A Penalized Principal Components using Probabilistic PCA

Chongsun Park[1] and Morgan Wang[2]

Abstract

Variable selection algorithm for principal component analysis using penalized likelihood method is proposed. We will adopt a probabilistic principal component idea to utilize likelihood function for the problem and use HARD penalty function to force coefficients of any irrelevant variables for each component to zero. Consistency and sparsity of coefficient estimates will be provided with results of small simulated and illustrative real examples.

Key Words : Principal Component Analysis, Variable Selection, Penalized Likelihood, Hard Thresholding Penalty Function

1. Introduction

Principal component analysis (PCA; Jolliffe, 1986) is clearly one of the most frequently used method in statistics and related fields and often giving relatively small number of linear combinations of variables which can effectively explain the large portion of a given data set. However, each component still include all non-zero coefficients on all variables and having problem in interpretation of the linear combination especially when the number of variables is large.

A number of methods are available to aid interpretation. A common approach is ignoring any coefficients less than some threshold value. Jolliffe (1972, 1973) examines some of possible methods which discard irrelevant variables using multiple correlation, PCA itself, and clustering, etc. More formal ways of making some of the coefficients zero are to restrict the coefficients to a smaller number of possible values in the derivation of the linear functions like -1, 0, 1 (Hausman, 1982) and variation (Vines, 2000) on this theme is also possible. Rotation method used in factor analysis is also applicable but has its drawbacks (Jolliffe, 1989, 1995). McCabe (1984) introduced a new strategy to select a subset of the variables themselves and called it 'principal variables.'

Other possible way would be introducing penalty function as in regression analysis. Recently, Jolliffe (2002) applied L_1 penalty function method to maximization problem of PCA in order to force any irrelevant coefficients in the principal components. He included L_1

[1] Associate Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, KOREA

[2] Associate Professor, Department of Statistics & Actuarial Science, University of Central Florida, Orlando, Florida, U.S.A.

penalty function as an extra constraint to maximization problem of variance of linear combination of variables and showed that it is more preferable to rotation methods and several others.

We further extend idea of introducing penalty function to PCA problem by using probabilistic PCA of Tipping and Bishop (1997). It enables us to use and utilize likelihood idea so that consistency and sparsity of coefficients estimates are possible. We have seen that using L_1 penalty function could result in relatively severe bias for the coefficient estimates and found that hard thresholding penalty function (Antoniadis; 1997, and Fan; 1997) is better in preserving original directions after adding penalty function in the model.

2. Probabilsitic PCA with Latent Variable Model

It is well-known that PCA is closely related with factor analysis (Young, 1940; Whittle, 1952; Anderson, 1963). Also it is known that factor analysis can be expressed as a latent variable model (Lawley, 1953; Anderson and Rubin, 1956). And further work by Tipping and Bishop (1999) has shown how PCA may be viewed as a ML procedure based on a probability density model of the observed data.

Suppose that we have p-dimensional data vectors x_n, $n \in \{1,\dots,N\}$ and sample covariance matrix S of x with N observations. Usual PCA becomes solving eigenvalue problem

$$Sw_j = \delta_j w_j \text{ for } j=1, \dots, q.$$

Then the q principal components of the observed vector x_n are

$$c_n = W^T(x_n - x) \text{ with } W=(w_1, w_2, \dots, w_q)$$

such that q principal components of the observed vector x_n are those orthonormal axes onto which the retained variance under projection is maximal. The components c_n are then uncorrelated such that the covariance matrix $\Sigma_n c_n c_n^T/N$ is diagonal with elements δ_j.

The above PCA can be expressed as a latent variable model which relates p-dimensional observation vector x to a corresponding q-dimensional vector of latent variable c as

$$x = Wc + \mu + \epsilon \tag{1}$$

with conventional assumption of $c \sim N(0, I)$. Now, additional use of the isotropic noise model $N(0, \sigma^2 I)$ for ϵ in conjunction with equation (1) implies that the c-conditional probability distribution over x-space is given by

$$x|\, c \sim N(Wc + \mu, \sigma^2 I).$$

3. Penalized PCA

We can consider problem of extending penalized likelihood idea to the PCA for variable selection in each component. A form of penalized likelihood becomes

$$l(W,\sigma^2) - N\sum_{i=1}^{p}\sum_{j=1}^{q} p_\lambda(|w_{ij}|)$$

with w_{ij} as the elements of W in its ith row and jth column and $p_\lambda(\cdot)$ as a penalty function. Fan and Li (2001) argued that unbiased, sparsity, and continuity as three properties that a good penalty function should have, and suggested Smoothly Clipped Absolute Deviation (SCAD) penalty function as the best one for regression problems.

Several well-known penalty functions including SCAD penalty function are as follows.

- L_p: $p_\lambda(|w_{ij}|) = \lambda |w_{ij}|^p$ and it becomes LASSO with $p=1$ for least squares case.
- Hard Thresholding (HARD) Penalty: $p_\lambda(|w_{ij}|) = \lambda^2 - (|w_{ij}| - \lambda)^2 I(|w_{ij}| < \lambda)$
- Smoothly Clipped Absolute Deviation (SCAD) Penalty:

$$p_\lambda(w_{ij}) = \begin{cases} \lambda w_{ij} & \text{if } w_{ij} < \lambda \\ -\dfrac{w_{ij}^2 - 2aw_{ij} + \lambda^2}{2(a-1)} & \text{if } \lambda \le w_{ij} < a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } w_{ij} \ge a\lambda \end{cases}$$

Unfortunately, none of three penalty functions satisfy above all three properties simultaneously. L_p penalty function is biased and this cause some serious problem especially when applied to PCA problems. We have seen from small simulations that bias problem in L_1 penalty is so serious that including penalty function usually resulted in domination of one or few variables with relatively large coefficients compared to other variables. By the way, hard thresholding (HARD) penalty function is unbiased and has sparsity but it is not continuous. SCAD behaves like something between L_1 and HARD and need two dimensional GCV (Generalized Cross-Validation) or usual CV to find optimal values for two parameters, a, and λ.

Overall, it looks reasonable to use HARD for the PCA problem since it looks best in forcing coefficients of irrelevant variables to zero and at the same time in preserving original directions after introducing penalty function in the ML procedure.

Now, we establish the consistency and sparsity of HARD penalty function for our non-concave penalized PCA estimator. Let's assume that the selected one component of W can be divided as $w_0 = (w_{10}, \ldots, w_{p0}) = (w_{10}^T, w_{20}^T)^T$ and without loss of generality, assume that $w_{20} = 0$. And let $Q(w)$ be the marginal penalized likelihood function of w with respect to the selected component only. In the first theorem we show that there exists a penalized likelihood estimator that converges at the rate $O_p(n^{-1/2} + a_n)$, where $a_n = max_i\{p'_{\lambda_n}(|w_{ij}|): w_{ij} \neq 0\}$. This implies that for the hard threshholding and SCAD penalty functions, the penalized likelihood estimator is root-$n$consistent if $\lambda_n \to 0$. Furthermore, we demonstrate that such a root-n consistent estimator must satisfy $\widehat{w_2 = 0}$ and this implies that the penalized likelihood estimator performs as well as if $\widehat{w_2 = 0}$ were known.

Theorem 1 *If* $max_i\{|p'_{\lambda_n}(|w_{ij}|)| : w_{ij} \neq 0\} \to 0$, *then there exists a local maximizer* $\hat{w}$ *of* $Q(w)$ *such that* $||\hat{w} - w|| = O_p(n^{-1/2} + a_n)$, *where* a_n *is given previously.*

Theorem 2 *Assume that*

$$\lim inf_{n \to \infty} \lim inf_{\theta \to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0.$$

If $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to 1, for any given* $\hat{w}_1$

1. $Q\left\{\binom{\tilde{w}_1}{0}\right\} = max_{||w_2|| \le Cn^{-1/2}} Q\left\{\binom{\tilde{w}_1}{\tilde{w}_2}\right\}$

2. Sparsity: $\widehat{w_2} = 0$.

MLE can be obtained via EM algorithm as treating c_n as missing so complete data set as (x_n, c_n) (Tipping and Bishop, 1999)

4. Simulation Results and Illustrations

We compared our method with small set of simulated and real data. Even though coefficient estimates from ordinary PCA are orthonormal, we reported original coefficient estimates from our method after standardizing only.

4.1 Simulation Results

The data sets are simulated as follows. It is based on the observation that x is marginally distributed as normal with mean μ and covariance matrix $\psi = WW^T + \sigma^2 I$. Further we can set μ as zero without loss of generality.

The following sets of data are generated 100 times for each combination.

- N: the number of observations (20, 50, 100, 300)
- P: the number of variables 6
- Q: the number of components considered: (1, 1, 0, 0, 0, 0), (1, 0, -2, 0, 0, 0)
- Largest eignenvalue: 2.5 (out of 6)

We will look at estimated directions and their standardized values for comparison with true W and σ^2. Especially, number of zero estimates for true zero (T0) coefficients for each case, and zero estimates for non-zero coefficients (F0) are our concern. We tried only two preset values for λ of 0.5 and 1.0.

N	$\lambda=0.5$	$\lambda=1.0$
20	T0: 174 F0: 4	T0: 264 F0: 28
50	T0: 259 F0: 2	T0: 303 F0: 5
100	T0: 267 F0: 0	T0: 306 F0: 0
300	T0: 294 F0: 0	T0: 316 F0: 0

표 1 CASE I: $W=(1,1,0,0,0,0)$

N	$\lambda=0.5$	$\lambda=1.0$
20	T0: 174 F0: 4	T0: 264 F0: 28
50	T0: 259 F0: 2	T0: 303 F0: 5
100	T0: 267 F0: 0	T0: 306 F0: 0
300	T0: 294 F0: 0	T0: 316 F0: 0

표 2 CASE II:$W=(1,0,-2,0,0,0)$

4.2 A Real Example

We combine three kinds of iris data sets into one and applied penalized PCA. We look at standardized coefficient estimates for first three components of PCA. There are four variables, Sepal Length, Sepal Width, Petal Length, and Petal Width. Each species have 50 observations so total of 150 cases.

Variable	Comp 1	Comp 2	Comp 3
Sepal Length	0.701	0.096	0.294
Sepal Width	0.045	0.887	-0.088
Petal Length	0.692	-0.448	0.867
Petal Width	0.166	0.051	0.392

표 3 CASE I: $\lambda=0.0$

Variable	Comp 1	Comp 2	Comp 3
Sepal Length	0.675	0.084	0.305
Sepal Width	0.034	0.889	-0.082
Petal Length	0.716	-0.451	0.867
Petal Width	0.175	0.000	0.386

표 4 CASE II: $\lambda=0.5$

Variable	Comp 1	Comp 2	Comp 3
Sepal Length	0.681	0.000	0.291
Sepal Width	0.000	0.918	-0.061
Petal Length	0.719	-0.397	0.876
Petal Width	0.139	0.000	0.380

표 5 CASE III: $\lambda=1.0$

Sepal Width in the first component and two variables Sepal Length, and Petal Width becomes 0 with $\lambda=0.5$ and 1.0. When λ becomes larger it tends to force bigger estimates bigger and smaller estimates smaller.

5. Discussions

We introduced a variable selection algorithm for principal component analysis using penalized likelihood method. From the results from small simulation we could have strong feeling that our method would be quite effective in forcing coefficients related to irrelevant variables in PCA problems to zero. Hence the proposed method can be successfully applied to high-dimensional PCA problems with relatively large portion of irrelevant variables included in the data set. And also it is straightforward to extend our likelihood method in handling problems with missing observations by using EM algorithms. Further extension of the penalized PCA method to any problem which need to solve eigenvalue or general eigenvalue problems like sliced inverse regression (SIR) or so.

References

Anderson, T. W. (1963), Asymptotic Theory for Prinicipal Component Analysis, Annals of Mathematical Statistics, 34, 122-148.

Anderson, T. W., and Rubin, H. (1956), Statistical Inference in Factor Analysis, Annals of Mathematical Statistics, 34, 122-148.

Antoniadis, A. (1997), Wavelets in Statistics: A Review (with discussion), Journal of the Italian Staitistical Association, 6, 97-144.

Fan, J. (1997), Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis, Journal of the Italian Statistical Association, 6, 131-138.

Fan, J., and Li, R. (2001), Variable Selction via Nonconcave Penalized Likelihood and its Oracle Properties, Journal of the American Statistical Association, 96, 1348-1360.

Hausman, R. (1982), Constrained Multivariate Analysis. In Zanckis, S. H. and Rustagi, J. S. (Eds.). Optimisation in Statistics, pp. 137-151, North Holland: Amsterdam.

Jolliffe, I. T. (1972), Discarding Variables in a Principal Component Analysis. I: Artificial Data, Applied Statistics, 21, 160-173.

Jolliffe, I. T. (1973), Discarding Variables in a Principal Component Analysis. II: Real Data,, Applied Statistics, 22, 21-31.

Jolliffe, I. T. (1986), Principal Component Analysis, Springer-Verlag, New-York.

Jolliffe, I. T. (1989), Rotation of Ill-defined Principal Components, Applied Statistics, 38, 139-147.

Jolliffe, I. T. (1995), Rotation of Principal Components: Choice of Normalization Constraints, Journal of Applied Statistics, 22, 29-35.

Jolliffe, I. T., Trendafilov, N., and Uddin, M. (2002), A modified principal component technique based on the lasso, PostScript preprint, Department of Mathematical Sciences, University of Aberdeen.

Lawley, D. N. (1953), A Modified Method of Estimation in Factor Analysis and Some Large Sample Results, In Uppsala Symposium on Psychological Factor Analysis, Number 3 in Nordisk Psykologi Monograph Series, pp. 35-42, Uppsala: Almqvist and Wiksell.

McCabe, G. P. (1984), Principal Variables, Technometrics, 26, 137-144.

Tipping, M. E., Bishop, C. M. (1997), Probabilisitic Principal Component Analysis, Journal of the Royal Statistical Society, Series B, 61, 611-622.

Vines, S. K. (2000), Simple Principal Components, Applied Statistics, 49, 441-451.

Whittle, P. (1952), On Principal Components and Least Square Methods of Factor Analysis, Skandinavisk Aktuarietidskrift, 36, 223-239.

Young, G. (1940), Maximum Likelihood Estimation and Factor Analysis, Psychometrika, 6, 49-53.