2013

DIMENSION REDUCTION FOR POWER SYSTEM MODELING USING PCA METHODS CONSIDERING INCOMPLETE DATA READINGS

Ting Zhao
Michigan Technological University

DIMENSION REDUCTION FOR POWER SYSTEM MODELING USING PCA METHODS CONSIDERING INCOMPLETE DATA READINGS

By

Ting Zhao

A REPORT

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Mathematical Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY

2013

This report has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Mathematical Sciences.

Department of Mathematical Sciences

Report Advisor: *Jianping Dong*

Committee Member: *Thomas D. Drummer*

Committee Member: *Chee-Wooi Ten*

Department Chair: *Mark Gockenbach*

1. Abstract

Principal Component Analysis (PCA) is a popular method for dimension reduction that can be used in many fields including data compression, image processing, exploratory data analysis, etc. However, traditional PCA method has several drawbacks, since the traditional PCA method is not efficient for dealing with high dimensional data and cannot be effectively applied to compute accurate enough principal components when handling relatively large portion of missing data. In this report, we propose to use EM-PCA method for dimension reduction of power system measurement with missing data, and provide a comparative study of traditional PCA and EM-PCA methods. Our extensive experimental results show that EM-PCA method is more effective and more accurate for dimension reduction of power system measurement data than traditional PCA method when dealing with large portion of missing data set.

2. Introduction

The most important property of PCA is that it can achieve the optimal results in terms of mean squared error (MSE) via performing linear transformation of high dimensional vectors. As a result, a new set of much lower dimensional vectors can be obtained, while the original data set can be reconstructed approximately.

The objective of this project is to reduce the cost in building response surface performance models for power system control and optimization. The dimension reduction techniques of parameter space have been applied by integrated circuit modeling researchers in the past decade, where PCA plays an important role [7]. It has been shown that high-dimensional circuit parameters will first be reduced by dimension reduction techniques, such as PCA method. Then the top few principal components (linear combinations of the original parameter set) will be used for building the quadratic (second order) response surface models for circuit performances. In [7],

researchers are working on dimension reduction methods for integrated circuit applications, whereas in our project dimension reduction for power distribution system modeling is concerned. However, traditional PCA method has several drawbacks. One of them is that the traditional PCA method is not suitable for dealing with high dimensional data, since computing the sample covariance required by PCA method can be very costly. Consequently, it is desirable to avoid computing sample covariance for high dimensional data set. Another drawback is that it is not obvious how to compute the principal components properly with missing data.

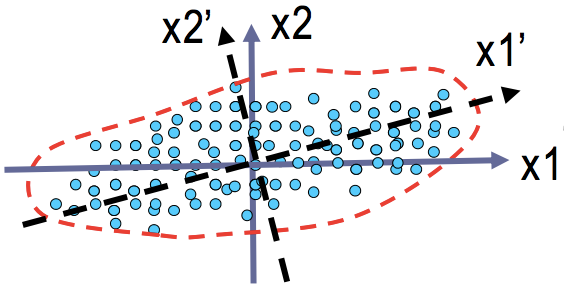
To address these drawbacks of traditional PCA method, a variety of the alternative PCA-like dimension reduction methods have been proposed in many research fields. The goal of such research is to improve classical PCA method such that the computational efficiency can be improved and missing data in the high dimensional data set can be also handled.

In the following sections, we will introduce alternative PCA-based dimension reduction methods and also demonstrate experimental results on realistic high dimensional power system measurement data set.

3. Classic Principal Component Analysis

We will start with a simple example for introducing traditional PCA method. Consider using a straight line to best represent scattered points in two-dimensions (with X1-X2 coordinates), which is shown in the following figure. The key is to find a new coordinate (X1'), such that data variance observed along X1' is maximized. Similarly, for m-dimensional data set with m orthogonal coordinates, it is usually very important to find a much fewer new coordinates such that the maximum variance of the original data set can be well preserved. The classical PCA method can be applied to solve the above problems by using Least Squared Error minimization method.

Figure 1.



The general derivation of PCA is in terms of a standardized linear projection that maximizes the variance in the projected space [3]. For n observed m-dimensional vectors (n samples) $X = [x_1 x_2 ... x_n]^T$, PCA can find the top-one principal component (axe) that maximize the data variance preserved from the original data set, which can be achieved by finding an m-dimension vector (x_0) such that

$$J_0(x_0) = \sum_{k=1}^{n} \|x_0 - x_k\|^2 \qquad (1)$$

is the smallest, where $J_0(x_0)$ is Mean Squared Error criterion function. Let

$$m = \frac{1}{n} \sum_{k=1}^{n} x_k \qquad (2)$$

So we have: [4]

$$J_0(x_0) = \sum_{k=1}^{n} \|(x_0 - m) - (x_k - m)\|^2 = \sum_{k=1}^{n} \|x_0 - m\|^2 - 2(x_0 - m)^t \sum_{k=1}^{n} (x_k - m) + \sum_{k=1}^{n} \|x_k - m\|^2$$

$$= \sum_{k=1}^{n} \|x_0 - m\|^2 + \sum_{k=1}^{n} \|x_k - m\|^2 \qquad (3)$$

where $\sum_{k=1}^{n} \|x_k - m\|^2$ is independent of x_0. So if $x_0 = m$, $J_0(x_0)$ will be the smallest.

Suppose that we want to find a coordinate X1' and the goal is still to minimize the squared-error. Let every x_k has a corresponding element x_k' in the X1'. If we move it in the vector e direction, then [4]

$$x_k^{'} = m + a_k e \tag{4}$$

where a_k is a scalar. Now we redefine the squared-error criterion function as follows:

$$J_1(a_1,...,a_n,e) = \sum_{k=1}^{n} \| (x_k^{'} - x_k) \|^2 = \sum_{k=1}^{n} \| ((m + a_k e) - x_k) \|^2 = \sum_{k=1}^{n} \| (a_k e - (x_k - m)) \|^2$$

$$= \sum_{k=1}^{n} a_k^2 \| e \|^2 - 2\sum_{k=1}^{n} a_k e^t (x_k - m) + \sum_{k=1}^{n} \| x_k - m \|^2 \tag{5}$$

where e is unit vector and $\| e \|^2 = 1$, hence

$$J_1^{'}(a_1,...,a_n,e) = \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k e^t (x_k - m) \tag{6}$$

Solving partial differential for a_k, we can get

$$a_k = e^t (x_k - m) \tag{7}$$

It means if we have already known vector e, any vector x_k projected to the line X1' can be computed by taking the inner product with e^t. Then we can obtain the new coordinate after linear transformation ($x_k^{'} = a_k$). Hence, we can rewrite $J_1(e)$, such that [4]

$$J_1(e) = \sum_{k=1}^{n} a_k^2 \| e \|^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \| x_k - m \|^2 = -\sum_{k=1}^{n} [e^t (x_k - m)]^2 + \sum_{k=1}^{n} \| x_k - m \|^2$$

$$= -\sum_{k=1}^{n} e^t (x_k - m)(x_k - m)^t e + \sum_{k=1}^{n} \| x_k - m \|^2 = -e^t S + \sum_{k=1}^{n} \| x_k - m \|^2 \tag{8}$$

where $S = \sum_{k=1}^{n} e^t (x_k - m)(x_k - m)^t e$, is called a scatter matrix. Using method of Lagrange Multipliers, we will obtain vector e so that $e^t Se$ is maximized. Let

$$u = e^t Se - \lambda (e^t e - 1) \tag{9}$$

and obtain the partial differential for vector e,

$$\frac{\partial u}{\partial e} = 2Se - 2\lambda e \tag{10}$$

Let it equal 0, so

$$Se = \lambda e \tag{11}$$

This is a classic problem that is related to eigendecomposition and the vector e is exactly one of the eigenvectors.

Suppose we need to reduce m-dimensional vectors to k-dimensional vectors, we will select the top k eigenvectors with their corresponding eigenvalues and project every x_i onto them, then we can get $x_i^{'} = [x_{i1}^{'} x_{i2}^{'} ... x_{ik}^{'}]$, such that [4]

$$x_{ij}^{'} = e_j^T (x_i - m) \tag{12}$$

where m is the data sample mean.

Therefore, the classic PCA algorithm can reduce dimension successfully and make data analysis and system modeling easier.

4. EM-PCA Algorithm

In this section, we will introduce the expectation maximization (EM) algorithm for principal component analysis of data set [2][5][8]. The EM-PCA algorithm can handle high dimensional data more efficiently than traditional PCA method since it doesn't need to calculate the sample covariance matrix explicitly.

4.1 Probabilistic Model of PCA

Principal component analysis can be used as a limitation case of linear Gaussian models. Linear Gaussian model is to assume y as a linear transformation of some k dimensional latent variable x plus additive Gaussian noise. Let G be the m*k matrix, and v is the m-dimensional error vector (with covariance matrix R), so the general model can be written as

$$y = Gx + v \tag{13}$$

where $x \sim N(0, I)$, the error $v \sim N(0, R)$. So the y is a corresponding

Gaussian-distribution vector for the observations

$$y \sim N(0, GG^T + R) \tag{14}$$

We know the probabilistic model is the following according to [5]:

$$p(x) = (2\pi)^{-m/2} \exp(-\frac{1}{2} x^T x) \tag{15}$$

For the case of error $R = \sigma^2 I$, the probability distribution over y space for a given x is

$$p(y \mid x) = (2\pi\sigma^2)^{-m/2} \exp(-\frac{1}{2\sigma^2} \| y - Gx \|^2) \tag{16}$$

So, the distribution of y is [5]

$$p(y) = \int p(y \mid x) p(x)\, d\, x = (2\pi)^{-m/2} \mid W \mid^{-1/2} \exp[(-\frac{1}{2}) y^T W^{-1} y] \tag{17}$$

where W is a $m \times m$ matrix and $W = \sigma^2 I + GG^T$.

Using Bayes' rule, $p(x \mid y) = \frac{p(x) p(y \mid x)}{p(y)}$, the latent variables x given the observed y can be calculated as follows: [5]

$$p(x \mid y) = (2\pi)^{-m/2} \mid \sigma^{-2} M \mid^{-1/2} \exp[-\frac{1}{2}(x - M^{-1} G^T m) \sigma^{-2} M (x - M^{-1} G^T y)] \tag{18}$$

where M is a $k \times k$ matrix and $M^{-1} = (\sigma^2 I + G^T G)^{-1}$.

Therefore, according to [5], the log function of y is

$$L = \sum_{n=1}^{N} \ln[p(y_n)] = -\frac{N}{2}[m \ln(2\pi) + \ln \mid W \mid + tr(W^{-1} S)] \tag{19}$$

where S is the sample covariance matrix of the observations $\{y_n\}$, $S = \frac{1}{N} \sum_{n=1}^{N} y_n y_n^T$.

Estimates for G and σ^2 can be obtained by iterative maximization of L using the EM algorithm.

4.2 The EM-PCA Algorithm

We use the EM algorithm to obtain the parameter G and σ^2 in the probabilistic model. We will use the latent variables $\{x_n\}$ to be the missing data and consider the complete data for composing the observations together with these latent variables. The complete-data log-likelihood can be derived as follows [5]:

$$L_C = \sum_{n=1}^{N} \ln\{p(y_n, x_n)\} \tag{20}$$

Using formula (15)、(16),we can obtain [5]

$$p(y_n, x_n) = (2\pi\sigma^2)^{-m/2} \exp(-\frac{\| y_n - Gx_n \|^2}{2\sigma^2})(2\pi)^{-k/2} \exp\{-\frac{1}{2} x_n^T x_n\} \tag{21}$$

In the *E-step*, we take the expectation of L_C with respect to the distribution $p(x_n \mid y_n, G, \sigma^2)$: [5]

$$< L_C >= -\sum_{n=1}^{N} [\frac{m}{2} \ln \sigma^2 + \frac{1}{2} tr(< x_n x_n^T >) + \frac{1}{2\sigma^2} y_n^T y_n - \frac{1}{\sigma^2} < x_n >^T G^T y_n + \frac{1}{2\sigma^2} tr(G^T G < x_n x_n^T >)] \tag{22}$$

where $< x_n >= M^{-1} G^T y_n$, $< x_n x_n^T >= \sigma^2 M^{-1} + < x_n >< x_n >^T$, $M = \sigma^2 I + G^T G$.

In the *M-Step*, $< L_C >$ is maximized with respect to G and σ^2 giving the new parameter estimates [5]

$$G_{new} = (\sum_n y_n < x_n >^T)(\sum_n < x_n x_n^T >)^{-1} \tag{23}$$

$$\sigma^2{}_{new} = \frac{1}{Nm} \sum_{n=1}^{N} [\| y_n \|^2 - 2 < x_n >^T G_{new}^T y_n + tr(< x_n x_n^T > G_{new}^T G_{new})] \tag{24}$$

These equations are iterated in sequence until the algorithm is considered to have converged [5].

5. PCA with Missing Data

In the above sections, we introduced a traditional PCA algorithm and described an existing EM algorithm for calculating principal components. In the next section, we will discuss PCA methods for data sets with missing points. In general, the

real-world data is not always complete. If the number of missing data points is very small compared to the full data set, we can directly use the sample mean to substitute the missing data points. On the other hand, if the number of missing data points is relatively large, the following EM formulation of PCA is applied for handling missing values [6].

5.1 EM-PCA with Missing Data

We consider the same problem when the data matrix Y has missing values and suppose that values are missing randomly. For example, Y matrix can be written in the following form:

Figure 2

$$Y = \begin{bmatrix} y_{11} & y_{12} & y_{13} & * & * \\ y_{21} & y_{22} & * & y_{24} & y_{25} \\ y_{31} & * & * & y_{34} & y_{35} \end{bmatrix}$$

where every $*$ indicates the missing value. For such kind of data set where relatively large portion of data is missing, the EM-PCA with missing data can be used [6]. Now, our goal is to find the parameters G and σ that maximize the likelihood of observed data (vectors y is partially observed). We use EM algorithm that estimates in the E-step the missing values (the vector x and the missing parts of the y which is denoted by y_h). In the M-step, we fix these estimates, and maximize the expected joint log-likelihood of x and y.

We assume that the distribution over x and y_h factors so that we get the lower-bound log-likelihood as follows [6]:

$$\log p(y_0) = \frac{1}{2}\log |M| + [q(x) + q(y_h)] + E_q[\log p(x) + \log p(y \mid x)] \qquad (25)$$

We will maximize this equation with respect to the distribution of q in the E-step, and respect to the parameters in the M-step.

E-step: From the above we find the optimal distributions q as [6]

$$q(y_h) \propto \exp \int q(x) \log(y_h \mid x) \sim N(y_h; G_h \bar{x}, \sigma^2 I) \tag{26}$$

$$q(x) \propto p(x \mid y_0) \exp \int q(y_h) \log(y_h \mid x) \sim N(x; \sigma^{-2} M G^T \bar{y}, M) \tag{27}$$

where $\bar{y}$ is the mean of $q(y_h)$ for the missing data, y_0 for the observed part, and $\bar{x}$ is the mean of $q(x)$.

M-step: Based on equation (28), it follows as [6]

$$\sum_{n=1}^{N} E_{qn}[\log p(x_n) + \log p(y_n \mid x_n)] = -\frac{ND}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\sum_n \| y_n - G\, \bar{x}_n \|^2 - tr(GMG^T))$$

$$-\frac{D_h}{2\sigma^2} \sigma_{old}^2 - \frac{1}{2} \sum_n \| x_n \|^2 - \frac{N}{2} tr(M) \tag{28}$$

where D_h is the number of missing data points, σ_{old} is the current value for σ that was used in the E-step to compute the q.

Maximizing the equation (31) over G and σ, we can obtain [6]

$$G_{new} = \bar{X} \bar{Y}^T (NM + \bar{X} \bar{X}^T)^{-1} \tag{29}$$

$$\sigma^2 = \frac{1}{Nm} [Ntr(GMG^T) + \sum_n \| \bar{y}_n - G\bar{x}_n \| + D_h \sigma_{old}^2] \tag{30}$$

where $\bar{X}$, $\bar{Y}$ are the matrices that collect all $\bar{x}$, $\bar{y}$ as columns, separately.

6. Experimental Results

As mentioned at the beginning of this project report, our goal is to perform dimension reductions on the measured data set, and subsequently build distribution performance models that will be utilized by MTU's power distribution control center to predict future electricity usages in December such that resources can be optimized accordingly. The parameter dimension reduction methods demonstrated in this work are implemented in Matlab programming language, and will significantly reduce the performance model characterization cost, thereby allowing for real time processing of huge amount of data obtained by the-state-of-art smart meters.

We demonstrate the results obtained by traditional PCA and EM-PCA methods for analyzing power distribution data sets in EERC building of MTU during last December. Three smart meters have collected the data sets, and each of them records a measurement every ten minutes. It should be noted that the meter readings are just input parameters for building the response surface models that will predict activities happened inside a building, though the detailed model extraction procedures have not started yet. We use the meter reading of each hour as a variable, since the events during two consecutive hours are typically not strongly correlated. For instance, typical undergraduate courses will not last longer than 50 minutes. So we consider the sum of meter readings in every hour as a variable, so we have totally 24*3=72 variables. Consequently, for 31 days in December, we have a data set with 72*31=2232 measurements. We also want to emphasize that the meter readings are not always independent. Since some facilities in the EERC building are using more than one power supply sources, multiple meter readings will be changing if such facilities are started. For instance, air heater will be started with the fan for air circulation at the same time. If they are using two separate power sources, the readings of two meters will be simultaneously affected.

Table 1:Example of the original data set

Timestamp	Trend Flags	Status	Value
12:00:08 AM 12/1/12 EST	0	0	235.3
12:10:00 AM 12/1/12 EST	0	0	245.3
12:20:00 AM 12/1/12 EST	0	0	247
12:30:08 AM 12/1/12 EST	0	0	234.3
12:40:00 AM 12/1/12 EST	0	0	233.3
12:50:00 AM 12/1/12 EST	0	0	235.4
1:00:07 AM 12/1/12 EST	0	0	233.8
1:10:00 AM 12/1/12 EST	0	0	237.2
1:20:00 AM 12/1/12 EST	0	0	244.2
1:30:07 AM 12/1/12 EST	0	0	230
1:40:00 AM 12/1/12 EST	0	0	233.3
1:50:00 AM 12/1/12 EST	0	0	233.5
2:00:04 AM 12/1/12 EST	0	0	232.4
2:10:00 AM 12/1/12 EST	0	0	245.5
2:20:00 AM 12/1/12 EST	0	0	233

We also want to note that this is not a traditional time series problem, since the meter measurements of one time point (hour) is not likely to be influenced by the previous many time points (hours) measurements. The missing data points analyzed in our problem are typically due to the instability of smart meter devices manufactured by different vendors. Although the meters in the EERC building do not show obvious instabilities, the meters operating in extreme conditions (extremely high temperature environment, extreme humidity conditions, etc) can produce incorrect/missing results. Since this research project is targeting general power systems modeling that can involve smart meters operating under any environments, different missing data rates (10% to 50%) are considered.

1. Results obtained by using traditional PCA and EM-PCA with full data set have been demonstrated as follows.

Figure 3.Singular values after performing traditional PCA with full data set

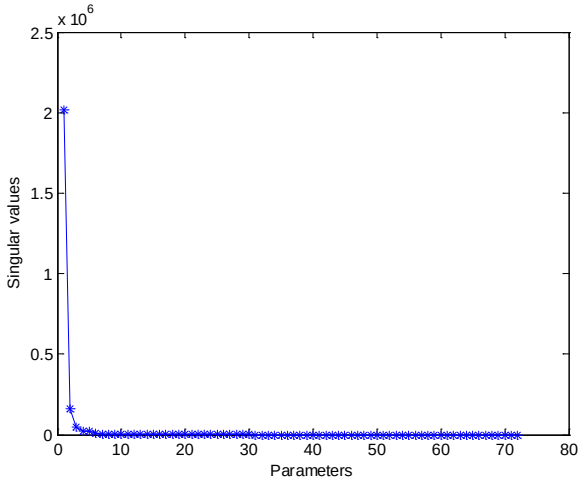


Table 2. Top 7 Singular values

	Singular value
1	2016500
2	158700
3	49800
4	22800

5	19700
6	8300
7	6000

It is observed that keeping only top three principle components will be sufficient since they can already explain more than 99% variability of the original data set.

Table 3. Top 3 Principle Component Mapping Vectors (PCMVs)

PCMV1	PCMV2	PCMV3
0.0234	-0.0134	-0.1333
0.0246	-0.0048	-0.1157
0.0137	0.0003	-0.1134
0.0165	0.0115	-0.1459
0.0152	0.0087	-0.1446
0.0174	0.0242	-0.1391
0.0217	0.0089	-0.1173
0.0343	0.0283	-0.1108
0.0731	0.0631	-0.0662
0.0975	0.0747	-0.0277
0.1057	0.08	-0.0299
0.1133	0.0729	-0.0428
0.1113	0.0114	-0.0258
0.1118	0.0286	-0.0492
0.1144	0.025	-0.1082
0.1077	0.0212	-0.0465
0.1003	-0.0287	-0.0753
0.0693	-0.0432	-0.0884

0.0589	-0.0643	-0.088
0.0502	-0.0591	-0.0779
0.045	-0.073	-0.1309
0.0432	-0.0644	-0.1336
0.032	-0.0628	-0.1393
0.0281	-0.0553	-0.1104
0.048	-0.03	-0.1631
0.0466	-0.0234	-0.1448
0.0379	-0.017	-0.1405
0.0418	0.0026	-0.1638
0.0381	-0.0031	-0.156
0.0395	0.0132	-0.1622
0.0461	-0.0014	-0.1217
0.0627	0.0238	-0.1207
0.1186	0.0788	-0.0507
0.1587	0.0946	0.0158
0.1764	0.1118	0.0251
0.1886	0.1086	0.0136
0.1879	0.0324	0.0286
0.1899	0.0432	0.0005
0.1862	0.0328	-0.0766
0.1762	0.0222	-0.0161
0.1612	-0.0433	-0.0474
0.1153	-0.0759	-0.0662
0.0957	-0.1013	-0.0828
0.083	-0.0998	-0.0749
0.0751	-0.1141	-0.1374

0.0711	-0.1016	-0.141
0.0593	-0.0998	-0.1507
0.0525	-0.0851	-0.1313
0.0612	-0.0432	-0.2577
0.0582	-0.0274	-0.1536
0.0509	-0.0523	-0.1322
0.0531	-0.0144	-0.1607
0.0462	-0.0118	-0.1508
0.0498	0.0002	-0.1253
0.0522	-0.0544	-0.1263
0.1124	0.0988	0.0517
0.193	0.1958	0.2485
0.1985	0.1585	0.1158
0.2133	0.1477	0.089
0.2238	0.1415	0.032
0.2237	0.0602	0.0854
0.2228	0.0485	-0.0127
0.2248	0.0265	-0.0555
0.2364	0.05	0.2136
0.2148	-0.0157	0.1939
0.1734	-0.0866	0.1378
0.1301	-0.3295	0.1045
0.1144	-0.3303	0.1516
0.11	-0.3642	0.0994
0.1018	-0.3366	0.0864
0.0771	-0.3376	0.13
0.06	-0.3391	0.1087

Comparing principal component mapping vectors (PCMVs) obtained by traditional PCA and EM-PCA with full data set, we obtained the following results. If the scatter plot shows that all dots fall on the line y=x, the elements inside the two vectors are perfectly matching with each other.

Figure 4. PCMV1

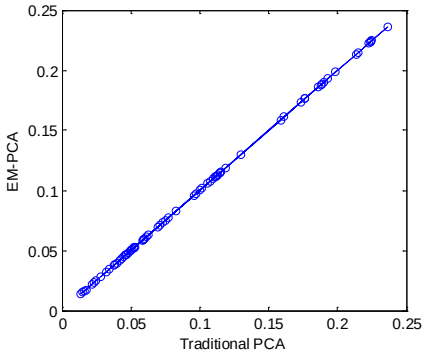
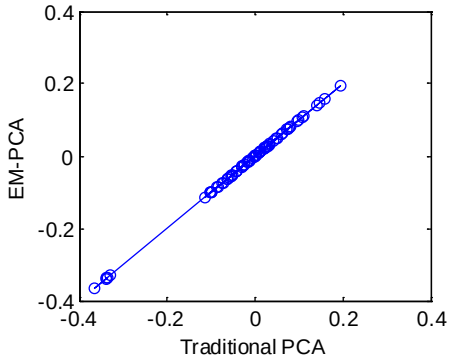


Figure 5. PCMV 2



2. Results of the principal component mapping vectors obtained by using traditional PCA (that replaces missing data by mean values) and EM-PCA with partial data have been demonstrated as follows. We consider various missing data rates that range from 10% to 50%.

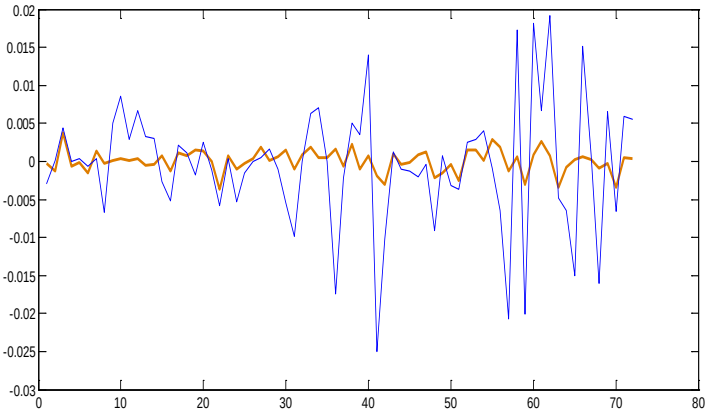
(1) When missing data rate is 10%, results are shown below.

a) *Error of finding the PCMV1*

Figure 6

Traditional PCA: Range (0.5%-2.5%)

EM-PCA: Range (0.2%-0.6%)



---- Traditional PCA

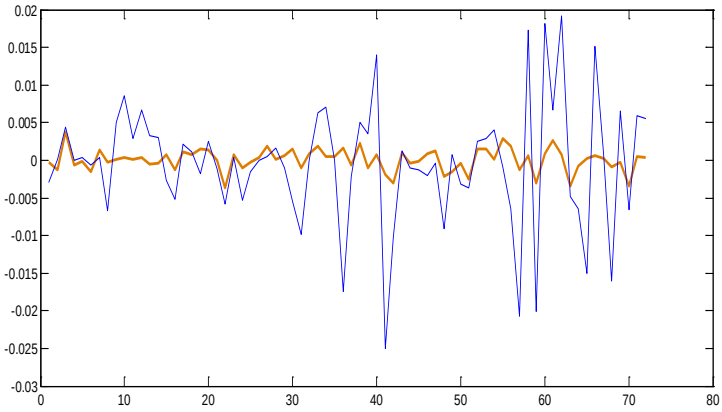
---- EM-PCA

(b) *Error of finding the PCMV2:*

Figure 7

Traditional PCA: Range (4%-8%)

EM-PCA: Range (0.5%-1.5%)



---- Traditional PCA

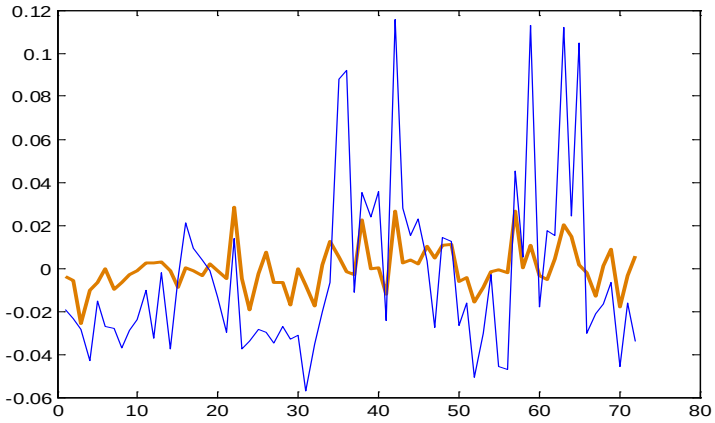
---- EM-PCA

(c) *Error of finding the PCMV3:*

Figure 8

Traditional PCA: Range (5%-20%)

EM-PCA: Range (1%-5%)



---- Traditional PCA

---- EM-PCA

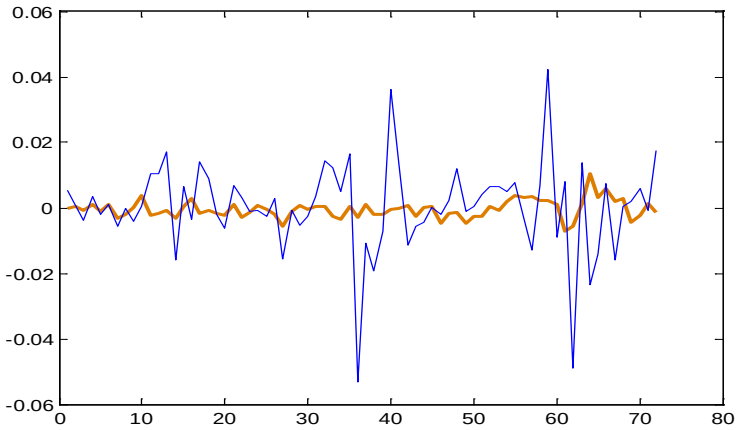
(2) When missing data rate is 20%, results are shown below.

a) *Error of finding the PCMV1:*

Figure 9

Traditional PCA: Range (1%-4%)

EM-PCA: Range (0.2%-0.5%)



---- Traditional PCA

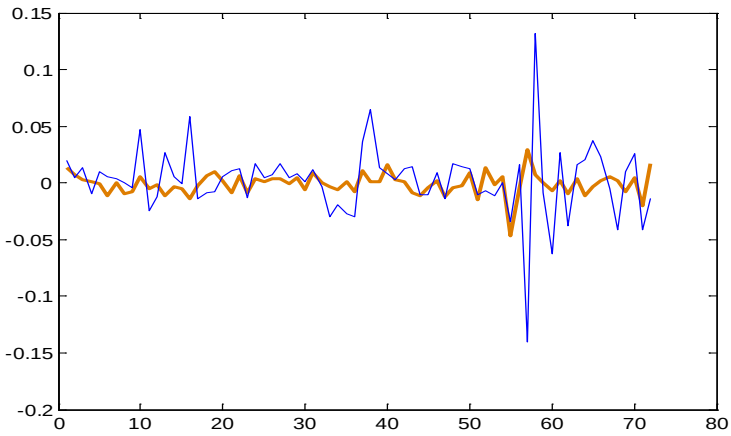
---- EM-PCA

b) *Error of finding the PCMV2:*

Figure 10

Traditional PCA: Range (2%-15%)

EM-PCA: Range (1%-4%)



---- Traditional PCA

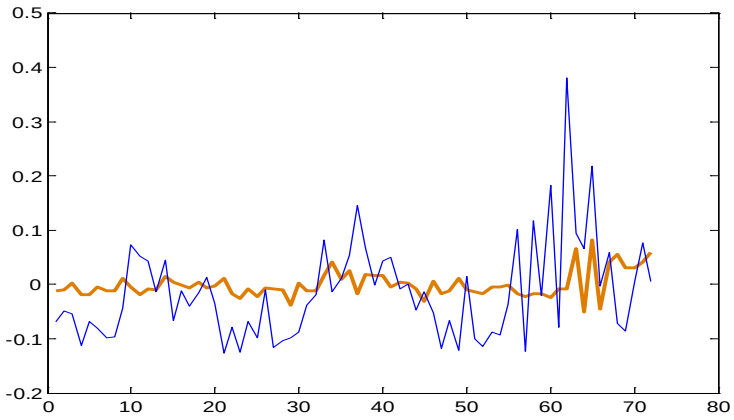
---- EM-PCA

c) *Error of finding the PCMV3:*

Figure 11

Traditional PCA: Range (5%-25%)

EM-PCA: Range (1%-4%)



---- Traditional PCA

---- EM-PCA

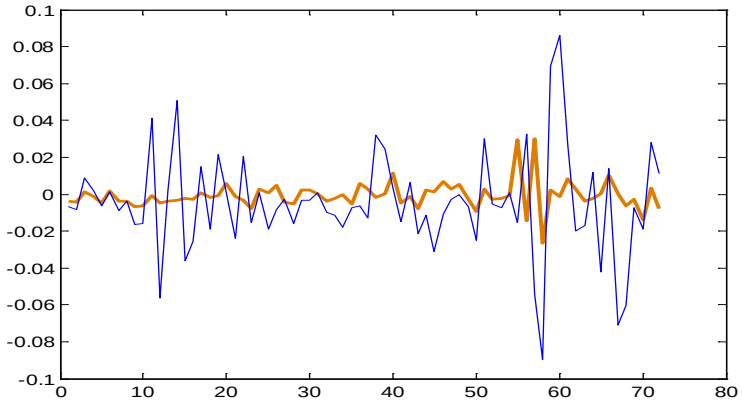
(3) When missing data rate is 50%, results are shown below.

a) *Error of finding the PCMV1:*

Figure 12

Traditional PCA: Range (5%-10%)

EM-PCA: Range (1%-2.5%)



---- Traditional PCA

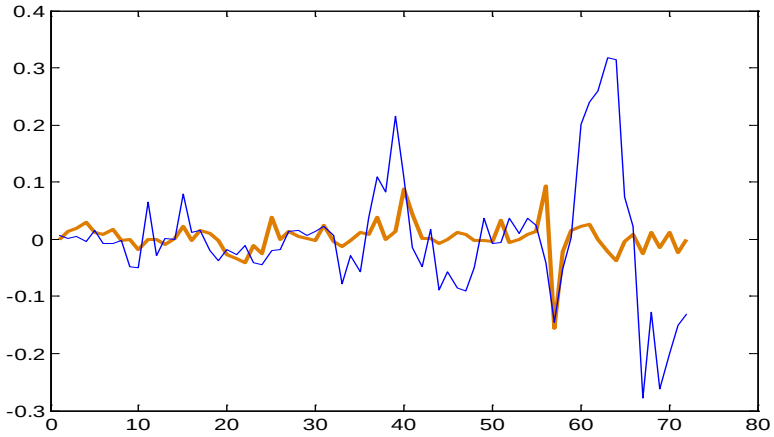
---- EM-PCA

b) *Error of finding the PCMV2:*

Figure 13

Traditional PCA: Range (5%-30%)

EM-PCA: Range (0.5%-10%)



---- Traditional PCA

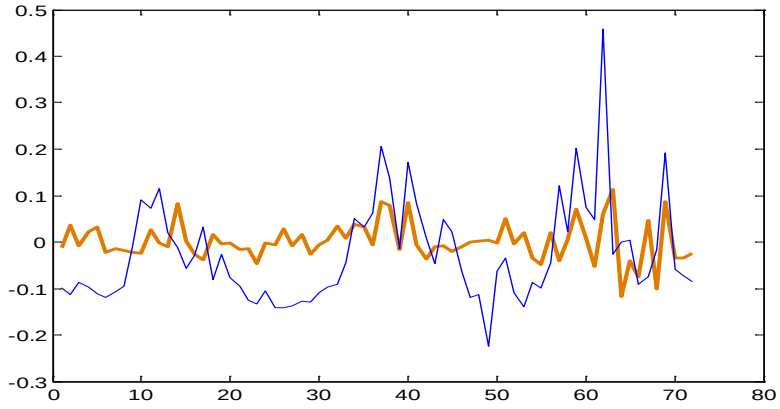
---- EM-PCA

c) *Error of finding the PCMV3:*

Figure 14

Traditional PCA: Range (5%-40%)

EM-PCA: Range (5%-10%)



---- Traditional PCA

---- EM-PCA

Since inner products of the three principle component mapping vectors obtained by using the traditional PCA with full data set and the EM-PCA with partial data (the traditional PCA with partial data set) can indicate the quality of the computed principles, we demonstrate the following inner products results.

Table 4. PCMV 1

	10% missing data	20% missing data	50% missing data
Traditional PCA	0.9982	0.9958	0.9757
EM-PCA	0.9999	0.9997	0.9985

Table 5. PCMV 2

	10% missing data	20% missing data	50% missing data
Traditional PCA	0.9788	0.9333	0.5169
EM-PCA	0.9981	0.9959	0.9688

Table 6. PCMV 3

	10% missing data	20% missing data	50% missing data
Traditional PCA	0.8717	0.3677	0.1794
EM-PCA	0.9956	0.9781	0.8590

Inner product of two unit-length vectors equals to cosine value of their angels. ($\langle v \cdot w \rangle = \| v \| \cdot \| w \| \cdot \cos\theta \Rightarrow \cos\theta = \frac{\langle v \cdot w \rangle}{\| v \| \cdot \| w \|}$, where θ is the angel of vector v and w). Obviously, the EM-PCA method is consistently much better than tradition PCA method in handling missing data, since the inner product values obtained by EM-PCA are closer to 1, indicating an almost perfectly matching result with the true principle components obtained by traditional PCA with full data set.

7. Conclusion

In this work, we have presented the fundamentals of traditional PCA method and EM-PCA method, and their applications in handling missing data. By analyzing power distribution data sets in EERC building of MTU during last December, we have performed dimension reductions on the measured data set for future research on power distribution system modeling by using the traditional PCA and EM-PCA methods. In the last, we conclude that the EM-PCA method is consistently more effective and accurate in finding principal components than tradition PCA method when missing data set has to be considered.

References

[1] T. Wiberg. Computation of principal components when data is missing. In *Proc. Second Symp. Computational Statistics*, pages 229-236, 1976

[2] Sam Roweis. EM algorithm for PCA and SPCA. *Neural Information Processing Systems*,1997

[3] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*,1993.

[4] Richard O.duda, Peter E. hart, David GStork. Pattern Classification

[5] Tipping M E, Bishop C M. Probabilistic principle component analysis. *Journal of the Royal Statistical Society*. 1999.61(3):611-622

[6] Jakob Verbeek, Note on Probabilistic PCA with Missing Values.

[7] Xin Li, Jiayong Le, Lawrence T. Pileggi: Statistical Performance Modeling and Optimization. *Foundations and Trends in Electronic Design Automation* 1(4) (2006).

[8] Tipping M E, Bishop C M. Mixture of Probabilistic Principle Component Analyses. Neural Computation, 1999, 11(2):443-482.