

MATH3290 Mathematical Modeling

Tutorial 4

11th October 2017

Outline

1 Principal Components Analysis

- Framework of PCA
- Example of PCA
- Details of Implementation

2 Review

- Revision
- Examples

PCA Method

Given a set of data points $x_1, \dots, x_n \in \mathbb{R}^d$, define the $d \times d$ matrix:

$$Q = \frac{1}{n} \sum_{j=1}^n \tilde{x}_j \tilde{x}_j^T,$$

$$\tilde{x}_j = x_j - m,$$

$$m = \frac{1}{n} \sum_{j=1}^n x_j.$$

Compute the eigenvectors u_k and eigenvalues λ_k of Q ,

$k = 1, \dots, d$. Assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$.

Remark

1. In some applications, the statisticians may first apply the **normalization** to their data set that dividing by the standard deviation to each vector \tilde{x}_j .
2. In the setting of our course, one can perform PCA without any normalization.
3. In the example below, we will see how to normalize the data.

PCA Method (Cont.)

- Choose k principal directions u_1, \dots, u_k . Define the following rate of contribution as follows:

$$w_i = \frac{\lambda_i}{\sum_{j=1}^d \lambda_j}, \quad i = 1 \dots, d.$$

Choose the smallest k such that

$$w_1 + \dots + w_k \geq \delta,$$

where δ is in the interval $[0, 1]$. Usually, $\delta \approx 90\%$.

PCA Method (Cont.)

- Compute the projection of \tilde{x}_j to these eigenvectors, where $j = 1, \dots, n$, that is

$$c_{js} := \tilde{x}_j^T u_s, \quad s = 1, \dots, k.$$

Storage is $n \times k + d \times k + d$.

- Recover the data using

$$\hat{x}_j = \sum_{s=1}^k c_{js} u_s + m, \quad j = 1, \dots, n$$

- Compute the relative error e_i as follow:

$$e_i = \frac{\|x_j - \hat{x}_j\|}{\|x_j\|}, \quad j = 1, \dots, n.$$

Example of PCA

Consider the following data for the 34 cities in China (2002):

- x_1 : Gross Domestic Product (thousand million CNY);
- x_2 : General Expected Income (thousand million CNY);
- x_3 : Fixed Property Investment (thousand million CNY);
- x_4 : Total Import-Export Value from Foreign trading (thousand million USD);
- x_5 : Per Capita Disposable Income (CNY);
- x_6 : Per Capita GDP (CNY);

Example of PCA

Exmple of PCA (Cont.)

City	x_1	x_2	x_3	x_4	x_5	x_6
Shanghai	5408.8	717.8	2158.4	726.6	13250	36206
Beijing	3130	534	1814.3	872.3	12464	24077
Guangzhou	3001.7	245.9	1001.5	525.1	13381	38568
Shenzhen	2239.4	303.3	478.3	279.3	24940	136071
Tianjin	2022.6	171.8	811.6	228.3	9338	20443
Chongqing	1971.1	157.9	995.7	17.9	7238	9038
Hangzhou	1780	118.3	769.4	131.1	11778	38247
Chengdu	1663.2	78.3	702.1	20.8	8972	20111
Qingdao	1518.2	100.7	367.8	169.3	8721	26961
Ningbo	1500.3	111.8	747.2	122.7	12970	35446
Wuhan	1493.1	85.8	570.4	22	7820	16206

Example of PCA

Example of PCA (Cont.)

City	x_1	x_2	x_3	x_4	x_5	x_6
Dalian	1406	98.7	601.3	146	8200	29706
Nanjing	1295	144.1	602.9	10.1	9157	27128
Haerbin	1232.1	67.7	361.1	17.1	7004	18244
Jinan	1200	66.3	404.7	14.9	8982	25192
Shijiazhuang	1184	44.5	412.3	11.4	7230	25476
Fuzhou	1160.2	60.2	284	61	9191	31582
Changchun	1150	37.8	320.5	28.9	6900	21336
Zhengzhou	926.8	54.2	340	10.4	7772	16028
Xian	823.5	60.1	338.2	18.7	7184	15493
Changsha	810.9	46.1	362.6	16.6	9021	23942
Kunming	730	54.7	290	13.4	7381	24109

Example of PCA

Example of PCA (Cont.)

City	x_1	x_2	x_3	x_4	x_5	x_6
Xiamen	648.3	64.3	211.7	151.9	11768	38567
Nanchang	552	25.7	137	9.1	7021	18388
Taiyuan	432.2	26.8	147.6	15.1	7376	12821
Hefei	412.4	29.1	168.6	23	7144	17770
Lanzhou	386.8	21.1	194.5	5.1	6555	15051
Nanning	356	26.2	122.9	5.5	8796	16121
Urumqi	354	37.3	147.9	6.4	8653	17655
Guiyang	336.4	33	187.4	5.7	7306	11728
Hohhot	300	16.6	131.3	3.4	6996	11789
Haikou	157.9	8.5	82.6	11.3	8004	23920
Yinchuan	133	11.1	73	2.3	6848	11975
Xining	121.3	7.2	77.4	1	6444	6676

Example of PCA (Cont.)

Our tasks here are the following:

1. Perform **PCA** on the data set. Solving the corresponding eigen-pairs.
2. Calculate the rate of contributions and determine how many principal components should be taken.
3. Sorting these cities with the overall competitiveness.

Example of PCA (Cont.)

1. The dimensions of the data set are $n = 34$ and $d = 6$.
First, we form the data matrix $D \in \mathbb{R}^{n \times d}$ as follows:

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}.$$

Example of PCA (Cont.)

- Normalize the data. Given the data matrix D , do the following:

$$X = \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{s_1} & \dots & \frac{x_{1d} - \bar{x}_d}{s_d} \\ \frac{x_{21} - \bar{x}_1}{s_1} & \dots & \frac{x_{2d} - \bar{x}_d}{s_d} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{s_1} & \dots & \frac{x_{nd} - \bar{x}_d}{s_d} \end{pmatrix},$$

where

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ji}, \quad i = 1, \dots, p,$$

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2.$$

Example of PCA (Cont.)

- Form the matrix Q as follows:

$$Q = \frac{1}{n} X^T X.$$

- Calculate the eigenvalues λ_i and the corresponding eigenvectors v_i of Q . Assume that

$$\lambda_1 \geq \cdots \geq \lambda_d.$$

Example of PCA (Cont.)

2. Calculate the rate of contributions w_i ($i = 1, \dots, d$) and choose the smallest k such that

$$w_1 + \dots + w_k \geq 90\%.$$

Here, we found that $k = 2$ is suitable in our case.

3. For the i^{th} city, define the score of competitiveness as follows:

$$p_i = \left(w_1(x_{i1} \cdots x_{id})v_1 + w_2(x_{i1} \cdots x_{id})v_2 \right) / 10000,$$

where v_1 and v_2 are the eigenvectors corresponding to λ_1 and λ_2 , respectively.

Example of PCA (Cont.)

- Finally, we can sort the city with the score of competitiveness. First five of the cities are

City	Score
Shenzhen	5.8764
Shanghai	2.0389
Guangzhou	2.0131
Hangzhou	1.8853
Xiamen	1.8502

Example of PCA (Cont.)

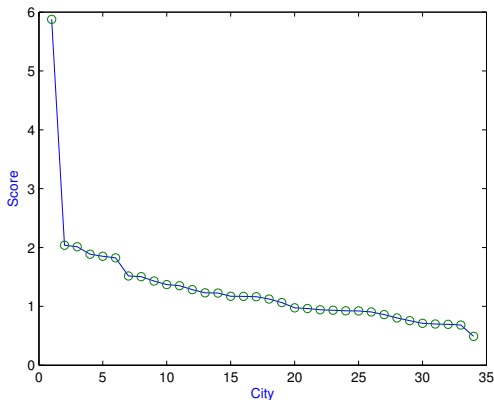


Figure: The score of competitiveness for all cities.

Summary

So far, we have studied five chapters of this course and I think the following topics are important among them:

1. Least-square Method, especially the linearized techniques.
2. Spline Model, e.g. natural cubic spline. In addition, we can let our admire function to be any other polynomial accompanied some conditions (say quadratic polynomial, at that moment we call it quadratic spline).

Summary (cont.)

3. Interpolation and Divided difference techniques. (Basic calculation required)
4. The computation of eigenvalue and eigenvector. Solving eigen-problem involves in our course very much, such as the part of predicting the long-term behavior of a linear model or in the computation of PCA method.

Example I

Given a data set as follow:

x	1.00	1.25	1.50	1.75	2.00
y	5.10	5.79	6.53	7.45	8.46

Assume that our desired model function is of the form $f(x) = Ce^{Kx}$, where C and K are parameters to be determined. Find such parameters C and K that best fit the data set given above.

Solution of Example I

Since this model function f is not a linear model with respect to the parameters C and K , we convert this f into the form:

$$g(x) = \log(f(x)) = \log(C) + Kx,$$

and $g(x)$ is linear with respect to the **new** parameter $T = \log(C)$ and K . By taking logarithm on our data set y , we introduce $Y = \log(y)$ and obtain the following **new** data set:

x	1.00	1.25	1.50	1.75	2.00
Y	1.629	1.756	1.876	2.008	2.135

Solution of Example I (Cont.)

Then, we define the matrix A and vector b to be

$$A = \begin{pmatrix} 1 & 1.00 \\ 1 & 1.25 \\ 1 & 1.50 \\ 1 & 1.75 \\ 1 & 2.00 \end{pmatrix}, \quad b = \begin{pmatrix} 1.629 \\ 1.756 \\ 1.876 \\ 2.008 \\ 2.135 \end{pmatrix}.$$

Thus, we can obtain the **normal equation** as follow:

$$A^T A \begin{pmatrix} T \\ K \end{pmatrix} = \begin{pmatrix} 5 & 7.50 \\ 7.50 & 11.875 \end{pmatrix} \begin{pmatrix} T \\ K \end{pmatrix} = \begin{pmatrix} 9.404 \\ 14.422 \end{pmatrix} = A^T b.$$

Solution of Example I (Cont.)

Finally, we obtain $T = 1.122$ and $K = 0.5056$. By the definition of T , we obtain $C = 3.071$ and the model function f that best fits the data set is

$$f(x) = 3.071e^{0.5056x}.$$

We can also compare the error between y_i and $f(x_i)$:

y	5.10	5.79	6.53	7.45	8.46
$f(x)$	5.09	5.78	6.56	7.44	8.44

The the least-square error $\sum_{i=1}^5 (y_i - f(x_i))^2 = 0.0016$.

Example II

Given a data set as follow:

x	-1	0	1
y	3	4	6

We will use a **quadratic spline**, called Q , defined on $-1 \leq x \leq 1$ to fit the given data, satisfying the conditions below:

- It is a quadratic polynomial on each subinterval $([-1, 0]$ or $[0, 1])$;
- It passes through all the data points;
- The first derivative is continuous at all interior points.

Example II (Cont.)

1. Derive a set of equations arising from the above conditions.
2. Determine the quadratic spline, so that the following quantity

$$\int_{-1}^1 \left(Q'(x) \right)^2 dx$$

is minimized.

Solution of Example II

We assume the quadratic spline $Q(x)$ satisfies

$$Q(x) = a_0 + a_1x + a_2x^2 \quad \text{in } [-1, 0],$$

$$Q(x) = b_0 + b_1x + b_2x^2 \quad \text{in } [0, 1].$$

Next, we derive the linear system as follows:

$$3 = a_0 - a_1 + a_2$$

$$4 = a_0$$

$$4 = b_0$$

$$6 = b_0 + b_1 + b_2$$

$$a_1 = b_1$$

Examples

Solution of Example II (Cont.)

Then, we get $a_0 = b_0 = 4$, $a_2 = a_1 - 1$ and $b_2 = 2 - a_1$. Here a_1 can be viewed as a parameter. Next we apply the minimization condition, obtain

$$\begin{aligned}
 \int_{-1}^1 (Q'(x))^2 dx &= \int_{-1}^0 (Q'(x))^2 dx + \int_0^1 (Q'(x))^2 dx \\
 &= \int_{-1}^0 (a_1 + 2a_2x)^2 dx + \int_0^1 (b_1 + 2b_2x)^2 dx \\
 &= \int_{-1}^0 (a_1 + (2a_1 - 2)x) dx \\
 &\quad + \int_0^1 (a_1 + (4 - 2a_1)x) dx \\
 &= \frac{2}{3} \left(a_1^2 + a_1 + 8 \right) = \frac{2}{3} \left(a_1 + \frac{1}{2} \right)^2 + \frac{31}{6}.
 \end{aligned}$$

Examples

Solution of Example II (Cont.)

Since

$$\frac{2}{3} \left(a_1 + \frac{1}{2} \right)^2 \geq 0,$$

we conclude that

$$\min \int_{-1}^1 \left(Q'(x) \right)^2 dx = \frac{31}{6}, \quad a_1 = -\frac{1}{2}.$$

Hence, we have

$$a_0 = b_0 = 4, \\ a_1 = -\frac{1}{2} = b_1, \quad a_2 = -\frac{3}{2} \quad \text{and} \quad b_2 = \frac{5}{2}.$$

Example III

Consider the following model:

$$P_{n+1} = 0.75P_n + 0.4Q_n$$

$$Q_{n+1} = 0.25P_n + 0.6Q_n$$

This model can be written in matrix form as $u_{n+1} = Au_n$, where

$$A = \begin{pmatrix} 0.75 & 0.4 \\ 0.25 & 0.6 \end{pmatrix} \quad \text{and} \quad u_n = \begin{pmatrix} P_n \\ Q_n \end{pmatrix}.$$

Example III (Cont.)

- (a) Find the eigenvalues λ_1, λ_2 of A , assuming $\lambda_1 \geq \lambda_2$.
- (b) Find the eigenvectors of A corresponding to λ_1, λ_2 , denoted by x_1 and x_2 , respectively. Further, we assume $\|x_1\|_2 = \|x_2\|_2 = 1$.
- (c) Let $u_0 = (2 \ 5)^T$, and let x_1, x_2 be the eigenvectors of A obtained in part (b). Find the constants c_1 and c_2 , such that

$$u_0 = \begin{pmatrix} 2 \\ 5 \end{pmatrix} = c_1 x_1 + c_2 x_2.$$

- (d) Predict the long-term behavior with the initial conditions $P_0 = 2$ and $Q_0 = 5$.

Solution of Example III

In the previous tutorial, we mentioned that to predict the long-term behavior of the linear model, one can solve the eigen-problem within A and the initial condition P_0 and Q_0 .

- (a) To obtain the eigenvalues, we solve the eigen-polynomial of matrix A : $\det(\lambda I - A)$. Then, the eigenvalues of A are $\lambda_1 = 1$ and $\lambda_2 = 0.35$.
- (b) The eigenvector x_1 corresponding to λ_1 is

$$x_1 = \begin{pmatrix} \frac{8}{\sqrt{89}} \\ \frac{5}{\sqrt{89}} \end{pmatrix}.$$

The eigenvector x_2 corresponding to λ_2 is

$$x_2 = \begin{pmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Solution of Example III (Cont.)

- (c) Take x_1 and x_2 as above. In order to obtain the coefficients c_1 and c_2 , we solve the following linear system:

$$\begin{pmatrix} \frac{8}{\sqrt{89}} & \frac{-1}{\sqrt{2}} \\ \frac{5}{\sqrt{89}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}.$$

Thus, the coefficients are

$$c_1 = \frac{7\sqrt{89}}{13} \quad \text{and} \quad c_2 = \frac{30\sqrt{2}}{13}.$$

Solution of Example III (Cont.)

(d) From the matrix form of the model, we obtain

$$u_n = \cdots = A^n u_0 = A^n \begin{pmatrix} 2 \\ 5 \end{pmatrix} = \frac{7\sqrt{89}}{13} x_1 + \frac{30\sqrt{2}}{13} (0.35)^n x_2$$

Since $0.35 < 1$, we can conclude that the long-term behavior (limit) with the given initial conditions is

$$P = \frac{56}{13} \quad \text{and} \quad Q = \frac{35}{13}.$$

Remark of Example III

Note that the inverse of $I - A$ does not exist (since $\lambda_1 = 1$ is the eigenvalue of A) and it implies that the following system

$$(I - A)u = \begin{pmatrix} 0.25 & -0.4 \\ -0.25 & 0.4 \end{pmatrix} u = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has infinitely many solutions. Hence, every point on the line (which is the same direction as x_1)

$$Q = 0.625P$$

is the equilibrium point of this system. Also, one can check that the limit

$$P = \frac{56}{13}, \quad Q = \frac{35}{13}$$

is also one of the equilibrium point of the system.

Remark of Example III (Cont.)

Therefore, the long-term behavior of the system will go to the direction of x_1 with any given initial condition u_0 . The system in some sense is also stable.

Remark of Example III (Cont.)

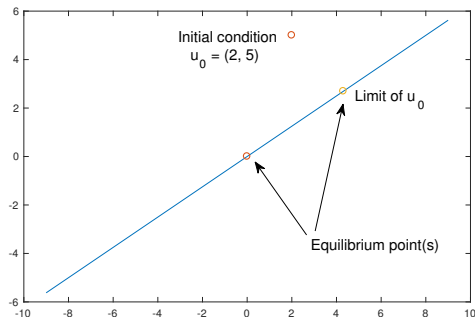


Figure: Equilibrium points, initial condition and limit.