Probabilistic predictive principal component analysis for spatially misaligned and high-dimensional air pollution data with missing observations

Phuong T. Vu[1], **Timothy V. Larson**[2], **Adam A. Szpiro**[1]

[1]Department of Biostatistics, University of Washington

[2]Department of Civil & Environmental Engineering, University of Washington

Abstract

Accurate predictions of pollutant concentrations at new locations are often of interest in air pollution studies on fine particulate matters ($PM_{2.5}$), in which data is usually not measured at all study locations. $PM_{2.5}$ is also a mixture of many different chemical components. Principal component analysis (PCA) can be incorporated to obtain lower-dimensional representative scores of such multi-pollutant data. Spatial prediction can then be used to estimate these scores at new locations. Recently developed predictive PCA modifies the traditional PCA algorithm to obtain scores with spatial structures that can be well predicted at unmeasured locations. However, these approaches require complete data, whereas multi-pollutant data tends to have complex missing patterns in practice. We propose probabilistic versions of predictive PCA which allow for flexible model-based imputation that can account for spatial information and subsequently improve the overall predictive performance.

Keywords

air pollution; multi-pollutant analysis; missing data; dimension reduction

1 Introduction

In recent years, there has been a growing interest in studying the role and health impact of $PM_{2.5}$, which is fine particulate matter with aerodynamic diameter less than 2.5 μm (Brook et al., 2004). $PM_{2.5}$ is a complex mixture of many components, and its chemical profile may vary drastically across time and space (Brook et al., 2004; Bell et al., 2007; Dominici et al., 2010). Obtaining a lower-dimensional representation of $PM_{2.5}$ multi-pollutant data is often necessary, as including many correlated pollutants in a statistical model is problematic.

Corresponding author: Phuong T. Vu, Department of Biostatistics, University of Washington, F-600, Health Sciences Building, Box 357232, Seattle, WA 98195-7232. phuongvu@uw.edu.

Principal component analysis (PCA) (Jolliffe, 1986) is an unsupervised dimension reduction technique that has gained popularity in multi-pollutant analysis (Dominici et al., 2003).

Examples of environmental studies utilizing $PM_{2.5}$ data include studies on the associations between various health outcomes and long-term (Pope III et al., 2002; Künzli et al., 2005; Miller et al., 2007; Chan et al., 2015; Kaufman et al., 2016) or short-term (Gold et al., 2000; Tolbert et al., 2007; Pascal et al., 2014; Achilleos et al., 2017; Hsu et al., 2017; Tian et al., 2017) exposures to $PM_{2.5}$. For instance, Chan et al. (2015) found significant associations between long-term exposure to $PM_{2.5}$ and higher systolic blood pressure, pulse pressure, and mean arterial pressure in the Sister Study. Kaufman et al. (2016) showed evidence of strong association between ambient concentration of $PM_{2.5}$ and accelerated atherosclerosis in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. In a recent systematic review of epidemiological studies, Achilleos et al. (2017) found substantial increases in all-cause, cardiovascular, and respiratory mortalities due to acute exposure of $PM_{2.5}$. Many studies have suggested that the associations between $PM_{2.5}$ total mass and various health outcomes can be modified by some specific constituents or the overall chemical composition (Franklin et al., 2008; Bell et al., 2009; Krall et al., 2013; Zanobetti et al., 2014; Dai et al., 2014; Kioumourtzoglou et al., 2015; Wang et al., 2017; Keller et al., 2018).

In the United States, $PM_{2.5}$ studies often rely on data collected from regulatory monitoring networks managed by the Environmental Protection Agency (EPA). Unfortunately, for many pollution-health association studies, these fixed monitoring sites are usually not at the same locations where health outcomes are available. Such *spatial misalignment* motivates an exposure modeling stage in which a spatial prediction model, such as land-use regression or universal kriging, is often used to estimate the exposure at unmeasured locations where pollutant data is not observed (Brauer et al., 2003; Künzli et al., 2005; Crouse et al., 2010; Bergen et al., 2013; Chan et al., 2015).

Derivation of a lower-dimensional representation of $PM_{2.5}$ multivariate data prior to making these spatial predictions is necessary, as predicting chemically and spatially correlated pollutant surfaces is challenging and intractable in most cases. As PCA is capable of performing dimension reduction without meddling with the health outcomes, it can be easily integrated in the analysis of spatially-misaligned data. Using PCA, a lower-dimensional scores of the multi-pollutant data at monitoring locations can be obtained. These monitoring scores, along with geographic covariates, can then be used in a spatial prediction model to estimate the corresponding scores at unmeasured locations. However, PCA does not account for exogenous geographic information and spatial correlations across neighboring locations. Hence, PCA may produce scores that summarize the monitoring data well but are difficult to be predicted at unmeasured locations. A spatially predictive PCA algorithm (Jandarov et al., 2017) was developed to mitigate this issue by producing scores with spatial patterns that can be subsequently predicted well at new locations.

An additional challenge arises in practice where there is often a large amount of missing data, especially for multi-pollutant monitoring data. For example, not all $PM_{2.5}$ components are measured at all monitoring sites, either due to environmental considerations, logistic constraints or lack of resources. The missing patterns can sometimes be complex or spatially

informative. Neither traditional PCA nor predictive PCA is well-equipped to deal with missing data, and thus a separate imputation step is required prior to dimension reduction. Existing non-parametric imputation schemes, ranging from simple mean imputation to sophisticated matrix completion, do not account for external spatial information. They may therefore distort the underlying spatial structure in the original data even before the dimension reduction stage, and thus negatively impact the predictive performance in the final stage.

In this paper, our goal is to enhance the dimension reduction procedure under the presence of missing data by proposing a probabilistic framework in place of the deterministic algorithm of predictive PCA. Similar to Jandarov et al. (2017), our methods seek to produce principal components that can be well predicted at new locations. The added probabilistic assumptions allow for flexible model-based imputation that takes into account the embedded geographic and spatial information, and thus eliminates the need for a preprocessing stage with non-parametric imputation.

2   Motivating example

To illustrate the merit of our proposed methods, we use data collected nationally by the Air Quality System (AQS) network of monitors managed by the EPA. Measurements of annually averaged $PM_{2.5}$ total mass and its components are only collected at a few sub-networks of AQS. For consistency with previous related work (Keller et al., 2017; Jandarov et al., 2017), we choose to use the 2010 data from the Chemical Speciation Network (CSN), of which monitoring sites are located strategically in various urban areas. Data is available for 21 components of $PM_{2.5}$: elemental carbon (EC), organic carbon (OC), sulfate ion (SO_4^{2-}), nitrate ion (NO_3^-), aluminum (Al), arsenic (As), bromine (Br), cadmium (Cd), calcium (Ca), chromium (Cr), copper (Cu), iron (Fe), potassium (K), magnesium (MN), sodium (Na), sulfur (S), silicon (Si), selenium (Se), nickel (Ni), vanadium (V), and zinc (Zn).

Geographic covariates are obtained for all available sites through the Exposure Assessment Core Database by the MESA Air team at the University of Washington. Data on roughly 600 Geographic Information System (GIS) covariates are available, including distances from roads, distances from major pollution sources, land-use information, vegetation indices, etc. The specific sources and attributions of these geographic covariates are carefully described in Bergen et al. (2013).

Data for 2010 is available for 221 CSN sites, with only 130 of those sites having complete data on all 21 components. Overall the amount of missing data in 2010 is roughly 32.1%. Not only do we compare the predictive performances following the application of different PCA methods, but we also examine how different the chemical profiles are when considering only complete sites versus all available data. The data processing, analysis procedures, and results are discussed in Section 6.

3 Review of PCA and predictive PCA

3.1 Traditional PCA

We denote $X \in \mathbb{R}^{n \times p}$ as the exposure data with p pollutants observed at n monitoring sites with spatial coordinates $s_1, \ldots, s_n$. The exposure data X may contain missing elements as some pollutants are not measured at all monitoring site. Let r_i be a vector of k geographic covariates pertaining to the i-th monitoring site. Variables corresponding to locations where exposure data are of interest but not measured are distinguished by an asterisk, i.e. n^*, X^*, $s_1^*, \ldots, s_{n^*}^*, r_1^*, \ldots, r_{n^*}^*$.

The data of interest, X^*, is high-dimensional but inaccessible. If X^* were observed, dimension reduction could be applied directly to obtain a lower-dimensional representation $U^* \in \mathbb{R}^{n^* \times q}$ where $q < p$. Because of spatial misalignment, a spatial prediction model is required to estimate the unobserved exposures. Modeling highly correlated surfaces is challenging and inefficient given the final aim of recovering only the lower-dimensional U^*. Thus, a sensible modeling procedure under the presence of spatially misaligned multi-pollutant data with missing observations may consist of several steps: (1) imputation for missing data, (2) dimension reduction to derive scores at monitoring sites, and (3) spatial prediction to estimate corresponding scores at new locations. In this paper, we focus on dimension reduction using PCA, an unsupervised technique that is suitable for handling spatially-misaligned data.

Traditional PCA provides a mapping from the original p-dimensional exposure surface to a corresponding q-dimensional representation where $X \approx UV^\top$ for $q < p$. We refer to the orthogonal columns of $V \in \mathbb{R}^{p \times q}$ as the loadings or principal directions. The columns of $\mathbb{R}^{n \times q}$, $\{u_1, \ldots, u_q\}$, are the principal component (PC) scores. These PC scores can be thought of as linear combinations of the original features of X. These newly transformed variables are considered uncorrelated due to orthogonality of the loadings, which is an attractive feature of PCA. The PCA algorithm is also optimal in the sense that the derived PC scores are conveniently ordered by the amount of variability explained in X.

3.2 Spatially predictive PCA

While PCA provides a unique solution in the reduced dimensions, the algorithm can be reformulated into a series of biconvex optimization problems, in which the loading and corresponding score of each PC can be solved in an iterative fashion (Shen and Huang, 2008),

$$\min_{u,v} \|X - uv^\top\|_F^2 \quad \text{s.t. } \|v\|_2 = 1.$$

Utilizing such optimization framework, Jandarov et al. (2017) develop a spatially predictive PCA algorithm (PredPCA hereafter) by directly incorporating spatial information in the objective function:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{v}} \left\| X - \left(\frac{Z\boldsymbol{\alpha}}{\|Z\boldsymbol{\alpha}\|_2} \right) \boldsymbol{v}^\mathsf{T} \right\|_F^2,$$

where $Z = \begin{bmatrix} R & \widetilde{R} \end{bmatrix}$, in which $R \in \mathbb{R}^{n \times k}$ contains k GIS covariates, and $\widetilde{R} \in \mathbb{R}^{n \times \widetilde{k}}$ contains $\widetilde{k}$ thin-plate spline basis functions. The induced PC score, $Z\boldsymbol{\alpha}/\|Z\boldsymbol{\alpha}\|_2$, is constrained to have an underlying smooth spatial structure guided by geographic and spatial information encoded in Z. An advantage of PredPCA over PCA is the capability to identify principal directions that lead to spatially predictable PC scores at unmeasured locations. Recent work by Bose et al. (2018) further improves PredPCA by adaptively selecting information to be included in Z for each PC.

3.3 Challenge with missing data

When monitoring data is incomplete, both PCA and PredPCA in their current forms cannot handle missing data directly. Simply omitting locations with missing data may reduce the usable sample size substantially; thus, imputation is often required.

Folch-Fortuny et al. (2015) proposed new methods for building PCA models with missing data, using known-data regression. These proposed methods offer decent performance in simulated data and some chemical data applications. However, some methods either were strongly time-consuming due to multiple imputation (MI) or unfeasible with a larger number of variables. Liu and Brown (2013) compared five iterative imputation methods, including general iterative PC imputation, singular value decomposition imputation, regularized expectation-maximization (EM) with multiple ridge regression, regularized EM with truncated total least squares, and MI by chained equations. No single imputation scheme emerged as the overall best method in both simulated and real datasets. Focusing on air pollution datasets, Gómez-Carracedo et al. (2014) also evaluated various methods to fill in missing data, including single imputation techniques and MI. Under mild to moderate missing data conditions, these methods performed similarly, however, MI led to imputed values that had more variability. It is important to note that these techniques proposed or reviewed in these papers are based on only observed pollutant values but without additional spatial information. When the missingness is spatially informative, such imputation schemes may bias the results of these techniques. Furthermore, these methods were not geared towards spatially-misaligned data, where the ultimate goal is to obtain reliable and meaningful predictions of the exposure and its chemical profile at new locations.

In the next section, we propose a probabilistic framework that aims to derive spatially predictive PC scores, with the ability to handle incomplete monitoring data and induce flexible model-based imputation that accounts for spatial and geographic information. We refer to this framework as probabilistic predictive PCA, or ProPrPCA. In particular, we propose two versions of ProPrPCA. The first model, ProPrPCA-Krige, uses a latent variable structure similar to probabilistic PCA (Tipping and Bishop, 1999), with the addition of spatial patterns in the latent variable space. We discuss the connection of this model to some existing methods in multivariate analysis and highlight its contribution to handling missing data. The second proposed model, ProPrPCA-Spline, is motivated by the optimization

problem of PredPCA. This model utilizes thin-plate spline basis functions to capture spatial patterns in the data and is less computationally intensive than the Krige version.

4 Probabilistic predictive PCA

4.1 Probabilistic formulation with latent variable structure: the Krige model

Tipping and Bishop (1999) proposed a probabilistic formulation of PCA based on a Gaussian latent variable model. Their model assumes $X = uv^\top + E$, where $u \sim \mathcal{N}(0, I_n)$, $v \in \mathbb{R}^p$, $\|v\|_2 = 1$, and the elements of E are independently and identically distributed (i.i.d.) with mean zero and variance γ^2. We extend this framework by directly imposing a spatial mean and covariance structure on the latent variable space. That is, given a desired number of PCs, q, our model assumes

$$X = \sum_{l=1}^{q} (u_l v_l^\top + E_l),$$
$$u_l = R\beta_l + \eta_l,$$

where $\beta_l \in \mathbb{R}^k$ includes the coefficients corresponding to the geographic covariates in R, while $\eta_l \in \mathbb{R}^n$ has zero mean and spatial covariance $\Sigma(\xi_l)$, with ξ_l denoting the spatial covariance parameters of the latent space. We use similar constraint $\|v_l\|_2 = 1$, and assume that $\Sigma(\xi_l)$ has no nugget effect. The latent score u_l is stochastic with a full spatial distribution.

Let Θ_l be the collection of the model parameters, $\{v_l, \beta_l, \gamma_l^2, \xi_l\}$, corresponding to the l-th PC. When the monitoring data is complete, estimate of the first loading, $\hat{v}_1$, can be obtained using the original data matrix X. The corresponding score $\hat{u}_1$ at monitoring locations can then be calculated by projecting X onto the direction of $\hat{v}_1$. In later steps, Θ_l can be estimated using $X_l = X_{l-1} - \hat{u}_{l-1}\hat{v}_{l-1}^\top$, where $X_1 = X$. The PC score $\hat{u}_l$, can then be derived by projecting X_l onto $\hat{v}_l$. Note that we use projection of the data matrix to obtain the PC score in each step instead of using model estimate of the latent mean $R\beta_l$. When some elements of X are missing, estimation of Θ_l is based only on the observed elements of X_l. Estimated PC score $\hat{u}_l$ can then be made by projecting the model-based imputed exposure data onto the direction of $\hat{v}_l$.

Our approach to estimate Θ_l in each step is similar to the EM algorithm employed by Tipping and Bishop (1999). We consider the latent variable u_l to be the "missing" portion, and thus the "complete" data consists of the observed X_l and the latent variable u_l. The goal is then to maximize the joint likelihood of X_l and u_l. The mathematical details and algorithms for both complete and missing monitoring data are described in the Supplemental Materials. We refer to this model as ProPrPCA-Krige due to the kriging formulation in the model assumptions.

Our ProPrPCA-Krige model is closely related to the SupSVD model recently proposed by Li et al. (2016). The SupSVD model is expressed as $X = UV^\mathsf{T} + E$ where $U = YB + F$. Here U is a the latent score matrix, V is a full-rank loading matrix, $\mathbf{F}$ and $\mathbf{E}$ are error matrices. Li et al. (2016) also propose an EM approach to estimate the model parameters. The ProPrPCA-Krige model is also related to the envelope model proposed in Cook et al. (2010), which is a more general version compared to SupSVD. As discussed in Li et al. (2016), the SupSVD model attempts to extract a low-rank representation of the original data based on some auxiliary data, while the envelope model aims to reduce variation in regression coefficient estimation. We note that our model is motivated by spatial misalignment where data are not observed at cohort locations, but some geographic information is available. The end goal is also different from the SupSVD and envelope models, as we seek to accurately predict a low-rank representation of the data at unmeasured locations. Thus, our model is designed such that patterns of available covariates and spatial structure are properly induced in the latent scores at locations where we have data, so that we can easily predict them at new locations. An additional contribution is that we develop EM algorithms for parameter estimation for both complete and missing data scenarios.

4.2 Probabilistic formulation within thin-plate splines: the Spline model

While the ProPrPCA-Krige algorithm is cohesive with a prediction stage using universal kriging, the parameter estimation appears to be computational burdensome. In general, the EM algorithm is often computationally expensive and convergence is not always guaranteed. We propose a more simplified version of ProPrPCA,

$$X = \sum_{l=1}^{q} ((Z\boldsymbol{\beta}_l)\boldsymbol{v}_l^\mathsf{T} + E_l),$$

where Z contains thin-plate spline functions similar to PredPCA. Compared to the ProPrPCA-Krige model, the latent score u_l no longer has a stochastic component. Instead, u_l is now a smooth structure enriched with spatial patterns included in Z.

The overall procedure to obtain PC scores is similar to the Krige algorithm. The algorithm with complete monitoring data is shown in Table 1. When some elements of X_l are missing, estimation of $\widehat{\Theta}_l = \{\boldsymbol{v}_l, \boldsymbol{\beta}_l, \gamma_l^2\}$ is based on the observed elements of X_l, and estimated PC score $\hat{u}_l$ can be derived by projecting the model-based imputed exposure matrix onto the direction of $\hat{v}_l$. When the monitoring data is complete, the algorithm for parameter estimation at each step is straightforward. The mathematical derivations and the algorithm for missing data are described in the Supplemental Materials. We refer to this model as ProPrPCA-Spline due to the use of thin-plate spline basis functions.

5 Simulations

We conduct two sets of simulations to compare the different PCA approaches. The first set involves a low-dimensional setting with three-pollutant exposure surfaces. The second set

illustrates a higher-dimensional setting with 15 generated pollutant surfaces. In both cases, the multi-pollutant data is generated on a 100×100 grid ($N = 10,000$).

In each simulation, we randomly choose 400 training locations and 100 testing locations. We then apply the four competing methods (PCA, PredPCA, ProPrPCA-Krige, and ProPrPCA-Spline) to the training data, X^{train}, to obtain the corresponding loading $\hat{v}_l^{train}$ and score $\hat{u}_l^{train}$, for $I = 1, \ldots, q$ where q is a desired number of PCs. We then use $\hat{u}_l^{train}$ and relevant covariate information to obtain $\hat{u}_l^{test}$, predicted scores at testing locations, in a universal kriging model with an exponential covariance assumption. Finally, we compare the predicted scores to the known scores, u_l^{test}, which are defined by projecting X^{test} onto the direction of $\hat{v}_l^{train}$.

We also consider various scenarios in which some training data is missing. These scenarios include missing completely at random (MCAR), with 30%, 35%, and 40% of missing data, and missing at random (MAR), in which the missing patterns are associated with the generated spatial covariates. When there is missing data, we apply low-rank matrix completion (LRMC) via the SoftImpute algorithm (Mazumder et al., 2010) to fill in the missing entries prior to PCA and PredPCA.

There are several metrics to evaluate the predictive performance. The metric of interest is the prediction R^2 adapted from Szpiro et al. (2011), which reflects the correlation between $\hat{u}_l^{test}$ and $\hat{u}_l^{test}$. We also look at the reconstruction error (RE), defined as $\|X^{test} - \hat{X}^{test}\|_F$ where $\hat{X}^{test} = \hat{U}^{test}(\hat{V}^{train})^\top$, $\hat{U}^{test} = \left[\hat{u}_1^{test} \ldots \hat{u}_q^{test}\right]$, and $\hat{V}^{train} = \left[\hat{v}_1^{train} \ldots \hat{v}_q^{train}\right]$.

5.1 Three-dimensional exposure surfaces

We simulate three-dimensional surfaces with $\{x_1, x_2, x_3\}$, and three independent covariates $[r_1, r_2, r_3]$. Only $r_1 \sim \mathcal{N}(0, I_N)$ is "observed" and thus used in the universal kriging model. Both $r_2 \sim \mathcal{N}(0, I_N)$ and $r_3 \sim \mathcal{N}(0, I_N)$ are unobserved and primarily used to induce correlations across $[x_1, x_2, x_3]$. We generate data such that $x_1 = 4r_1 + 2r_3 + \epsilon_1$, $x_2 = 3r_2 + \epsilon_2$, and $x_3 = 2r_1 + 4r_2 + \epsilon_3$, where $\epsilon_1, \epsilon_2, \epsilon_3 \sim \mathcal{N}(0, \Sigma)$, where Σ has an exponential structure with partial sill $\sigma^2 = 3.5^2$, nugget $\tau^2 = 1$, and range $\phi = 50$. Under this setting, only x_1 and x_3 are predictable by r_1. While not dependent on r_1, x_2 is moderately correlated with x_3 via r_2. We also generate a second set of data in which the errors $\epsilon_1, \epsilon_2, \epsilon_3 \sim \mathcal{N}(0, 1)$. For MAR scenarios, x_1 is missing at training locations where r_1 values are larger than its 80th sample percentile, while x_2 and x_3 have 20% MCAR. We look at the first PC for these simulations, i.e. $q = 1$.

Figure 1 shows the prediction R^2's and REs across 1,000 simulations for data generated with spatially correlated noise. Table 2 displays the means and standard deviations of the estimated loadings from each method when the training data is complete. The principal direction produced by PCA is loaded heavily on x_3 and only moderately on both x_1 and x_2. This leads to poor predictive performance for PCA (median $R^2 = 0.40$). Meanwhile, loadings from the other three methods put the most weight on x_1 and some on x_3, thus they have higher prediction R^2's (median R^2's are about 0.75) and lower REs.

Under MCAR scenarios, prediction R^2's substantially decrease and REs increase for both PCA and PredPCA as the amount of missing data increases. Median R^2 of PredPCA drops to as low as 0.64 when the training data are 35% MCAR. On the other hand, there are only some subtle reductions in the predictive performances of both ProPrPCA approaches. Under MAR, the performances of both PCA and PredPCA are significantly worse. While ProPrPCA-Krige performs better than PredPCA on average, the variability of ProPrPCA-Krige in performance is high across simulations. Despite not achieving the same predictive level as it had under complete data, ProPrPCA-Spline has the highest predictive performance among the four competing methods.

Table 3 shows the estimated loadings with complete data, while Figure 2 shows the prediction R^2's and REs across 1,000 simulations for data generated with independent noise. Similar trends, where ProPrPCA outperforms the rest when missing data is more severe, are also observed in this set of generated data.

5.2 High-dimensional exposure surfaces

We also demonstrate the performance of ProPrPCA algorithms via simulations with 15 generated pollutants. The full setup is described in the Supplemental Materials. Overall, the high-dimensional exposure surfaces are generated from three underlying scores, u_1, u_2, and u_3. The data generating mechanism is such that u_1 is the most spatially predictable, u_2 is moderately predictable, and u_3 is not predictable by any covariates used in the universal kriging model. The loadings used to generate the data are sparse, in order to clearly identify the behaviors of the PCA methods. That is, the first five pollutants, $(x_1, x_2, x_3, x_4, x_5)$, are generated from u_1. Meanwhile, $(x_6, x_7, x_8, x_9, x_{10})$ are generated from u_2, and $(x_{11}, x_{12}, x_{13}, x_{14}, x_{15})$ are generated from u_3. For MAR scenario, we induce a mild spatial pattern in the missing data for the first five pollutants. In these simulations, we evaluate the predictive performance based on two PCs, i.e. $q = 2$.

We create two scenarios: scenario 1 with $Var(u_1) = 10$, $Var(u_2) = 7.5$, and $Var(u_3) = 5$, and scenario 2 with $Var(u_3) = 10$, $Var(u_1) = 7.5$, and $Var(u_2) = 5$. In scenario 1, where the order of variance contribution is the same as the order of spatial predictability, we expect all methods to identify linear combinations of u_1 and u_2 as the first two PCs when training data is complete. In scenario 2, the non-predictable score u_3 has the highest variance contribution. Thus we expect PCA to identify linear combinations of u_3 and u_1 for the first two PCs, with a large contribution of u_3 for the first PC. Meanwhile, we anticipate the other predictive methods to still pick linear combinations of u_1 and u_2.

Table 4 shows the results for the prediction R^2's across 1,000 simulations under scenario 1. As expected under scenario 1, all methods perform comparably when the training data is complete. While the results for MCAR 30% and 40% are not shown in this chapter, we observed similar patterns to the three-dimensional simulations where the performance of PCA and PredPCA decreases steadily as the amount of MCAR missing data increases. Under MCAR 35% setting, ProPrPCA-Spline has the best median R^2's for both PCs.

Under MAR, data among the first five pollutants are more likely to be missing at locations with extreme geographic covariate values. This setup effectively has an impact on the actual

variance contributions of the underlying scores in a given sample, and particularly lowers the variability contributed by u_1. As a result, for PC1, PCA is likely to produce loadings with higher contribution from u_2 than before. As the predictive methods (PredPCA and ProPrPCA) attempt to balance out the trade-off between data representativeness and spatial predictability, these methods will also likely to obtain linear combinations with more weights from u_2 for PC1 than before. Subsequently, linear combinations obtained for PC2 will have more weights from u_1 than before. This explains the decreases in median R^2's of PC1 for all methods but slight increases for PC2. ProPrPCA-Spline notably has the best median R^2 for PC1.

We further compare the differences in R^2 values between ProPrPCA-Spline and PredPCA in Figure 3. With complete training data, ProPrPCA-Spline only outperforms PredPCA for less than 60% of the simulations, and the magnitude of the difference between the two methods is rather negligible. Under MCAR 35%, ProPrPCA-Spline outperforms PredPCA for both PCs in 69.7% of the 1,000 simulations, and, for 28.5% of the time, ProPrPCA-Spline is better in one of the PCs. Finally, under MAR, there are only 2.5% of the simulations in which ProPrPCA-Spline is worse than PredPCA for both PCs. There are 38.7% of the simulations where ProPrPCA-Spline is better for only PC1 (blue top-left quadrant). Particularly for points lying in this quadrant, the greater spread along the y-axis implies that a higher increase in R^2 for PC1 is often accompanied by a smaller decrease in R^2 for PC2. Thus ProPrPCA-Spline shows more prominent benefits than PredPCA for PC1 without trading off too much in predictability of PC2.

Table 5 and Figure 4 show the corresponding results under scenario 2. In this scenario, as expected, PCA often identifies linear combinations of u_3 and u_1 as the first two PCs, and thus the predictive performance is generally poor, especially for PC1. ProPrPCA-Krige severely underperforms compared to PredPCA and ProPrPCA-Spline, even with complete data. Both PredPCA and ProPrPCA-Spline produce similar median R^2's with complete data. Similar to scenario 1, ProPrPCA-Spline performs consistently well with an increasing amount of MCAR, while the performance of PredPCA deteriorates. ProPrPCA-Spline shows clear benefits under MAR, particularly for PC1 (0.72) compared to PredPCA (0.63). The visualization of the differences in prediction R^2's between ProPrPCA-Spline and PredPCA in Figure 4 further supports similar conclusions to those of scenario 1, when the order of spatial predictability is the same as the order of variance contributed.

6 Data application

6.1 Methods

In this section, we first compare the pollutant profiles obtained by different dimension reduction methods to the annual average 2010 CSN data. Prior to our analysis, we take a similar approach to Keller et al. (2017) and convert the mass concentrations of $PM_{2.5}$ components to proportions by dividing by the total mass of $PM_{2.5}$, and then log-transform these proportions. We also apply a similar preprocessing procedure as described in Keller et al. (2017) and Jandarov et al. (2017) to the GIS covariates to be used in the predictive algorithms and the spatial prediction model. That is, we remove covariates that are missing at all chosen sites, that have the same values in at least 80% of the sites, or that have at least

2% of their values being more than five standard deviations away from the sample mean. We also remove land-use covariates whose maximal value is only 10% among all chosen sites. Finally, we apply PCA on the processed GIS data and use the first five PCs in later stages.

After the preprocessing procedure, we end up with a total of 221 CSN sites, only 130 of which have complete data on all 21 $PM_{2.5}$ components. We first apply three methods, PCA, PredPCA, and ProPrPCA-Spline, on the 130 sites with complete data (the "complete" set). We then proceed to apply these methods on all 221 CSN sites (the "full" set), where LRMC is applied prior to PCA and PredPCA. The goal is to assess how the estimated loadings and PC scores change when using only sites with complete data compared with using all available sites. The design matrix, Z, used in PredPCA and ProPrPCA-Spline includes the five PCs of GIS covariates and thin-plate spline basis functions generated from the spatial coordinates, similar to Jandarov et al. (2017). We do not use ProPrPCA-Krige in our comparison because of its inferior and unstable performance compared to ProPrPCA-Spline in our previously described simulations. In addition, the computational burden of the Krige version is exponentially larger than the Spline version, as described in the Supplemental Material.

We also conduct leave-one-site-out cross-validation to compare the predictive performances among these methods. In each round of cross-validation, we leave out one site among the complete sites as test data. We then perform dimension reduction and fit a universal kriging model on training data comprised of either only the remaining complete sites (the "complete" training data), or all remaining sites (the "full" training data), while the testing data in each round stays the same. The goal is to assess the predictive performance of different methods with both complete and missing data.

6.2 Results

6.2.1 The multi-pollutant profile—Figure 5 shows the estimated loadings and the spatial distributions of corresponding scores of the first PC for four combinations of method and dataset: PCA applied to the complete set, PredPCA applied to the complete set, imputation followed by PredPCA applied to the full set, and ProPrPCA-Spline applied to the full set. The results for ProPrPCA-Spline when using the complete set (not shown here) are essentially identical to PredPCA results.

The estimated PC1 loadings are similar across PredPCA applied to either set and to ProPrPCA-Spline, with highly positive weights on SO_4^{2-} and S and highly negative weights on Al, Ca, Na, and Si. Highly positive scores are observed in the east and part of the Midwest, probably due to sulfur emissions from coal combustion (Thurston et al., 2011; Hand et al., 2012). Negative scores are observed in the west and southwest, and have a classic resuspended soil profile (Thurston et al., 2011; Tong et al., 2012; Clements et al., 2017). While the spatial distribution of PCA scores looks similar to other methods, loadings obtained by PCA applied to the complete set are fundamentally different than the rest, with much weaker positive weights on SO_4^{2-} and S, and strongly negative weights on many additional elements, including Cr, Cu, Fe, Mn, Ni, Zn.

Figure 6 shows the estimated loadings and the score distributions for the PC that has a highly positive composition of Na, Ni, and V. This feature corresponds to PC3 obtained by PCA or PredPCA applied to the complete set, and PC2 obtained by PredPCA or ProPrPCA-Spline applied to the full set. ProPrPCA-Spline results in highly positive scores along the west coast, the east coast, and southeast region, possibly due to residual oil combustion (Thurston et al., 2011), and marine aerosol (Thurston et al., 2011; Kotchenruther, 2017). ProPrPCA-Spline also identifies pronounced negative loadings on Zn and NO_3^-. The remaining three combinations of methods and datasets are able to produce fairly similar maps with strongly positive scores along the west coast and across the northern east coast, although they fail to highlight some relevant coastal locations in the southeast region.

Figure 7 shows the results for features highly positive in NO_3^- and Zn, which corresponds to PC2 obtained by PCA or PredPCA applied to the complete set, and PC3 obtained by PredPCA or ProPrPCA-Spline applied to the full set. For all methods, highly positive scores are observed in the northern Midwest, possibly due to nitrate hazes (Coutant et al., 2003; Pitchford et al., 2009; Hand et al., 2012). Additionally, loadings produced by ProPrPCA-Spline are also strongly positive in Ni, V, and negative in Al, Si, with greater magnitude compared to other methods. Thus, moderately positive scores are also observed along the west coast. ProPrPCA-Spline also results in highly positive scores in the southeast region due to the calcium poor soils in that region compared to Al and Si content (Shacklette and Boerngen, 1984).

6.2.2   Cross-validation results—Finally, we look at the predictive performances in leave-one-site-out cross-validations. While having decent performance for PC2 and PC3 (R^2 = 0.51), using PCA applied to the complete training data yields a poor result for PC1 (R^2 = 0.24). PredPCA has similar performances for PC1 with either complete or full training data. However, there is a substantial trade-off in performances between PC2 and PC3, which can potentially be explained by the switching between PC2 and PC3 observed in the pollutant profile. ProPrPCA-Spline applied on the full training data shows the highest predictive performance for PC1 (R^2 = 0.57) and PC3 (R^2 = 0.69), but suffers from a decrease in the ability to predict PC2 well (R^2 = 0.35). In the Supplemental Materials, these results are given with further details, and we also discuss an alternate approach to evaluate the predictive performance using the variance explained to reorder the PCs under each iteration of the cross-validation procedure.

A possible explanation to the overall relatively low R^2's for all methods is that we use the same pre-specified spatial information encoded in Z to characterize the spatial variability across all PCs, which may not be effective. A potential solution, which is beyond the scope of this paper, is adaptive selection of features to be included in Z, which is proposed and discussed in Bose et al. (2018).

7   Discussion

We propose a probabilistic extension to the PredPCA algorithm developed by Jandarov et al. (2017). The proposed ProPrPCA algorithms can be applied to misaligned multi-pollutant

data with missing observations. The ultimate goal is to improve the predictive performance of the exposure modeling stage that is often required in air pollution epidemiology studies that rely on fixed site monitoring data. In spite of its simplicity, these probabilistic extensions are nontrivial and effective in mitigating the impact of missing data on the predictive performance of the exposure model. The proposed methods also eliminate the necessity of a separate imputation procedure prior to dimension reduction. The scientific motivation, especially in health-pollution studies on $PM_{2.5}$ and its components, includes the ability to use estimated PC scores at study locations as effect modifiers for the main health associations of interest.

We have demonstrated via simulations that ProPrPCA-Spline consistently outperforms its competitors under various missing observation scenarios. Its computational speed is on par with both PCA and PredPCA, which are non likelihood-based methods. The complex version, ProPrPCA-Krige, assumes a universal kriging formulation for the latent variable, with the mean model enriched by spatial covariates, and spatial correlations among the residuals. ProPrPCA-Spline incorporates thin-plate spline basis functions, which can be regarded as an alternative to a fixed low-rank kriging model (Kammann and Wand, 2003). Intuitively, the latent specification of ProPrPCA-Krige would have been cohesive with the later prediction stage using universal kriging. Possible explanations for the inferior performance of the Krige algorithm in our simulations include the difficult nature of the numerical optimization for spatial variance parameters, the number of parameters to estimate, and no guaranteed convergence to the global optima using the EM algorithm.

PCA is closely related to factor analysis (Harman, 1976), k-mean clustering (MacQueen, 1967), or positive matrix factorization (Paatero and Tapper, 1994), which have recently been used as source apportionment or dimension reduction for exposure data prior to health analyses (Sarnat et al., 2008; Ostro et al., 2011; Zanobetti et al., 2014; Ljungman et al., 2016). These applications, however, have been limited to time-series analysis in specific regions, without the challenge of spatial misalignment and severe missing data. Recent work by Keller et al. (2017) and Jandarov et al. (2017) has modified the traditional clustering and PCA methods, respectively, to the setting of spatially-misaligned multi-pollutant data, where the products of the dimension reduction procedure are desired to be spatially predictable. We further extend these frameworks by considering the realistic challenge of missing monitoring data. Our proposed framework essentially performs model-based imputation, which is cohesive and complementary to the spatial prediction stage. While one can impute the original data with sophisticated low-rank matrix completion techniques, which also operate based on the assumption of a latent variable structure, such methods only rely on observed measures. Therefore, if the missing patterns depend on external geographic covariates, such imputation schemes cannot recover the correct data structure.

In the literature, spatial latent variable models have been explored under the Bayesian framework. For example, Wang and Wall (2003) proposed a generalized common spatial factor model using MCMC techniques. Hogan and Tchernis (2004) formulated a Bayesian factor analysis model, which was later extended by Liu et al. (2005) to motivate a generalized spatial structural equations model, and by Zhu et al. (2005) to deal with spatiotemporal data. These rich modeling approaches have not been utilized in the setting of

multi-pollutant analysis with spatial misalignment. The main goal of these models is often to explain the associations between the original variables and the underlying factors. Here the goal of an improved PCA algorithm is to obtain a lower-dimensional representation of the data in a spatially predictive way for subsequent use in spatial prediction and health regression.

The multi-stage procedure in analyzing health-pollution association under spatial misalignment is a common and pragmatic approach (Crouse et al., 2010; Bergen et al., 2013; Chan et al., 2015). However, it is important to be mindful of the potential implications of measurement errors and model uncertainty of the spatial prediction stage on the health inference model, a topic which has been discussed extensively in Szpiro and Paciorek (2013). Additionally, these authors emphasized that the spatially structured components of the covariates used in the health model should be included in the exposure modeling stage to guarantee a consistent estimation of the health effects. In the multi-pollutant setting with missing observations, additional stages of imputation and dimension reduction lead to more complicated layers of uncertainty. Our proposed methods eliminate the need for a separate imputation step prior to dimension reduction, as these two steps are handled simultaneously using a model-based approach. A possible alternative to the multi-stage paradigm is a unified approach where both exposure and health data are considered simultaneously in a joint model, while leveraging the factor analysis framework to perform dimension reduction. Szpiro and Paciorek (2013) point out several disadvantages of such joint model, including sensitivity to influential or outlying health data, vulnerability to model mis-specifications, and computational burden, especially with multi-pollutant data.

While we focus our discussion in this paper exclusively on studies involving data on $PM_{2.5}$ and its components, our proposed method is both appropriate for other multi-pollutant studies and applicable to other fields in general where spatial misalignment necessitates an exposure modeling procedure. Future work should include further understanding and improvement of the ProPrPCA-Krige algorithm, and a possible extension to spatiotemporal data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Achilleos S, Kioumourtzoglou M-A, Wu C-D, Schwartz JD, Koutrakis P, and Papatheodorou SI (2017). Acute effects of fine particulate matter constituents on mortality: A systematic review and meta-regression analysis. Environment International, 109:89–100. [PubMed: 28988023]

Bell ML, Dominici F, Ebisu K, Zeger SL, and Samet JM (2007). Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. Environmental Health Perspectives, 115(7):989. [PubMed: 17637911]

Bell ML, Ebisu K, Peng RD, Samet JM, and Dominici F (2009). Hospital admissions and chemical composition of fine particle air pollution. American Journal of Respiratory and Critical Care Medicine, 179(12):1115–1120. [PubMed: 19299499]

Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, Kaufman JD, and Szpiro AA (2013). A national prediction model for $PM_{2.5}$ component exposures and measurement error–

corrected health effect inference. Environmental Health Perspectives, 121(9):1017. [PubMed: 23757600]

Bose M, Larson T, and Szpiro AA (2018). Adaptive predictive principal components for modeling multivariate air pollution. Environmetrics, 29(8).

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(2):228–239. [PubMed: 12606891]

Brook RD, Franklin B, Cascio W, Hong Y, Howard G, Lipsett M, Luepker R, Mittleman M, Samet J, Smith SC, and Tager I (2004). Air pollution and cardiovascular disease. Circulation, 109(21):2655–2671. [PubMed: 15173049]

Chan SH, Van Hee VC, Bergen S, Szpiro AA, DeRoo LA, London SJ, Marshall JD, Kaufman JD, and Sandler DP (2015). Long-term air pollution exposure and blood pressure in the Sister Study. Environmental Health Perspectives, 123(10):951. [PubMed: 25748169]

Clements AL, Fraser MP, Upadhyay N, Herckes P, Sundblom M, Lantz J, and Solomon PA (2017). Source identification of coarse particles in the Desert Southwest, USA using positive matrix ractorization. Atmospheric Pollution Research, 8(5):873–884. [PubMed: 30505154]

Cook RD, Li B, and Chiaromonte F (2010). Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica, pages 927–960.

Coutant BW, Engel-Cox J, and Swinton KE (2003). Compilation of existing studies on source apportionment for $PM_{2.5}$ Technical report of the Office of Air Quality planning and Standards, Washington, D.C.: USEPA.

Crouse DL, Goldberg MS, Ross NA, Chen H, and Labrèche F (2010). Postmenopausal breast cancer is associated with exposure to traffic-related air pollution in Montreal, Canada: a case–control study. Environmental Health Perspectives, 118(11):1578. [PubMed: 20923746]

Dai L, Zanobetti A, Koutrakis P, and Schwartz JD (2014). Associations of fine particulate matter species with mortality in the United States: a multicity time-series analysis. Environmental Health Perspectives, 122(8):837. [PubMed: 24800826]

Dominici F, Peng RD, Barr CD, and Bell ML (2010). Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. Epidemiology (Cambridge, Mass.), 21(2):187.

Dominici F, Sheppard L, and Clyde M (2003). Health effects of air pollution: A statistical review. International Statistical Review, 71(2):243–276.

Folch-Fortuny A, Arteaga F, and Ferrer A (2015). Pca model building with missing data: New proposals and a comparative study. Chemometrics and Intelligent Laboratory Systems, 146:77–88.

Franklin M, Koutrakis P, and Schwartz J (2008). The role of particle composition on the association between $PM_{2.5}$ and mortality. Epidemiology (Cambridge, Mass.), 19(5):680.

Gold DR, Litonjua A, Schwartz J, Lovett E, Larson A, Nearing B, Allen G, Verrier M, Cherry R, and Verrier R (2000). Ambient pollution and heart rate variability. Circulation, 101(11):1267–1273. [PubMed: 10725286]

Gómez-Carracedo M, Andrade J, López-Mahía P, Muniategui S, and Prada D (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems, 134:23–33.

Hand J, Schichtel B, Pitchford M, Malm W, and Frank N (2012). Seasonal composition of remote and urban fine particulate matter in the United States. Journal of Geophysical Research: Atmospheres, 117(D5).

Harman HH (1976). Modern factor analysis. University of Chicago Press.

Hogan JW and Tchernis R (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. Journal of the American Statistical Association, 99(466):314–324.

Hsu C-Y, Chiang H-C, Chen M-J, Chuang C-Y, Tsen C-M, Fang G-C, Tsai Y-I, Chen N-T, Lin T-Y, Lin S-L, and Chen Y-C (2017). Ambient $PM_{2.5}$ in the residential area near industrial complexes: Spatiotemporal variation, source apportionment, and health impact. Science of the Total Environment, 590:204–214. [PubMed: 28279531]

Jandarov RA, Sheppard LA, Sampson PD, and Szpiro AA (2017). A novel principal component analysis for spatially misaligned multivariate air pollution data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(1):3–28.

Jolliffe IT (1986). Principal component analysis and factor analysis In Principal component analysis, pages 115–128. Springer.

Kammann E and Wand MP (2003). Geoadditive models. Journal of the Royal Statistical Society: Series C (Applied Statistics), 52(1):1–18.

Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, Daviglus ML, Roux AVD, Gassett AJ, Jacobs DRJ, Kronmal R, Larson TV, Navas-Acien A, Olives C, Sampson PD, Sheppard L, Siscovick DS, Stein JH, Szpiro AA, and Watson KE (2016). Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. The Lancet, 388(10045):696–704.

Keller JP, Drton M, Larson T, Kaufman JD, Sandler DP, and Szpiro AA (2017). Covariate-adaptive clustering of exposures for air pollution epidemiology cohorts. The Annals of Applied Statistics, 11(1):93. [PubMed: 28572869]

Keller JP, Larson TV, Austin E, Barr RG, Sheppard L, Vedal S, Kaufman JD, and Szpiro AA (2018). Pollutant composition modification of the effect of air pollution on progression of coronary artery calcium: The Multi-Ethnic Study of Atherosclerosis. Environmental Epidemiology, 2(3):e024. [PubMed: 30854505]

Kioumourtzoglou M-A, Austin E, Koutrakis P, Dominici F, Schwartz J, and Zanobetti A (2015). $PM_{2.5}$ and survival among older adults: effect modification by particulate composition. Epidemiology (Cambridge, Mass.), 26(3):321.

Kotchenruther RA (2017). The effects of marine vessel fuel sulfur regulations on ambient $PM_{2.5}$ at coastal and near coastal monitoring sites in the US. Atmospheric Environment, 151:52–61.

Krall JR, Anderson GB, Dominici F, Bell ML, and Peng RD (2013). Short-term exposure to particulate matter constituents and mortality in a national study of US urban communities. Environmental Health Perspectives, 121(10):1148. [PubMed: 23912641]

Künzli N, Jerrett M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, and Hodis HN (2005). Ambient air pollution and atherosclerosis in Los Angeles. Environmental Health Perspectives, 113(2):201. [PubMed: 15687058]

Li G, Yang D, Nobel AB, and Shen H (2016). Supervised singular value decomposition and its asymptotic properties. Journal of Multivariate Analysis, 146:7–17.

Liu X, Wall MM, and Hodges JS (2005). Generalized spatial structural equation models. Biostatistics, 6(4):539–557. [PubMed: 15843593]

Liu Y and Brown SD (2013). Comparison of five iterative imputation methods for multivariate classification. Chemometrics and Intelligent Laboratory Systems, 120:106–115.

Ljungman PL, Wilker EH, Rice MB, Austin E, Schwartz J, Gold DR, Koutrakis P, Benjamin EJ, Vita JA, Mitchell GF, Vasan RS, Hamburg NM, and Mittleman MA (2016). The impact of multi-pollutant clusters on the association between fine particulate air pollution and microvascular function. Epidemiology (Cambridge, Mass.), 27(2):194.

MacQueen J (1967). Some methods for classification and analysis of multivariate observations In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA.

Mazumder R, Hastie T, and Tibshirani R (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11(8):2287–2322. [PubMed: 21552465]

Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, and Kaufman JD (2007). Long-term exposure to air pollution and incidence of cardiovascular events in women. New England Journal of Medicine, 2007(356):447–458.

Ostro B, Tobias A, Querol X, Alastuey A, Amato F, Pey J, Pérez N, and Sunyer J (2011). The effects of particulate matter sources on daily mortality: a case-crossover study of Barcelona, Spain. Environmental Health Perspectives, 119(12):1781. [PubMed: 21846610]

Paatero P and Tapper U (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 5(2):111–126.

Pascal M, Falq G, Wagner V, Chatignoux E, Corso M, Blanchard M, Host S, Pascal L, and Larrieu S (2014). Short-term impacts of particulate matter (PM_{10}, $PM_{10-2.5}$, $PM_{2.5}$) on mortality in nine French cities. Atmospheric Environment, 95:175–184.

Pitchford ML, Poirot RL, Schichtel BA, and Malm WC (2009). Characterization of the winter midwestern particulate nitrate bulge. Journal of the Air & Waste Management Association, 59(9):1061–1069. [PubMed: 19785273]

Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, and Thurston GD (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association, 287(9):1132–1141. [PubMed: 11879110]

Sarnat JA, Marmur A, Klein M, Kim E, Russell AG, Sarnat SE, Mulholland JA, Hopke PK, and Tolbert PE (2008). Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods. Environmental Health Perspectives, 116(4):459. [PubMed: 18414627]

Shacklette HT and Boerngen JG (1984). Element concentrations in soils and other surficial materials of the conterminous United States Geological Survey Professional Paper 1270, Washington, D.C.

Shen H and Huang JZ (2008). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, 99(6):1015–1034.

Szpiro AA and Paciorek CJ (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics, 24(8):501–517. [PubMed: 24764691]

Szpiro AA, Paciorek CJ, and Sheppard L (2011). Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology (Cambridge, Mass.), 22(5):680.

Thurston GD, Ito K, and Lall R (2011). A source apportionment of US fine particulate matter air pollution. Atmospheric Environment, 45(24):3924–3936. [PubMed: 24634604]

Tian L, Zeng Q, Dong W, Guo Q, Wu Z, Pan X, Li G, and Liu Y (2017). Addressing the source contribution of $PM_{2.5}$ on mortality: an evaluation study of its impacts on excess mortality in China. Environmental Research Letters, 12(10):104016.

Tipping ME and Bishop CM (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622.

Tolbert PE, Klein M, Peel JL, Sarnat SE, and Sarnat JA (2007). Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta. Journal of Exposure Science and Environmental Epidemiology, 17(S2):S29. [PubMed: 18079762]

Tong DQ, Dan M, Wang T, and Lee P (2012). Long-term dust climatology in the western United States reconstructed from routine aerosol ground monitoring. Atmospheric Chemistry and Physics, 12(11):5189–5205.

Wang F and Wall MM (2003). Generalized common spatial factor model. Biostatistics, 4(4):569–582. [PubMed: 14557112]

Wang Y, Shi L, Lee M, Liu P, Di Q, Zanobetti A, and Schwartz JD (2017). Long-term exposure to $PM_{2.5}$ and mortality among older adults in the southeastern US. Epidemiology (Cambridge, Mass.), 28(2):207–214.

Zanobetti A, Austin E, Coull BA, Schwartz J, and Koutrakis P (2014). Health effects of multi-pollutant profiles. Environment International, 71:13–19. [PubMed: 24950160]

Zhu J, Eickhoff J, and Yan P (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. Biometrics, 61(3):674–683. [PubMed: 16135018]
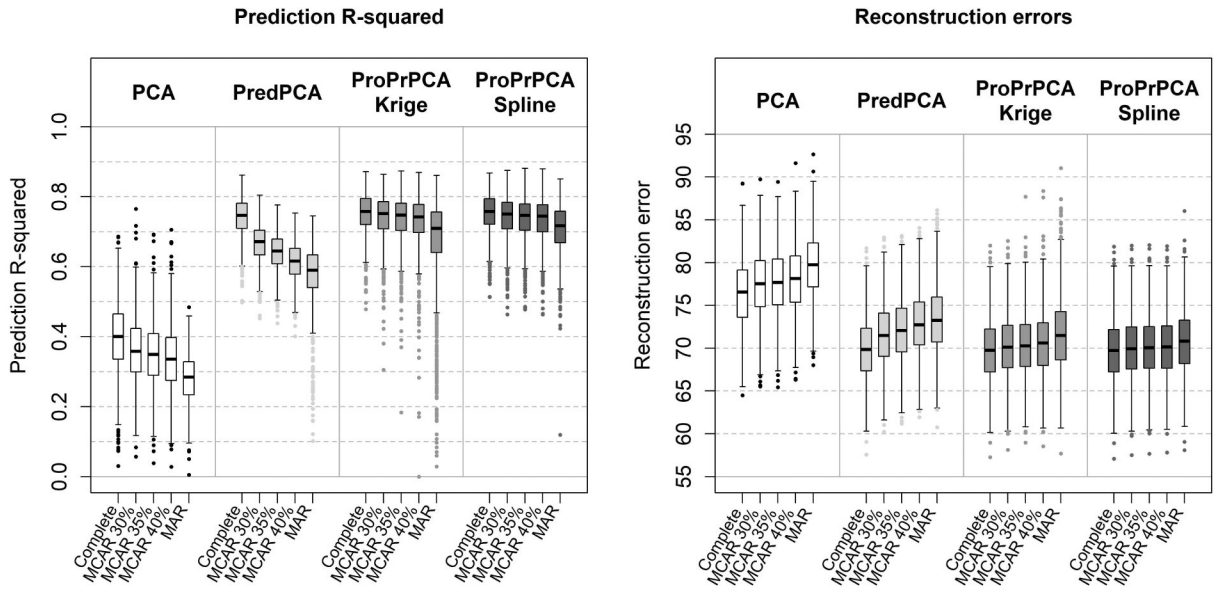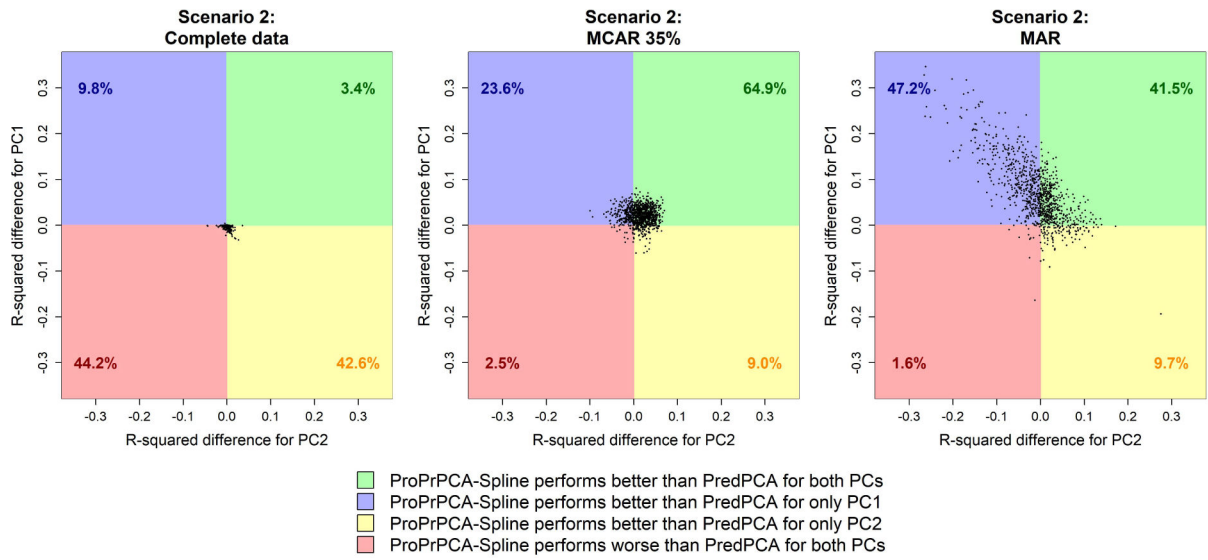
Figure 1:

Prediction R^2's and reconstruction errors across 1,000 replications with a three-dimensional surface generated with spatially correlated noises. Under missing data scenarios, LRMC is used prior to the application of either PCA or PredPCA.
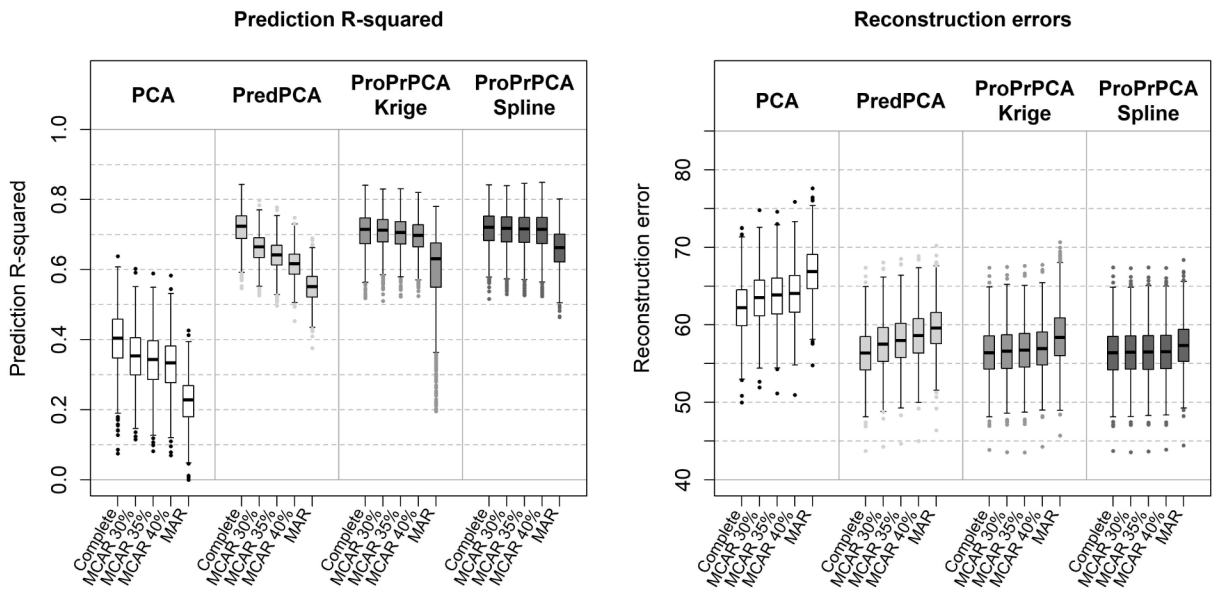
Figure 2:
Prediction R^2's and reconstruction errors across 1,000 replications with a three-dimensional surface generated with independent noises. Under missing data scenarios, LRMC is used prior to the application of either PCA or PredPCA.
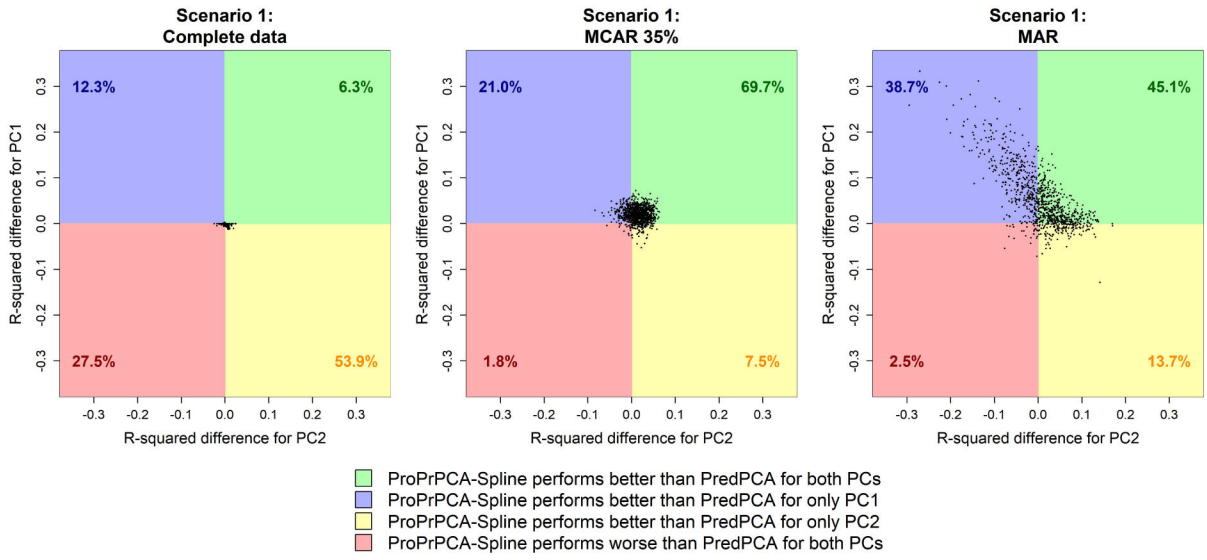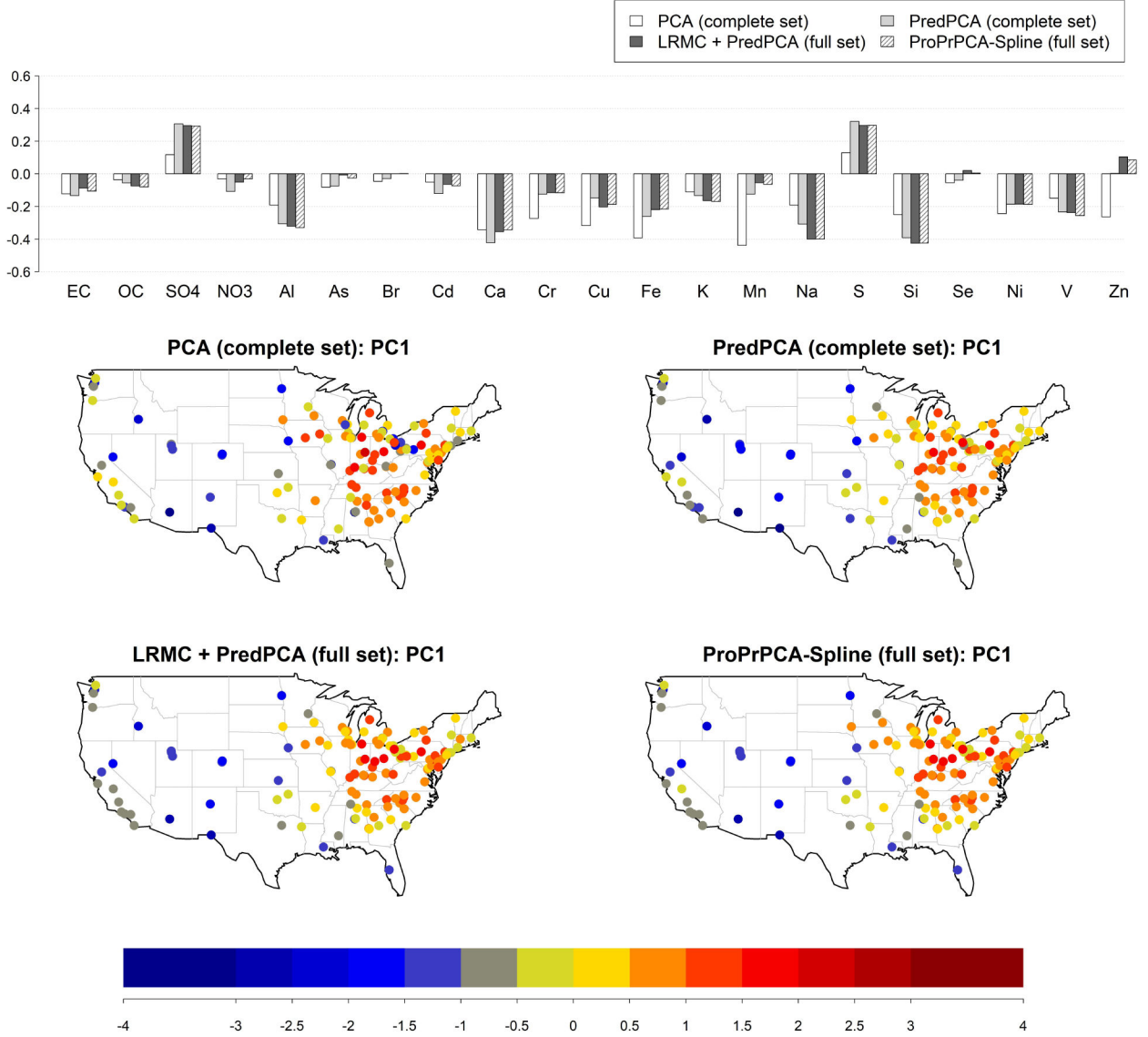
Figure 3:

Differences in prediction R^2 values between ProPrPCA-Spline and PredPCA for high-dimensional scenario 1. Each dot represents result from one simulation. Percentages indicate the proportion out of 1,000 simulations.
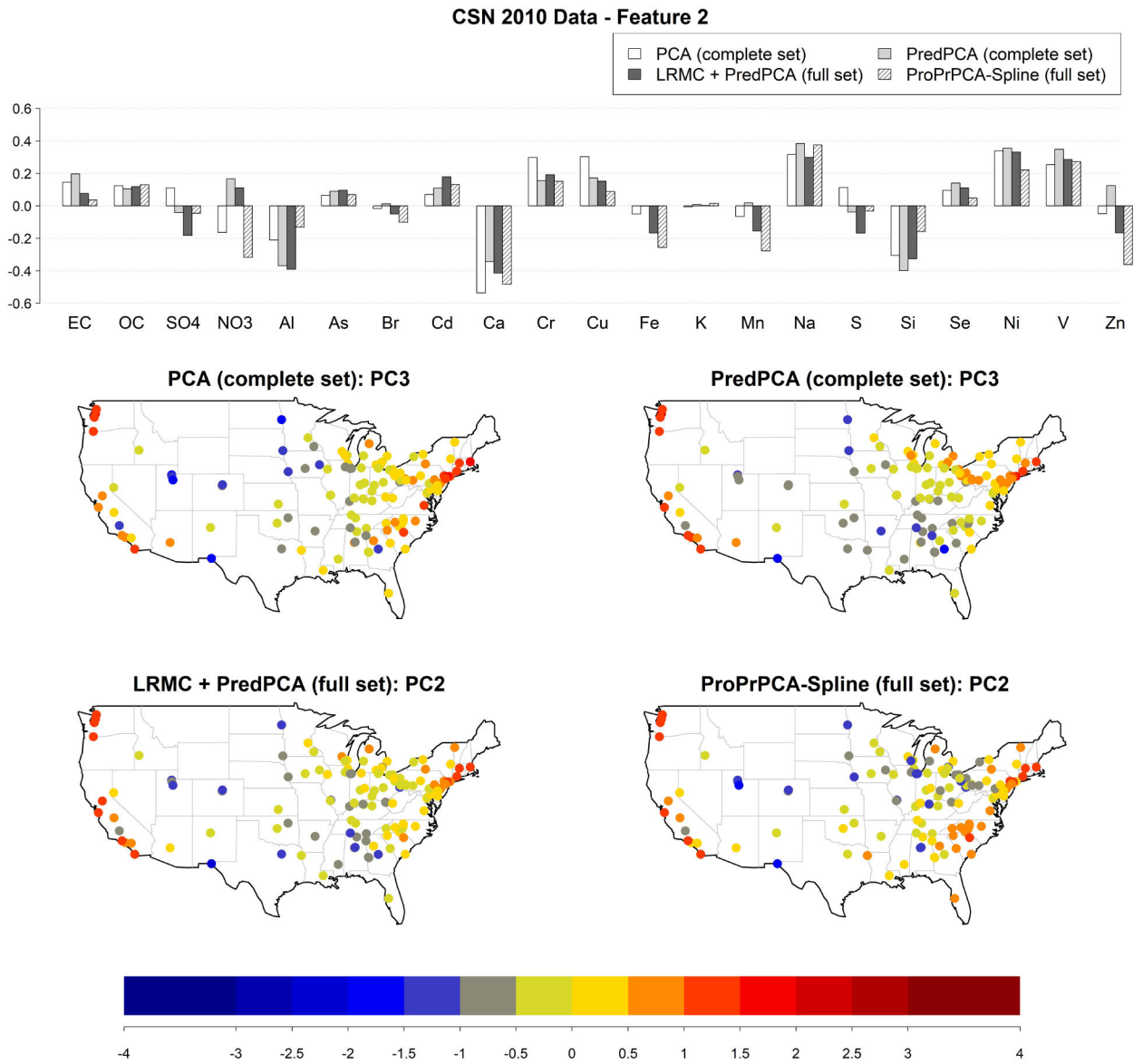
Figure 4:
Differences in prediction R^2 values between ProPrPCA-Spline and PredPCA for high-dimensional scenario 2. Each dot represents the result from one simulation. Percentages indicate the proportion out of 1,000 simulations.
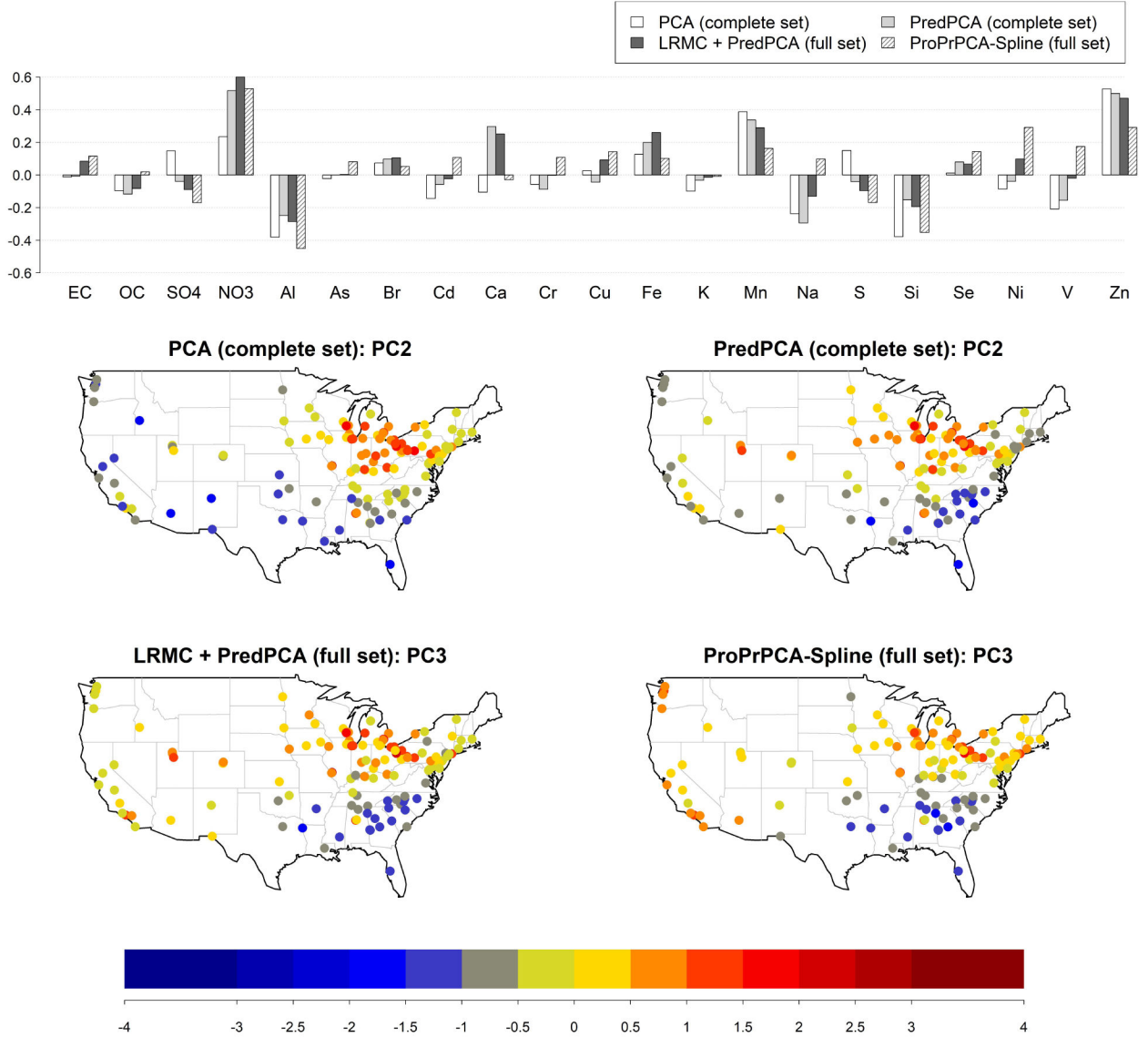
Figure 5:

Estimated loadings for the feature with highly positive weights on SO_4^{2-} and S, and corresponding scores, obtained from different PCA algorithms applied to 2010 CSN data: PCA and PredPCA applied to the complete set (130 sites with complete data), PredPCA and ProPrPCA-Spline applied to the full set (all 221 available sites).

Figure 6:
Estimated loadings for the feature with highly positive weights on Na, Ni, and V, and corresponding scores, obtained from different PCA algorithms applied to 2010 CSN data: PCA and PredPCA applied to the complete set (130 sites with complete data), PredPCA and ProPrPCA-Spline applied to the full set (all 221 available sites).

Figure 7:

Estimated loadings for the feature with highly positive weights on NO_3^- and Zn, and corresponding scores, obtained from different PCA algorithms applied to 2010 CSN data: PCA and PredPCA applied to the complete set (130 sites with complete data), PredPCA and ProPrPCA-Spline applied to the full set (all 221 available sites).

Table 1:

The algorithm for ProPrPCA-Spline with complete monitoring data

---

Input X, Z, q, and t_{max}

for l in $\{1, ..., q\}$ **do**

$\quad X_l \leftarrow X_{l-1} - \hat{u}_{l-1}\hat{v}_{l-1}^\mathsf{T}$ where $X_0 = X$, $\hat{u}_0 = \mathbf{0}$, and $\hat{v}_0 = \mathbf{0}$

$\quad$ **Initialize** $v_l^{(0)}, (\gamma_l^{(0)})^2, \beta_l^{(0)}$, and $t = 1$

$\quad\quad$ **while** not converged **or** $t < t_{max}$ **do**

$\quad\quad\quad v_l^{(t+1)} \leftarrow \tilde{v}_l / \|\tilde{v}_l\|_2$ where $\tilde{v}_l \leftarrow X_l^\mathsf{T} Z\beta_l^{(t)} / \|Z\beta_l^{(t)}\|_2^2$

$\quad\quad\quad \beta_l^{(t+1)} \leftarrow (Z^\mathsf{T} Z)^{-1}\left(Z \otimes v_l^{(t+1)}\right)^\mathsf{T} \mathrm{vec}(X_l)$

$\quad\quad\quad (\gamma_l^{(t+1)})^2 \leftarrow (np)^{-1}\|\mathrm{vec}(X_l) - (I_n \otimes v_l^{(t+1)})Z\beta_l^{(t+1)}\|_2^2$

$\quad\quad\quad t \leftarrow t + 1$

$\quad\quad$ **end while**

$\quad \hat{v}_l \leftarrow v_l^{(t)}, \hat{\gamma}_l^2 \leftarrow (\gamma_l^{(t)})^2, \hat{\beta}_l \leftarrow \beta_l^{(t)}$

$\quad \hat{u}_l = X_l\hat{v}_l$

end for

Output $\{\hat{v}_1, ..., \hat{v}_q\}, \{\hat{u}_1, ..., \hat{u}_q\}, \{\hat{\beta}_1, ..., \hat{\beta}_q\}, \{\hat{\gamma}_l^2, ..., \hat{\gamma}_q^2\}$

Table 2:

Means (standard deviations) of estimated PC1 loadings across 1,000 replications with a three-dimensional surface with spatially correlated noise and complete training data.

	X_1	X_2	X_3
PCA	0.40 (0.11)	0.41 (0.09)	0.80 (0.07)
PredPCA	0.88 (0.04)	−0.07 (0.04)	0.46 (0.09)
ProPrPCA-Krige	0.85 (0.04)	−0.11 (0.08)	0.50 (0.08)
ProPrPCA-Spline	0.86 (0.03)	−0.12 (0.07)	0.49 (0.07)

Table 3:

Means (standard deviations) of estimated PC1 loadings across 1,000 replications with a three-dimensional surface with independent noise and complete training data.

	X_1	X_2	X_3
PCA	0.53 (0.06)	0.39 (0.04)	0.75 (0.03)
PredPCA	0.89 (0.02)	0.01 (0.02)	0.45 (0.04)
ProPrPCA-Krige	0.88 (0.02)	0.03 (0.04)	0.47 (0.04)
ProPrPCA-Spline	0.89 (0.02)	0.01 (0.03)	0.46 (0.04)

Table 4:

The median prediction R^2's across 1,000 simulations for high-dimensional scenario 1. Under missing data scenarios, LRMC is used prior to either PCA or PredPCA.

PC1	Complete	MCAR 35%	MAR
PCA	0.83	0.80	0.61
PredPCA	0.84	0.81	0.63
ProPrPCA-Krige	0.83	0.83	0.64
ProPrPCA-Spline	0.84	0.83	0.69
PC2	**Complete**	**MCAR 35%**	**MAR**
PCA	0.60	0.58	0.67
PredPCA	0.60	0.58	0.68
ProPrPCA-Krige	0.60	0.60	0.69
ProPrPCA-Spline	0.60	0.60	0.68

Table 5:

The median prediction R^2's across 1,000 simulations for high-dimensional scenario 2. Under missing data scenarios, LRMC is used prior to either PCA or PredPCA.

PC1	Complete	MCAR 35%	MAR
PCA	0.01	0.01	0.00
PredPCA	0.81	0.78	0.63
ProPrPCA-Krige	0.70	0.66	0.41
ProPrPCA-Spline	0.81	0.80	0.72
PC2	**Complete**	**MCAR 35%**	**MAR**
PCA	0.78	0.74	0.60
PredPCA	0.56	0.54	0.62
ProPrPCA-Krige	0.30	0.26	0.23
ProPrPCA-Spline	0.56	0.56	0.59