A novel dimension reduction approach for spatially-misaligned multivariate air pollution data

Roman Jandarov

CLARC WIP Webinar

February 2014

Objective

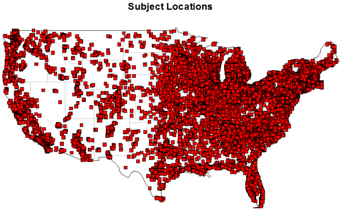
- Identify important spatially varying air pollution mixtures and quantify long-term health effects in cohort study data
- Initial application:
  - Cohort: NIEHS Sister Study
  - Health endpoint: Systolic blood pressure
  - Pollution data: Annual average concentration of PM2.5 components from national CSN and IMPROVE networks
- Future planned application:
  - Cohort: MESA Air
  - Pollution data: Mobile monitoring data from UW CCAR

Challenges

- Dimension reduction
  - Problem: Difficult to fit health model and interpret coefficients with multi-pollutant exposure (e.g. 15-20 components of PM2.5)
  - Solution: Principal component analysis
- Spatial misalignment
  - Problem: Concentration data is available at monitor locations but not where study subjects live
  - Solution: Predict exposures at subject locations using spatial prediction model that incorporates geographic covariates and spatial smoothing
- Solving two challenges above together

NIEHS Sister Study

- Y - blood pressure
- Data on Y and subject-specific covariates from NIEHS Sister Study data (cohort study on risk factors for breast cancer)
  - $> 50,000$ sisters of women with breast cancer
  - Statistically significant association between PM2.5 exposure and Y (Chan et al. (under review))



Need for dimension reduction

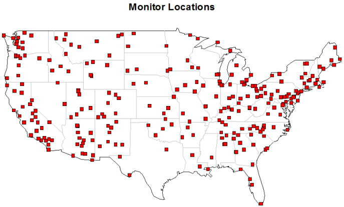
- Dimensionality of multi-pollutant data
  - General health model is not practical

$$Y = \alpha_0 + \sum_{l=1}^{m} \alpha_l \widehat{P}_l + \text{interactions} + \text{covariates} + \ldots$$

  - Pollutant concentrations are potentiality correlated
  - Large number of main effects and interactions: hard to estimate and interpret

Monitor locations and covariates

- 17 pollutants and 277 monitors (CSN and IMPROVE networks)
  - **PM25, EC, OC, Al, As, Br, Ca, Cr, Cu, Fe, K, Mn, Na, S, Si, V, Zn**
  - Annual averages from November, 2009 to October, 2010.



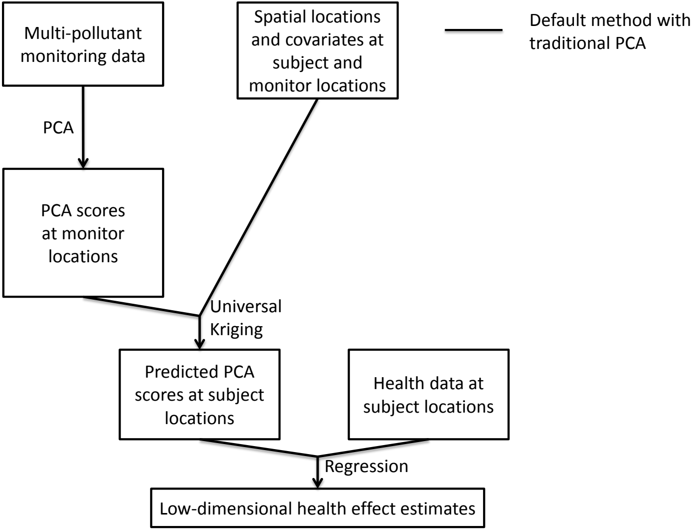
- GIS covariates from MESA Air geographic database
  - Let **Z** be a matrix of transformed geographical covariates and thin-plate spline basis functions
  - Available at all monitor and subject locations

A possible solution: Sequential approach

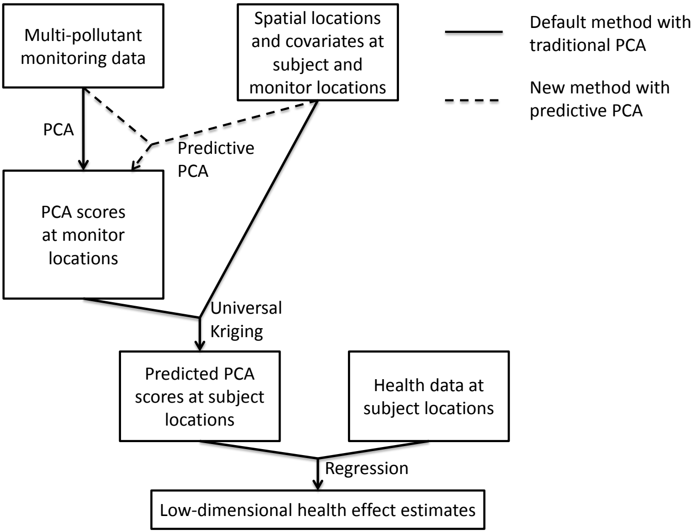
1. Dimension reduction
   - Compute first few principal components of multi-pollutant data
2. Predict scores obtained from principal components at participant locations using GIS covariates and splines
3. Fit a health model with smaller number of variables
   - Interpret coefficient of the model to identify important mixtures

This talk: Steps 1 and 2: Dimension reduction and prediction

Combining PCA with spatial prediction



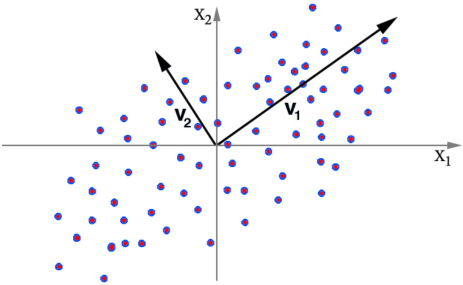
Combining PCA with spatial prediction



Review of principal component analysis

- Principal component analysis (PCA) is a popular dimension reduction technique
- A version of unsupervised learning
- Goal of PCA: Reduce the number of variables of interest into a smaller set of component scores
- PCA transforms the original variables into a set of component scores (linear combinations of originals) equal to the number of original variables

PCA: Example



- Let $\mathbf{X}$ be a $n \times p$ matrix with standardized columns
- PCA finds direction $\mathbf{v}_1$ and $\mathbf{v}_2$ (also called loadings)
- Principal component scores: PC1 = $\mathbf{X}\mathbf{v}_1$, PC2 = $\mathbf{X}\mathbf{v}_2$

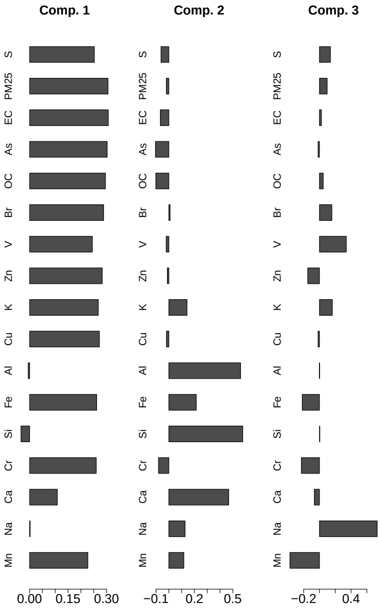
PCA algorithm

- First, find $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$, s.t. $\|\widetilde{\mathbf{u}}\| = 1$ that minimizes

$$\|\mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T\|_F$$

  Define PC loadings by $\mathbf{v} = \widetilde{\mathbf{v}}/\|\widetilde{\mathbf{v}}\|$. Define PC1 by $\mathbf{u} = \mathbf{X}\mathbf{v}$.
- Subsequently, find $(\widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$ by approximating the corresponding residual matrices. Define corresponding PC scores and loadings.

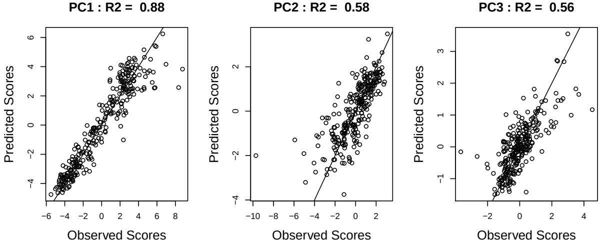
Application of PCA to PM2.5 data: Loadings



Spatial prediction for scores

- Let: $u_{(s)}$ - value of PC score at location s (available only at monitor locations)
- Let: $z_{(s)}$ - vector of geographical covariates at location s (available at all locations)
- Goal: Predict $u_{(s)}$ at subject locations
- Universal Kriging model:
  - $u_{(s)} = \mu_{(s)} + \epsilon_{(s)}$, where $\mu_{(s)} = \mu(z_{(s)})$
  - $\{\epsilon_{(s)}\}$ is a Gaussian process with mean 0 and spatial covariance function $c = c()$
  - After estimates of $\mu()$ and $c()$ are obtained from data at monitor locations, one can predict $u_{(s^*)}$ at new locations s^* using $z_{(s^*)}$

Application of PCA to PM2.5 data: Predictability

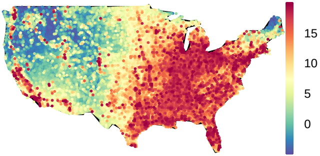


Observed scores: PC scores calculated at monitor locations with known pollutants X and fixed loadings v

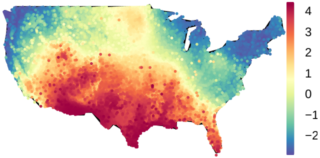
Predicted scores: Predictions of PC scores at locations where X is unknown

Application of PCA to PM2.5 data: Heat maps

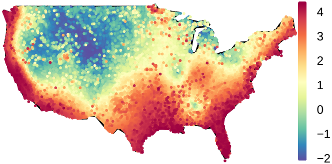
Predictions of PC1: R^2 = 0.88



Predictions of PC2: R^2 = 0.58



Predictions of PC3: R^2 = 0.56



Motivation for a new PCA approach

- Can we improve predictability of principal component scores?
- Can we simplify interpretability of the component loadings?

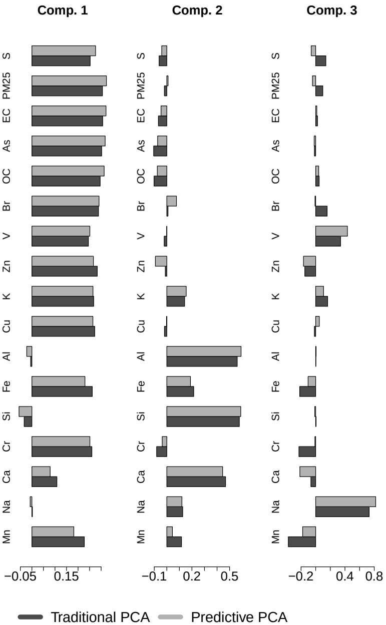
New approach: Idea

- Focus on predictability of principal component scores first
- We want a PCA algorithm that results in PC scores that can be predicted well
  - Develop an algorithm that forces PC scores to be close to spatial covariates
- Work with interpretability by adding a penalty to component loadings later

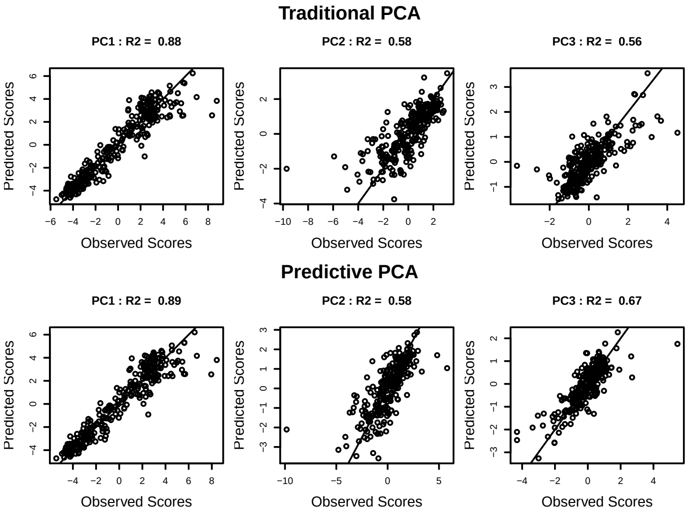
Motivation: Predictive PCA

- Recall: **Z** - matrix derived from geographical covariates and spline basis functions
  - Modify PCA so that the scores can be predicted well by **Z**
- At first step of the algorithm, minimize the following with respect to β and $\widetilde{\mathbf{v}}$:
  $\|\mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T\|_F$ with constraint $\|\mathbf{Z}\beta\|^2 = 1$, rather than
  $\|\mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T\|_F$ with constraint $\|\widetilde{\mathbf{u}}\|^2 = 1$
- Define loadings: $\mathbf{v} = \frac{\widetilde{\mathbf{v}}}{\|\widetilde{\mathbf{v}}\|}$ and PC scores: $u = X\mathbf{v}$ (not observable at subject locations)
- Subsequently, optimization using residual matrices $(\mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T)$

Application of predictive PCA to PM2.5 data: Loadings



Application of predictive PCA to PM2.5 data: Predictability



Summary of the talk so far

1. Predictive PCA improves predictability of PC scores
2. Loadings from both traditional and predictive PCA are difficult to interpret

How to improve interpretability: Sparse PCA (sPCA)

- Principal components scores and loadings can sometimes be difficult to interpret
- Sparse PCA produces modified PCs with sparse loadings: loadings with only a few nonzero elements
- In sparse PCA, penalty parameter, $\lambda \geq 0$, controls sparsity of loadings:
  - Large λ results in very sparse loadings
  - Different λ can be used for different PCs

How to introduce sparsity to PCA?

- For a fixed value of penalty parameter λ:
  - Recall: In traditional PCA, we minimize

    $$\|\mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T\|_F$$

    with respect to $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$, s.t. $\|\widetilde{\mathbf{u}}\| = 1$
  - In Sparse PCA (Shen and Huang, 2008): we minimize

    $$\|\mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T\|_F + P_\lambda(\widetilde{\mathbf{v}})$$

    with respect to $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$, s.t. $\|\widetilde{\mathbf{u}}\| = 1$, where $P_\lambda(\widetilde{\mathbf{v}}) := \lambda \sum_{l=1}^{m} |v_l|$, an L_1 (LASSO) penalty function

Sparse PCA vs non-sparse PCA: Example

	Traditional PCA			Traditional Sparse PCA		
	Comp. 1	Comp. 2	Comp. 3	Comp. 1	Comp. 2	Comp. 3
S	0.25	-0.06	0.14	0.28	0	0
PM25	0.31	-0.02	0.1	0.34	0	0
EC	0.31	-0.07	0.02	0.34	0	0
As	0.3	-0.1	-0.02	0.33	0	0
OC	0.3	-0.1	0.05	0.33	0	0
Br	0.29	0.01	0.16	0.3	0	0
V	0.24	-0.02	0.34	0.24	0	0.38
Zn	0.28	-0.01	-0.15	0.26	-0.01	-0.14
K	0.27	0.14	0.16	0.26	0.08	0.06
Cu	0.27	-0.02	-0.02	0.26	0	0
Al	-0.01	0.56	0	0	0.62	0.02
Fe	0.26	0.21	-0.22	0.21	0.13	-0.02
Si	-0.03	0.58	0	0	0.62	0
Cr	0.26	-0.08	-0.23	0.24	0	0
Ca	0.11	0.47	-0.07	0	0.45	-0.1
Na	0	0.13	0.73	0	0	0.9
Mn	0.23	0.12	-0.37	0.14	0	-0.12

Sparse predictive PCA

- Recall: In predictive PCA, we minimize

$$\|\mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T\|_F$$

  with constraint $\|\mathbf{Z}\beta\|^2 = 1$

- Analogous to sparse PCA (Shen and Huang, 2008), we can introduce sparsity to predictive PCA by minimizing

$$\|\mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T\|_F + P_\lambda(\widetilde{\mathbf{v}})$$

  with constraint $\|\mathbf{Z}\beta\|^2 = 1$

Candidate PCA algorithms

	Unpenalized	Sparse *
Traditional PCA	$\min \lVert \mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T \rVert_F$	$\min(\lVert \mathbf{X} - \widetilde{\mathbf{u}}\widetilde{\mathbf{v}}^T \rVert_F + P_\lambda(\widetilde{\mathbf{v}}))$
Predictive PCA	$\min \lVert \mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T \rVert_F$	$\min(\lVert \mathbf{X} - \mathbf{Z}\beta\widetilde{\mathbf{v}}^T \rVert_F + P_\lambda(\widetilde{\mathbf{v}}))$

$*$ **Maximize pollutants**: λ selected to maximize spatial predictability of pollutants

Maximize scores: λ selected to maximize spatial predictability of principal scores

Simulation study

- Two simulated scenarios with 17 pollutants
- **Simulated Scenario 1:** Predictability is HIGH - most pollutants can be predicted well
- **Simulated Scenario 2:** Predictability is LOW - most pollutants cannot be predicted well

Simulation study - Scenario 1 (Predictability is HIGH)

			Predictability of Scores (R^2)			Abs. Correlation (Average)	Sparseness (%)
			PC1	PC2	PC3		
Without Penalty		Trad.PCA	0.96	0.87	0.7	0.04	0.00%
		Pred.PCA	0.96	0.88	0.71	0.05	0.00%
With Penalty	Max. Scores	Trad.PCA	0.97	0.91	0.84	0.38	35.22%
		Pred.PCA	0.97	0.92	0.84	0.43	35.53%
	Max. Pollutants	Trad.PCA	0.96	0.87	0.66	0.11	19.96%
		Pred.PCA	0.96	0.87	0.74	0.18	20.04%

Simulation study - Scenario 2 (Predictability is LOW)

			Predictability of Scores (R^2)			Abs. Correlation (Average)	Sparseness (%)
			PC1	PC2	PC3		
Without Penalty		Trad.PCA	0.76	0.6	0.27	0.02	0.00%
		Pred.PCA	0.85	0.76	0.56	0.08	0.00%
With Penalty	Max. Scores	Trad.PCA	0.93	0.79	0.28	0.31	52.47%
		Pred.PCA	0.95	0.9	0.83	0.42	69.96%
	Max. Pollutants	Trad.PCA	0.79	0.66	0.28	0.08	19.14%
		Pred.PCA	0.88	0.8	0.61	0.13	25.69%

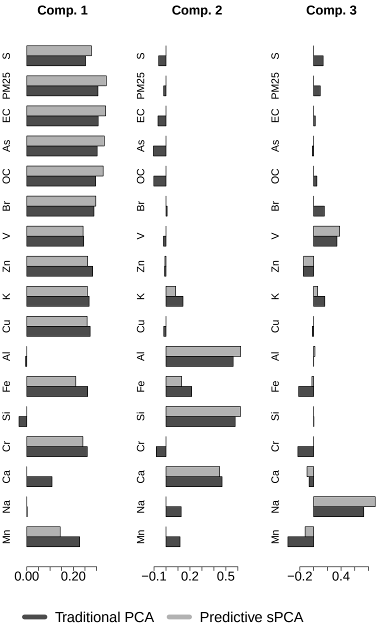
Simulation study: Conclusions

- Predictive sPCA results in improved predictability of PC scores:
  - Difference between approaches increases with increase in # of unpredictable pollutants
- Effect of penalty parameter:
  - Simplifies interpretability of loadings
  - If penalty maximizes predictability of scores:
    - PC scores are highly predictable
    - PC scores are highly correlated
  - **If penalty maximizes predictability of pollutants:**
    - **Predictability of PC scores is still high**
    - **PC scores are not correlated**

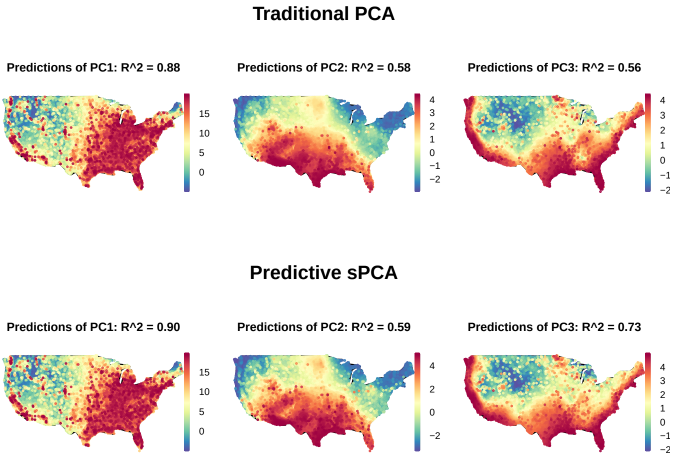
Application of sparse PCA to PM2.5 data: Comparison

			Predictability of Scores (R^2)			Correlations			Sparseness (%)
			PC1	PC2	PC3	PC1vsPC2	PC1vsPC3	PC2vsPC3	
Without Penalty		Trad.PCA	0.88	0.58	0.56	0.02	0	0	0.00%
		Pred.PCA	0.89	0.58	0.67	-0.05	0.02	-0.03	0.00%
With Penalty	Max. Scores	Trad.PCA	0.91	0.86	0.78	0.93	0.7	0.8	70.60%
		Pred.PCA	0.92	0.93	0.9	0.88	0.99	0.83	80.40%
	Max. Pollutants	Trad.PCA	0.89	0.64	0.57	0.08	0.55	-0.71	47.10%
		Pred.PCA	0.9	0.59	0.73	0.11	0.08	0.06	47.10%

Application of predictive sPCA to PM2.5 data: Loadings



Application of predictive sPCA to Data: Heat maps



Summary of the approach

- Developed by adding a constraint to traditional sparse PCA
- Predictive sPCA results in improved predictability of PC scores
- Penalty can be optimally selected by maximizing predictibility of pollutants
  - Simplifies interpretation of loadings (and increases predictability of scores)
  - Obtained PC scores are uncorrelated

Future work: Applications of current method

- Scientific interpretation of obtained principal component scores
- Number of principal components to use in health analysis
- Analysis of systolic blood pressure in Sister Study
- Application to MESA Air and mobile monitoring data from CCAR

Future work: Extensions of current method

- Additional penalty parameter to penalize regression coefficients β can be added
- Accounting for measurement error
- Spatial all-at-once dimension reduction approach (reduced rank regression)

Thank you!