이 학 석 사 학 위 논 문

PCA with Extreme Value Data

극값자료와 주성분 분석

2013년 2월

서울대학교 자연과학대학원

통계학과

김 용 민

PCA with Extreme Value Data

지도교수 오 희 석

이 논문을 이학석사 학위논문으로 제출함
2013년 2월

서울대학교 자연과학대학원
통계학과
김 용 민

김용민의 이학석사 학위논문을 인준함
2012년 10월

위 원 장	(인)
부위원장	(인)
위 원	(인)

PCA with Extreme Value Data

by

Yongmin Kim

A Thesis
submitted in fulfillment of the requirement
for the degree of
Master of Philosophy
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2013

Abstract

Principal component analysis (PCA) is very important to analyze multivariate data set. PCA method makes understanding data set easier because PCA can reduce the dimension of the data set. But the assuming of normal distributed data, PCA can not apply the non-normal data set like skewed data. In this paper, we suggest a quantile PCA that can apply the PCA methodology to the skewed data set.

Keywords: Multiscale, principal component analysis, extreme value, quantile.
Student Number: 2011-20242

Contents

1 Introduction **1**

2 Review of PCA **3**

2.1 Definition of Principal Components 3

2.2 Derivation of Principal Components 4

3 Quantile PCA **6**

3.1 Definition of Quantile PCA 6

3.2 Example of Quantile PCA 7

4 Conclusions **11**

List of Figures

3.2.1 (a) Four leading PCs obtained from the ordinary PCA for maximum daily precipitation in August. (b) Four leading PCs from 50% quantile PCA for the same data. (c) Four leading PCs from 1% quantile PCA for the same data. (d) Four leading PCs from 99% quantile PCA for the same data. 8

3.2.2 RMSE values between real maximum precipitation data and the reconstructions by one to four PCs obtained from the ordinary PCA (black), 50% quantile PCA (red), 1% quantile PCA (green) and 99% quantile PCA (blue). 10

Chapter 1

Introduction

The Principal component analysis (PCA) is probably the oldest, and best known of the methods of multivariate analysis. PCA has been made for dimension reduction, while principal components (PCs) have most of variation of original data set. Like many multivariate techniques, PCA was not often used before because calculating many matrix product are complicated and time-consuming work. But nowadays thanks to technical development of computer systems, we can simply get the answer of matrix product. So PCA is now make a good use of analyzing multivariate data set.

The purpose of classical PCA is to find a transformed data set that has no correlation each other and explain most of the variation of the original data, also has fewer dimension than original data set. PCA method is explained in more detail in the next chapter.

In this paper we propose a quantile PCA. The key components of the proposed method are two-fold: First, the original PCA can be expressed as a least squares framework so it uses quadratic functions as a loss function, and second we replace quadratic loss function with a convex loss function. We use check function to get a quantile information, widely used for fine a sample quantile

in the data set. We use $\tau = 0.01,\ 0.5,\ 0.99$ quantiles to calculate the results of the quantile PCA and compare with the result of ordinary PCA.

However, it can not differentiate the tip point of the check function is a fatal deficiency of ordinary check function. So we also propose a modified check function. We can easily overcome this defection only changing the neighborhood of tip point of the check function with other differentiated function. In this paper, we use quadratic function rather than absolute function only the neighborhood of the tip point.

Chapter 2

Review of PCA

2.1 Definition of Principal Components

The key point of PCA is to reduce the dimensionality of a data set, while keeping as much as possible of the variation present in the data set. This purpose is accomplished by transforming to a new set of variables, the PCs, which are uncorrelated and ordered so that the first few PCs have most of the variation in all of the original variables.

Suppose that $\mathbf{x}$ is a vector of p random variables, and we are interested that the variances of the p random variables and the structure of the correlations between p variables. Although PCA does not ignore correlations, it focused on variances of data set.

First, we need to find a linear function $\boldsymbol{\alpha}_1'\mathbf{x}$ having maximum variance, where $\boldsymbol{\alpha}_1$ is a vector of p constants $\alpha_{11},\ \alpha_{12},\ \ldots,\ \alpha_{1p}$. So that,

$$\boldsymbol{\alpha}_1'\mathbf{x} = \sum_{j=1}^{p} \alpha_{1j} x_j.$$

Next, find a linear function $\boldsymbol{\alpha}_2'\mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$ having maximum variance. Similarly, we can find a kth stage a linear function $\boldsymbol{\alpha}_k'\mathbf{x}$ that has

maximum variance subject to being uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$, $\boldsymbol{\alpha}_2'\mathbf{x}$, ..., $\boldsymbol{\alpha}_{k-1}'\mathbf{x}$. The kth variable, $\boldsymbol{\alpha}_k'\mathbf{x}$ is the kth PC.

2.2 Derivation of Principal Components

To derive the PCs, we will use Lagrange multiplier method. Consider first $\boldsymbol{\alpha}_1'\mathbf{x}$; the vector $\boldsymbol{\alpha}_1$ maximizes $\text{var}[\boldsymbol{\alpha}_1'\mathbf{x}] = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$. The maximum will not be achieved for finite $\boldsymbol{\alpha}_1$ so a normalization constraint is needed. The constraint used in the derivation is $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, that is, the sum of squares of elements of $\boldsymbol{\alpha}_1$ equals 1.

To maximize $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximize

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1),$$

where λ is a Lagrange multiplier. Differentiation with respect to $\boldsymbol{\alpha}_1$ gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = \mathbf{0},$$

or

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_1 = \mathbf{0},$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. Then, λ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector. To decide which of the p eigenvectors gives $\boldsymbol{\alpha}_1'\mathbf{x}$ with maximum variance. Note that the quantity to be maximized is

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda,$$

so λ must be as large as possible. Thus, $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$, and $\text{var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$, the largest eigenvalue.

In general, the kth PC of $\mathbf{x}$ is $\boldsymbol{\alpha}_\mathbf{k}'\mathbf{x}$ and $\text{var}(\boldsymbol{\alpha}_\mathbf{k}{}'\mathbf{x}) = \boldsymbol{\alpha}_\mathbf{k}{}'\boldsymbol{\Sigma}\boldsymbol{\alpha}_\mathbf{k} = \lambda_k$, where λ_k is the kth largest eigenvalue of $\boldsymbol{\Sigma}$. Also, $\boldsymbol{\alpha}_\mathbf{k}$ is the corresponding eigenvetor.

The second PC, $\boldsymbol{\alpha}_\mathbf{2}'\mathbf{x}$ maximizes $\boldsymbol{\alpha}_\mathbf{2}'\boldsymbol{\Sigma}\boldsymbol{\alpha}_\mathbf{2}$ subject to being uncorrelated with $\boldsymbol{\alpha}_\mathbf{1}'\mathbf{x}$. By calculating similar way, we can get $(\boldsymbol{\Sigma} - \lambda\mathbf{I}_\mathbf{p})\boldsymbol{\alpha}_\mathbf{2} = \mathbf{0}$.

As stated above, it can be shown that for the third, fourth, ..., pth PCs, the vectors of coefficients are the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to $\lambda_3,\ \lambda_4, \ldots,\ \lambda_p$ and also $\text{var}[\boldsymbol{\alpha}_k'\mathbf{x}] = \lambda_k$ for $k = 1,\ 2, \ldots,\ p$.

Chapter 3

Quantile PCA

3.1 Definition of Quantile PCA

The point of quantile PCA is to use a PCA method well even if the data set are not symmetric. In ordinary PCA, if data set is multivariate normally distributed, we can germ a new data set with uncorrelated and maximal variances.

Then, if data set is not symmetric-distributed, we can now use quantile PCA.

This method is simply changing its loss function, quadratic loss function to quantile check function. To complement the non-differentiate point of the check function, we now adapt a square function of the neighborhood into the non-differentiate point of the check function. We now call this function as a modified check function.

Definition 3.1.1. (Modified Check Function) *Modified check function is defined as :*

$$\rho_{k,c}(u) = \begin{cases} (\tau_k - 1)(u + 0.5c) & \text{for} \quad u < c \\ 0.5(1-\tau_k)u^2/c & \text{for} -c \leq u < 0 \\ 0.5\tau_k u^2/c & \text{for} \quad 0 \leq u < c \\ \tau_k(u - 0.5c) & \text{for} -c \leq u \end{cases}$$

for some τ_k. In this paper, we use three quantiles : 0.01, 0.5, 0.99. This is a modified check function so that it is differentiable at zero. As c goes to zero, the function $\rho_{k,c}$ converges to ρ_k. Same as Lim and Oh (2012), we set $c = 10^{-6}$.

We think that if $\tau = 0.01$, the result of quantile PCA represents a low 1% pattern of the entire data set. Similarly, if $\tau = 0.99$, the result of quantile PCA represents a high 99% pattern of the entire data set.

3.2 Example of Quantile PCA

In this section, the proposed quantile PCA and other ordinary PCA methods are applied to the maximum daily precipitation data in August from the CPC merged analysis of precipitation (CMAP). This data set is analyzed by the Climate Research Unit, UK during year 1997–2008. We average daily values to get yearly data, so the number of observation is 12. These are the maximum precipitation on 360 $\times$ 180 grids that cover the entire globe with a 1° interval. We now focus on the East Asia region that covers 30–50°N and 120–140°E (number of variables is 441).

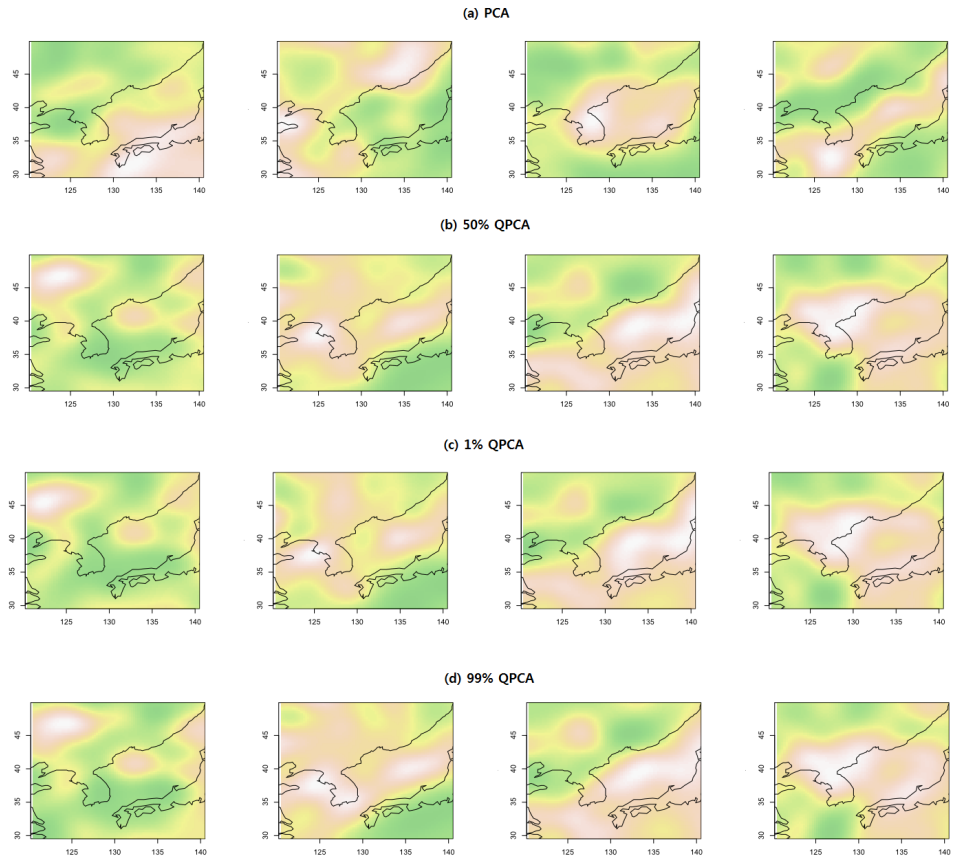

Figure 3.2.1: (a) Four leading PCs obtained from the ordinary PCA for maximum daily precipitation in August. (b) Four leading PCs from 50% quantile PCA for the same data. (c) Four leading PCs from 1% quantile PCA for the same data. (d) Four leading PCs from 99% quantile PCA for the same data.

We obtain four leading PCs from the conventional PCA, 1% quantile PCA, 50% quantile PCA, and 99% quantile PCA, which are displayed in Figure 3.2.1.

To evaluate the performance of the methods, we reconstruct data with one to four PCs and see how well they approximate the real data in the sense of root mean square error (RMSE);

$$\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\sum_{i=1}^{12} \sum_{j=1}^{441} (\mathbf{X}_{(i,j)} - \hat{\mathbf{X}}_{(i,j)})^2},$$

where $\mathbf{X}_{(i,j)}$ and $\hat{\mathbf{X}}_{(i,j)}$ denote observed maximum daily precipitation and reconstructed data at j grid point on i year, respectively. This is shown in Figure 3.2.2.

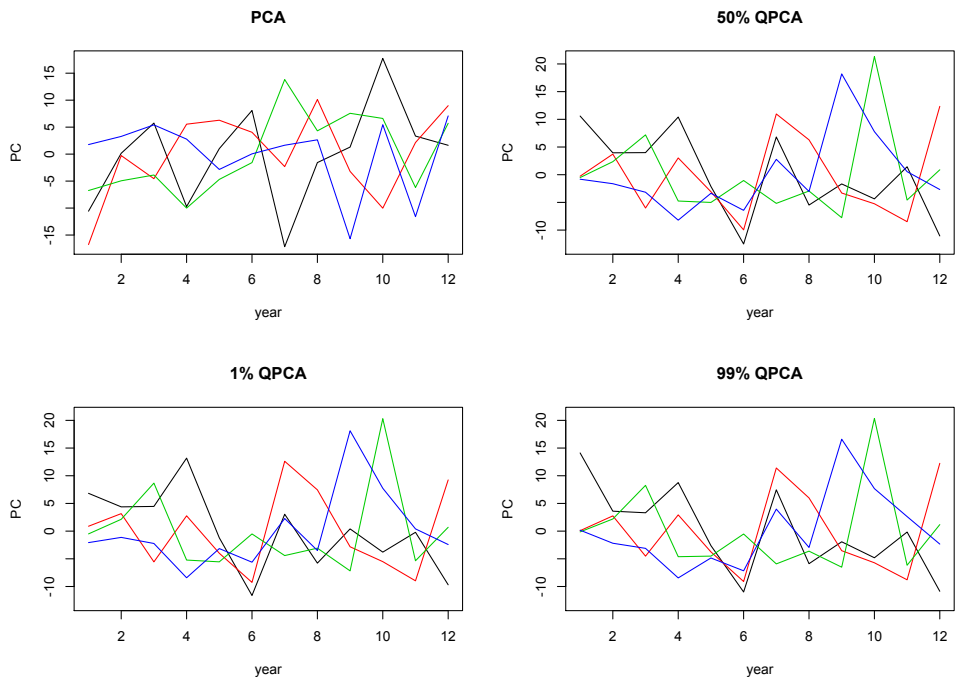

Figure 3.2.2: RMSE values between real maximum precipitation data and the reconstructions by one to four PCs obtained from the ordinary PCA (black), 50% quantile PCA (red), 1% quantile PCA (green) and 99% quantile PCA (blue).

Chapter 4

Conclusions

In this paper, we have proposed a quantile PCA method that works well even though the data distribution is skewed. THen, the distribution assumption of data for PCA is extended from Normality to a more general distribution. In real data example, we can find a extreme quantile PCs and RMSE for these four quantile PCA. We found it useful to check the extreme quantile of the sample data set using quantile PCA that replaces the quadratic loss function.

References

P. Bickel, P. diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zeger. (2011). *Principal Component Analysis*, Springer.

Y. Lim, H. Oh. (2012). A Data-Adaptive Principal Component Analysis, Technical Report, Department of Statistics, Seoul National University.

국문초록

주성분 분석은 다변량 자료를 분석하는 데에 있어서 굉장히 유용한 방법이다. 주성분 분석은 자료값들의 차원을 줄여 자료를 이해하기 더욱 쉽게 만들어 준다. 하지만 이러한 주성분 분석은 자료가 다변량 정규분포를 따른다는 것을 가정하고 있기 때문에, 이러한 가정이 만족되지 않으면 주성분 분석을 할 수 없게 된다. 이 논문에서 우리는 분위수 주성분 분석을 제안하고 있다. 이 분위수 주성분 분석은 기존의 주성분 분석을 적용할 수 없는, 다변량 정규분포를 만족하지 않는 자료값들에 대해서도 주성분 분석을 행할 수 있게 만들어 준다. 대칭으로 이루어져 있지 않고 한 쪽으로 치우친 다변량 분포에서 추출된 자료값들을 이 분위수 주성분 분석으로 효과적으로 분석할 수 있음을 알 수 있다.

주 요 어 : 다중척도, 주성분 분석, 극값 자료, 분위수
학 번 : 2011-20242