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] and Adam A. Szpiro[1]

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

Supplemental Materials

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

1 The ProPrPCA-Krige model and algorithm

1.1 The model

The ProPrPCA-Krige assumes that $\boldsymbol{X} = \sum_{l=1}^{q} \left(\boldsymbol{u}_l \boldsymbol{v}_l^\mathsf{T} + \boldsymbol{E}_l\right)$ and $\boldsymbol{u}_l = \boldsymbol{R}\boldsymbol{\beta}_l + \boldsymbol{\eta}_l$, where $\boldsymbol{u}_l \in \mathbb{R}^n$, $\boldsymbol{v}_l \in \mathbb{R}^p$, $\boldsymbol{\beta}_l \in \mathbb{R}^k$, $E_{ij} \sim \mathcal{N}(0, \gamma^2)$, and $\boldsymbol{\eta}_l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\xi}_l))$. For notation simplification, we are going to ignore the subscript l for the following mathematical derivation. The parameter estimation is the same for all PC. For each PC, the model becomes:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{u}\boldsymbol{v}^\mathsf{T} + \boldsymbol{E}, \\ \boldsymbol{u} &= \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{\eta}, \end{aligned}$$

Denote $\Theta = \{\boldsymbol{v}, \boldsymbol{\beta}, \gamma^2, \boldsymbol{\xi}\}$ as the collection of the model parameters. To solve for Θ, we first rewrite the model in the conventional vectorized version. Denote $\boldsymbol{W} \in \mathbb{R}^N$, for $N = np$, as the vectorized version of $\boldsymbol{X}$, i.e.

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{1:} \\ \vdots \\ \boldsymbol{X}_{n:} \end{bmatrix},$$

where $\boldsymbol{X}_{i:}$ is the i-th row of $\boldsymbol{X}$. The model assumes that $\boldsymbol{W}_i = \boldsymbol{X}_{i:} = u_i\boldsymbol{v} + \boldsymbol{\epsilon}_i$, for $i = 1, ..., n$. Here $\boldsymbol{v}$ represents the transformation from the latent variable space to the multi-pollutant exposure space, and $\boldsymbol{\epsilon}_i$'s are i.i.d. Gaussian noises distributed with mean zero and variance γ^2. The full model can then be written as $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{V} = \boldsymbol{I}_n \otimes \boldsymbol{v}$ and $\otimes$ denotes the Kronecker product. The model also assumes that the latent variables are normally distributed with a spatial mean model and covariance structure. That is, $\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{R}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\xi})\right)$. In this paper, we assume $\boldsymbol{\Sigma}(\boldsymbol{\xi})$ has an exponential structure with no nugget effect. For identifiability, we assume that $\|\boldsymbol{v}\|_2 = 1$. When every element of $\boldsymbol{X}$ is observed, we have the following hierarchical model:

$$\begin{aligned} \boldsymbol{W} \mid \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{V}\boldsymbol{u}, \gamma^2\boldsymbol{I}_N\right), \\ \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{R}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\xi})\right). \end{aligned}$$

In practice, not all pollutants are measured at every monitoring location. Denote $\boldsymbol{W}_o \in \mathbb{R}^{N_o}$ as the collection of all observed elements of $\boldsymbol{W}$, and $\boldsymbol{W}_m \in \mathbb{R}^{N_m}$ as the collection of all missing entries, where $N_o + N_m = N$. Algebraically, there exists a linear transformation $\boldsymbol{G}$ such that

$$\boldsymbol{G}\boldsymbol{W} = \begin{bmatrix} \boldsymbol{G}_o \\ \boldsymbol{G}_m \end{bmatrix} \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_o \\ \boldsymbol{W}_m \end{bmatrix},$$

where $\boldsymbol{G}_o \in \mathbb{R}^{N_o \times N}$ and $\boldsymbol{G}_m \in \mathbb{R}^{N_m \times N}$. Each row and column of $\boldsymbol{G}$ contains exactly one element of one and $(N-1)$ zeros. Thus by construction, $\boldsymbol{G}_o$ and $\boldsymbol{G}_m$ are both full row rank, as well as $\boldsymbol{G}_o\boldsymbol{G}_o^\mathsf{T} = \boldsymbol{I}_{N_o}$ and $\boldsymbol{G}_m\boldsymbol{G}_m^\mathsf{T} = \boldsymbol{I}_{N_m}$. The hierarchical model for the observed

elements become:

$$\begin{aligned}\boldsymbol{W}_o \mid \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u}, \gamma^2\boldsymbol{I}_{N_o}\right),\\ \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{R}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\xi})\right).\end{aligned}$$

1.2 Estimation of model parameters when monitoring data is complete

Our approach to estimate the model parameters is similar to the EM algorithm employed by Tipping and Bishop (1999). We consider the latent variable $\boldsymbol{u}$ to be the "missing" portion. Thus the "complete" data consists of the observed data $\boldsymbol{W}$, and the latent variable $\boldsymbol{u}$. The goal is then to maximize the joint likelihood of $(\boldsymbol{W}, \boldsymbol{u})$, i.e. $\mathcal{L} = f(\boldsymbol{W}, \boldsymbol{u}) = f(\boldsymbol{W}|\boldsymbol{u})f(\boldsymbol{u})$. The "complete" log-likelihood, up to a constant, is:

$$\ell_c = -\frac{N}{2}\log\gamma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\gamma^2}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u}) - \frac{1}{2}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta}).$$

Because this log-likelihood involves $\boldsymbol{u}$ which is unobserved, in each E step, we find the expectation of ℓ_c with respect to the conditional distribution of $\boldsymbol{u}|\boldsymbol{W}$. We first derive this distribution as follows:

$$\begin{aligned}f(\boldsymbol{u}|\boldsymbol{W}) &\propto \exp\left[-\frac{1}{2\gamma^2}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u})\right] \times \exp\left[-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta})\right]\\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{u}^\mathsf{T}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u} - \boldsymbol{u}^\mathsf{T}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{W}+\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\beta}\right)\right.\right.\\ &\qquad\left.\left. - \left(\frac{1}{\gamma^2}\boldsymbol{W}^\mathsf{T}\boldsymbol{V}+\boldsymbol{\beta}^\mathsf{T}\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u}\right]\right\}.\end{aligned}$$

Thus the distribution of $\boldsymbol{u}|\boldsymbol{W}$ is $\mathcal{N}(\boldsymbol{M}, \boldsymbol{S})$, where:

$$\begin{aligned}\boldsymbol{S} &= \left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}+\boldsymbol{\Sigma}^{-1}\right)^{-1},\\ \boldsymbol{M} &= \boldsymbol{S}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{W}+\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\beta}\right).\end{aligned}$$

We can further simplify these expressions by noticing that $\boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{I}_n$ and $\boldsymbol{V}^\mathsf{T}\boldsymbol{W} = \boldsymbol{X}\boldsymbol{v}$, using properties of Kronecker products. The conditional covariance and mean become

$$\begin{aligned}\boldsymbol{S} &= \left(\frac{1}{\gamma^2}\boldsymbol{I}_n+\boldsymbol{\Sigma}^{-1}\right)^{-1},\\ \boldsymbol{M} &= \boldsymbol{S}\left(\frac{1}{\gamma^2}\boldsymbol{X}\boldsymbol{v}+\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\beta}\right).\end{aligned}$$

Given a current estimate $\tilde{\Theta}$, the expectation of ℓ_c with respect to $\boldsymbol{u}|\boldsymbol{W}$ is:

$$E\left[\ell_c \mid \boldsymbol{W}, \tilde{\Theta}\right] = -\frac{N}{2}\log\gamma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\gamma^2}E\left[(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u}) \mid \boldsymbol{W}, \tilde{\Theta}\right] - \frac{1}{2}E\left[(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta}) \mid \boldsymbol{W}, \tilde{\Theta}\right] \quad (1)$$

The conditional distribution $\boldsymbol{u}|\boldsymbol{W}, \tilde{\Theta}$ is $\mathcal{N}(\tilde{\boldsymbol{M}}, \tilde{\boldsymbol{S}})$, where $\tilde{\boldsymbol{M}} = \boldsymbol{M}(\tilde{\Theta})$ and $\tilde{\boldsymbol{S}} = \boldsymbol{S}(\tilde{\Theta})$. This implies that

$$\boldsymbol{V}\boldsymbol{u} - \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}, \boldsymbol{V}\tilde{\boldsymbol{S}}\boldsymbol{V}^\mathsf{T}\right),$$
$$\boldsymbol{u} - \boldsymbol{R}\boldsymbol{\beta} \sim \mathcal{N}\left(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta}, \tilde{\boldsymbol{S}}\right).$$

Thus the first expectation term of (1) is

$$\begin{aligned} E\left[(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}-\boldsymbol{V}\boldsymbol{u}) \mid \boldsymbol{W}, \tilde{\Theta}\right] &= \mathrm{Tr}\left(\boldsymbol{V}\tilde{\boldsymbol{S}}\boldsymbol{V}^\mathsf{T}\right) + (\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}) \\ &= \mathrm{Tr}(\tilde{\boldsymbol{S}}) + (\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}). \end{aligned}$$

The second expectation term of (1) is

$$E\left[(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}-\boldsymbol{R}\boldsymbol{\beta}) \mid \boldsymbol{W}, \tilde{\Theta}\right] = \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) + (\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})$$

Hence (1) can be simplified as

$$\begin{aligned} E\left[\ell_c \mid \boldsymbol{W}, \tilde{\Theta}\right] = &-\frac{N}{2}\log\gamma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\gamma^2}\mathrm{Tr}(\tilde{\boldsymbol{S}}) - \frac{1}{2\gamma^2}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}) \\ &- \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) - \frac{1}{2}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta}). \end{aligned}$$

To solve for $\boldsymbol{v}$, we effectively maximize $\{-\frac{1}{2\gamma^2}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})\}$, which can be rewrite as follows:

$$\begin{aligned} &-\frac{1}{2\gamma^2}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}) \\ =&-\frac{1}{2\gamma^2}\left\|\boldsymbol{W} - \boldsymbol{V}\tilde{\boldsymbol{M}}\right\|_2^2 = -\frac{1}{2\gamma^2}\left\| \begin{bmatrix} \boldsymbol{X}_{1:} \\ \boldsymbol{X}_{2:} \\ \vdots \\ \boldsymbol{X}_{n:} \end{bmatrix} - \begin{bmatrix} \boldsymbol{v} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{v} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{v} \end{bmatrix} \begin{bmatrix} \tilde{M}_1 \\ \tilde{M}_2 \\ \vdots \\ \tilde{M}_n \end{bmatrix} \right\|_2^2 \\ =&-\frac{1}{2\gamma^2}\sum_{i=1}^{n}\left\|\boldsymbol{X}_{i.} - \tilde{M}_i\boldsymbol{v}\right\|_2^2 = -\frac{1}{2\gamma^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \tilde{M}_i v_j\right)^2. \end{aligned}$$

Differentiate this expression with respect to each v_j, we get

$$\breve{v}_j = \frac{\sum_{i=1}^n X_{ij}\tilde{M}_i}{\sum_{i=1}^n \tilde{M}_i^2} = \frac{\sum_{i=1}^n X_{ij}\tilde{M}_i}{\left\|\tilde{\boldsymbol{M}}\right\|_2^2}.$$

Thus, the solution for $\boldsymbol{v}$ can be written in closed-form as

$$\hat{\boldsymbol{v}} = \frac{\breve{\boldsymbol{v}}}{\|\breve{\boldsymbol{v}}\|_2}, \qquad \text{where } \breve{\boldsymbol{v}} = \frac{\boldsymbol{X}^\mathsf{T}\tilde{\boldsymbol{M}}}{\left\|\tilde{\boldsymbol{M}}\right\|_2^2}$$

To solve for γ^2, we maximize $\Big\{ -\frac{N}{2}\log\gamma^2 - \frac{1}{2\gamma^2}\mathrm{Tr}(\tilde{\boldsymbol{S}}) - \frac{1}{2\gamma^2}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})^\mathsf{T}(\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W})\Big\}$. The closed-form solution for γ^2 is simply

$$\hat{\gamma}^2 = \frac{1}{N}\big[\mathrm{Tr}(\tilde{\boldsymbol{S}}) + \left\|\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}\right\|_2^2.\big]$$

The solution for $\boldsymbol{\beta}$ by maximizing $\Big\{ -\frac{1}{2}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})\Big\}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{M}}.$$

Finally, to solve for $\boldsymbol{\xi}$, we maximize $\Big\{ -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) - \frac{1}{2}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})\Big\}$ numerically, where $\boldsymbol{\Sigma}$ is a function of $\boldsymbol{\xi}$. In this paper, we adopt the exponential covariance structure for $\boldsymbol{\Sigma}$. For identifiability, we assume that $\boldsymbol{\Sigma}$ has no nugget effect.

Thus, parameter estimation of ProPrPCA-Krige with complete monitoring data can be summarized as:

Algorithm: ProPrPCA-Krige with complete monitoring data

Input $\boldsymbol{X}$, $\boldsymbol{R}$, q, and t_{max}
  for l in $\{1, ..., q\}$ **do**
    $\boldsymbol{X}_l \leftarrow \boldsymbol{X}_{l-1} - \hat{\boldsymbol{u}}_{l-1}\hat{\boldsymbol{v}}_{l-1}^\mathsf{T}$ where $\boldsymbol{X}_0 = \boldsymbol{X}$, $\hat{\boldsymbol{u}}_0 = \boldsymbol{0}$, and $\hat{\boldsymbol{v}}_0 = \boldsymbol{0}$
    Initialize $\boldsymbol{v}_l^{(0)}$, $(\gamma_l^{(0)})^2$, $\boldsymbol{\beta}_l^{(0)}$, $\boldsymbol{\xi}_l^{(0)}$, and $t = 1$
    $\boldsymbol{\Sigma}_l^{(0)} \leftarrow \boldsymbol{\Sigma}(\boldsymbol{\xi}_l^{(0)})$
      while not converged **or** $t < t_{max}$ **do**
        $\tilde{\boldsymbol{S}}_l \leftarrow \left[(\gamma_l^{(t)})^{-2}\boldsymbol{I}_n + (\boldsymbol{\Sigma}_l^{(t)})^{-1}\right]^{-1}$
        $\tilde{\boldsymbol{M}}_l \leftarrow \tilde{\boldsymbol{S}}_l \left[(\gamma_l^{(t)})^{-2}\boldsymbol{X}_l\boldsymbol{v}_l^{(t)} + (\boldsymbol{\Sigma}_l^{(t)})^{-1}\boldsymbol{R}\boldsymbol{\beta}_l^{(t)}\right]$
        $\tilde{\boldsymbol{v}}_l \leftarrow \boldsymbol{X}_l^\mathsf{T}\tilde{\boldsymbol{M}}_l / \left\|\tilde{\boldsymbol{M}}_l\right\|_2^2$
        $\boldsymbol{v}_l^{(t+1)} \leftarrow \tilde{\boldsymbol{v}}_l / \|\tilde{\boldsymbol{v}}_l\|_2$
        $(\gamma_l^{(t+1)})^2 \leftarrow (np)^{-1}\left[\mathrm{Tr}(\tilde{\boldsymbol{S}}_l) + \left\|(\boldsymbol{I}_n \otimes \boldsymbol{v}_l^{(t+1)})\tilde{\boldsymbol{M}}_l - \mathrm{vec}(\boldsymbol{X}_l)\right\|_2^2\right]$
        $\boldsymbol{\xi}_l^{(t+1)} \leftarrow \underset{\boldsymbol{\xi}_l}{\arg\max}\{ -\log|\boldsymbol{\Sigma}_l| - \mathrm{Tr}\left(\boldsymbol{\Sigma}_l^{-1}\tilde{\boldsymbol{S}}_l\right) - (\tilde{\boldsymbol{M}}_l - \boldsymbol{R}\boldsymbol{\beta}_l^{(t)})^\mathsf{T}\boldsymbol{\Sigma}_l^{-1}(\tilde{\boldsymbol{M}}_l - \boldsymbol{R}\boldsymbol{\beta}_l^{(t)})\}$
          where $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}(\boldsymbol{\xi}_l)$
        $\boldsymbol{\beta}_l^{(t+1)} \leftarrow \left(\boldsymbol{R}^\mathsf{T}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}_l^{(t+1)})^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^\mathsf{T}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}_l^{(t+1)})^{-1}\tilde{\boldsymbol{M}}_l$
        $t \leftarrow t + 1$
      end while
    $\hat{\boldsymbol{v}}_l \leftarrow \boldsymbol{v}_l^{(t)}$, $\hat{\gamma}_l^2 \leftarrow (\gamma_l^{(t)})^2$, $\hat{\boldsymbol{\beta}}_l \leftarrow \boldsymbol{\beta}_l^{(t)}$, $\hat{\boldsymbol{\xi}}_l \leftarrow \boldsymbol{\xi}_l^{(t)}$
    $\hat{\boldsymbol{u}}_l = \boldsymbol{X}_l\hat{\boldsymbol{v}}_l$
  end for
Output $\{\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_q\}$, $\{\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_q\}$, $\{\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_q\}$, $\{\hat{\gamma}_1^2, ..., \hat{\gamma}_q^2\}$, $\{\hat{\boldsymbol{\xi}}_1, ..., \hat{\boldsymbol{\xi}}_q\}$

1.3 Parameter estimation and model-based imputation with missing monitoring data

The hierarchical model in the case of missing monitoring data can be written as

$$\begin{aligned}\boldsymbol{W}_o \mid \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u}, \gamma^2\boldsymbol{I}_{N_o}\right),\\ \boldsymbol{u} &\sim \mathcal{N}\left(\boldsymbol{R}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\xi})\right).\end{aligned}$$

Thus the "complete" log-likelihood becomes

$$\begin{aligned}\ell_c = &-\frac{N_o}{2}\log\gamma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\gamma^2}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u})\\ &-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u} - \boldsymbol{R}\boldsymbol{\beta}).\end{aligned}$$

Similar to the case with complete data, we first derive the conditional distribution of

$\boldsymbol{u}|\boldsymbol{W}_o$ as follows

$$
\begin{aligned}
f(\boldsymbol{u}|\boldsymbol{W}_o) &\propto \exp\left[-\frac{1}{2\gamma^2}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u})^\mathsf{T}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{u})\right] \times \exp\left[-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u} - \boldsymbol{R}\boldsymbol{\beta})\right] \\
&\propto \exp\left\{ -\frac{1}{2}\left[\boldsymbol{u}^\mathsf{T}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{G}_o\boldsymbol{V} + \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u} - \boldsymbol{u}^\mathsf{T}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{W}_o + \boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\beta}\right)\right.\right. \\
&\qquad \left.\left. - \left(\frac{1}{\gamma^2}\boldsymbol{W}_o^\mathsf{T}\boldsymbol{G}_o\boldsymbol{V} + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u}\right]\right\}.
\end{aligned}
$$

Thus the distribution of $\boldsymbol{u}|\boldsymbol{W}_o$ is $\mathcal{N}(\boldsymbol{M}, \boldsymbol{S})$, where:

$$
\begin{aligned}
\boldsymbol{S} &= \left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{G}_o\boldsymbol{V} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \\
\boldsymbol{M} &= \boldsymbol{S}\left(\frac{1}{\gamma^2}\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{W}_o + \boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\beta}\right).
\end{aligned}
$$

Hence the expectation of ℓ_c with respect to the distribution of $\boldsymbol{u}|\boldsymbol{W}_o, \tilde{\Theta}$ is:

$$
\begin{aligned}
E\left[\ell_c \mid \boldsymbol{W}_o, \tilde{\Theta}\right] = &-\frac{N_o}{2}\log\gamma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\gamma^2}\mathrm{Tr}(\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{V}\tilde{\boldsymbol{S}}) \\
&- \frac{1}{2\gamma^2}(\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}_o)^\mathsf{T}(\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}_o) \\
&- \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) - \frac{1}{2}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta}) \qquad (2)
\end{aligned}
$$

The solutions for $\gamma^2, \boldsymbol{\beta}$, and $\boldsymbol{\xi}$ that maximize (2), given current estimates, are relatively similar to the complete case:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{M}}, \\
\hat{\boldsymbol{\xi}} &= \arg\max_{\boldsymbol{\xi}}\left\{-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) - \frac{1}{2}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{M}} - \boldsymbol{R}\boldsymbol{\beta})\right\}, \\
\hat{\gamma}^2 &= \frac{1}{N_o}\left[\mathrm{Tr}(\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{S}}) + \left\|\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}_o\right\|_2^2\right].
\end{aligned}
$$

To solve for $\boldsymbol{v}$, we maximize a slightly different function,

$$\begin{aligned}
h(\boldsymbol{v}) &= -\frac{1}{2\gamma^2}\mathrm{Tr}(\boldsymbol{V}^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{G}_o^\mathsf{T}\boldsymbol{V}\tilde{\boldsymbol{S}}) - \frac{1}{2\gamma^2}(\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}_o)^\mathsf{T}(\boldsymbol{G}_o\boldsymbol{V}\tilde{\boldsymbol{M}} - \boldsymbol{W}_o) \\
&= -\frac{1}{2\gamma^2}\left[\mathrm{Tr}\left(\begin{bmatrix} \sum_{j\in\Omega_{1.}} v_j^2 & 0 & ... & 0 \\ 0 & \sum_{j\in\Omega_{2.}} v_j^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sum_{j\in\Omega_{n.}} v_j^2 \end{bmatrix}\tilde{\boldsymbol{S}}\right) + \sum_{i=1}^{n}\sum_{j\in\Omega_{i.}}\left(X_{ij} - \tilde{M}_i v_j\right)^2\right] \\
&= -\frac{1}{2\gamma^2}\left[\sum_{i=1}^{n}\sum_{j\in\Omega_{i.}}\tilde{S}_{ii}v_j^2 + \sum_{i=1}^{n}\sum_{j\in\Omega_{i.}}\left(X_{ij} - \tilde{M}_i v_j\right)^2\right] \\
&= -\frac{1}{2\gamma^2}\sum_{i=1}^{n}\sum_{j\in\Omega_{i.}}\left[\tilde{S}_{ii}v_j^2 + \left(X_{ij} - \tilde{M}_i v_j\right)^2\right] \\
&= -\frac{1}{2\gamma^2}\sum_{i=1}^{n}\left[\tilde{S}_{ii}v_j^2 + \left(X_{ij} - \tilde{M}_i v_j\right)^2\right]\mathbf{1}_{[j\in\Omega_{i.}]}.
\end{aligned}$$

Here $\Omega_{i.}$ denotes the set of observed elements across the i-th row of $\boldsymbol{X}$ and $\mathbf{1}_{[.]}$ denotes the indicator function. Taking derivative of $h(\boldsymbol{v})$ with respect to each v_j and setting it equal to zero, we can find the closed-form unscaled solution

$$\breve{v}_j = \frac{\sum_{i=1}^{n} X_{ij}\tilde{M}_i\mathbf{1}_{[j\in\Omega_{i.}]}}{\sum_{i=1}^{n}\left(\tilde{S}_{ii} + \tilde{M}_i^2\right)\mathbf{1}_{[j\in\Omega_{i.}]}}.$$

The final solution for $\boldsymbol{v}$ is then $\hat{\boldsymbol{v}} = \frac{\breve{\boldsymbol{v}}}{\|\breve{\boldsymbol{v}}\|_2}$, where $\breve{\boldsymbol{v}} = (\breve{v}_1, ..., \breve{v}_p)$.

When some elements of the exposure data are missing, parameter estimation for each PC is based only on the observed elements $\boldsymbol{W}_o$. Estimate for PC score can then be made by projecting the model-based imputed exposure data onto the direction of $\boldsymbol{v}$. The joint distribution of $\boldsymbol{W}_o$ and $\boldsymbol{W}_m$ can be written as

$$\begin{bmatrix}\boldsymbol{W}_o \\ \boldsymbol{W}_m\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_o \\ \mathbf{m}_m\end{bmatrix}, \begin{bmatrix}\mathbf{C}_{oo} & \mathbf{C}_{om} \\ \mathbf{C}_{mo} & \mathbf{C}_{mm}\end{bmatrix}\right) = \mathcal{N}(\boldsymbol{M}, \mathbf{C})$$

where $\boldsymbol{M} = \boldsymbol{G}\boldsymbol{V}\boldsymbol{R}\boldsymbol{\beta}$ and $\mathbf{C} = \gamma^2\boldsymbol{G}\boldsymbol{G}^\mathsf{T} + \boldsymbol{G}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}\boldsymbol{G}^\mathsf{T}$. The missing elements, $\boldsymbol{W}_m$, can then be imputed by the conditional mean,

$$E(\boldsymbol{W}_m \mid \boldsymbol{W}_o) = \mathbf{m}_m + \mathbf{C}_{mo}\mathbf{C}_{oo}^{-1}(\boldsymbol{W}_o - \mathbf{m}_o).$$

Thus, the parameter estimation of ProPrPCA-Krige with missing monitoring data can be summarized as:

Algorithm: ProPrPCA-Krige with missing monitoring data

Input $\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{G}_o$ q, and t_{max}

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

$\boldsymbol{X}_l \leftarrow \boldsymbol{X}_{l-1}^{zero} - \hat{\boldsymbol{u}}_{l-1}^{zero}\hat{\boldsymbol{v}}_{l-1}^{\mathsf{T}}$ where $\boldsymbol{X}_0^{zero} = \boldsymbol{X}$ imputed with zeros, $\hat{\boldsymbol{u}}_0^{zero} = \boldsymbol{0}$, and $\hat{\boldsymbol{v}}_0 = \boldsymbol{0}$

$\boldsymbol{W}_o \leftarrow \boldsymbol{G}_o \text{vec}(\boldsymbol{X}_l)$

Initialize $\boldsymbol{v}_l^{(0)}$, $(\gamma_l^{(0)})^2$, $\boldsymbol{\beta}_l^{(0)}$, $\boldsymbol{\xi}_l^{(0)}$, and $t = 1$

$\boldsymbol{\Sigma}_l^{(0)} \leftarrow \boldsymbol{\Sigma}(\boldsymbol{\xi}_l^{(0)})$

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

$\tilde{\boldsymbol{S}}_l \leftarrow \left[(\gamma_l^{(t)})^{-2}\boldsymbol{V}_l^{(t)\mathsf{T}}\boldsymbol{G}_o^{\mathsf{T}}\boldsymbol{G}_o\boldsymbol{V}_l^{(t)} + (\boldsymbol{\Sigma}_l^{(t)})^{-1}\right]^{-1}$ where $\boldsymbol{V}_l^{(t)} = \boldsymbol{I}_n \otimes \boldsymbol{v}_l^{(t)}$

$\tilde{\boldsymbol{M}}_l \leftarrow \tilde{\boldsymbol{S}}_l\left[(\gamma_l^{(t)})^{-2}\boldsymbol{V}_l^{(t)\mathsf{T}}\boldsymbol{G}_o^{\mathsf{T}}\boldsymbol{W}_o + (\boldsymbol{\Sigma}_l^{(t)})^{-1}\boldsymbol{R}\boldsymbol{\beta}_l^{(t)}\right]$

$\boldsymbol{v}_l^{(t+1)} \leftarrow \tilde{\boldsymbol{v}}_l/\|\tilde{\boldsymbol{v}}_l\|_2$ where the $j-$th element of $\tilde{\boldsymbol{v}}_l$ (for $j = 1, ..., p$) is calculated as:

$$\frac{\sum_{i=1}^{n}(\boldsymbol{X}_l)_{ij}(\tilde{\boldsymbol{M}}_l)_i\boldsymbol{1}_{[j\in\Omega_{i.}]}}{\sum_{i=1}^{n}\left((\tilde{\boldsymbol{S}}_l)_{ii} + (\tilde{\boldsymbol{M}}_l)_i^2\right)\boldsymbol{1}_{[j\in\Omega_{i.}]}}$$

$(\gamma_l^{(t+1)})^2 \leftarrow (N_o)^{-1}\left[\text{Tr}(\boldsymbol{V}_l^{(t+1)\mathsf{T}}\boldsymbol{G}_o^{\mathsf{T}}\boldsymbol{G}_o\boldsymbol{V}_l^{(t+1)}\tilde{\boldsymbol{S}}_l) + \left\|\boldsymbol{V}_l^{(t+1)\mathsf{T}}\boldsymbol{G}_o^{\mathsf{T}}\tilde{\boldsymbol{M}}_l - \boldsymbol{W}_o\right\|_2^2\right]$

where $\boldsymbol{V}_l^{(t+1)} = \boldsymbol{I}_n \otimes \boldsymbol{v}_l^{(t+1)}$

$\boldsymbol{\xi}_l^{(t+1)} \leftarrow \underset{\boldsymbol{\xi}_l}{\arg\max}\{-\log|\boldsymbol{\Sigma}_l| - \text{Tr}\left(\boldsymbol{\Sigma}_l^{-1}\tilde{\boldsymbol{S}}_l\right) - (\tilde{\boldsymbol{M}}_l - \boldsymbol{R}\boldsymbol{\beta}_l^{(t)})^{\mathsf{T}}\boldsymbol{\Sigma}_l^{-1}(\tilde{\boldsymbol{M}}_l - \boldsymbol{R}\boldsymbol{\beta}_l^{(t)})\}$

where $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}(\boldsymbol{\xi}_l)$

$\boldsymbol{\beta}_l^{(t+1)} \leftarrow \left(\boldsymbol{R}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}_l^{(t+1)})^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\xi}_l^{(t+1)})^{-1}\tilde{\boldsymbol{M}}_l$

$t \leftarrow t + 1$

end while

$\hat{\boldsymbol{v}}_l \leftarrow \boldsymbol{v}_l^{(t)}$, $\hat{\gamma}_l^2 \leftarrow (\gamma_l^{(t)})^2$, $\hat{\boldsymbol{\beta}}_l \leftarrow \boldsymbol{\beta}_l^{(t)}$, $\hat{\boldsymbol{\xi}}_l \leftarrow \boldsymbol{\xi}_l^{(t)}$

$\boldsymbol{X}_l^{zero} \leftarrow \boldsymbol{X}_l$ with elements at missing indices replaced with zero

$\boldsymbol{X}_l^{imp} \leftarrow \boldsymbol{X}_l$ with elements at missing indices replaced with conditional means

$\hat{\boldsymbol{u}}_l^{zero} = \boldsymbol{X}_l^{zero}\hat{\boldsymbol{v}}_l$

$\hat{\boldsymbol{u}}_l = \boldsymbol{X}_l^{imp}\hat{\boldsymbol{v}}_l$

end for

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

2 The ProPrPCA-Spline model and algorithm

2.1 The model

For each PC, the ProPrPCA-Spline algorithm assumes the following model

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\beta}\boldsymbol{v}^\mathsf{T} + \boldsymbol{E}$$

Here $\boldsymbol{Z}$ contains both geographic covariates and the thin-plate spline basis functions. The collection of model parameters Θ now includes $\{\boldsymbol{v}, \boldsymbol{\beta}, \gamma^2\}$. Using the same vectorization established in previous section, the model assumes $\boldsymbol{W}_i = \boldsymbol{X}_{i.} = (\boldsymbol{Z}\boldsymbol{\beta})_i \boldsymbol{v} + \boldsymbol{\epsilon}_i$, for $i = 1, ..., n$. We can then rewrite this model in the vectorized form as

$$\boldsymbol{W} \mid \Theta \sim \mathcal{N}\left(\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}, \gamma^2 \boldsymbol{I}_N\right),$$

where $\boldsymbol{V} = \boldsymbol{I}_n \otimes \boldsymbol{v}$. When there are missing data, the distribution of interest becomes

$$\boldsymbol{W}_o \mid \Theta \sim \mathcal{N}\left(\boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}, \gamma^2 \boldsymbol{I}_{N_o}\right).$$

2.2 Estimation of model parameters when monitoring data is complete

To solve for the parameters, we maximize the log-likelihood (up to a constant) directly:

$$\ell(\Theta \mid \boldsymbol{W}) = -\frac{1}{N}\log\gamma^2 - \frac{1}{2\gamma^2}\left(\boldsymbol{W} - \boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)^\mathsf{T}\left(\boldsymbol{W} - \boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)$$

To solve for $\boldsymbol{v}$, we effectively maximize the following function:

$$-\frac{1}{2\gamma^2}\left(\boldsymbol{W} - \boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)^\mathsf{T}\left(\boldsymbol{W} - \boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)$$

Denote $\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{\beta} \in \mathbb{R}^{n\times 1}$, we can rewrite this function similarly to the function involved $\boldsymbol{v}$ with complete data for ProPrPCA-Krige. Thus, the solution for $\boldsymbol{v}$ can be written in closed-form as

$$\hat{\boldsymbol{v}} = \frac{\breve{\boldsymbol{v}}}{\|\breve{\boldsymbol{v}}\|_2}, \qquad \text{where } \breve{\boldsymbol{v}} = \frac{\boldsymbol{X}^\mathsf{T}\boldsymbol{K}}{\|\boldsymbol{K}\|_2^2} = \frac{\boldsymbol{X}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\beta}}{\|\boldsymbol{Z}\boldsymbol{\beta}\|_2^2}.$$

The closed-form solution for $\boldsymbol{\beta}$ is straightforwardly a result of ordinary least squares, $\hat{\boldsymbol{\beta}} = \left[(\boldsymbol{V}\boldsymbol{Z})^\mathsf{T}(\boldsymbol{V}\boldsymbol{Z})\right]^{-1}(\boldsymbol{V}\boldsymbol{Z})^\mathsf{T}\boldsymbol{W}$. This can be further simplified thanks to the constraint on $\boldsymbol{v}$ and noticing that $\boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{I}_n$. Thus we have, $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\right)^{-1}(\boldsymbol{Z}\otimes\boldsymbol{v})^\mathsf{T}\boldsymbol{W}$. Finally, the solution for γ^2 is simply $\hat{\gamma}^2 = N^{-1}\left\|\boldsymbol{W} - \boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2$. Thus parameter estimation of ProPrPCA-Spline with complete monitoring data can be summarized as:

Algorithm: ProPrPCA-Spline with complete monitoring data

Input $\boldsymbol{X}$, $\boldsymbol{Z}$, q, and t_{max}
  for l in $\{1, ..., q\}$ **do**
    $\boldsymbol{X}_l \leftarrow \boldsymbol{X}_{l-1} - \hat{\boldsymbol{u}}_{l-1}\hat{\boldsymbol{v}}_{l-1}^{\mathsf{T}}$ where $\boldsymbol{X}_0 = \boldsymbol{X}$, $\hat{\boldsymbol{u}}_0 = \boldsymbol{0}$, and $\hat{\boldsymbol{v}}_0 = \boldsymbol{0}$
    Initialize $\boldsymbol{v}_l^{(0)}$, $(\gamma_l^{(0)})^2$, $\boldsymbol{\beta}_l^{(0)}$, and $t = 1$
      while not converged **or** $t < t_{max}$ **do**
        $\boldsymbol{v}_l^{(t+1)} \leftarrow \tilde{\boldsymbol{v}}_l / \|\tilde{\boldsymbol{v}}_l\|_2$ where $\tilde{\boldsymbol{v}}_l \leftarrow \boldsymbol{X}_l^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\beta}_l^{(t)} / \left\|\boldsymbol{Z}\boldsymbol{\beta}_l^{(t)}\right\|_2^2$
        $\boldsymbol{\beta}_l^{(t+1)} \leftarrow \left(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z} \otimes \boldsymbol{v}_l^{(t+1)}\right)^{\mathsf{T}} \mathrm{vec}(\boldsymbol{X}_l)$
        $(\gamma_l^{(t+1)})^2 \leftarrow (np)^{-1}\left\|\mathrm{vec}(\boldsymbol{X}_l) - (\boldsymbol{I}_n \otimes \boldsymbol{v}_l^{(t+1)})\boldsymbol{Z}\boldsymbol{\beta}_l^{(t+1)}\right\|_2^2$
        $t \leftarrow t + 1$
      end while
    $\hat{\boldsymbol{v}}_l \leftarrow \boldsymbol{v}_l^{(t)}$, $\hat{\gamma}_l^2 \leftarrow (\gamma_l^{(t)})^2$, $\hat{\boldsymbol{\beta}}_l \leftarrow \boldsymbol{\beta}_l^{(t)}$
    $\hat{\boldsymbol{u}}_l = \boldsymbol{X}_l\hat{\boldsymbol{v}}_l$
  end for
Output $\{\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_q\}$, $\{\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_q\}$, $\{\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_q\}$, $\{\hat{\gamma}_1^2, ..., \hat{\gamma}_q^2\}$

2.3 Parameter estimation and model-based imputation with missing monitoring data

The observed log-likelihood with missing monitoring data is

$$\ell(\Theta \mid \boldsymbol{W}_o) = -\frac{1}{N_o}\log\gamma^2 - \frac{1}{2\gamma^2}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right).$$

The solutions for $\boldsymbol{\beta}$ and γ^2 are trivial and fairly similar to those with complete data. To solve for $\boldsymbol{v}$, we maximize $\{-\frac{1}{2\gamma^2}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta})\}$. We can rewrite the function of $\boldsymbol{v}$ as

$$\begin{aligned}
&-\frac{1}{2\gamma^2}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{Z}\boldsymbol{\beta}\right)\\
=&-\frac{1}{2\gamma^2}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{K}\right)^{\mathsf{T}}\left(\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}\boldsymbol{K}\right)\\
=&\sum_{i=1}^{n}\sum_{j\in\Omega_{i.}}\left(X_{ij} - K_i v_j\right)^2 = \sum_{i=1}^{n}\left(X_{ij} - K_i v_j\right)^2 \mathbf{1}_{[j\in\Omega_{i.}]},
\end{aligned}$$

Taking derivative with respect to each v_j and setting it equal to zero, we can find the closed-form unscaled solution

$$\breve{v} = \frac{\sum_{i=1}^{n} X_{ij}K_i\mathbf{1}_{[j\in\Omega_{i.}]}}{\sum_{i=1}^{n} K_i^2\mathbf{1}_{[j\in\Omega_{i.}]}}.$$

The final solution for $\boldsymbol{v}$ is then $\hat{\boldsymbol{v}} = \frac{\breve{\boldsymbol{v}}}{\|\breve{\boldsymbol{v}}\|_2}$, where $\breve{\boldsymbol{v}} = (\breve{v}_1, ..., \breve{v}_p)$.

When some elements of the exposure data are missing, parameter estimation for each PC is based only on the observed elements $\boldsymbol{W}_o$. Estimate for PC score can then be made by projecting the model-based imputed exposure data onto the direction of $\boldsymbol{v}$. The missing elements, $\boldsymbol{W}_m$, can then be imputed by its estimate $\boldsymbol{G}_m\hat{\boldsymbol{V}}\boldsymbol{Z}\hat{\boldsymbol{\beta}}$. Thus the parameter estimation of ProPrPCA-Spline with missing monitoring data can be summarized as:

Algorithm: ProPrPCA-Spline with missing monitoring data

Input $\boldsymbol{X}$, $\boldsymbol{G}_o$ $\boldsymbol{Z}$, q, and t_{max}

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

$\boldsymbol{X}_l \leftarrow \boldsymbol{X}_{l-1}^{zero} - \hat{\boldsymbol{u}}_{l-1}^{zero}\hat{\boldsymbol{v}}_{l-1}^{\mathsf{T}}$ where $\boldsymbol{X}_0^{zero} = \boldsymbol{X}$ imputed with zeros, $\hat{\boldsymbol{u}}_0^{zero} = \boldsymbol{0}$, and $\hat{\boldsymbol{v}}_0 = \boldsymbol{0}$

$\boldsymbol{W}_o \leftarrow \boldsymbol{G}_o \text{vec}(\boldsymbol{X}_l)$

Initialize $\boldsymbol{v}_l^{(0)}$, $(\gamma_l^{(0)})^2$, $\boldsymbol{\beta}_l^{(0)}$, and $t = 1$

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

$\boldsymbol{v}_l^{(t+1)} \leftarrow \tilde{\boldsymbol{v}}_l / \|\tilde{\boldsymbol{v}}_l\|_2$ where the $j-$element of $\tilde{\boldsymbol{v}}_l$ for $j = 1, ..., p$ is calculated as:

$$\frac{\sum_{i=1}^{n}(\boldsymbol{X}_l)_{ij}(\boldsymbol{K}_l)_i \mathbf{1}_{[j\in\Omega_{i.}]}}{\sum_{i=1}^{n}(\boldsymbol{K}_l)_i^2 \mathbf{1}_{[j\in\Omega_{i.}]}}, \text{ and } \boldsymbol{K}_l = \boldsymbol{Z}\boldsymbol{\beta}_l^{(t)}$$

$$\boldsymbol{\beta}_l^{(t+1)} \leftarrow \left[\left(\boldsymbol{G}_o\boldsymbol{V}_l^{(t+1)}\boldsymbol{Z}\right)^{\mathsf{T}}\left(\boldsymbol{G}_o\boldsymbol{V}_l^{(t+1)}\boldsymbol{Z}\right)\right]^{-1}\left(\boldsymbol{G}_o\boldsymbol{V}_l^{(t+1)}\boldsymbol{Z}\right)^{\mathsf{T}}\boldsymbol{W}_o$$

where $\boldsymbol{V}_l^{(t+1)} = \boldsymbol{I}_n \otimes \boldsymbol{v}_l^{(t+1)}$

$(\gamma_l^{(t+1)})^2 \leftarrow {N_o}^{-1}\left\|\boldsymbol{W}_o - \boldsymbol{G}_o\boldsymbol{V}_l^{(t+1)}\boldsymbol{Z}\boldsymbol{\beta}_l^{(t+1)}\right\|_2^2$

$t \leftarrow t + 1$

end while

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

$\boldsymbol{X}_l^{zero} \leftarrow \boldsymbol{X}_l$ with elements at missing indices replaced with zero

$\boldsymbol{X}_l^{imp} \leftarrow \boldsymbol{X}_l$ with elements at missing indices replaced with conditional means

$\hat{\boldsymbol{u}}_l^{zero} = \boldsymbol{X}_l^{zero}\hat{\boldsymbol{v}}_l$

$\hat{\boldsymbol{u}}_l = \boldsymbol{X}_l^{imp}\hat{\boldsymbol{v}}_l$

end for

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

3 Simulations

3.1 Data generating mechanism for high-dimensional simulations

To further demonstrate the performance of ProPrPCA, we simulate multi-pollutant exposure surfaces with $p = 15$. We first generate three underlying PC scores on the 100×100 grid ($N = 10,000$), such that

$$\begin{aligned}
&\boldsymbol{u}_j \sim \mathcal{N}\left(\boldsymbol{R}_j \boldsymbol{b}_j, \boldsymbol{S}_j\right), \quad \text{where } j = 1, 2, 3, \\
&\boldsymbol{R}_1 = \begin{bmatrix} \boldsymbol{r}_{1o} & \boldsymbol{r}_{1u} \end{bmatrix}, \quad \text{where } \boldsymbol{r}_{1o}, \boldsymbol{r}_{1u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}), \ \boldsymbol{b}_1^\mathsf{T} = \begin{bmatrix} 5 & 1 \end{bmatrix}, \\
&\boldsymbol{R}_2 = \begin{bmatrix} \boldsymbol{r}_{2o} & \boldsymbol{r}_{2u} \end{bmatrix}, \quad \text{where } \boldsymbol{r}_{2o}, \boldsymbol{r}_{2u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}), \ \boldsymbol{b}_2^\mathsf{T} = \begin{bmatrix} 5 & 2 \end{bmatrix}, \\
&\boldsymbol{R}_3 = \begin{bmatrix} \boldsymbol{r}_{3u} \end{bmatrix}, \quad \text{where } \boldsymbol{r}_{3u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}), \ \boldsymbol{b}_3 = 1.
\end{aligned}$$

In this setting, $\boldsymbol{r}_{jo}$'s are GIS covariates observed for the model, while $\boldsymbol{r}_{ju}$'s are unobserved covariates, and used primarily to generate the scores themselves. That is, only $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{r}_{1o} & \boldsymbol{r}_{2o} \end{bmatrix}$ is used in the spatial prediction model. Here $\boldsymbol{S}_1$ has exponential structure with no nugget effect, partial sill of 5, and range of 50. Meanwhile, $\boldsymbol{S}_2 = 7.5\boldsymbol{I}_N$ and $\boldsymbol{S}_3 = 2\boldsymbol{I}_N$. This setup is created so that $\boldsymbol{u}_1$ is the most spatially predictable, $\boldsymbol{u}_2$ is moderately predictable in space, and $\boldsymbol{u}_3$ is not spatially predictable. Here spatial predictability refers to how well the quantity can be predicted at new locations using relevant and available covariates.

We then create two scenarios in which we scale the variance of $\boldsymbol{u}_j$'s differently,

$$\begin{aligned}
&\text{Scenario 1: } Var(\boldsymbol{u}_1) = 10, Var(\boldsymbol{u}_2) = 7.5, Var(\boldsymbol{u}_3) = 5, \\
&\text{Scenario 2: } Var(\boldsymbol{u}_1) = 7.5, Var(\boldsymbol{u}_2) = 5, Var(\boldsymbol{u}_3) = 10.
\end{aligned}$$

In both scenarios, the multi-pollutant exposure surface is generated as

$$\begin{aligned}
&\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V} + \boldsymbol{E}, \quad \text{where } E_{ij} \sim \mathcal{N}(0, 1) \\
&\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{bmatrix}, \quad \text{where } \boldsymbol{v}_j = \frac{\breve{\boldsymbol{v}}_j}{\left\| \breve{\boldsymbol{v}}_j \right\|_2}, \quad \text{for } j = 1, 2, 3 \\
&\breve{\boldsymbol{v}}_1^\mathsf{T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
&\breve{\boldsymbol{v}}_2^\mathsf{T} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
&\breve{\boldsymbol{v}}_3^\mathsf{T} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}
\end{aligned}$$

The use of such sparse loadings is to clearly identify the behavior of each dimension reduction method. Because of the variance contribution setup, in scenario A we expect all three methods to pick $\boldsymbol{u}_1$ as PC1, $\boldsymbol{u}_2$ as PC2, and $\boldsymbol{u}_3$ as PC3. In scenario B, however, we expect TradPCA to pick $\boldsymbol{u}_3$ as PC1, $\boldsymbol{u}_1$ as PC2, and $\boldsymbol{u}_2$ as PC3, as $\boldsymbol{u}_3$ has the largest variance contribution. Meanwhile, PredPCA and ProPrPCA-Spline will still pick $\boldsymbol{u}_1$ as PC1, $\boldsymbol{u}_2$ as PC2.

For these high-dimensional simulations, we consider three MCAR scenarios (30%, 35%, and 40%), and one MAR scenario. In the MAR scenario, we identify training locations with

$\boldsymbol{r}_{1o}$ value larger than its sample 60th percentile, and among $\boldsymbol{x}_1$ through $\boldsymbol{x}_5$, 75% of these training locations become missing data. For the rest of the pollutants, from $\boldsymbol{x}_6$ to $\boldsymbol{x}_{15}$, each has 25% of its locations missing completely at random. This setup guarantees a mild spatial pattern in the missing data, as $\boldsymbol{x}_1$ to $\boldsymbol{x}_5$ are generated entirely by $\boldsymbol{u}_1$, which is the most predictable score based on $\boldsymbol{r}_{1o}$.

3.2 Evaluation of computational burden

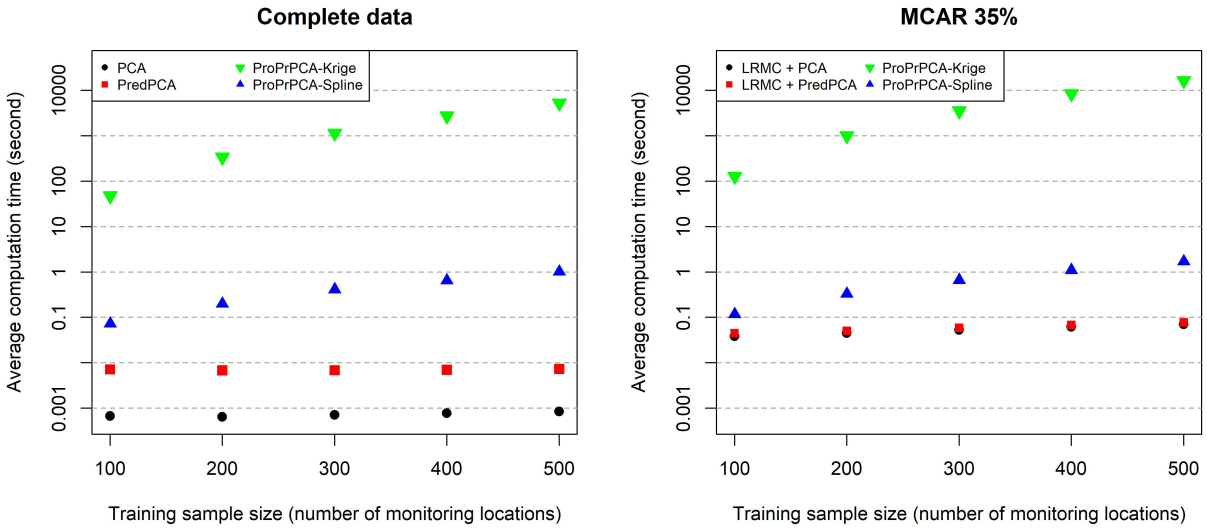

Figure 1: Computational time (average over 1,000 simulations under high-dimensional scenario 1) of PCA, PredPCA, ProPrPCA-Krige and ProPrPCA-Spline, with complete and MCAR 35% missing data by training sample size.

Figure 1 compares the computation burden among PCA, PredPCA, ProPrPCA-Krige, and ProPrPCA-Spline as the sample size increases. The results were averaged over 1,000 simulations under high-dimensional scenario 1, for complete data and MCAR 35% scenarios. The results were mostly under 1 second, on average, for PCA, PredPCA, and the Spline model. The computational burden of the Krige model is exponentially larger than the rest. While there maybe computational and programming tricks to alleviate the time cost, the Krige model would be likely to still take longer than other methods, given the nature of difficult optimization and EM algorithm.

4 Data Application

Table 1: Prediction R^2's from leave-one-site-out cross-validation on 2010 CSN data. Sites with complete $PM_{2.5}$ component data are used as test data. Training data may include only complete sites (complete training data), or all available sites (full training data). The PCs are defined in the order of which they are obtained, i.e. PC1, PC2, PC3 are the first, second, and third PCs returned by the methods, respectively.

	PC1	PC2	PC3
PCA (complete training data)	0.24	0.51	0.51
PredPCA (complete training data)	0.52	0.44	0.62
LRMC + PredPCA (full training data)	0.54	0.53	0.45
ProPrPCA-Spline (full training data)	0.57	0.35	0.69

In this section, we evaluate the predictive performance in leave-one-site-out cross-validations. Table 1 shows the results when the PCs are simply defined in the order of which they are obtained. As discussed in the main text of the manuscript, while having decent performance for PC2 and PC3, using PCA applied to the complete training data yields a poor result for PC1. PredPCA has similar performances for PC1 with either complete or full training data. However, there is a 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 and PC3, but suffers from a decrease in the ability to predict PC2 well.

Table 2: Prediction R^2's from leave-one-site-out cross-validation on 2010 CSN data. Sites with complete $PM_{2.5}$ component data are used as test data. Training data may include only complete sites (complete training data), or all available sites (full training data). The PCs are defined in the order of variance explained in the training data, i.e. PC1 is the component with the largest variance contributed, and so on.

	PC1	PC2	PC3
PCA (complete training data)	0.24	0.51	0.51
PredPCA (complete training data)	0.52	0.44	0.62
LRMC + PredPCA (full training data)	0.54	0.45	0.53
ProPrPCA-Spline (full training data)	0.57	0.41	0.65

Similar to Shen and Huang (2008), we also order the PCs by the variance explained in the training data, as given in Table 2. That is, in each round of the cross-validation procedure, out of the three PCs obtained by each PCA method, PC1 is defined as the component with the largest variance explained in the training data, and so on for PC2 and PC3. For spatially predictive methods, especially PredPCA, the order of variance explain was not necessary the same as the order by which the PCs were produced. Thus, while the results for PCA (and PredPCA) with complete training data were the same in both

Tables, the results with full training data were slightly different for the last two PCs. With this reordering approach, the results were more similar for PC2, and there was an apparent advantage of using ProPrPCA-Spline compared to PredPCA after imputation.

References

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

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