | 0     | 0     | 0.38  | 0     | 0     | 0.33  |
| 9         | 0.18  | 0.48  | 0     | 0     | 0     | 0.15  | 0     |
| 10        | 0.26  | 0.08  | 0     | 0.51  | 0     | 0     | 0.44  |
| 11        | 0.82  | 0     | 0.63  | 0     | 0.68  | 0     | 0     |
| 12        | 0.4   | 0     | 0.6   | 0     | 0.59  | 0     | 0.02  |
| 13        | 0.39  | 0     | 0.22  | 0     | 0.17  | 0     | −0.06 |
| 14        | 0.19  | 0.5   | 0     | 0     | 0     | 0.15  | 0     |
| 15        | 0     | 0     | 0.29  | 0     | 0.21  | 0     | 0     |
| 16        | 0.11  | −0.02 | 0     | 0     | 0     | −0.09 | 0     |
| 17        | 0.23  | −0.02 | −0.23 | 0     | −0.17 | −0.24 | −0.01 |
Table 5. Inference for the simulated health model coefficient

| End point PC | Results for traditional sparse PCA | Results for predictive sparse PCA |
|--------------|------------------------------------|----------------------------------|
|              | beta pval                           | beta pval                        |
| 1 1          | 0.29 0.08                           | 0.01 0.89                        |
| 2            | 0.04 0.66                           | 0.22 0.04                        |
| 3            | 0.02 0.80                           | 0.02 0.86                        |
| 2 1          | 0.27 0.11                           | 0.04 0.72                        |
| 2            | 0.06 0.53                           | 0.20 0.07                        |
| 3            | 0.22 0.02                           | 0.21 0.03                        |

Table 5 gives the results of the health analyses. For $Y_1$, we know that pollutant 1 is a predictor of $Y_1$ with true effect size equal to 0.35. The effect size estimate of the predictive PC 2 is significant (at the significance level of 0.05) and equal to 0.22. From Table 4, we see that pollutant 1 is an important part of this mixture (with the loading equal to 0.85). This can be interpreted as follows: when we increase PC 2 by 1 unit, the estimated change in $Y_1$ is equal to 0.22. However, since a 1-unit increase in PC 2 corresponds to a 0.85-unit increase in pollutant 1, the true change in $Y_1$ should be around $0.85 \times 0.35 = 0.297$. The effect size estimate of the traditional PC 1 is marginally significant and equal to 0.29. Pollutant 1 is also a part of the traditional PC 1 with the loading equal to 0.14. This means that, for a 1-unit increase in PC 1 in traditional PCA, we increase pollutant 1 by 0.14 units, which results in an estimated increase in $Y_1$ equal to 0.29, whereas the true increase is $0.14 \times 0.35 = 0.049$.

For $Y_2$, the true predictors are pollutant 5 and pollutant 8 with both of the effects sizes equal to 0.6. Here, the effect size estimate of the predictive PC 3 is significant and equal to 0.21. This mixture contains pollutant 8 with the loading equal to 0.33. This means that, when we increase the predictive PC 3 by 1 unit, we increase pollutant 8 by 0.33 units. This results in an estimated increase of $Y_2$ in 0.21 units, whereas the true increase in $Y_2$ is equal to $0.33 \times 0.6 = 0.198$. The effect size estimate of the predictive PC 2 is marginally significant and equal to 0.20. The predictive PC 2 contains pollutant 5 with the loading equal to 0.35. Again, the effect size equal to 0.20 can be interpreted as the change in $Y_2$ per 1-unit increase in the predictive PC 2, which is equivalent to a 0.35-unit change in pollutant 5. The true change in $Y_2$ that corresponds to a 0.35-unit change in pollutant 5 is equal to $0.35 \times 0.6 = 0.21$. Finally, the traditional PC 2 is a mixture that contains pollutant 8 (loading equal to 0.38) and is marginally associated with $Y_2$ with the effect size estimate equal to 0.27. This means that, when we increase the traditional PC 2 by 1 unit, we increase pollutant 8 by 0.38 units, and this results in an estimated increase in $Y_2$ of 0.27 units. The true change in $Y_2$ per 0.38-unit increase in pollutant 8 is equal to $0.38 \times 0.6 = 0.228$.

These results show that we obtain a more complete picture of the true health effects from the predictive PCA than we do from the traditional PCA when the goal is to study the effects of principal mixtures on the health end points. We observe this for both $Y_1$, where the signal for the score from the predictive PCA was stronger, and for $Y_2$, where the traditional PCA did not capture the effect of pollutant 5. It is also easy to see that the effect size estimates for the predictive PCs are more accurate after adjusting for the contributions of the individual pollutants compared with the effect sizes estimated for the traditional PCs. We explain this by the fact that the health effects in the true data-generating mechanism are attributable to pollutants 1, 5 and 8, and these have different loadings in predictive and traditional PCA. We note that,
when we ran the simulated health analyses multiple times, the predictive PCA performed better than the traditional PCA consistently (the results are not shown). We note, however, that the interpretation of the effects of individual pollutants is still a challenge for both methods since PCA cannot tell us specifically which pollutant is responsible for the observed effect, even though we think that the predictive PCA represents an improvement in that it provides more insight into the truth than traditional PCA does.

5. Application to multipollutant monitoring data

We now apply traditional and predictive (sparse) PCA to the multipollutant monitoring data that were described in Section 2. As an exploratory analysis, in Section 2, Table 1, we observed that, whereas some pollutants are highly predictable, there are others that have poor cross-validated $R^2$. This suggests that PC scores obtained from projecting the multipollutant data onto traditional PC loadings may not have good spatial predictability.

First, we apply traditional and predictive PCA without a penalty. Fig. 1 shows scatter plots obtained from tenfold cross-validation. From Fig. 1, we conclude that the cross-validated $R^2$ for the PC 1 and PC 2 scores from the two methods are very similar. This is also clear from looking at the bar plot of the corresponding loadings in Fig. 2, where we see that loadings from traditional and predictive PCA for the first two PC scores are close to each other. We note, however, that predictive PCA increases the predictability of the PC 3 scores from 0.45 to 0.68.

From Fig. 2, where pollutants are ordered from the smallest individual $R^2$ to the highest, we see that the different PCA methods have different loadings for PC 3.

![Fig. 1. Scatter plots of the predictions of the scores without a penalty (obtained by using cross-validation and UK):](image)

(a) traditional PCA, PC 1 $R^2 = 0.88$; (b) traditional PCA, PC 1 $R^2 = 0.58$; (c) traditional PCA, PC 1 $R^2 = 0.45$; (d) predictive PCA, PC 1 $R^2 = 0.89$; (e) predictive PCA, PC 1 $R^2 = 0.59$; (f) predictive PCA, PC 1 $R^2 = 0.68$
Now we apply both traditional and predictive PCA with a penalty to induce sparseness. A summary of the results from this analysis is given in Table 6. If we compare Table 6 with Tables 2 and 3, we note that Table 6 contains information about the exact correlations between the PC scores, whereas in Tables 2 and 3 we present only the average absolute correlations between the scores. We do this because, in simulated examples, the order of the PCs varies. In other words, the role of PC 1, PC 2 and PC 3 changes from simulation to simulation, making the interpretation...
Table 6. Summary of results†

| Method                  | Predictability of scores ($R^2$) | Absolute correlation (average) | Sparseness (%) |
|-------------------------|----------------------------------|--------------------------------|----------------|
|                         | PC 1    | PC 2    | PC 3    | PC 1 versus PC 2 | PC 1 versus PC 3 | PC 2 versus PC 3 |
| Without penalty         |         |         |         |                  |                  |                 |
| Traditional PCA         | 0.88    | 0.58    | 0.45    | 0.00             | 0.00             | 0.00             |
| Predictive PCA          | 0.89    | 0.59    | 0.67    | −0.05            | 0.03             | −0.04            |
| Maximized scores, with penalty |         |         |         |                  |                  |                 |
| Traditional PCA         | 0.92    | 0.88    | 0.64    | 0.96             | −0.14            | −0.08            |
| Predictive PCA          | 0.93    | 0.90    | 0.89    | 0.97             | 0.97             | 0.99             |
| Maximized pollutants, with penalty |         |         |         |                  |                  |                 |
| Traditional PCA         | 0.88    | 0.59    | 0.57    | 0.17             | −0.15            | −0.19            |
| Predictive PCA          | 0.89    | 0.59    | 0.66    | −0.04            | 0.01             | 0.01             |

†$R^2$ is obtained by using UK. Sparseness is the fraction zero valued. Here, ‘maximized scores’ means that the penalty was selected to maximize the predictability of the PC scores, and ‘maximized pollutants’ means that the penalty was selected to maximize the predictability of the pollutants.

of the average individual correlations more difficult and less informative. In Table 6, sparseness is again defined as the fraction of 0s in the PC loadings vector. Sparseness is equal to 0 for non-sparse PCA for both traditional and predictive PCA, as can be seen in the first part of Table 6. When the penalty is chosen to maximize predictability of the scores, we see that the cross-validated $R^2$s for all PC scores are high for both methods. In sparse predictive PCA, these $R^2$s are above 0.90. However, we note that both methods result in scores that are highly correlated.

When the penalty parameter is chosen to maximize predictability of the pollutants, Fig. 3 and Table 6 show that, whereas the PC 1 and PC 2 scores from the traditional and predictive versions of sparse PCA have similar $R^2$s, the predictability of the PC 3 score from the predictive method is higher than from the traditional method. The bar plots of the corresponding principal loadings for the predictive and traditional sparse PCA are given in Fig. 4. Here, we see that the noticeable differences in the loadings are again in PC 3. As expected, the loadings now are sparse. From these, it is also clear that the predictive sparse PCA results in lower correlations between scores when compared with the traditional method, which is consistent with the conclusions that were obtained in simulated examples in Section 4. Additionally, we see that traditional PCA has higher sparseness compared with predictive PCA in Table 6. However, we speculate that this is not a common occurrence as we did not observe the phenomenon that the sparseness was substantially different for the two methods in simulated examples. Looking at Fig. 4, we see that the difference in sparseness of the loadings in Table 6 is mostly because of the differences in the third principal loadings. We speculate that, although the loadings values for positive weights are similar, predictive PCA is adding elements to the mixture with negative loading values to increase the overall predictability of the final principal score. These elements are given loading values equal to 0 in traditional PCA. In Fig. 5, we present heat maps of the predictions of the scores from predictive sparse PCA. These heat maps show clearly how different mixtures derived from different scores vary spatially. When we apply a UK model to predict the sparse PC scores in Table 6, the estimated range for all the scores was around 600 km for both traditional and predictive sparse PCA. The estimate of the partial sill and nugget effect for the first principal
Fig. 3. Scatter plots of the predictions of the scores with a penalty to maximize predictability of the pollutants: (a) traditional PCA, PC 1 $R^2 = 0.88$; (b) traditional PCA, PC 1 $R^2 = 0.59$; (c) traditional PCA, PC 1 $R^2 = 0.57$; (d) predictive PCA, PC 1 $R^2 = 0.89$; (e) predictive PCA, PC 1 $R^2 = 0.59$; (f) predictive PCA, PC 1 $R^2 = 0.66$

score was 40 and 0.92 for traditional PCA, and 37 and 0.82 for predictive PCA; for the second principal score, it was 79 and 0.76 for traditional PCA, and it was 73 and 0.81 for predictive PCA; and, for the third principal score, 69 and 0.21 for traditional PCA, and 65 and 0.19 for predictive PCA.

In future work, we plan to utilize predictive sparse PCA to estimate multipollutant health effects in cohort studies where single-pollutant associations have been previously observed (see Chan (2015)). This is beyond the scope of the present paper, as we shall need to deal carefully with additional statistical and scientific issues including exposure measurement error (Szpiro and Paciorek, 2013), preferential sampling of monitoring locations (Lee et al., 2015) and effect modifiers in the health model (Analitis et al., 2014; Delfino et al., 2014; Park et al., 2008).

6. Discussion

In this paper, we describe a novel predictive (sparse) PCA approach that can be used to obtain exposure estimates in air pollution cohort studies. The approach can be used with or without a penalty. In contrast with traditional PCA, we demonstrate that predictive PCA may lead to PC scores that can be predicted well by using geographical covariates and splines in the exposure prediction stage. We have also shown that penalty parameters in predictive sparse PCA can be selected so that the correlations between the resulting scores are small. On the basis of the simulated examples and data analysis, we observed that, when the penalty parameter is selected to maximize predictability of the principal scores, the resulting loadings are nearly orthogonal and the scores are almost uncorrelated. We also find in simulated health effect studies that incorporating predictive sparse PCA rather than traditional sparse PCA in the exposure modelling leads to improved estimation of the true underlying associations.
In our data analysis, we apply the traditional and predictive (sparse) PCA approaches to multipollutant data from the EPA’s regulatory monitors from 2010. In Table 1, we present a summary of how individual pollutants can be predicted by using a UK model with GIS covariates. In Table 6 and Figs 1 and 3, we show how the different mixtures obtained from the PC scores can be predicted. In Figs 2 and 4, we present the component loadings of the PC scores.
Fig. 5. Heat maps based on the predictions of the scores from predictive sparse PCA: (a) predictions of PC 1, $R^2 = 0.89$; (b) predictions of PC 2, $R^2 = 0.59$; (c) predictions of PC 3, $R^2 = 0.66$
obtained by using the predictive and traditional (sparse) PCA. Finally, Fig. 5 provides further insights on how the PC scores vary spatially across the USA.

In predictive (sparse) PCA, we do not use the health data when applying the algorithm to multipollutant data. Therefore, we may think of our predictive PCA as an unsupervised dimension reduction approach. However, unlike traditional (sparse) PCA, the approach is ‘supervised’ by geographical data to increase predictability of the PC scores. Alternatively, one can potentially develop a dimension reduction approach that is fully supervised by both geographical and health data. This, for example, can be undertaken by using a Bayesian framework by introducing lower dimensional latent variables which govern the joint distribution between pollutants and the health variable. In joint models, in other words, the first-stage exposure model is a part of a unified model for the exposure and health data (see Sinha et al. (2010) in nutritional epidemiology and Gryparis et al. (2009) in air pollution epidemiology). We argue against such a joint approach for four reasons. First, using a two-step approach in air pollution studies offers a way of ‘cutting feedback’ that may drastically influence estimation and prediction in the exposure model in unexpected ways, especially when the prediction model is misspecified. Second, when simpler two-step approaches are used, performing a sensitivity analysis using different exposure models in the first step can be performed easily to understand the effects of these models. Third, using two-step approaches allows use of the same predicted exposures across studies with multiple health end points, which makes such approaches more practical and scientifically desirable. Fourth, an inherent disadvantage of a unified approach is that it requires specifying high dimensional joint distributions (and the associated priors in a Bayesian framework). By simple virtue of the high dimensionality, this task is computationally much more expensive than a two-stage approach, particularly in the context of air pollution studies where there are fewer than 1000 monitors that are used in the prediction step and a number of subjects that are in the tens of thousands in the health analysis step. For more details on discussion between unified and two-step approaches, readers are referred to Szpiro and Paciorek (2013).

In the literature, when both response variables (as in our multipollutant data) and covariates (as in our GIS covariates and thin plate splines) are multivariate, methods known as two-block partial least squares can be used for analysing the covariance between the two sets of variables (see Sampson et al. (1989), Streissguth et al. (1993) and Rohlf and Corti (2000)). These methods are similar to the well-known method of canonical correlation analysis (Jackson, 2005). Even though two-block partial least squares is related to our PCA approach, in predictive PCA, our goal is quite distinct: we want to improve predictability of the PC scores by using the same GIS covariates and splines and to explain most of the variability in the multipollutant data. In contrast with our goal, in two-block partial least squares and canonical correlation analysis, the two sets of variables are treated symmetrically in an attempt to find possible relationships between the underlying latent variables derived from the sets. These approaches are useful for exploring the combinations of variables in the two sets that account most for whatever covariation there is between them.

In this paper, we considered predictive PCA both with and without a lasso-type penalty on the loading vectors to induce sparseness. In principle, we could also add an additional penalty parameter to the $\alpha$-coefficients in equation (8), which could be particularly useful for incorporating richer spatial prediction models by allowing for a way to eliminate insignificant geographical covariates. This may, however, substantially slow down our iterative optimization algorithm described in Section 3.1. As future research, a related idea is to exploit the link between random effects and penalization (see Ruppert et al. (2003) and Verbyla et al. (1999)) to develop an algorithm that uses a random-effects model for the $\alpha$. Another promising avenue for future research is to develop a predictive dimension reduction approach that finds all principal
loadings and scores in one step. In contrast with our sequential approach, this would facilitate borrowing information between scores potentially to improve predictability further.

When the variability of high dimensional correlated variables is of interest in the context of reducing the dimensionality of the data, factor analysis is another widely used statistical technique (see Harman (1967)). In factor analysis, the original variables are defined as linear combinations of the unobserved variables, called factors, with error terms by using a statistical model, whereas in PCA the component scores are calculated as linear combinations of the original variables. Therefore, the main goal in factor analysis is to explain the correlations between the variables with these lower dimensional factors to understand what constructs underlie the data. In PCA our goal is to extract the lower dimensional scores as important independent composite variables in an interpretable way for subsequent use in health analyses.

In our approach, we achieve sparsity by adding a penalty term to the optimization. This allows an easy way to incorporate the GIS covariates in a new optimization in predictive PCA. In both PCA and factor analysis, to make the interpretation of the scores or factors easier, one can also achieve sparsity via rotation of the loadings. Two main types of rotation can be used: orthogonal, where orthogonality of the loadings is preserved, and oblique, where the new loadings are not necessarily required to be orthogonal to each other. The most commonly used technique based on rotation of the loadings is called varimax rotation (see Kaiser (1958)). Similar to sparse PCA, varimax rotation can result in scores or factors that have a small number of large loadings and a large number of zero loadings. We note that it may also be possible to develop an alternative predictive sparse PCA approach based on rotation techniques.

Acknowledgements

This paper was made possible by US EPA grant RD-83479601-0 and National Institutes of Health–National Institute of Environment Health Sciences grants R01-ES020871-03 and R21-ES024894. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA or National Institutes of Health–National Institute of Environment Health Sciences. Further, the US EPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Reference

Abdi, H. (2003) Partial least squares regression (pls-regression). In Encyclopedia for Research Methods for the Social Sciences (eds M. Lewis-Beck, A. Bryman and T. Futing), pp. 792–795. Thousand Oaks: Sage.

Analitis, A., Michelozzi, P., D’Ippoliti, D., de Donato, F., Menne, B., Matthies, F., Atkinson, R. W., Iñiguez, C., Basagaña, X., Schneider, A., Lefranc, A., Paldy, A., Bisanti, L. and Katsouyanni, K. (2014) Effects of heat waves on mortality: effect modification and confounding by air pollutants. Epidemiology, 25, 15–22.

Anderson, T. W. (2003) An Introduction to Multivariate Statistical Analysis. New York: Wiley.

Bell, M. L. and Davis, D. L. (2001) Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. Environ. Hlth Perspect., 109, suppl. 3, 389–394.

Bergen, S., Sheppard, L., Sampson, P. D., Kim, S.-Y., Richards, M., Vedal, S., Kaufman, J. D. and Szpiro, A. A. (2013) A national prediction model for PM2.5 component exposures and measurement error–corrected health effect inference. Environ. Hlth Perspect., 121, 1017–1025.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M. and Brunekreef, B. (2003) Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology, 14, 228–239.

Brook, J. R., Burnett, R. T., Dann, T. F., Cakmak, S., Goldberg, M. S., Fan, X. and Wheeler, A. J. (2007) Further interpretation of the acute effect of nitrogen dioxide observed in Canadian time-series studies. J. Expos. Sci. Environ. Epidem., 17, suppl., S36–S44.

Bryant, F. B. and Yarnold, P. R. (1995) Principal components analysis and exploratory and confirmatory factor analysis. In Reading and Understanding Multivariate Statistics (eds L. G. Grimm and P. R. Yarnold), pp. 99–136. Washington DC: American Psychological Association.

Cattell, R. B. (1966) The scree test for the number of factors. Multiv. Behav. Res., 1, 245–276.
Chan, S. H., Van Hee, V. C., Bergen, S., Szpiro, A. A., De Roo, L. A., London, S. J., Marshall, J. D., Kaufman, J. D. and Sandler, D. P. (2015) Long-term air pollution exposure and blood pressure in the Sister Study. *Environ. Hlth Perspect.*, 123, 951–958.

Crouse, D. L., Goldberg, M. S., Ross, N. A., Chen, H. and Labrèche, F. (2010) Post-menopausal breast cancer is associated with exposure to traffic-related air pollution in Montreal, Canada: a case–control study. *Environ. Hlth Perspect.*, 118, 1578–1583.

Delfino, R. J., Wu, J., Tjoa, T., Gullesserian, S. K., Nickerson, B. and Gillen, D. L. (2014) Asthma morbidity and ambient air pollution: effect modification by residential traffic-related air pollution. *Epidemiology*, 25, 48–57.

Dubrule, O. (1984) Comparing splines and kriging. *Comput. Geosci.*, 10, 327–338.

Eldred, R. A., Cahill, T. A. and Pitchford, M. (1988) IMPROVE: a new remote area particulate monitoring system for visibility studies. In *Proc. 81st A. Meet. Air Pollution Control Association*, Dallas: Air Pollution Control Association.

Environmental Protection Agency (2009) Integrated science assessment for particulate matter. US Environmental Protection Agency, Washington DC.

Gorsuch, R. L. (1983) *Factor Analysis*, 2nd edn. Hillsdale: Erlbaum.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2009) Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10, 258–274.

Harman, H. H. (1967) *Modern Factor Analysis*, 2nd edn. Chicago: University of Chicago Press.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. and Briggs, D. (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmosph. Environ.*, 42, 7561–7578.

Hutchinson, M. and Gessler, P. (1994) Splines—more than just a smooth interpolator. *Geoderma*, 62, 45–67.

Jackson, J. E. (2005) *A User’s Guide to Principal Components*. Hoboken: Wiley.

Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. and Thun, M. J. (2005) Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, 16, 727–736.

Jolliffe, I. T. (1986) *Principal Component Analysis*. New York: Springer.

Kaiser, H. F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.

Kaiser, H. F. (1960) The application of electronic computers to factor analysis. *Educ. Psychol. Measurmt.*, 20, 141–151.

Kim, J.-O. and Mueller, C. W. (1978) *Factor Analysis: Statistical Methods and Practical Issues*. New York: Sage.

Kim, S.-Y., Sheppard, L., and Kim, H. (2009) Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*, 20, 442–450.

Känzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J. and Hodis, H. N. (2005) Ambient air pollution and atherosclerosis in Los Angeles. *Environ. Hlth Perspect.*, 113, 201–206.

Lee, A., Szpiro, A., Kim, S. and Sheppard, L. (2015) Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, 26, 255–267.

Logan, W. (1953) Mortality in the London fog incident, 1952. *Lancet*, 261, 336–338.

Matheron, G. (1981) Splines and kriging: their formal equivalence. *Down-to-earth Statistics: Solutions Looking for Geological Problems* (ed. D. Merriam), vol. 8, pp. 77–95. Syracuse: Syracuse University Geology Contributions.

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, E. L., Oron, A. P., Larson, T., Liu, L.-J. S. and Kaufman, J. D. (2011) Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmosph. Environ.*, 45, 4412–4420.

Miller, G. J. (1998) *Handbook of Research Methods in Public Administration*. Boca Raton: CRC Press.

Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L. and Kaufman, J. D. (2007) Long-term exposure to air pollution and incidence of cardiovascular events in women. *New Engl. J. Med.*, 356, 447–458.

Nemery, B., Hoet, P. H. and Nemmar, A. (2001) The Meuse Valley fog of 1930: an air pollution disaster. *Lancet*, 357, 704–708.

Park, S. K., O’Neill, M. S., Vokonas, P. S., Sparrow, D., Wright, R. O., Coull, B., Nie, H., Hu, H. and Schwartz, J. (2008) Air pollution and heart rate variability: effect modification by chronic lead exposure. *Epidemiology*, 19, 111–120.

Peres-Neto, P. R., Jackson, D. A. and Somers, K. M. (2005) How many principal components?: stopping rules for determining the number of non-trivial axes revisited. *Computat. Statist. Data Anal.*, 49, 974–997.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K. and Thurston, G. D. (2002) Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Ass.*, 287, 1132–1141.

Pope III, C. A. and Dockery, D. W. (2006) Health effects of fine particulate air pollution: lines that connect. *J. Air Waste Mangmnt Ass.*, 56, 709–742.

Rohlf, F. J. and Corti, M. (2000) Use of two-block partial least-squares to study covariation in shape. *Syst. Biol.*, 49, 740–753.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I. and Zeger, S. L. (2000) Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *New Engl. J. Med.*, 343, 1742–1749.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V. and Kaufman, J. D. (2013) A regionalized national universal kriging model using partial least squares regression for estimating annual pm 2.5 concentrations in epidemiology. *Atmosph. Environ.*, 75, 383–392.

Sampson, P. D., Streissguth, A. P., Barr, H. M. and Bookstein, F. L. (1989) Neurobehavioral effects of prenatal alcohol: Part ii, partial least squares analysis. *Neurotoxicol. Teratol.*, 11, 477–491.

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J. and Kaufman, J. D. (2011) Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmosph. Environ.*, 45, 6593–6606.

Shen, H. and Huang, J. Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.*, 99, 1015–1034.

Sinha, S., Mallick, B. K., Kipnis, V. and Carroll, R. J. (2010) Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, 66, 444–454.

Streissguth, A. P., Bookstein, F. L., Sampson, P. D. and Barr, H. M. (1993) *The Enduring Effects of Prenatal Alcohol Exposure on Child Development: Birth through Seven Years, a Partial Least Squares Solution*. Ann Arbor: University of Michigan Press.

Szpiro, A. A. and Paciorek, C. J. (2013) Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, 24, 501–517.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58, 267–288.

Vedal, S., Kaufman, J., Larson, T., Sampson, P., Sheppard, L., Simpson, C., Szpiro, A., McDonald, J., Lund, A. and Campen, M. (2012) University of Washington/Lovelace Respiratory Research Institute National Particle Component Toxicity (NPACT) Initiative: integrated epidemiological and toxicological cardiovascular studies to identify toxic components and sources of fine particulate matter (draft). Health Effects Institute, Boston.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Appl. Statist.*, 48, 269–311.

Wood, S. N. (2003) Thin plate regression splines. *J. R. Statist. Soc. B*, 65, 95–114.