Dynamic PCA for Multiple Air Pollutants

Oleg Melnikov[1,2], Loren H. Raun[3], Katherine B. Ensor[4]
Department of Statistics, MS 138, Rice University, Houston, TX 77251-1892, USA

Abstract

The dynamic nature of air quality chemistry and transport makes it difficult to identify the mixture of air pollutants for a region. In this study of air quality in the Houston metropolitan area we apply dynamic principal component analysis (DPCA) to a normalized multivariate time series of daily concentration measurements of five pollutants (O3, CO, NO2, SO2, PM2.5) from January 1, 2009 through December 31, 2011 for each of the 24 hours in a day. The resulting dynamic components are examined by hour across days for the 3 year period. Diurnal and seasonal patterns are revealed underlining times when DPCA performs best and two principal components (PCs) explain most variability in the multivariate series. DPCA is shown to be superior to static principal component analysis (PCA) in discovery of linear relations among transformed pollutant measurements. DPCA captures the time-dependent correlation structure of the underlying pollutants recorded at up to 34 monitoring sites in the region. In winter mornings the first principal component (PC1) (mainly CO and NO2) explains up to 70% of variability. Augmenting with the second principal component (PC2) (mainly driven by SO2) the explained variability rises to 90%. In the afternoon, O3 gains prominence in the second principal component. The seasonal profile of PCs' contribution to variance loses its distinction in the afternoon, yet cumulatively PC1 and PC2 still explain up to 65% of variability in ambient air data. DPCA provides a strategy for identifying the changing air quality profile for the region studied.

Key Words: dynamic principal component analysis, moving window PCA, multi-pollutant analysis, time series

1 Introduction

Chemical processes are complex and nonlinear. Their dependency structures are contaminated with cross and auto correlations, seasonality, diurnal cycles, outliers, and noise. Direct data visualization or even basic statistical summaries are unable to reveal the key underlying patterns and distributions of the mixtures of air pollutants. Multivariate data analysis (MDA) has been effectively utilized in discovering these latent structures. Principal component analysis (PCA) is one such tool that can identify linearly related variables that describe most of the variability in the data.

Recently PCA has gained traction in the study of air quality (AQ). Buhr [7] used PCA to examine sources of nitrogen oxides (NO_x) and carbon monoxide (CO) from

[1]Corresponding author. Tel.: +1 323 205 6534. Address: Duncan Hall 2125, Department of Statistics, Rice University, 6100 Main St., Houston, TX 77005, USA
[2]xisreal@gmail.com, http://Oleg.Rice.edu
[3]Loren.Raun@CityOfHouston.net
[4]Ensor@Rice.edu, http://www.stat.rice.edu/%7Ekathy

other pollutants. Trainer [38] studied formation and loss of ozone (O_3) through PCA and bivariate regression of pollutants. Gonçalves [12] related child morbidity and meteorology patterns to ambient AQ through PCA of pollutants and meteorological factors.

Many authors recognize the seasonal characteristics of environmental data and analyze winter and summer observations separately. Pissimanis [33] applied PCA to examine spatial distribution of $\max(O_3)$ concentrations in the summer months. Álvarez [2] applied rotated PCA to assess spatio-temporal variability in winter and summer. Statheropoulos [27] related key principal components (PC) to emissions and ozone via PCA on winter and summer data.

Some other authors recognize the diurnal pattern of the air pollution data. Buhr [6] contrasted air pollution to emission ratios with the help of PCA performed on morning data. Abdul-Wahab [1] employed PCA to construct uncorrelated components based on air pollution and environmental data separately aggregated for the day and night hours. Lengyel [24] examined day and night AQ via PCA of air pollution and meteorological observations. Sousa [37] exploited hourly air pollution data and meteorology to construct the components.

Still most analyses ignore the non-stationary structure of environment AQ data [43]. Since PCA assumes fixed distribution parameters, an application of static PCA on observations from a distribution with time-dependent parameters is improper and deficient. While dynamic PCA variants have been applied to chemical processes ([22]), climatology ([20]), (to our knowledge) it has not been used to study air pollution until now.

We construct DPCA components on a two dimensional time domain (hours of a day $\times$ days of studied time period) and investigate the organization of principal components and their contribution to overall variability. We define DPCA as a moving window static PCA. Such form of DPCA was studied by [17], [25], [39] and applied to electroencephalography in [41].

The novelty of this paper is its application of DPCA to air pollution observations with the objective to

1. Demonstrate a proper application of PCA technique to cyclostationary time-series
2. Approximate non-linear dependence with a linear technique
3. Assess absolute and relative performance of such application
4. Interpret linearity between PCA input variables and translate it to original AQ indicators
5. Reveal diurnal and seasonal patterns of strong and weak linear dependence among PCA input variables

This paper stops short of use of the identified dynamic PCs in forecasting, construction of air quality indicators (AQI), dimension reduction, etc. Some of the aforementioned papers (and references therein) have already demonstrated such extensions to PCA. Also, we do not account for spatial information, which has been investigated by other authors (e.g. [2, 33]). Instead, we construct spatially-averaged observations (SAO) to achieve a greater degree of robustness.

The paper is organized as follows. Section 2 discusses PCA assumptions, methodology and interpretation. It also defines DPCA and a notation helpful when referencing dynamic factors. Section 3 describes data pre-treatment, determines a suitable transformation, and verifies dynamic correlations to justify the use of DPCA. In Section 4 we apply DPCA to construct an informative 3D profile, identifying the contribution to the explained variability of each PC. We then evaluate contributions averaged across each hour and study dynamic PC loadings for two times of a day, namely 7am and 2am. Finally, we compare our DPCA efforts to employment of static PCA on air pollution data. We close with a short section of concluding remarks.

2 Methodology

2.1 Assumptions

PCA assumes the distribution of a data matrix X is characterized by constant mean and covariance parameters. In other words, since PCs are linear combinations of input variables (columns of X), the latter must be linearly related on the full observational interval [36]. This condition is problematic, since most observed processes are not linearly related and their distribution parameters may change with state, space, or time (even if the distribution family remains the same). For example, environmental and meteorological data often exhibit trend non-stationarity as the process mean exhibits seasonal and diurnal patterns. Fortunately, this behavior, termed cyclostationarity, still exhibits stationarity on a neighborhood of any point of a cycle. This local stationarity can be tested and local observations can be further explored with the usual PCA [18, p.314], [20, 8]. Similarly, in this paper, we perform static PCA on a fixed-size window, sliding in time along observations. This yields time-changing (dynamic) PCs on samples that are sufficiently small to remain weakly stationary, but still seize the local dependence structure.

Still, there is a body of literature discussing the assumption of whether X must have independent and identically distributed (iid) rows, each of which are multivariate normal (MVN) for PCA to make sense [18, p.19], [28, p.229], [3, p.488], [16, p.102]. The authors determine that theoretical derivations, descriptive use of PCA, and most results of a sample PCA do not require normality. In the case of time series, a weak stationarity of X is usually sufficient for consideration of the consistent estimates of the first two moments of the distribution of X [35, p.485], [16, p.365]. The assumption of normality adds an additional meaning to the inferred PCs. An interested reader may not that in some disagreement, a few authors imply that MVN assumption is important [10, p.558], [21, p.151], some claim that MVN assumption can be omitted altogether [18, p.39], [35, p.490], some develop alternative approximations to overcome the MVN assumption [34], and most simply proceed with PCA without explicitly noting any assumptions. We use and test normality only to determine the robustness point at which data outliers become insignificant.

2.2 Robustness

PCA, as a least squares method, is dangerously sensitive to outliers. These "atypical" observations may significantly affect estimation of the components of the analysis, such as the eigenvectors and eigenvalues of the covariance matrix of X. PCA ro-

bustness can be achieved in a variety of ways, ranging from the least-recommended removal of peripheral observations (or even variables) and transformation of the input data to robustifying the intermediate covariance matrix or the terminal PC components [18, p.233], [16, p.365], [40].

In practice, real world environmental data often exhibits ill-suited skewness and can be "symmetricized" with several favored non-linear transformations, such as logarithms, roots, powers (e.g. Box-Cox transform), ratios, log differences, reciprocals, logit transforms (of proportions), and alike [11, 32]. This data pre-treatment often coincides with *normalization* (herein defined as aligning data to MVN), which can, in turn, be checked with a battery of statistical tests. Among popular MVN tests are those developed by Mardia, Henze-Zirkler's and Royston. Mardia's skewness and kurtosis tests give a greater insight on the shape fit to MVN [14, 21]. We use one such MVN test to identify a suitable transformation for our data. With air pollution data, in particular, natural logarithm of some or all variables helps stabilize asymmetric variability and diminish the effect of extreme events [11, 1, 6, 8, 30].

2.3 Definition and interpretation

Consider a centered data matrix $X = [x_{np}] \in \mathbb{R}^{\mathfrak{n}\times\mathfrak{p}}$ with $\mathfrak{n}$ observations and $\mathfrak{p}$ variables, where each row follows the same multivariate, but not necessarily normal, distribution with fixed mean and variance parameters, estimated as $(\mathbf{0}, \Sigma)$. A (static) principal component analysis (PCA) is defined as a linear transformation of these correlated variables to uncorrelated principal components, PC_p, $k = 1..\mathfrak{p}$,

$$\begin{aligned} z_{.k} &:= \mathrm{PC}_k = v_{1k}x_{.1} + \cdots + v_{k\mathfrak{p}}x_{.\mathfrak{p}} \\ &= [x_{.1} \ldots x_{.\mathfrak{p}}]\, v_{k.} = X v_{k.} \end{aligned} \tag{2.1}$$

where $v'_{k.} = [v_{1k} \ldots v_{\mathfrak{p}k}] \in \mathbb{R}^{1\times\mathfrak{p}}$ are the suitable loading coefficients, and $x'_{.p} = [x_{1p} \ldots x_{\mathfrak{n}p}] \in \mathbb{R}^{1\times\mathfrak{n}}$. In other words, PCA decomposes X into two component (or factor) matrices, latent values (PC scores) and latent vectors (PC loadings). For convenience, the components are ordered by their contribution to the overall variability of the transformed data set. So, PC_1 has the largest contribution to variance, PC_2 - second largest, and so on.

One interpretation of PCA is that in the process of decorrelation of original variables it breaks up the entire variability of uncorrelated PCs into summable variances represented by squared eigenvalues of Σ. The largest eigenvalues identify principal components most relevant to the analysis since they contain most of variability. The smallest eigenvalues are thought to represent the noise in the data. Hence, if the noise components are identified, a reasonable approximation of X can be recovered from the surviving dominant patterns.

Since PCA is scale-dependent, disparate units and scales of input variables hinder interpretability of the results [28, p.219]. It is, thus, common to scale raw observations in some standardized way (usually, to mean 0 and variance 1), so that neither variable dominates the sample covariance matrix, and, consequently, the resulting components. Such standardization deems the input variables unitless, thereby clouding the subsequent inference. A good rule of thumb is to keep data in their original units, if PCA on a standardized dataset is not significantly different from that of PCA on raw data. Also, note that scaling up pure noise observations (with low variance) will enhance their impact in the analysis [40].

Still direct reading of PC loadings remains challenging since loading coefficients can take negative values (weights) and void the sum-of-parts interpretability that is prised in other popular factorization techniques, such as negative matrix decomposition (NMF). Hence, PC loadings may benefit from an additional transformations to ease interpretation [35, p.492].

2.4 Decomposition

The workhorse behind PCA is a singular value decomposition (SVD) of a data matrix $X_{\mathfrak{n}\times\mathfrak{p}}$, or, equivalently, the eigenvalue decomposition (EVD) of its sample covariance matrix, $\Sigma_{\mathfrak{p}\times\mathfrak{p}}$.

The former is a factorization

$$X_{\mathfrak{n}\times\mathfrak{p}} = U_{\mathfrak{n}\times\mathfrak{p}} \cdot \Lambda_{\mathfrak{p}\times\mathfrak{p}} \cdot V'_{\mathfrak{p}\times\mathfrak{p}}$$

where Λ is diagonal. U, V are orthogonal , i.e. $U'U = I_{\mathfrak{p}} = V'V$ (or $U' = U^{-1}$ and $V' = V^{-1}$) . These are left and right eigenvectors of X.

As with any symmetric positive semi-definite (PSD) matrix, the EVD of Σ is (up to a scaling factor)

$$\begin{aligned} \Sigma &\propto X'X = \left(U\Lambda V'\right)'\left(U\Lambda V'\right) = V\Lambda^2 V' \\ \Sigma V &\propto V\Lambda^2 V'V = V\Lambda^2 \end{aligned}$$

The components in both decompositions exist and are unique. Note that SVD eigenvalues are equal to EVD eigenvalues. Also, right eigenvectors of X are the eigenvectors of $X'X$, and left eigenvectors of X are the eigenvectors of XX'.

To summarize, PC transformation relates X to its *score* matrix Z as $Z_{\mathfrak{n}\times\mathfrak{p}} := U\Lambda = XV$ or $z_{\cdot k} := u_{\cdot k}\lambda_k = Xv_{\cdot k}$, where we index components by k and variables (pollutants herein) by p ($k, p = 1..\mathfrak{p}$), $v_{\cdot k}$ are loading coefficients from (2.1), and

- $\Lambda = \operatorname{diag}\{\lambda_k \mid 0 \leq \lambda_{k+1} \leq \lambda_k\}$ is diagonal matrix of (ordered) singular values of Σ (in other words, standard deviations of PCs). Off-diagonal zeros imply uncorrelated PCs.
- Λ^2 is a diagonal matrix of (ordered) eigenvalues of Σ and represent the variances of PCs.
- $V = [v_{\cdot 1} \ldots v_{\cdot\mathfrak{p}}] = [v_{pk}]$ is a standardized PC loading matrix with columns as standardized PC loadings of X, representing PC directions or eigenvectors of Σ. The elements of V, v_{pk}, are PC loading coefficients or weights; and, $\mathfrak{p} \times \mathfrak{p}$ matrix $V\Lambda = [\lambda_1 v_{\cdot 1} \ldots \lambda_{\mathfrak{p}} v_{\cdot\mathfrak{p}}]$ is V's non-standardized counterpart.
- $U = [u_{\cdot 1} \ldots u_{\cdot\mathfrak{p}}]_{\mathfrak{n}\times\mathfrak{p}}$ is a standardized PC score matrix with columns as standardized *PCs* of X and rows (transformed observations) as row scores, also termed factor scores or z-scores, of PCs. The nth element of $z_{\cdot k}$, z_{nk}, is the PC score (or factor score) of the pth PC for the nth observation. The matrix $Z := U\Lambda = [z_{\cdot 1} \ldots z_{\cdot\mathfrak{p}}]$ is its non-standardized analog.

An expanded matrix notation of PCA factorization is

$$\underbrace{\overset{\text{correlated variables}}{[x_{\cdot 1}...x_{\cdot \mathfrak{p}}]}}_{X}\underbrace{\overset{\text{eigenvectors (PC loadings)}}{[v_{\cdot 1}...v_{\cdot \mathfrak{p}}]}}_{V} = \overset{\text{observations}}{\begin{bmatrix} x'_{1\cdot} \\ \vdots \\ x'_{\mathfrak{n}\cdot} \end{bmatrix}} \quad V = \underbrace{\overset{\text{uncorrelated PCs}}{[z_{\cdot 1}...z_{\cdot \mathfrak{p}}]}}_{Z} = \overset{\text{PC scores}}{\begin{bmatrix} z'_{1\cdot} \\ \vdots \\ z'_{\mathfrak{n}\cdot} \end{bmatrix}} \tag{2.2}$$

where, in a time series context, multivariate observations $x_{n\cdot} \in \mathbb{R}^{\mathfrak{p}}$ and PC scores $z_{n\cdot} \in \mathbb{R}^{\mathfrak{p}}$ are chronologically indexed by time.

PCA offers some properties useful in interpretation of the results. Explained variance (EV) is the proportion of the total variability (of the PCs) accounted for by a specific PC. These are the diagonal values of $\Lambda^2/\text{Trace}\left(\Lambda^2\right)$ matrix. Variables of primary interest are EV and cumulative EV (CEV):

$$\begin{aligned} \text{EV}_k &= \lambda_k^2/\text{Trace}\left(\Lambda^2\right) \\ \text{CEV}_k &= \sum_{i=1..k} \text{EV}_i \end{aligned} \tag{2.3}$$

A more detailed discussion of PCA is established in [28, 10, 16, 31].

2.5 DPCA notation and diagram

Application of static PCA on a data with time-dependent structure is unreliable and improper, since the procedure attempts to linearly approximate the complex non-linear relations between variables [22]. Instead, dynamic PCA (DPCA), a simple extension of PCA, can reveal the dynamics of the underlying data structure. Our definition of DPCA is an application of the sample PCA on a sliding window of fixed width ℓ [17, 39]. For a cyclostationary time series, a local (in time) sample of observations is approximately weakly stationary with (some) fixed distribution [18]. PCA applied on a windowed data captures the linear relation of the variables. As the window slides forward at a constant rate of one observation at a time, the time-indexed PC loadings and scores express the overall non-linear relation.

Since we apply PCA on a window sliding across time, all resulting statistics are time dependent. For reasons discussed in Section 3.2, we consider time to be a two dimensional domain of hours $\times$ days. This avoids diurnal and seasonal non-stationarities and allows for separate diurnal and seasonal data analysis. Whenever notation EV_k may be ambiguous, we underline the specific time dependencies:

$$\begin{aligned} \underline{\text{EV}}_{h.k} &:= [\text{EV}_{hdk}]_{\forall d} \in \mathbb{R}_+^{\mathfrak{d}} \\ \underline{\text{EV}}_{..k} &:= [\text{EV}_{hdk}]_{\forall h,d} \in \mathbb{R}_+^{24\times\mathfrak{d}} \end{aligned} \tag{2.4}$$

where h is an hour of a day, d is a day of the time period, and k identifies the corresponding kth PC. In our dataset we have $\mathfrak{d} = \max\{d\}$ or 1095 days.

Similarly, dynamic PC loadings are defined via a 4 dimensional array $V = [v_{hdpk}] \in \mathbb{R}^{24\times\mathfrak{d}\times\mathfrak{p}\times\mathfrak{p}}$ with analogous definitions $v_{h.pk} \in \mathbb{R}^{\mathfrak{d}}$, $v_{hd.k}, v_{hdp\cdot} \in \mathbb{R}^{\mathfrak{p}}$, $v_{hd..} \in \mathbb{R}^{\mathfrak{p}^2}$, $v_{\cdot dp\cdot} \in \mathbb{R}^{24\times\mathfrak{p}}$, $v_{h...} \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{p}\times\mathfrak{p}}$, etc. A dot increments a dimension of the variable by the maximum of the corresponding index placeholder. One dot designates a vector, two - a matrix (first dot defines rows, second - columns), three - a 3D array (third dot defines the size of the third dimension). So $[v_{h.pk}]_{\forall p,k}$ is a $\mathfrak{p}\times\mathfrak{p}$ matrix of $\mathfrak{n}$-vectors as elements. This ameliorates visualization of dynamic loadings and other

variables. In the same way we assign notation for dynamic PCs: $[\mathrm{PC}_{hdk}] \in \mathbb{R}^{24\times \mathtt{n}\times \mathtt{p}}$, $\mathrm{PC}_{h.k} \in \mathbb{R}^{\mathtt{n}}$, $\mathrm{PC}_{..k} \in \mathbb{R}^{24\times \mathtt{n}}$, etc.

Schematically, our application of DPCA is exhibited in Figure 2.1, with an exception of forecasting.

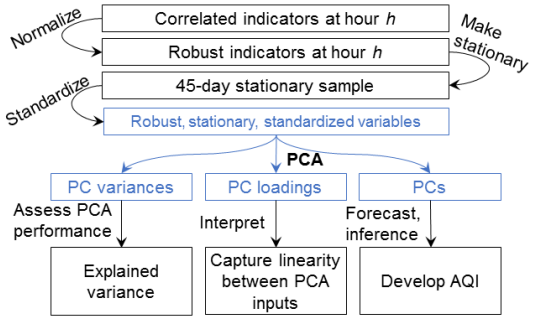

Figure 2.1: Application of DPCA. The correlated indicators are spatially-averaged observations (SAO) constructed in Section 3.2.

3 Data

Texas Commission on Environmental Quality (TCEQ) provides an access to measurements of air pollutant concentrations from Texas monitoring stations (sites). The dataset contains hourly observations of 5 pollutants: ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter less than 2.5 micrometers ($PM_{2.5}$), from 1/1/2009 00:00 CST to 12/31/2011 23:00 CST (that is 1095 days or 26,280 samples) collected from 35 monitoring sites throughout the area of Houston, Texas; see Figure 3.1. The study region excludes sites that are non-representative of air pollution profile of the Houston metropolitan area (HMA). For example, the Galveston Bay area is an oceanic coastal line with concentrations expected to differ from those in HMA. The Houston Ship Channel, unlike HMA, is an industrialized home to numerous petroleum refineries, and port and chemical manufacturing plants [4, 19]. Some other sites are considered too remote. Gas concentrations are measured in (dimensionless units of) parts per billion (ppb), whereas $PM_{2.5}$ is in $\mu g/m^3$.

Spatial information is lost once we construct spatially-averaged observations (SAO) in Section 3.2.

3.1 Missing observations

In this preferential sampling (i.e. chiefly surveying the areas of heightened concern, [26]) with high screening costs, not all pollutant concentrations are tracked at each monitoring site. Out of 35 sites, only C416 (black pin in Figure 3.1) measured all 5 pollutants. Also, TCEQ uses nearly 30 codes to identify invalid measurements resulting from downtimes, data losses, rejected measurements, equipment malfunctions, etc. Our dataset contained 16 such codes, which we consider to be missing data.

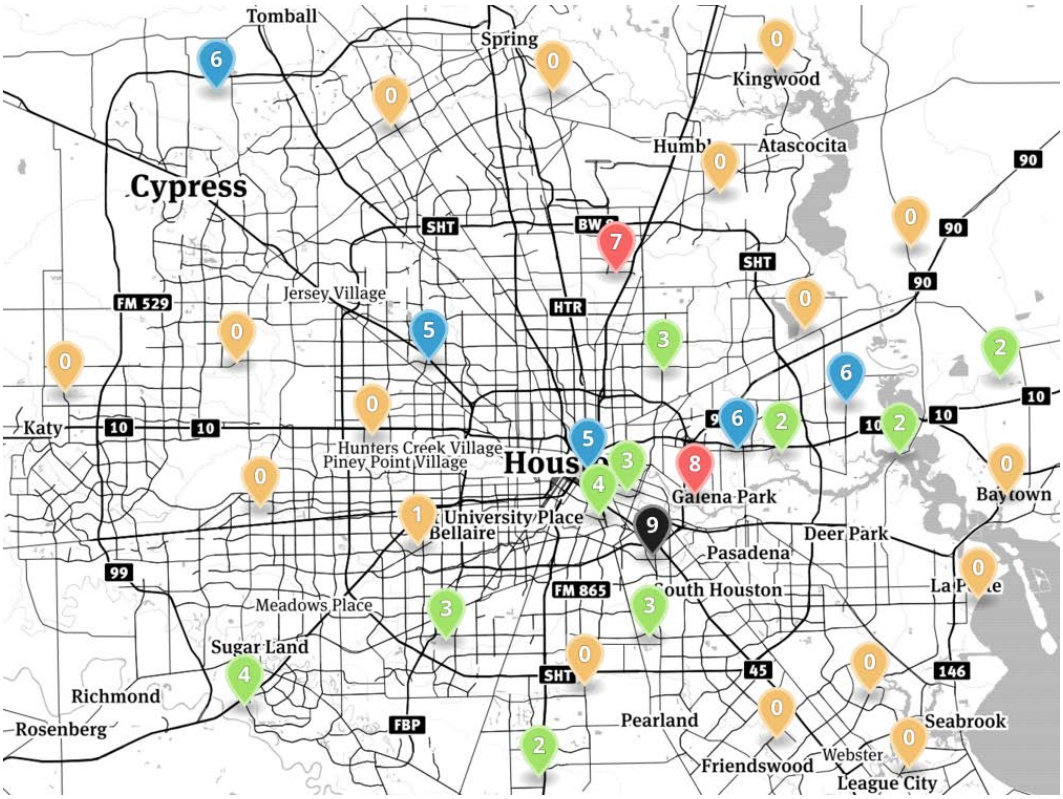

Figure 3.1: Monitoring sites in Houston, Harris County, TX, USA. Area of study excludes some sites located near Galveston Bay and Houston Ship Channel, or too distant from Houston. Pin colors: black - tracks all 5 pollutants, red - 4, blue - 3, green - 2, orange - 1. Pin numbers: 9 - tracks OCNSP (short for O_3, CO, NO_2, SO_2, $PM_{2.5}$, respectively), 8 - OCNS, 7 - OCNP, 6 - ONP, 5 - OCN, 4 - OC, 3 - OS, 2 - ON, 1 - N, 0 - O). Data source: www.tceq.state.tx.us

We impute short temporal stretches of NAs, defined as up to 4 contiguous hourly NAs from the same site within each air pollutant, with monotone Hermite splines [9]. The advantage of this method is that imputed observations stay within the bounds of starting and ending observed values, which prevents negative imputations near extreme observations noted with other methods. A similar approximation could have been achieved with linear approximates, but we feel that splines can better incorporate the nearby diurnal structure, if only a few consecutive observations are missing.

The larger gaps are replaced by the spatial averages within each pollutant, when we construct a spatially averaged observations (SAO) indicator in Section 3.2.

The summary of missing values and data imputations are given in Table1. Apparently, most sites are equipped to gauge ozone, while CO, SO_2, and $PM_{2.5}$ are quantified at only a handful of locations. The short NA gaps are least troublesome with $PM_{2.5}$ and O_3 observations. Notably, NO_2 and $PM_{2.5}$ stand out with larger proportion of missing data.

Adjustment for daylight savings time yield little improvement and we leave details to an appendix of the paper.

	O_3	CO	NO_2	SO_2	$PM_{2.5}$
Long runs	2.89	1.90	3.16	1.94	6.30
Short runs	0.95	1.44	2.13	1.36	0.58
Total	3.84	3.34	3.34	3.30	6.88
#sites	34	7	13	6	5

Table 1: Proportion of missing values (in %) attributed to Long and Short runs Off-line sites (non-contributing for over one month) are dropped from NA summary for the non-participation period: C695, C696 for CO and C555, C572, C695, C696 for O_3.

3.2 Spatially-averaged observations (SAO)

It is common to spatially average observations from multiple monitoring sites. While the true average estimator is unknown, a mean-based indicator, $\bar{x}_{hdp}$, is a popular choice in literature. Here we index our observations by hour $h = 0..23$, by day $d = 0..1095$, and by measured air pollutant $p = 1..5$, representing O_3, CO, NO_2, SO_2, and $PM_{2.5}$ respectively. Such equi-weighted measure of centrality assumes homogeneity among monitoring sites. In this paper we prefer a more robust, median-based, measure of spatially averaged observations (SAO), $\text{SAO}_{hdp} := \tilde{x}_{hdp}$, and the related matrix $\text{SAO}_h := \text{SAO}_{h..} = [\tilde{x}_{hdp}]_{\forall dp} =\in \mathbb{R}_+^{1095\times 5}$.

In our PCA median-based SAO_h performs better than its mean-based counterpart, yielding a clearer cyclostationary EV pattern, more stable $\text{PC}_{h.1}$ coefficients (see Section 4). Other indicators considered in practice and literature include the use of a maximum (i.e. aids in study of air pollution peaks and health), a combination of averaging functions, and a multi-level aggregation, such as spatial clustering of sites based on some notion of similarity. Bruno [5], Lee [23] and references therein present a good overview of various air quality indicators.

Raw (non-standardized) SAO are shown in Figure 3.2 along with rolling ($\ell = 45$-day) mean and standard deviation. Note the non-stationarity of the data expressed with time-dependent mean and variance. For example, the first two sample moments NO_2 and SO_2 are elevated in winters, those of $PM_{2.5}$ - in summers. The clustered behavior persists across all pollutants. Yet, covariance is more difficult to observe due to dissimilar scale and embedded noise.

3.3 Normalization and standardization

As part of robustifying SAO_h, we have assessed logarithmic and other non-linear transformations, which are common in the examination of AQ data. Since normalization (herein aligning data to MVN) is usually associated with robustifying PCA (see Section 2.2), it is reasonable to use an MVN test to target the desirable transform. The 45-day moving window p-values, p_{hd}, of Henze-Zirkler's MVN test are presented in Figure 3.3. That is $p_{h.} \in \mathbb{R}^{1095}$ is a non-local time series of (daily) p-values fixed at $h \in \{0..23\}$, hour of a day. The plot has a 5% significance level cut off; and, more blue indicates a greater likelihood of tested data following MVN. The summary observations from Figure 3.3 are:

1. The top panel shows that non-transformed data, $\text{SAO}_{7\text{am}}$, fails to exhibit normality at 7am. This time of a day is representative of daily traffic build up.

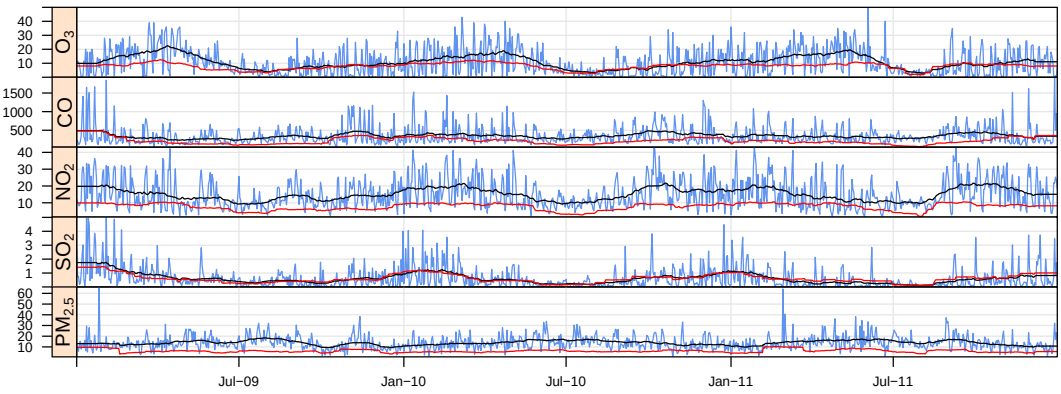

Figure 3.2: Spatially averaged observations at 7am, $\text{SAO}_{7\text{am}}$. Measurements ($\mu\text{g/m}^3$ for $\text{PM}_{2.5}$, ppb for others) are adjusted for DST. Overlaid curves are 45-day rolling statistics: simple mean (black), standard deviation (red).

2. The middle panel reflects a slight improvement in MVN test of log transform

$$\begin{aligned} \hat{x}_{hdp} &:= \log\left(1 + \tilde{x}_{hdp}\right) \\ \text{LSAO}_h &:= \hat{x}_{h..} \in \mathbb{R}^{1095\times 5} \end{aligned}$$

where $\tilde{x}_{hdp}$ is defined in Section 3.2.

1. The bottom panel illustrates log differencing as a considerably promising normalization. It is a routine method in financial models, which use log returns, or percent change, computed analogously from the observed stock prices. Similarly, we define normalized SAO as

$$\begin{aligned} y_{hdp} &:= \hat{x}_{hdp} - \hat{x}_{h,d-1,p} && (3.1) \\ \text{NSAO}_h &:= y_{h..} \in \mathbb{R}^{1095\times 5} && (3.2) \end{aligned}$$

Also, as expected, median-based SAO_h exhibits greater normality than a similar mean-based measure across the evaluated transformations. Other MVN tests (see Section 2.2) also support the use of the median-based transform defined in (3.1). Likewise, other transforms listed in Section 2.2 yield similar-to-slightly-inferior performance as that of log mapping (LSAO_h). When NSAO_h is assessed at other hours of a day (night time, traffic time, etc.), the MVN test's conclusions are similar.

Outliers, assessed for the same three transformations, are presented in Figure 3.4 and also support the use of use of log differencing. Hence, we proceed with the analysis on median-based NSAO_h data.

Raw concentrations use different scales and are not suited for PCA, as noted in Section 2.3. A common approach is to standardize the units to have mean 0 and variance 1 prior to application of PCA [6, 40]. We do so on each 45-day window. For instance, CO measurements dominate the results of PCA of $\text{SAO}_{h.2}$, if left unscaled (see data summary in Table 2). Similarly, non-standardized O_3 observations govern PCA of LSAO and NSAO because its variability is up to twice that of other variables. In fact, when PCA was tried on unscaled NSAO_7, $\text{PC}_{7.1}$ explained 80% of variability with dynamic loadings for the (normalized) O_3 quantities playing a prominent part.

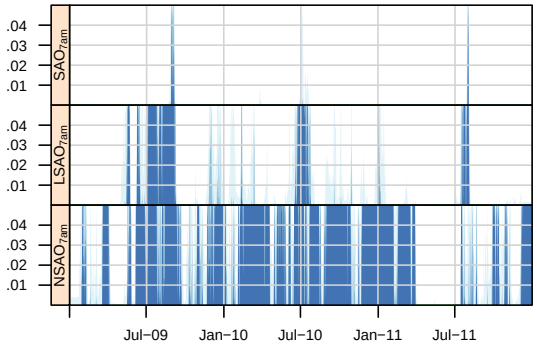

Figure 3.3: Horizon plots showing 45-day rolling p-values, $p_{7\cdot}$, of (Henze-Zirkler's) MVN tests on SAO_7, $LSAO_7$, and $NSAO_7$ datasets. More blue indicates higher likeliness of the underlying data following MVN, implying fewer outliers, and, thus, greater suitability of PCA. See [21].

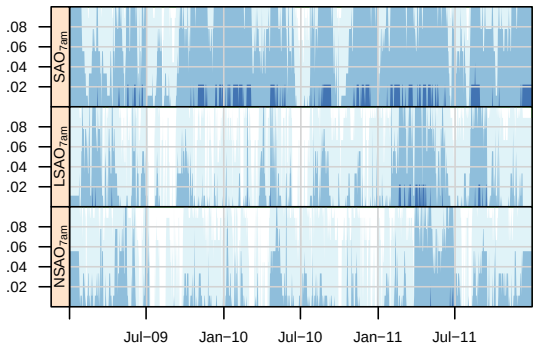

Figure 3.4: Horizon plots showing 45-day rolling proportions of outliers for SAO_7, $LSAO_7$, and $NSAO_7$ datasets. Detection is based on adjusted robust Mahalanobis distance, $\mathrm{rMD}(\cdot)$, with decision on parameter $\alpha \in [.5, 1]$ (we use $\alpha = .75$). We note a consistent relative outcome: SAO_7 contains most outliers (more blue) $NSAO_7$ has least. See [21]

In contrast, normalized O_3 participation is similar to that of CO and NO_2 in $PC_{7.1}$, when $NSAO_7$ is standardized. Since PCs are designed to capture and attribute variables' variability, the former $PC_{7.1}$ is likely inflated by variability of O_3.

Table 2 describes raw, log and log differenced pollutant indicators. Note that NSAO variable's mean and median are nearly identical, an expected property of data from MVN distribution. While SAO and LSAO exhibit dramatic differences in various statistic measures (across pollutants), NSAO pollutants' statistics (min, max, ...) are better aligned. In our analysis we do not require strict normality. Our primary goal is to prepare data for DPCA by minimizing the effect of outliers on each rolling subsample.

3.4 Dynamic correlation

PCA maps highly correlated variables to uncorrelated components. It would make little sense to apply PCA to uncorrelated variables. So, we quickly check the degree of association between normalized pollutants. Indeed, as shown in Figure 3.5, some

	SAO...					LSAO...					NSAO...				
	O3	CO	NO2	SO2	PM2.5	O3	CO	NO2	SO2	PM2.5	O3	CO	NO2	SO2	PM2.5
Min	0	6.6	.8	0	.1	0	2.0	.6	0	.1	-1.7	-1.6	-1.1	-1.3	-2.2
1Q	13.0	142.3	4.6	.1	7.3	2.6	5.0	1.7	.1	2.1	-0.1	-.1	-.1	-.1	-.1
Med	23.0	182.6	7.3	.3	10.2	3.2	5.2	2.1	.3	2.4	0	0	0	0	0
Mean	24.8	220.7	9.7	.6	11.3	3.0	5.3	2.2	.4	2.4	0	0	0	0	0
3Q	34.0	242.3	12.1	.7	14.2	3.6	5.5	2.6	.6	2.7	.1	.1	.1	.1	.1
Max	101.0	2076.5	50.5	19.8	81.4	4.6	7.6	3.9	3.0	4.4	2.4	1.5	1.2	1.4	1.6
SD	15.7	149.6	7.4	.9	5.8	.9	.5	.6	.4	.5	.3	.2	.2	.2	.2

Table 2: Summary statistics for SAO, LSAO and NSAO

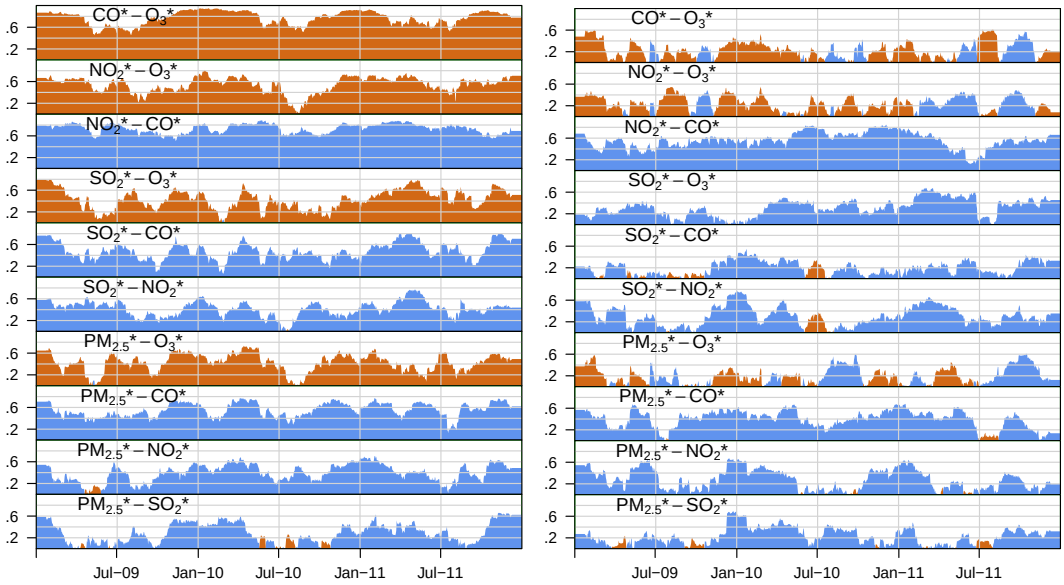

Figure 3.5: Pearson correlations for the paired NSAO_h variables are computed on a 45-day rolling window in the morning (7am, *left*) and afternoon (2pm, *right*). Positive/negative correlations are shown in blue/red, both on a positive axis.

variables of $\text{NSAO}_{h\in\{7,14\}}$ exhibit a high degree of contemporaneous dependency. O_3^* are strongly associated with CO^* and NO_2^* in the morning, but not in the afternoon (where * indicates a normalized observation). NO_2^* is correlated with CO^* in both samples.

In general, morning correlations are more substantial than those in the afternoon. Also, a seasonal pattern is observable in some correlations. For example, morning CO^* to O_3^* correlations are more negative in the winters and than in the summers. Thus, we have established that the issue of co-dependence is significant and the use of PCA is just. Also, the presence of seasonal cycles underlines the cyclostationary structure of the data and supports the use of DPCA.

3.5 Choice of a window size

Air pollution data carries clear seasonal and diurnal patterns. Its cyclostationarity allows us to assume a fixed mean and variance on a short (length ℓ) window of observations. We assume that $\ell = 45$ days carries sufficient information to grasp the approximately stationary structure at a particular time of a year.

4 Results and discussion

4.1 Explained variance (EV)

Examination of the dynamic nature of the explained variance for $\text{PC}_{h.1}$, $\text{PC}_{h.2}$ and the pointwise sum of the two for the NSAO_h, at each hour of a day, over the three-year study period yields key insights. Figure 4.1 depicts these components. In general, higher explained variance corresponds to a better PCA fit and stronger linear relations among PCA input variables that make up the PCs.

We use **R** (version 3.x) core (`base`, `stats`), `xts` and `lattice` packages for most of data scrubbing, imputation, PCA and visualization. Non-local NSAO_h are standardized on each 45-day window before PCA is applied and Figure 4.1 of dynamic EV is drawn. This 3D plot profiles EV components over a 2D time domain as a non-local (daily) pattern of $\text{EV}_{h.k}$ and local (hourly) pattern of $\text{EV}_{.dk}$.

Admirably, just two PCs explain up to 90% of variability in the components (in morning winters). But, more importantly, such profiling presents the EV pattern of the components ($\text{PC}_{..k\in\{1,2\}}$) dissected by time of day and day of the observed period.

The daily explained variability by the first principal component at hour h, $\text{EV}_{h.1}$, exhibits a strong seasonal trend, spiking in cool winters and sinking in hot and humid Texas summers, for any fixed hour of a day. A trend non-stationary $\text{EV}_{..1}$ ranges from about 30% to about 75% with overall mean, $\overline{\overline{\text{EV}}}_1$, of approximately 51% as shown in Table 3.

The seasonal form of $\text{CEV}_{h.2}$ follows that of $\text{EV}_{h.1}$ because the marginal difference, i.e. $\text{EV}_{h.2}$, is relatively too small and less variable. The mean of $\text{EV}_{h.2}$ is less than half of the mean of $\text{EV}_{h.1}$ (23% vs 51%, see Table 3). Overall mean variance explained by the first two principal components is $\overline{\overline{\text{CEV}}}_2 \approx 74\%$.

The measure $\text{EV}_{.d1}$ exhibits a strong diurnal pattern, when the figure panels are assessed vertically with changing hours of a day. The contributions are higher overnight, from late evening to early morning, peaking with sun rise at around 7am. These times of a day exhibit very little direct solar radiation. Contributions drop in the afternoons, reaching lowest points around 4-5 pm. Such diurnal pattern is strongest in the winters. Diurnal contributions from the second component, $\text{EV}_{.d2}$, slightly smooth out this diurnal pattern with elevated contribution mid-day and lower contributions at night. As a result, the patterns are less prominent in the right panel showing $\text{CEV}_{h.2}$.

Naturally, static EV_k fails to capture such complex diurnal and cyclostationary dynamics.

4.2 Mean explained variance

Eyeballing 3D EV (Figure 4.1 on page 14) is helpful as it reveals a great deal of detail. However, for a quick assessment of intraday contribution behavior, one may consider non-locally averaged EV, computed at a specific hour as

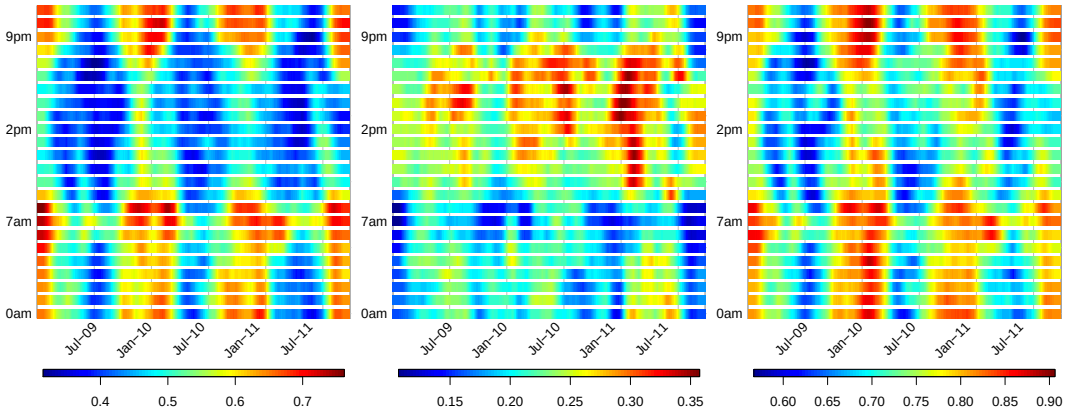

Figure 4.1: Heatmaps of the dynamic explained variance (EV) from $\mathrm{PC}_{h.1}$, $\mathrm{PC}_{h.2}$, and $\mathrm{PC}_{h.1} + \mathrm{PC}_{h.2}$ of the $\mathrm{NSAO}_h, h = 0..23$, *left*, *right* and *center*, respectively.

$$\begin{aligned} \overline{\mathrm{EV}}_{hk} &:= \frac{1}{\mathfrak{d}} \sum_{d} \mathrm{EV}_{hdk} \\ \overline{\mathrm{CEV}}_{hk} &:= \frac{1}{\mathfrak{d}} \sum_{i \le k} \overline{\mathrm{EV}}_{hi} \end{aligned} \tag{4.1}$$

where number of days $\mathfrak{d} = 1095$.

The plots in these section focus on analysis of quantities in (4.1) and their (somewhat limited due to aggregation) use as a measure of PCA performance.

To start off, we want to evaluate our choice of SAO averaging function and normalizing transformation. We briefly consider Figure 4.2 for such comparison. It exhibits $\overline{\mathrm{EV}}_{hk}$ based on SAO_h (identity transform), LSAO_h (log transform), and NSAO_h (log differencing transform), where input SAO_h is computed either via mean or median function, i.e. $\bar{x}_{hdp}$ and $\tilde{x}_{hdp}$, respectively (see Section 3.2). The overall shapes appear similar across all spatial averaging and normalizing methods. That is $\overline{\mathrm{EV}}_{hk}$ spikes at 7am and dips in the afternoon (1-5pm). Thus, at least with the $\overline{\mathrm{EV}}_{h1}$ measure, these methods do not grossly differ at representing the aggregate dynamics of underlying variables. Still $\bar{x}_{hdp}$ performs poorer (vs. $\tilde{x}_{hdp}$) around a peak (7am) and performs vaguely better in the afternoon (the bottom of the curve). Also, LSAO_h and NSAO_h of $\tilde{x}_{hdp}$ perform best near peak, but the former beats the latter at most hours of a day. If this aggregate was a single measure of performance of PCA analysis, then we would perform PCA on LSAO_h, as it is frequently done. However, the consideration of robustness in Figure 3.4 demands for PCA on NSAO_h, which produces clearer DPCA components. That is dynamic $\mathrm{EV}_{h.1}$ possess a coherent seasonal structure in Figure 4.1 and dynamic loadings in Figure 4.3 are more interpretable, as compared to those of LSAO_h, whose EV plots we added to supplemented material.

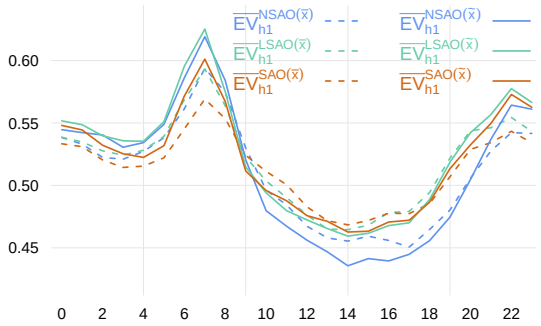

Figure 4.2: Comparison of $\overline{\mathrm{EV}}_{hk}$. Dotted lines use SAO_h based on spatial mean ($\bar{x}_{hdp}$), i.e. averaging of contemporaneous observations among monitoring sites. Solid lines uses SAO_h based on spatial median ($\tilde{x}_{hdp}$). Blue line computes $\overline{\mathrm{EV}}_{hk}$ based on NSAO_h, green - LSAO_h, brown - SAO_h. Ordinate units are in proportions, abscissa - in hours of a day (0 to 23).

We now return to examination of $\overline{\mathrm{EV}}_{hk}$ from PCA of NSAO_h based on $\tilde{x}_{hdp}$. Figure 4.3 reveals the relation between first three $\overline{\mathrm{EV}}_{hk}$ variables ($k = 1..3$, $h = 0..23$). It shows that $\overline{\mathrm{EV}}_{h1}$ is negatively correlated with $\overline{\mathrm{EV}}_{h2}$. So, when $\mathrm{PC}_{h.1}$ gains prominence in capturing variability (around 6-8am and midnight), $\mathrm{PC}_{h.2}$ gives up almost as much, and vise versa. The average explanatory power exceeds 60% at 7am and dives just below 45% in the afternoon (2-6pm).

The box-and-whisker plot is a compact way to describe a sample variability or its distribution's shape. These (static) descriptions are illustrated in Figure 4.3 for $\mathrm{EV}_{h.1}$ at each hour h. Greater number of outliers appear to coincide with poorer performance of DPCA (in terms of explained variability) around afternoon hours. Recall (from Figure 4.1) that afternoon hours were also blurring the seasonality in $\mathrm{EV}_{h.1}$.

Note that $\overline{\mathrm{EV}}_{hk}$ oversimplifies the results. It favors a clearer ("big picture") diurnal dynamics, while hides the seasonal structure of the underlying $\mathrm{EV}_{h.k}$. Still, the plots support the superiority of DPCA in the morning and near-midnight NSAO_h and inferiority of such analysis on data in the afternoon hours. If cyclostationarity of $\mathrm{EV}_{h.k}$ needs to be explicitly exemplified, then boxplots can be assessed on a windowed time interval (of, say, 45 days).

Further aggregation along the dimension of day hours is exhibited in Table 3. We compare these $\overline{\overline{\mathrm{EV}}}_k$ values to what other authors have achieved with static PCA, in Section 4.4.

$\overline{\overline{\mathrm{EV}}}_1$	$\overline{\overline{\mathrm{EV}}}_2$	$\overline{\overline{\mathrm{EV}}}_3$	$\overline{\overline{\mathrm{CEV}}}_2$	$\overline{\overline{\mathrm{CEV}}}_3$
.51	.23	.14	.74	.88

Table 3: Cumulative and non-cumulative overall mean explained variance, i.e. $\overline{\overline{\mathrm{EV}}}_k = \frac{1}{24}\sum_h \overline{\mathrm{EV}}_{hk}$.

4.3 Dynamic loading coefficients

Furthermore, we scrutinize the linearity of relationships and participation of NSAO_h variables (i.e. percent change in pollutants) in PCs. The two most remarkable hours

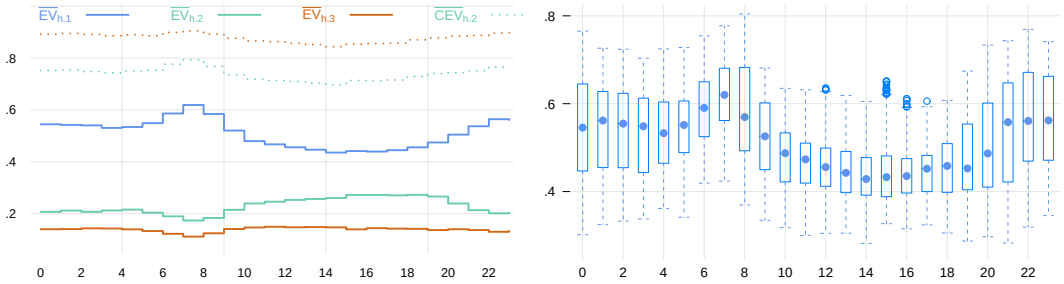

Figure 4.3: *Left*: Individual (solid) and cumulative (dotted) EV for the PC_1, PC_2, PC_3 by hour. *Right*: Distribution of $EV_{h.1}$ over 3 year period (ignoring non-stationarity). The centered bullet dots are 24 medians of the underlying samples. Note clustering of outliers near hours of poor explanatory power (low $\overline{EV}_{h1}$ values). Ordinate units are in proportions, abscissa - in hours of a day (0 to 23).

of a day are 7am and 2pm (see Figure 4.3), when $\overline{EV}_{hk}$ reaches its highest and lowest values, respectively. Figure 4.4 depicts corresponding PC loadings for $h = 7, 14$.

From the figure we observe that in the morning $PC_{7am.1}$ (i.e. first dynamic PC for $NSAO_{7am}$) is a fairly consistent linear function of all 5 variables with weights maintaining their approximate mean and relation to other variables. CO^* (i.e. normalized and standardized CO) is the largest driver behind $PC_{7.1}$ with weights averaging 0.53 and reaching 0.6 in summer 2010. Coefficients appear somewhat seasonal with CO^* playing a bigger part of $PC_{7.1}$ in hot summers. O_3^* and NO_2^* are also influential. O_3^* weights oppose those of all other variables, implying inverse relationship between log increments of O_3 and other pollutants.

Largest (yet unstable) contribution to $PC_{7.2}$ comes from SO_2^*. $PM_{2.5}^*$, second largest, has opposite sign weights, implying offsetting contribution to $PC_{7.2}$. In particular, $PM_{2.5}^*$ gains prominence in $PC_{7.2}$ during summers, reaching weights of -0.8. $PC_{7.3}$ largely depends on $PM_{2.5}^*$ and $PC_{7.3}$ - on O_3^*. $PC_{7.5}$ is overwhelmingly dependent on values of CO^* with mean of absolute coefficients (MAC) of 0.77. O_3^* and NO_2^* appear to weigh in seasonally in winters and summers respectively. Other variables appear to bring noise to the components.

In the afternoon (right figure) we note that the decomposition of $PC_{2pm.1}$ is more distorted. CO^* and NO_2^* are still significant (and positively) contributors, but their weights are now more variable (more rugged curve). Also, O_3^* is now a major contributor to $PC_{2pm.2}$, while appears as noise in $PC_{2pm.1}$. SO_2^* is a second major contributor to $PC_{2pm.2}$. However, its MAC dropped to 0.52 from 0.57. $PM_{2.5}^*$ dominates $PC_{2pm.3}$ and $PC_{2pm.5}$. The shapes of the remaining loading coefficients in other components are less discernible.

When evaluated at complementary hours (figures not shown), other dynamic loadings show similar trend in characteristics. That is higher $\overline{EV}_{1h}$ (peaking at 7am) correspond to greater linearity among loading, and vise versa.

Loading weights control variables' participation in the make up of the PCs. Hence, a greater (in absolute terms) loading coefficient of a variable implies greater contribution (from the associated variable) to the variance of the corresponding PC. So, when $v_{7..1}$ (see Figure 4.4) is juxtaposed with the corresponding $EV_{7.1}$ (see Figure 4.1), we notice the seasonal variability of $NSAO_7$ (see Figure 6.3 in Section 6.1) passing through the stable coefficients of $v_{7..1}$ yielding a seasonal variability of $PC_{7.1}$

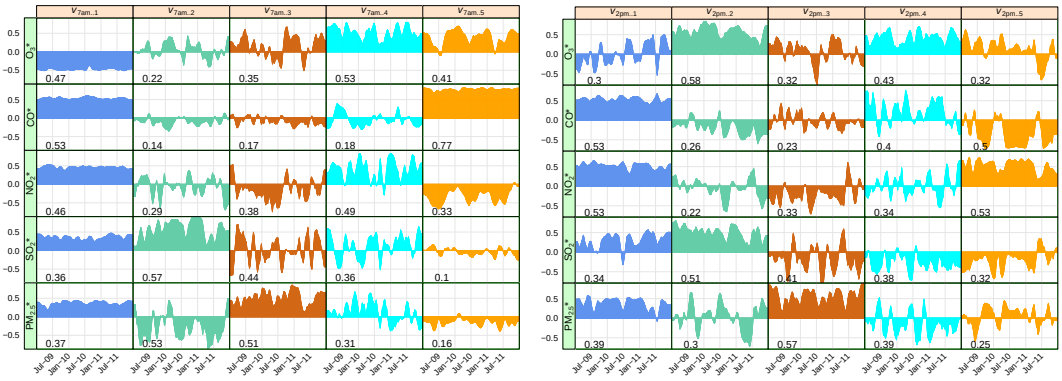

Figure 4.4: Dynamic loadings at 7am (*left*) and 2pm (*right*).

and $\text{EV}_{7.1}$. While we observe this in morning hours (near 7am, when MAC peaks), this relationship is weaker in the afternoon, especially 2pm.

Left panel of Figure 4.4 presents 5×5 loadings matrix, $\left[v_{7.pk} \in \mathbb{R}^{\mathfrak{d}}\right]_{p,k=1..5}$, computed from a PCA on a (standardized) 45-day window sliding in time along $\text{NSAO}_{7\text{am}}$. $\mathfrak{d} = 1095$ is number of days. Matrix columns, $[v_{7.pk}]_{\forall p}$, are dynamic PC loadings. Matrix elements, $[v_{7.pk}]$, are a daily TS of kth PC loading's coefficients (or weights), placed in plot panels. Linear combination (at a corresponding hour $\times$ day index) of kth PC loading and transformed observations results in a kth PC. For example, $\text{PC}_{7.1}$ is a linear combination of weighted (transformed) pollutant concentrations, i.e. $\text{PC}_{7.1} = \sum_p v_{7.p1} \odot \text{NSAO}_{7.p} \in \mathbb{R}^{\mathfrak{d}}$, where $\odot$ is a Hadamard product, and $v_{7.11}$ is a top left (daily TS in blue) element of loading matrix and so on. Refer to (2.2) for more info. Legend values (in gray on each panel) indicate the mean of absolute coefficients (MAC), i.e. $\overline{v}_{711} = \frac{1}{\mathfrak{d}} \sum_d |v_{7d11}| = .47$. Largest MAC, $\max_p \overline{v}_{hpk}$, of kth loading sets the direction, i.e. sign, of all kth loading's elements, since signs are *arbitrarily* set by many PCA computational packages (see `prcomp()` help manual in **R**). So, $(\mathring{p}, k) := \text{argmax}_p \overline{v}_{hpk}$ is largest MAC's location (panel). We flip signs of pointwise coefficients via $v_{hdpk} \cdot \text{sign}\left(v_{hd\mathring{p}k}\right)$, so as to keep $v_{hd\mathring{p}k} > 0$. Finally, we smooth coefficient series with a 45-day mean. Reflection and smoothing ease their visualization and interpretation. Horizontal units are days in a "mm/yy" format with vertical grid bars placed at 6 month increments.

When also tried varimax orthogonal rotations of loadings, but rotated coefficients were not materially more revealing.

4.4 Comparison to previous work

While application of PCA has recently gained traction in the perusal of environmental (and meteorological) data, unfortunately, most applications are still constrained to the static assessment. It is perspicuous that a time-invariant PCA is unable to seize the aforementioned two-dimensional ramifications of DPCA on a cyclostationary data, exampled with air pollution concentration series. Static PCA assumes that an observed sample is randomized, time ordering is unimportant and the underlying data patterns remain constant in time [22].

Some of the widely cited works of Statheropoulos and Abdul-Wahab ([27], [1], respectively) rely on employment of static PCA to dynamic air pollution data. Inter-

estingly, the former effort includes plots exhibiting non-stationary (seasonal) dynamics of daily time series of raw pollutant concentrations (and meteorological observations) and the latter discusses the diurnal dynamics of the pollutants. Both papers (and many other efforts) stepped in the direction of dynamic analysis by applying PCA separately to winter and summer seasons (Statheropoulos) and day and night time (Abdul-Wahab). Still this assumes that the data structure wobbles between two constant states, which is not the case with environmental and meteorological data. Moreover, there is limited discussion of PCA assumptions and robustness of the results. The latter paper utilizes standardized log (ozone) observations, but appears to leave other variables intact. The former publication does not mention any transformation of the notably cyclical observation series (see figures therein). Not surprisingly the $\text{EV}_k^{\text{static}}$ in both works remain low, under 35% for EV_1.

We consider our work an improved and proper extension of these two papers in application of PCA. In fact, when we employed their methods to our normalized set (with winter/summer and day/night observations identified analogously), we discovered a greatly improved $\text{EV}_k^{\text{static}}$, as shown below in Figure 4.5. Seasonal cycles appear much stronger in our work (see Figure 4.1) and summer/winter $\text{EV}_k^{\text{static}}$ appear to capture this with similar pattern strength in winter observations. Decomposition of day and night observations is less informative, likely due to the hours chosen by the authors (6am-5pm as day and remainder as night). Our analysis reveals the diurnal (local) dynamics among variables and suggests clustering night and morning hours separately from afternoon hours. In fact, it may be helpful to have three groups: night, morning and afternoon. Naturally, such discovery may go unnoticed without performing our DPCA technique on each hour of the day.

	EV_1	CEV_2	CEV_3	CEV_4
summer	.50	.70	.84	.96
winter	.58	.78	.88	.96
daytime	.47	.69	.86	.96
night time	.46	.68	.86	.97

Figure 4.5: Cumulative explained variance based on static PCA work of Statheropoulos (lagging O_3 observations; summer vs winter) and Abdul-Wahab (contemporaneous analysis; night vs day). We analogously aggregated hourly SAO data; then normalized with log differencing defined in (3.1) and employed PCA.

Finally, dynamic PCA yields a greater information, when compared to static PCA, about seasonal patterns in the variables, with $\text{EV}_{h.1}$ reaching $70 - 75\%$ (see Figure 4.1) in winter nights of our dataset. Our DPCA application enables a higher quality air pollution analysis targeted at a particular season or time of day. The components can further be used in regression or other statistical methods for the purposes of quality prediction and air pollution studies.

5 Conclusion

The objective of our study was to highlight the dynamic nature of air pollutants. We accomplished this objective by applying non-local DPCA at each of the 24 hours of a day to investigate Houston's air pollution profile. Thus, we constructed a two dimen-

sional analysis over hours $\times$ days domain, essentially separating diurnal and seasonal cycles. We have discovered that daylight savings have an insignificant impact on the analysis. We then chose and tested a suitable normalizer (log differencing) that transforms our data set SAO_h to an approximately multivariate normal, NSAO_h (percent change in averaged pollutant concentrations). Still, we briefly compared (at the aggregate level of MAC) DPCA done on NSAO_h versus those on the original SAO_h and (frequently used) LSAO_h datasets. We presented the dynamic explained variance and loadings at each hour.

The key finding was that the air pollution profile remains non-constant throughout a day and throughout a year. The best EV is achieved in the morning (around 7am), when loading coefficients exhibit linear and consistent (non-local) structure regardless of the season. $\text{PC}_{h.1}$ captures seasonal profile at any hour h, although its seasonal structure is poorest in the afternoon. This is when many of the dynamic loadings are least meaningful as well.

The novelty of this paper is a new and proper application of PCA to an air pollution dataset. We show that given the nature of complex pollutant associations with daily and annual cycles, it's not only important, but also highly worthwhile to apply PCA on a subset of cyclostationary data. Such practice identifies patterns of strengthening and weakening of correlations among studied variables throughout a day or a year.

We then compared our results to existing (static PCA) research efforts and concluded that DPCA unveils a much richer and more complete dynamics of the analyzed data.

This work does not attempt to build predictors, reduce dimensionality, or construct air pollution indicators. Yet, the determined uncorrelated PCs are suitable for application of further extensions such as regression, self organizing maps (SOM), artificial neural networks (ANN), and other techniques.

References

References

[1] Sabah A. Abdul-Wahab, Charles S. Bakheit, and Saleh M. Al-Alawi. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10):1263–1271, October 2005. 00228.

[2] E. Alvarez, F. de Pablo, C. Tomas, and L. Rivas. Spatial and temporal variability of ground-level ozone in Castilla-Leon (Spain). *International Journal of Biometeorology*, 44(1):44–51, May 2000. 00000.

[3] Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, Hoboken, N.J, 3 edition edition, July 2003. 00002.

[4] J. Brioude, G. Petron, G. J. Frost, R. Ahmadov, W. M. Angevine, E.-Y. Hsie, S.-W. Kim, S.-H. Lee, S. A. McKeen, M. Trainer, F. C. Fehsenfeld, J. S. Holloway, J. Peischl, T. B. Ryerson, and K. R. Gurney. A new inversion method to calculate emission inventories without a prior at mesoscale: Application to

the anthropogenic CO2 emission from Houston, Texas. *Journal of Geophysical Research: Atmospheres*, 117(D5):D05312, March 2012. 00018.

[5] Francesca Bruno and Daniela Cocchi. A unified strategy for building simple air quality indices. *Environmetrics*, 13(3):243–261, 2002. 00052.

[6] M. P. Buhr, M. Trainer, D. D. Parrish, R. E. Sievers, and F. C. Fehsenfeld. Assessment of pollutant emission inventories by principal component analysis of ambient air measurements. *Geophysical Research Letters*, 19(10):1009–1012, 1992. 00046.

[7] Martin Buhr, David Parrish, Jaimi Elliot, John Holloway, Jim Carpenter, Paul Goldan, William Kuster, Michael Trainer, Stephen Montzka, Stuart McKeen, and Fred Fehsenfeld. Evaluation of ozone precursor source types using principal component analysis of ambient air measurements in rural Alabama. *Journal of Geophysical Research: Atmospheres*, 100(D11):22853–22860, November 1995. 00000.

[8] G. M. Davis and K. B. Ensor. Outlier detection in environmental monitoring network data: an application to ambient ozone measurements for Houston, Texas. *Journal of Statistical Computation and Simulation*, 76(5):407–422, 2006. 00004.

[9] Randall L. Dougherty, Alan S. Edelman, and James M. Hyman. Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. *Mathematics of Computation*, 52(186):471–494, 1989. 00106.

[10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. 00671.

[11] Panos G. Georgopoulos and John H. Seinfeld. Statistical distributions of air pollutant concentrations. *Environmental Science & Technology*, 16(7):401A–416A, July 1982. 00152.

[12] F. L. T. Gonçalves, L. M. V. Carvalho, F. C. Conde, M. R. D. O. Latorre, P. H. N. Saldiva, and A. L. F Braga. The effects of air pollution and meteorological parameters on respiratory morbidity during the summer in São Paulo City. *Environment International*, 31(3):343–349, April 2005. 00000.

[13] Mark Gurevitz. Daylight saving time. Congressional Research Service, Library of Congress, 2005. 00002.

[14] Joseph F. Hair, William C. Black, Barry J. Babin, Rolph E. Anderson, Ronald L. Tatham, and others. *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2006. 44623.

[15] Walter Hecq, Youri Borisov, and Marc Totte. Daylight saving time effect on fuel consumption and atmospheric pollution. *Science of The Total Environment*, 133(3):249–274, June 1993. 00007.

[16] J. Edward Jackson. *A user's guide to principal components*, volume 587. John Wiley & Sons, 2005. 04958.

[17] Jyh-Cheng Jeng. Adaptive process monitoring using efficient recursive PCA and moving window PCA algorithms. *Journal of the Taiwan Institute of Chemical Engineers*, 41(4):475–481, July 2010. 00039.

[18] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002. 00000.

[19] Eugene Kim, Steven G. Brown, Hilary R. Hafner, and Philip K. Hopke. Characterization of non-methane volatile organic compounds sources in Houston during 2001 using positive matrix factorization. *Atmospheric Environment*, 39(32):5934–5946, October 2005. 00062.

[20] Kwang-Y. Kim and Qigang Wu. A Comparison Study of EOF Techniques: Analysis of Nonstationary Data with Periodic Statistics. *Journal of Climate*, 12(1):185–199, January 1999. 00107.

[21] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. MVN: An R Package for Assessing Multivariate Normality. *A peer-reviewed, open-access publication of the R Foundation for Statistical Computing*, page 151, 2014. 00013.

[22] Wenfu Ku, Robert H. Storer, and Christos Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1):179–196, November 1995. 00773.

[23] Duncan Lee, Claire Ferguson, and E. Marian Scott. Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1):109–126, January 2011. 00011.

[24] A. Lengyel, K. Heberger, L. Paksy, O. Banhidi, and R. Rajko. Prediction of ozone concentration in ambient air using multivariate methods. *Chemosphere*, 57(8):889–896, 2004. 00050.

[25] Xueqin Liu, Uwe Kruger, Tim Littler, Lei Xie, and Shuqing Wang. Moving window kernel PCA for adaptive monitoring of nonlinear processes. *Chemometrics and Intelligent Laboratory Systems*, 96(2):132–143, 2009. 00098.

[26] Nicola Loperfido and Peter Guttorp. Network bias in air quality monitoring design. *Environmetrics*, 19(7):661–671, November 2008. 00017.

[27] M. Statheropoulos, N. Vassiliadis, and A. Pappa. Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment*, 32(6):1087–1095, 1998. 00153.

[28] Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Academic press, 1979. 07877.

[29] Ricardo Muñoz. Morning peak of air pollutant concentrations in urban areas: Effect of time lag between emissions and turbulence. In *Seventh Symposium on the Urban Environment*, 2007. 00002.

[30] D. D. Parrish, M. Trainer, M. P. Buhr, B. A. Watkins, and F. C. Fehsenfeld. Carbon monoxide concentrations and their relation to concentrations of total

reactive oxidized nitrogen at two rural U.S. sites. *Journal of Geophysical Research: Atmospheres*, 96(D5):9309–9320, May 1991. 00121.

[31] Pedro R. Peres-Neto, Donald A. Jackson, and Keith M. Somers. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005. 00269.

[32] Walter W. Piegorsch and A. John Bailer. *Analyzing Environmental Data*. Wiley, Chichester, West Sussex, England ; Hoboken, NJ, 1 edition edition, March 2005. 00121.

[33] D. K. Pissimanis, V. A. Notaridou, N. A. Kaltsounidis, and P. S. Viglas. On the Spatial Distribution of the Daily Maximum Hourly Ozone Concentrations in the Athens Basin in Summer. *Theoretical and Applied Climatology*, 65(1-2):49–62, January 2000. 00009.

[34] G. Q. Qian, G. Gabor, and R. P. Gupta. Principal Components Selection by the Criterion of the Minimum Mean Difference of Complexity. *Journal of Multivariate Analysis*, 49(1):55–75, April 1994. 00000.

[35] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, 3rd edition, August 2010. 00000.

[36] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014. 01270.

[37] S. I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, and M. C. Pereira. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software*, 22(1):97–103, January 2007. 00232.

[38] M. Trainer, D. D. Parrish, P. D. Goldan, J. Roberts, and F. C. Fehsenfeld. Review of observation-based analysis of the regional factors influencing ozone concentrations. *Atmospheric Environment*, 34(12–14):2045–2061, 2000. 00106.

[39] Xun Wang, Uwe Kruger, and George W. Irwin. Process monitoring approach using fast moving window PCA. *Industrial & Engineering Chemistry Research*, 44(15):5691–5702, 2005. 00147.

[40] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987. 03367.

[41] Shengkun Xie and Sridhar Krishnan. Dynamic Principal Component Analysis with Nonoverlapping Moving Window and Its Applications to Epileptic EEG Classification. *The Scientific World Journal*, 2014, 2014:11, January 2014. 00000.

[42] Heidi G. Yacker. Daylight Saving Time. Congressional Research Service, Library of Congress, 1998. 00001.

[43] Hwa-Lung Yu, Yuan-Chien Lin, and Yi-Ming Kuo. A time series analysis of multiple ambient pollutants to investigate the underlying air pollution dynamics and interactions. *Chemosphere*, 2015. 00000.

6 Appendix

6.1 Daylight saving time (DST)

Most businesses operate in local time, setting pace for traffic hours and, hence, pollutant emissions [13, 42, 15]. Likewise, most of the Americas use DST to extend evening hours into daylight at the expense of morning hours. In particular, Houston, and the whole of Texas, are in the Central Time (CT) zone. This zone follows the Central Daylight Time (CDT) convention from a "jump" day in mid-March to a "compression" day in early November and the Central Standard Time (CST) convention for the remainder of the calendar year. CDT and CST are 5 and 6 hours (respectively) behind Coordinated Universal Time (UTC), which is Greenwich Mean Time (GMT), which does not observe DST.

Initially, our raw data is indexed with UTC-6:00 (i.e. ignores CST/CDT adjustments) uninterrupted (no jumps or compressions) hourly increments. However, the relation of pollutants to traffic and diurnal human activity prompts the investigation of the effect of DST [29] on PCA outcome. Apparently, the use of the CST/CDT index has made only a diminutive amelioration (of 0.01%) in EV_1. The whole improvement came from the PCA of a moving window over the jump and compression days.

Still we carry on the analysis in local (i.e. CST/DST) time zone. This results in one missing 2am observation when CDT goes into effect on jump day, and one duplicate when CST takes effect on compression day in each year. For simplicity, we interpolate the former and delete the later.

Figure 6.1 exemplifies a jump in observations when time shifts from CDT to CST. The left panel shows *non-local* observations, i.e. daily concentrations at a fixed time (at a 24 hour lag). The right panel shows *local* observations, i.e. consecutive hourly concentrations, as defined in [8]. Note that (averaged) non-local CO levels remain higher for the adjusted data at 8am, i.e. black curve is atop blue curve on the left panel. This is expected, since the CST/CDT-indexed concentrations reflect morning traffic's CO emissions faster than the UTC-6:00 indexed measurements. The right panel shows a shadow effect as unadjusted concentrations remain one hour behind the adjusted ones.

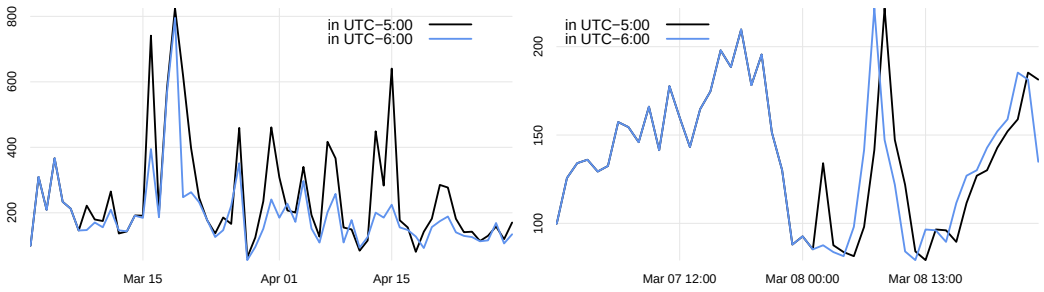

Figure 6.1: DST effect on CO measurements (in ppb). In 2009 CST/CDT jump occurred on March 8 at 2am. On jump day local time shifts forward by one hour from UTC-6:00 to UTC-5:00, i.e. 1am CST → 2am CDT and so on. *Left*: Daily measurements at 8am. *Right*: Hourly measurements around time change (jump event).

Supplemental material

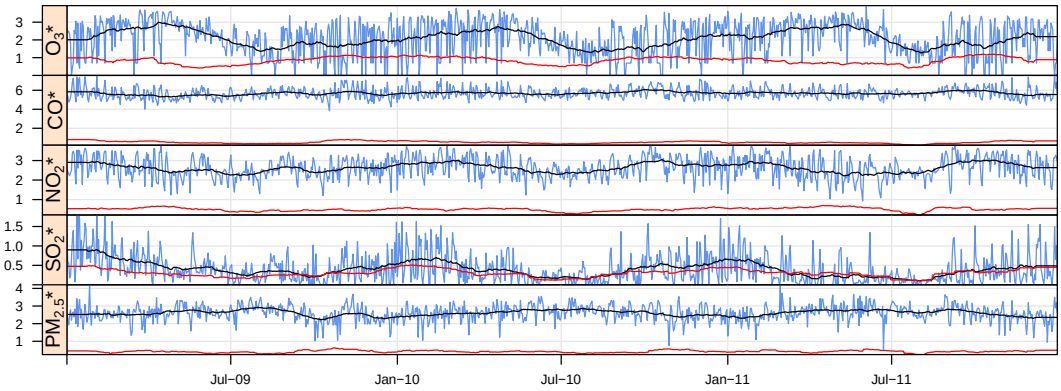

Figure 6.2: $LSAO_{7am}$. Measurements are adjusted for DST. Overlaid curves are 45-day rolling statistics: simple mean (black), standard deviation (red). Asterisk in O_3^* is the notation for the transformed O_3 concentrations.

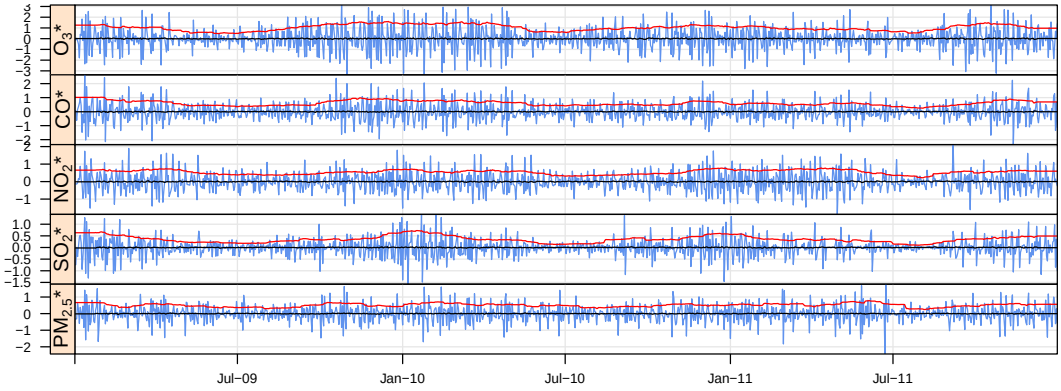

Figure 6.3: $NSAO_{7am}$. Measurements are adjusted for DST. Overlaid curves are 45-day rolling statistics: simple mean (black), standard deviation (red). Asterisk in O_3^* is the notation for the transformed O_3 concentrations.

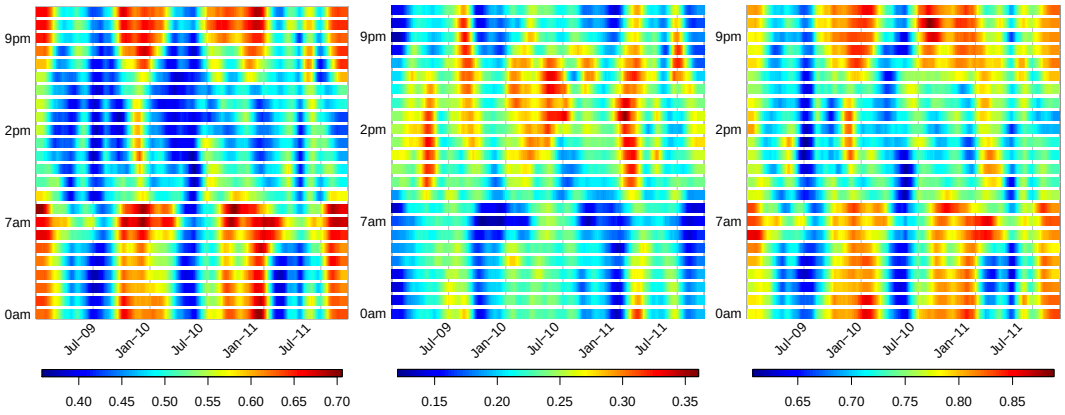

Figure 6.4: Heatmap of dynamic explained variance (EV) from $PC_{h.1}$ and $PC_{h.2}$ of SAO_h and SAO_h, $h = 0..23$. *Left* and *center*: non-cumulative for the first two components. *Right*: cumulative for both components.

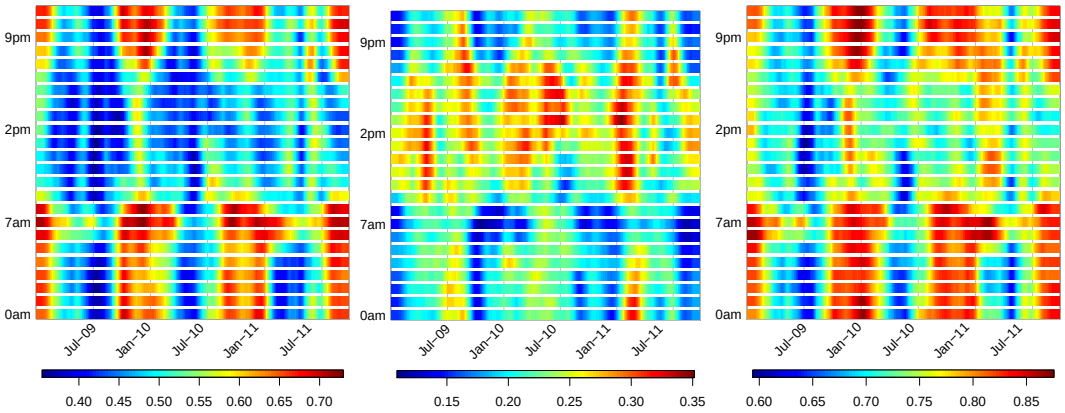

Figure 6.5: Heatmap of dynamic explained variance (EV) from $PC_{h.1}$ and $PC_{h.2}$ of SAO_h and $LSAO_h$, $h = 0..23$. *Left* and *center*: non-cumulative for the first two components. *Right*: cumulative for both components.

Dotted lines (in matching colors) represent the non-local means across the whole 3 year period. Vertical units are proportions on 0-1 scale (1 is 100% contribution to variance).

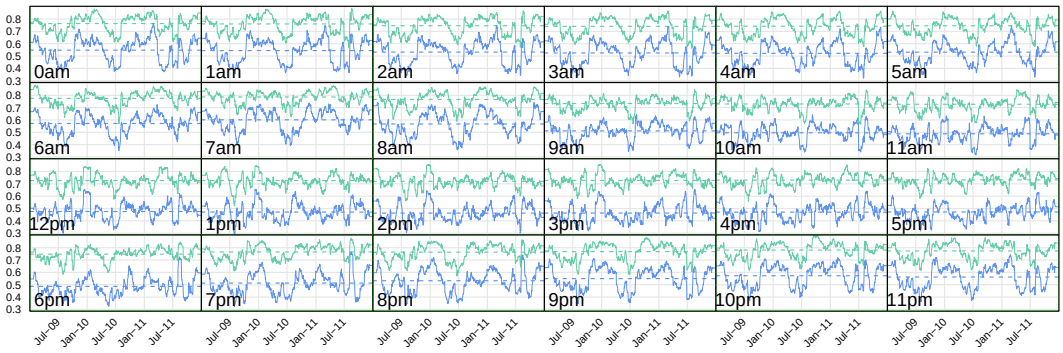

Figure 6.6: CEV from PC_1 and PC_2 of $SAO_h, h = 0..23$ hours.

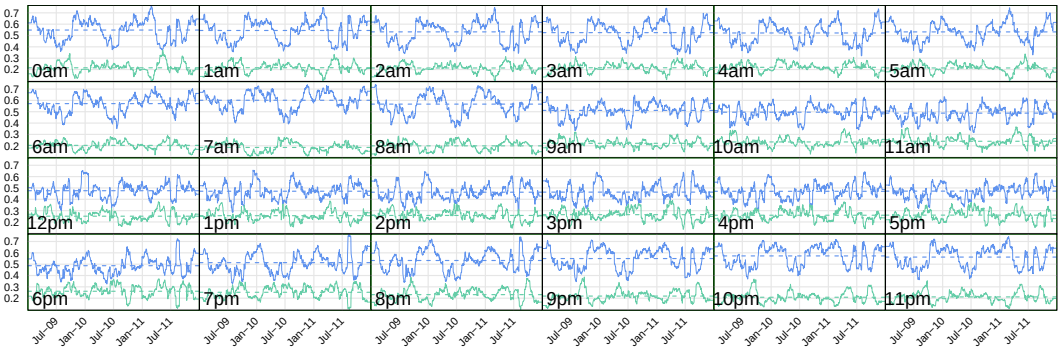

Figure 6.7: EV from PC_1 and PC_2 of $SAO_h, h = 0..23$ hours.

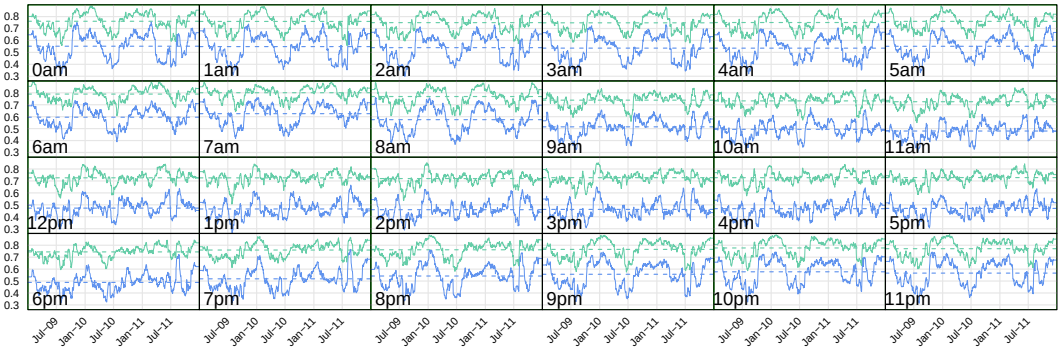

Figure 6.8: CEV for the first two components (EV_{ihd}, $i = 1, 2$) of $LSAO_h, h = 0..23$ hours.

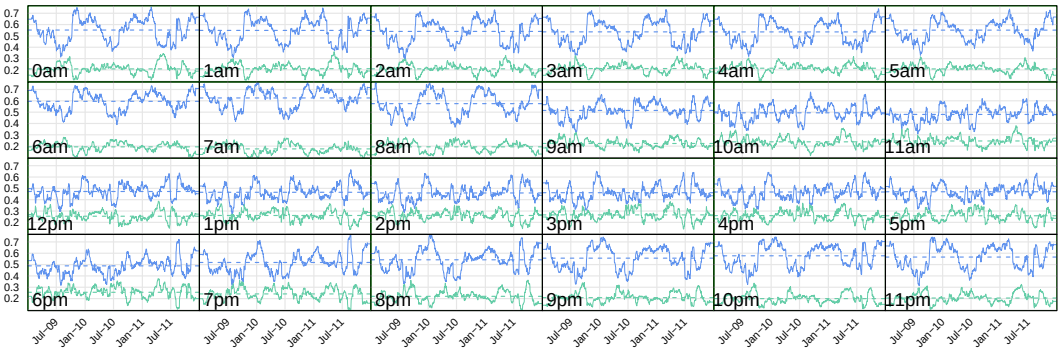

Figure 6.9: EV for the first two components (EV_{ihd}, $i = 1, 2$) of $LSAO_h, h = 0..23$ hours.

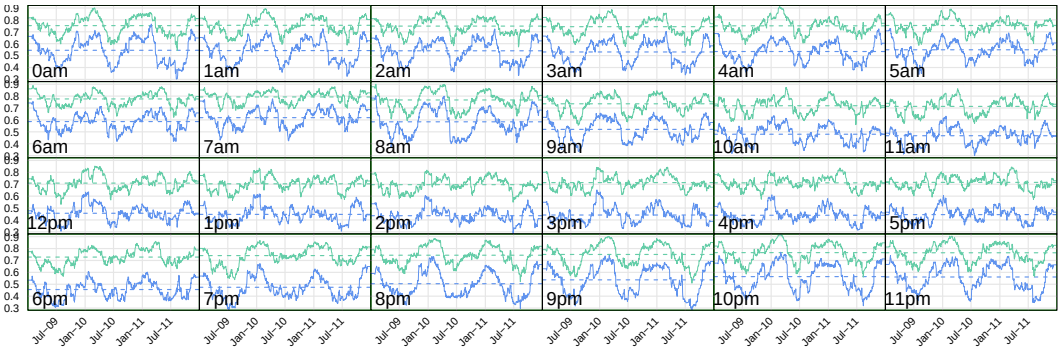

Figure 6.10: CEV for the first two components (EV_{ihd}, $i = 1, 2$) of $\mathrm{NSAO}_h, h = 0..23$ hours.

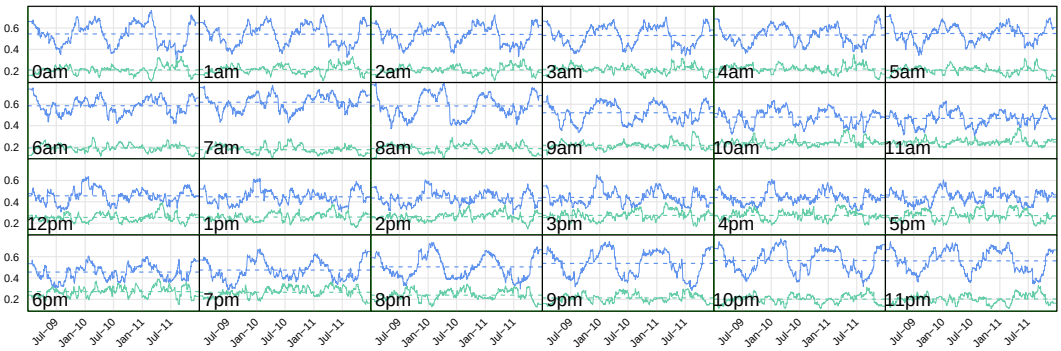

Figure 6.11: EV for the first two components (EV_{ihd}, $i = 1, 2$) of $\mathrm{NSAO}_h, h = 0..23$ hours.