International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.57-74
http://dx.doi.org/10.14257/ijdta.2015.8.6.06

A Literature Survey on High-Dimensional Sparse Principal Component Analysis

Shen Ning-min and Li Jing

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016 Nanjing, China
ningminshen@163.com, jingli@nuaa.edu.cn.

Abstract

Principal Component Analysis (PCA) is a classical method for dimensionality reduction, data pre-processing, compression and visualization of multivariate data for different applications in biology, social science and engineering. The limitation of PCA is lacking of interpretation due to the non-zero loadings and the inconsistence for high-dimensional data. Sparse principal component analysis (sparse PCA) is proposed mainly for the challenges of PCA above. For the past decades, many works of the development methods and theoretical analysis for sparse PCA have been presented. The goal of this paper is to give a comprehensive literatures review to recent progress in high-dimensional sparse PCA from algorithm and statistical theory. Firstly we give the overview for PCA and sparse PCA. Secondly the algorithms of sparse PCA are categorized into different classes and provide detailed descriptions for typical formulations and methods in each category, and the typical packages of sparse PCA are also given. Considering that statistical analysis in high dimension becomes more involved in sparse PCA, and then the survey of theoretical analysis of sparse PCA is also presented. Finally the future trends as well as challenges are given.

Keywords: *sparse principal component analysis; PCA; spiked-covariance model; deflation*

1. Introduction

Principal component analysis (PCA) has become a popular technique for dimension reduction and feature extraction and it is one of the most important techniques applied for multivariate analysis. From the observations of random variables, the objective of PCA is to estimate the leading eigenvectors of its covariance matrix and form new variables, called principal components (PCs), which are linear combinations of the original variables and the PCs are uncorrelated, the vector of coefficients (or loadings) are orthogonal. PCA has been used in widely areas such as biomedical problems, biology, social science and engineering.

PCA suffers from two major weaknesses. One weakness is that each PC is obtained by a linear combination of original variables and loadings are normally non-zero which makes the results in difficult to interpret. Such as in gene expression data which needs to obtain a small set of genes which contribute to the final results, the loadings is hoped to sparse. The other weakness is for high dimensional data with $d >> n$, PCA may be inconsistent in estimating the loadings u_1^* [1-5]. In order to make the estimation of high-dimensional PCA feasibly and improve the interpretability of results, sparsity is introduced to PCA which assumed that u_1^* is sparse, we call it sparse principal component analysis (sparse PCA). Moreover, besides the interpretability and high-dimensional estimation consistency would not be the only advantage of sparse PCA, we can discard

ISSN: 2005-4270 IJDTA

the variables with zero loadings in all of the PCs, so it will lead an automatic feature selection.

Sparse PCA has been paid a more attention over last several years, considerable work has been done to development of various algorithm [6-35] and theoretical analysis [1-2, 4, 36-61]. Moreover, sparse PCA has a widely applications in various fields which include bioinformatics [62-64], clustering and feature selection [49], multivariate time series analysis [65-66], large text data analysis [67], finance data analysis [68] and so on. But to the best of our knowledge, the survey of sparse PCA is very few, we observed that three most influential surveys on the sparse PCA published in the year of 2012 [69] and 2014 [70-71] respectively. Richtárik [69] unified 8-formulations of sparse PCA solving by an alternating maximization method. Considering the applications in in cancer research, Hsu [70] reviewed several popular approaches of sparse PCA, but only several typical methods are included. Trendafilov [71] reviewed the most popular methods and their performance, the theoretical analysis work of sparse PCA is omitted. Above reviews motivate our works because high-dimensional sparse PCA not only include the computational methods, but also the high-dimensional statistical consistency. Our paper thus aims to fill this gap, we will make a general and large survey of the most representative algorithms and theoretical analysis, but our paper avoids the challenges topics in optimization theory used in sparse PCA.

The remainder of this paper is organized as follows: firstly, the mathematic formulation of PCA and sparse PCA is presented. Then the details of algorithmic and statistical guarantees of foregoing categories of sparse PCA are elaborated. Finally, some open issues remained to be solved are discussed.

2. Overview of PCA and Sparse PCA

2.1. Notation

For convenience to the readers, some of the notations are introduced firstly in our paper. Given a vector $x \in R^k$ whose j^{th} coordinate is denoted as x_j . $u_1, u_2, ...u_n$ is a sequence of vectors. $< x, y >= \sum_i x_i y_i, \| x \| = < x, x >^{1/2} = \| x \|_2, \quad x, y \in R^n$, $\| x \|_q$ is the usual l_q norm with $\| x \|_0$ defined the number of nonzero entries of x (l_0 norm), $\| x \|_1 = \sum_i | x_i |$ (l_1 norm). $< X, Y >= TrX^T Y, \| X \| = < X, X >^{1/2} = \| X \|_F, X, Y \in R^{n \times d}$, $\| X \|_F$ is the squared Frobenius norm, the symbol Tr denotes the trace of its argument. The notation $X \pm 0$ means that X is positive semi-definite, $| X |$ is the matrix whose elements are the absolute values of the element of $\sum$, and $Tr(X) := X_{11} + X_{22} + ... + X_{nn}$,which is the sum of diagonal entries.

2.2. Formulations of PCA

Suppose the input data matrix as $X = [x_1, x_2, ..., x_n]^T \in R^{n \times d}$, where n and d are the size and the dimensionality of the given data, respectively. Assumes all the data are centered $\sum_i x_i = 0$ and $\sum = \frac{1}{n} X^T X \in R^{d \times d}$ be the data covariance matrix. PCA seeks to find a number of $p << d$ linear combinations of the n variables in the projected linear space as $\tilde{z}_k = X^T u_k = \sum_{i=1}^{d} u_{k,i} x_i$, where $\tilde{z}_k$ is the k -th principal component (PC) and u_k is the unit-length loadings vector. PCA can be performed by either an eigenvalue decomposition of the covariance matrix or by singular value decomposition (SVD). The formulations of PCA can be derived from three viewpoints:

(1) One way of looking at PCA is from the data-variance-maximization viewpoint. The goal is to find u where the input data variance X is maximized [26]. This leads to the following optimization model:

$$\max_{u} u^T \sum u \quad s.t. \| u \| = 1 \tag{1}$$

(2) Another way of looking at PCA is from reconstruction-error minimization viewpoint which can be viewed as a projection from a high dimensional space to a low dimensional subspace that minimizes the total squared reconstruction error $\min \| X - \bar{X} \|_F^2$ where X is the original data set, $\bar{X}$ is the new dataset obtained. The reconstruction error can be computed using two different methods as bellows.

a) Factor loadings model. First, Computes the principal component using $z_j = u_j^T x$, and then reconstruct $\bar{X}$ using $\bar{X} = Uz = \sum_{j=1}^{k} z_j u_j$ .The new formulation of PCA can be derived as:

$$\min \sum_{i=1}^{n} \| X_i - UU^T X_i \|^2 \quad s.t. \| u \| = 1 \tag{2}$$

b) Low-rank approximation. $\bar{X}$ can also be computed using rank-k approximation, the result of $\min \| X - \bar{X} \|_F^2$ is the top-k singular vectors of X minimizes the Frobenious norm of the difference with the matrix X . The SVD of X is $X = UDV^T$ where U is an $n \times r$ orthogonal matrix and the column vectors u_k are the PCs scaled to unit length. V is $p \times r$ orthogonal matrix which columns u_k are the loadings vectors, D is a diagonal matrix and the diagonal entries $d_1, ..., d_k$ are the singular values, where $d_k u_k = \tilde{u}_k$ is the k-th PC with the variance is d_k^2. The new PCA formulation using the low-rank approximation can be derived as:

$$(d, u, v) = \arg\min_{d,u,v} \| X - duv^T \|_F^2 \quad s.t. \| u \| = 1 \tag{3}$$

(3) The third way of looking PCA is from Probabilistic [82] viewpoint. It is most naturally expressed as a mapping from the latent space into the data space via $x = Uz + \mu + \varepsilon$ finding a lower-dimensional probabilistic description of the data. First, generates an independent, standard Gaussian random latent variable in an high dimensional space $p(z) = N(z \mid 0, I)$, then generates the observed variable x in a low dimensional space from $p(x \mid z) = N(x \mid Uz + \mu, \sigma^2 I)$, where $W \in R^{D \times M}, \mu \in R^D$, $\varepsilon \in N(0,1)$ is a Gaussian random noise , the columns of U span a linear subspace corresponding to the principal subspace. Based on this model, we have $p(x) = N(x \mid \mu, WW^T + \sigma^2 I)$, the probabilistic PCA parameter estimation is to estimate the parameter of μ, U and σ^2 .

2.3. Basic Formulations of Sparse PCA

The objective of sparse PCA is to force a number of n to be zero which derives the eigenvector to be sparse. In order to obtain the sparsity on the extracted components, most of methods find the PC's of the covariance matrix through adding a constraint or penalty term from the PCA formulations (1) and (2), A constrained l_0 -norm minimization problem is usually firstly considered as the basic sparse PCA problem as:

$$u = \arg\max_{u} u^T \sum u \quad s.t. \quad \| u \|_2 = 1, \| u \|_0 \le k \tag{4}$$

Where k is the nonzero number of loadings. Sparse PCA problem is non-convex and NP-hard. All of the formulations and algorithms can be categorized into three classes [36] from the viewpoint of data-variance-maximization, minimal-reconstruction-error and probabilistic modeling. So we can classify the sparse PCA problems in two three classes as shown in Figure1.

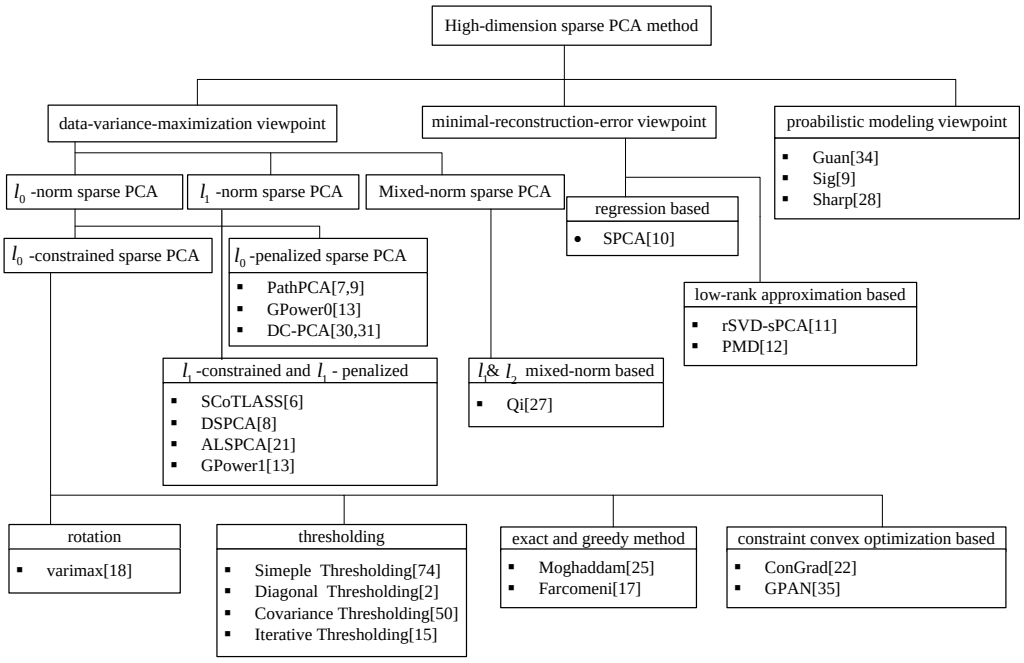

Figure 1. The Categorization of Sparse PCA Algorithms

3. Sparse PCA: Formulations and Algorithms

3.1. Sparse PCA from Data-Variance-Maximization View

From data-variance-maximization view, a direction with at most k non-zero number of coefficients and with maximal variance is needed to search. The l_0-sparse PCA problem is defined as (4). Some papers use $card(u) \le k$ substitute for $\| u \|_0 \le k$ Where $card(u)$ denotes the cardinality of u, that is the number of non-zero coefficients. They are equivalent. Given a covariance matrix $\sum$, most current approaches to sparse PCA can be categorized as solving one of the modified optimization problem based on constraint, penalization and relaxation.

According to the constraint, penalization and relaxation adding to the modified optimization problem, the sparse PCA has several following formulations from data-variance-maximization viewpoint. The first one is l_0-norm sparse PCA which includes l_0-constrained and l_0-penalized problem. l_0-constrained sparse PCA as (4). l_0-penalized sparse PCA is:

$$L_0(\sum) = \arg\max_u \{u^T \sum u - \lambda \| u \|_0 : \quad \| u \|_2 = 1\} \tag{5}$$

The second one is l_1-constrained and l_1-penalized sparse PCA. The formulas are as (6) and (7):

$$L_1(\sum) = \arg\max_u \{u^T \sum u : \quad \| u \|_2 = 1, \| u \|_1 \le \sqrt{k}\} \tag{6}$$

$$L_1(\sum) = \arg\max_u \{u^T \sum u - \lambda \| u \|_1 : \quad \| u \|_2 = 1\} \tag{7}$$

The last one is a kind of mixed-norm sparse PCA

$$L_\lambda(\sum) = \arg\max_u \{u^T \sum u : \quad \| u \|_2 = 1, \| u \|_\lambda \le 1\} \tag{8}$$

We call (4) and (5) are l_0-norm sparse PCA, and (6) and (7) are l_1-norm sparse PCA.

These formulations always focused on the deriving of the first principal component and the additional components can be obtained by the iterative deflating technique [73], the shortcoming of this technique always lead the sparse PCA lacking of non-orthogonality, sub-optimality, and multiple parameters needed to be tuned[21].

A. l_0-norm sparse PCA

(1) l_0 -constrained sparse PCA. l_0 -constrained sparse PCA is the fundamental formulation of sparse PCA, it needs the algorithm tackling (4) directly, not need any reformulation.

(a) Rotation. Sparsity can be obtained through rotating the loading matrix such as $U_{unrotated}\Lambda = U_{rotated}$. The oldest rotation approach is varimax [18] which is proposed in 1958. After varimax's rotation, some coefficients of loading vectors could have bigger values than others, but it is very hard to quantify the distinction between small and large coefficients.

(b) Thresholding. Jeffer [74] proposed a simple thresholding method through setting the coefficients less than 70% of the greatest one are zero, no matter their signs. Vienes [75] proposed the simple principal component by restricting the loadings coefficients to have integer values as -1, 0 and 1. Cadima *et al.* noted that simple thresholding, even after a rotation, can be misleading and in general it does not produce an optimal solution [76]. Johnstone *et al.* [2] proposed a two-step method, using an pre-processing step to select relevant variables by thresholding the diagonal of the sample covariance matrix followed by ordinary PCA in the reduced space. They considered the case of a signal u that is sparse in a suitable basis. Motivated by the work of [55,50] proposed a covariance thresholding algorithm for this kind of sparse PCA which computed the leading eigenvector from the thresholding covariance matrix by:

$$\hat{\Sigma}_{ij} = \begin{cases} \Sigma_{ij} & if\ |\Sigma_{ij}| > t \\ 0 & otherwise \end{cases} \tag{9}$$

Ma *et al.* [15] proposed an iterative thresholding method is known as ITSPCA which is based on subspace iteration which is a straightforward block generalization of the iterative power method as $U_k = AQ_k, U_k = Q_{k+1}R_{k+1}$,where $k = 0,1,....$ until the convergence. They added a thresholding step between the Multiplication and QR factorization steps to derive the sparse principal subspaces. Moreover this method makes the orthogonal iteration adapts to the high-dimensional problem. Similar to thresholding method above, there also have similar works reducing the feature space. Yuan *et al.* [14] proposed Truncated Power method for l_0 -sparse PCA problem which is based on the power method [73]. TPower method added a truncation step before the normalization step. Wang *et al.* [32] proposed a new sparse PCA method removing the variable using the approximated minimal variance loss (AMVL) criterion with smallest loading and reconsidered the sparse PCA in the reduced space.

(c) Exact and greedy method. Moghaddam *et al.* [25] proposed a spectral bounds framework for sparse PCA which obtains good numerical results using a combinatorial greedy method. The exact method was to search over all possible support sets S $s.t.\ \ |S| \le k$ and pick the S with the maximum value. They continued to propose the greedy strategy for this problem. At each iteration, repeat choosing a new variable which maximizes the eigenvalue of the sub-matrix until the $|S|$ is k , other PCs can be obtained using power-iteration. The main shortcoming of their method can be slow on large covariance matrices. Farcomeni [17] also proposed an exact approach based on branch and bound algorithm for (4). They enumerated the possible solutions using branch and bound algorithms. Then splitted the possible solutions set into subsets using branching and bound the solutions into each branch using some criterion. Only branches bigger than the current maximum are continue explored. Experimental results showed that their

method can work on a higher-dimensional problems.

(d) Constrained convex optimization method. If formula (4) is reformulated, sparse PCA can be solved by first-order gradient based algorithm [13, 31]. In contrast, Luss [22] directly tackled l_0 -constrained sparse PCA called ConGradU (Conditional gradient algorithm with unit step size) using an efficient conditional gradient method, also known as Frank-Wolf algorithm. They pointed out the first-order gradient methods in [13] and [31] are identical to ConGradU. Motivated by the work of ConGradU, [35] proposed the method based on gradient projection algorithm and an approximate Newton algorithm for (4) where the constraint set may be non-convex.

(2) l_0 -penalized sparse PCA

(a) PathPCA. d'Aspremont [7, 9] proposed a PathSPCA algorithm that computed a full set of solutions for all target numbers of nonzero coefficients which formulated (4) to (5). They continue to consider the formula:

$$\phi(\rho) = \max_{\|u\|_2=1} \sum_{i=1}^{p} ((x_i^T u)^2 - \rho)_+ \tag{10}$$

Where $(\alpha)_+ := \max\{\alpha, 0\}, \sum = X^T X$ and x_i is the i -th column of $X \in R^{p\times p}$. (10) can be relaxed to:

$$\phi(\rho) = \max . \sum_{i=1}^{n} (x_i^T U x_i - \rho)_+ \ s.t. \quad Tr(U) = 1, Rank(U) = 1, U \pm 0 \tag{11}$$

Where $U = uu^T$, $._+$ operator denoted $\max\{0, \cdot\}$. (11) can be rewritten as a semi-definite program in the variables Z and P_i :

$$\varphi(\rho) = \max \sum_{i=1}^{p} Tr(P_i B_i) \ s.t. Tr(U) = 1, U \text{ 굇} 0, U \quad P_i ? \ 0 \tag{12}$$

With $B_i = u_i u_i^T - \rho I$. They derived a greedy algorithm as in [61] to compute a full set of solutions of (12).

(b) GPower0. Journee *et al.* [13]designed a series of algorithms respectively for sparse PCA by formulating the sparse PCA problem as maximization of a convex function or a compact set with l_0 -or l_1 -norm sparsity-inducing penalties and extracting single unit sparse PC sequentially or block units ones simultaneously. They first reformulated (4) as single-unit l_0 -penalization sparse PCA:

$$\phi_{l_1}(\gamma) = \max_{x\in R^p, u^T u=1} u^T \sum u - \gamma \| u \|_0 \tag{13}$$

Then it is continued to formulated as $\phi_{l_0}(\gamma) = \max_{x\in R^p, x^T x=1} \sum_{i=1}^{n} [(a_i^T x)^2 - \gamma]_+$ (14) then set

$$u_i = [sign((a_i^T x)^2 - \gamma] + a_i^T x, u^* = u / \| u \|_2 \tag{15}$$

Similarly, block $GPower - l_0$ considered the following formulation: $\phi_{l_0,m}(\gamma) = \max_{X\in R^p} \sum_{j=1}^{m} \sum_{j=1}^{n} [(\mu_j a_i^T x_j)^2 - \gamma_j]_+$, where μ_j is the positive entries on the diagonal. They continually proposed the simple gradient method which is actually a generalized power method to solve it. Based on GPower method, Kuleshov [77] presented a fast algorithm based on Rayeleigh quotient iteration which modified the power method by Rayleigh quotient iteration.

(c) DC-SPCA. Considering the penalized formulation (5), Srperumbudur *et al.* [30-31] approximated $\| u \|_0$ by $\sum_{i=1}^{n} \log(\varepsilon + | u_i |)$, Since $\| u \|_0 = \sum_{i=1}^{n} 1_{\{|x_i|\neq 0\}} = \lim_{\varepsilon\to 0} \sum_{i=1}^{n} \frac{\log(1+| u_i | / \varepsilon)}{\log(1+1/\varepsilon)}$, this program is approximated by neglecting the limit and choosing $\varepsilon > 0$, (5) is formulated as:

$$\max_{x}\left\{x^{T}Ax-\rho_{\varepsilon}\sum_{i=1}^{n}\log(|x_{i}|+\varepsilon:x^{T}x\le 1\right\} \quad (16)$$

Where $\rho_{\varepsilon}:=\rho/\log(1+\varepsilon^{-1})$. $\|u\|_{0}$, they used majorization-minimization approach to solve it.

B. l_1-norm sparse PCA

(a) SCoTLASS. Jolliffe *et al.* [6] proposed a method is known as Simplified Component Technique for Least Absolute Shrinkage and Selection (SCoTLASS) to solve sparse PCA problem (6). It is the first true algorithmic method to achieve the sparse loading since rotation and simple thresholding and it is also non-convex. They suggested a simple projected gradient descent approach to solve it. But the computational cost is often high, even if Trendafilov *et al.* [78] proposed a globally convergent algorithm to solve the optimization problem. It remains high computational cost. Recently, Witten *et al.* [12] proposed an efficient algorithm can be used for SCoTLASS to obtain the first PC.

(b) DSPCA and its variants. D'Aspremont *et al.* [8] proposed a convex relaxation method, called Direct sparse PCA (DSPCA) to (6) solving a sequence of semi-definite programming (SDP) relaxations of sparse PCA to search for the sparse PCs. Let $X=uu^{T}$. It is rank-one. The equivalent formula of (6) is considered as follows:

$$\arg\max_{X\in R} Tr(\sum X) \quad s.t. \quad Tr(X)=1, Card(X)\le k^{2}, X\pm 0, rank(X)=1 \quad (17)$$

Since $Card(X)\le k^{2}$ and $\|u\|_{2}=1$ is still non-convex, they continued to relax the non-convex constraint by replacing $card(U)\le k^{2}$ using $1^{T}|U|1\le k$ and dropped the rank constraint. (17) can be rewritten in:

$$\arg\max_{X\in S_{p}} Tr(\sum X)\ s.t.\ Tr(X)=1, 1^{T}|X|1\le k, X\pm 0 \quad (18)$$

Due to IP (Interior Point) [79] solvers only be useful for small problems. They continued to use a penalized version of the relaxed scheme as:

$$\max_{X} Tr(\sum X)-\rho 1^{T}|X|1: X\pm 0, TrX=1 \quad (19)$$

Then the optimal first-order minimization algorithm [80] is proposed to minimize the smooth approximation. DSPCA is computationally expensive. Zhang *et al.* [67] proposed a safe feature elimination method as a preprocessing step to reduce the problem space and proposed a block coordinate ascent algorithm to solve DSPCA. Moreover, because (19) is a non-smooth semi-definite programming problem, Ma *et al.* [23] proposed an alternating direction method to solve it. Different from above methods, Vu *et al.* [16] extended the DSPCA formulation from rank-1 to rank-k case, considering a new formulation of sparse principal subspace problem as a novel semi-definite programming with a Fantope constraint as:

$$\max_{U} Tr(\sum X)-\rho\sum_{ij}|X_{ij}| \quad s.t.\ 0\ 같X \quad I, Tr(X)=k \quad (20)$$

The constraint set is called Fantope which is solved by an efficient ADMM algorithm [81].

(c) ALSPCA. Lu *et al.* [21] developed an augmented Lagrangian method (ALSPCA) for sparse PCA by solving a class of non-smooth constrained optimization problems .They considered the next formulation:

$$\max_{V\in R^{p\times r}} Tr(u^{T}\sum u)-\rho\|u\|_{1}\ subject\ to|u_{i}^{T}\sum u_{j}|\le\Delta_{ij}, \forall i\ne j, u^{T}u=I \quad (21)$$

Where each column of u corresponds to a loading vector of the sample covariance matrix $\sum$ and $\Delta_{ij}\ge 0(i\ne j)$ are the parameters that control the correlation of the PCs. It is solved by an augmented Lagrangian approach. This method can obtain near orthogonality of PCs.

(d) GPower1. As the introduction of GPower0 above, single-unit GPower l_1 reformulated (4) as single-unit l_1-penalization sparse PCA:

$$\phi_{l_1}(\gamma) = \max_{x \in R^p, u^T u=1} \sqrt{u^T \sum u} - \gamma \| u \|_1 \quad (22)$$

And reformulated it as:

$$\phi_{l_1}^2(\gamma) = \max_{u \in R^p, x^T x=1} \sum_{i=1}^{n} [| a_i^T x | - \gamma]_+^2 \quad (23)$$

Similarly, the block $GPower_{l_1}$ considered the following reformulation $\phi_{l_{1,m}}^2(\gamma) = \max_{x \in R^p} \sum_{j=1}^{m} \sum_{i=1}^{n} [\mu_j | a_i^T x_j | - \gamma_j]_+^2$, where μ_j is the positive entries on the diagonal. The leading PC can be obtained as GPower l_0 .The shortcoming of GPower is that their formulation and algorithm does not directly related to the given cardinality, so must require several runs.

C. Mixed-norm sparse PCA

Qi *et al.* [27] proposed a new convex combination of l_1 and l_2 norms, which can efficiently obtain the uncorrelated principal components. They replaced the l_2-norm of u with a mixed norm $\| \cdot \|_\lambda$ by $\| u \|_\lambda = \sqrt{[(1-\lambda) \| u \|_2^2 + \lambda \| u \|_1^2]}, \forall u \in R^p$, $\lambda \in [0,1]$, $\| u \|_1 = \sum_{i=1}^{p} | u_i |$ be the l_1-norm of u . If λ =0, this norm is the l_2 norm, if $\lambda = 1$, it is the l_1 norm. Using this norm, the optimization problem solving the coefficient vector u of the first and the higher order sparse PCs as follows:

$$\max_{v \in R^p, \|v\|_2=1} \frac{u^T \sum u}{(1-\lambda_1) \| u \|_2^2 + \lambda_1 \| u \|_1^2} = \max_{u \in R^p, \|u\|_2=1} \frac{u^T \sum u}{\| u \|_{\lambda_1}^2} \quad (24)$$

Where $0 \le \lambda_1 \le 1$ is the tuning parameter for the first principal component. This formulation maked the optimization problem convex and had a unique solution.

3.2. Sparse PCA from Data-Variance-Maximization View

(a) SPCA. Zou *et al.* [10] formulated the sparse PCA problem as a regression-type optimization as:

$$(\hat{U}, \hat{W}) = \arg\min_{U,W} \| X - XWU^T \|_2^2 + \lambda \sum_{j=1}^{q} \| w_j \|_2^2 + \sum_{j=1}^{q} \lambda_{1,j} \| w_j \|_1 \ s.t. \ \ U^T U = I_k \quad (25)$$

Where $u_j = w_j / \| w_j \|$, w is the lasso estimates. $\lambda_{1,j}$'$s (j = 1,...,q)$ regularization parameters with positive value and $\| \cdot \|_1$ is the l_1 norm of w . All q components share the same λ and different $\lambda_{1,j}$'s are allowed for penalizing the loadings of different principal components. They proposed a framework of l_1-penalized regression on regular principal components using Elastic Net [82], solved by least angle regression (LARS). The main drawback of SPCA is that the orthogonality of loadings is not guaranteed.

(b) rSVD-sPCA. Shen *et al.* [11] proposed sequential methods known as sPCA-rSVD which searched for the sparse PCs by solving a regularized low rank matrix approximation problem under multiple sparsity-including penalties. They considered the singular value decomposition from low-rank viewpoint. From the low rank approximation of SVD, (3) can be formulated as

$$\min_{a,b,\|b\|_2=1} \| X - vu^T \|_F^2 + P_\lambda(u) \quad (26)$$

Where $u \in R^p$ and $v \in R^n$, $P_\lambda(u)$ can be Lasso penalty, hard thresholding penalty and SCAD. Finally, they derived the sparse loadings using the iterative procedure alternating $\tilde{u}$ and $\tilde{v}$ is fixed.

(c) PMD. Witten *et al.* [12] proposed penalized matrix decomposition (PMD), which

was a framework for computing a rank- k approximation of a matrix. The PMD generalizes this decomposition by additional constraints on U and V . The rank-1 PMD can be formulated as the following optimization problem:

$$\min_{d,u,v} imize \frac{1}{2} \| X - duv^T \|_F^2 \; s.t. \| u \|_2^2 = 1, \| v \|_2^2 = 1, P_1(u) \le \alpha_1, P_2(v) \le \alpha_2, d \ge 0 \qquad (27)$$

Where u is a column of U , v is a column of V , d is a diagonal element of D , P_1 and P_2 are convex penalty functions that can take various form. This optimization problem can be formulated to the following optimization problem:

$$\max_{u,v} imize \, u^T Xv \; s.t. \| u \|_2^2 \le 1, \| v \|_2^2 \le 1, P_1(u) \le \alpha_1, P_2(v) \le \alpha_2 \qquad (28)$$

Where α_1 and α_2 are constants, $P_1()$ and $P_2()$ are penalty functions. A simple maximization strategy of iteratively maximizing with respect to u and v is developed. They also have extended their works to Fisher's linear discrimination [83].

3.3. Sparse PCA from Data-Variance-Maximization View

A probabilistic interpretation of PCA called PPCA has been introduced [72], In order to perform sparsity on loading coefficients of probabilistic PCA modeling, Guan *et al.* [34] assigned Laplacian prior to each element of loadings based on this framework due to the Laplacian prior is equivalent to l_1 regularization in the sparse modeling. The object of the sparse probabilistic PCA is to estimate the parameters. Variational expectation-maximization (EM) algorithm or Markov Chain Monte Carlo algorithm can be used to estimate the parameters. Based on expectation-maximization for PPCA, Sigg and Buhmann [11] derived EMPCA for sparse and non-negative principal component analysis. Since MCMC procedures are too slow for a very high-dimensional application Sharp [58] presented a dense message passing algorithm for more efficient approximate inference in sparse probabilistic PCA.

4. Sparse PCA Software Package

Most of the developed method of sparse PCA has released their software package, in order to compare the performance of difference sparse PCA for readers (you can reference [84] to review the algorithm performance of typical sparse PCA).Table 1 summarizes some sparse PCA which have been published on the internet. From this table, we notice most packages are based on Matlab/R language and become more and more since recent years.

Table 1. Summary of Available Codes of Sparse PCA

No	Name	Author	Language	Year	Source
1	DSPCA	d'Aspreomot[8]	Matlab	2004	http://www.di.ens.fr/~aspremon/ZIP/DSPCA.zip
2	PathPCA	d'Aspremont[9]	Matlab	2007	http://www.di.ens.fr/~aspremon/ZIP/PathSPCA.zip
3	Fantop	Vu[16]	R	2013	https://github.com/vqv/fps
4	GPower	Journee[13]	Matlab	2008	http://www.montefiore.ulg.ac.be/~journee/GPower.zip
6	24am	Richtárik[69]	C++	2012	https://24am.googlecode.com/files/24am-v1_0.zip
6	TPower	Yuan[14]	Matlab	2011	https://sites.google.com/site/xtyuan1980/TPower_1.0.zip?attredirects=0
7	GRQI	Kuleshov[77]	Matlab	2013	https://github.com/kuleshov/generalized-rayleigh-quotient
8	ALSPCA	Lu[21]	Matlab	2012	http://www.sfu.ca/~yza30/homepage/codes/ALSPCA1.0.tar.gz

9	ITSPCA	Ma[15]	Matlab	2013	http://www-stat.wharton.upenn.edu/~zongming/software/SPCALab/SPCALab.zip
10	SPCA	Zou[10]	R	2006	http://cran.r-project.org/web/packages/elasticnet/
11	rSVD-sPCA	Shen[41]	R	2008	http://www.unc.edu/~haipeng/publication/rsvd.spca.Rfun.R
12	PMA	Witten[83]	R	2009	http://cran.r-project.org/web/packages/PMA/
13	Nsprcomp	Sigg[29]	R	2008	http://cran.r-project.org/web/packages/nsprcomp/index.html
14	DMP	Sharp[28]	Matlab	2010	http://www.cs.man.ac.uk/~sharpk/Code/DMP_v_1_0_Code_Only.zip
15	SPCA-ALM	Naikal[4]	Matlab	2011	http://www.eecs.berkeley.edu/~yang/software/SPCA/SPCA_ALM.zip
16	SPCA-bioin	Bonner[62]	R	2014	http://beyene-sigma-lab.com/code/analyze_DEPCs.R
17	Sparse logPCA	Lee[50]	R	2010	https://github.com/andland/SparseLogisticPCA
18	Structure sPCA	Jenatton[86]	Matlab	2010	http://rodolphejenatton.com/software/SparseStructuredPCA_MatlabToolbox_V1.0_rjenatton.tar
19	sPCA-Ran-def	Asteris[19]	Matlab	2011	http://megasthenis.github.io/repository/sparsePC-matlab.zip
20	sPCA-Consrank	Asteris[20]	Matlab	2014	http://megasthenis.github.io/repository/sparsePC-matlab.zip

5. Theoretical Analysis of High-Dimensional Sparse PCA

Sparsity can not only enhance the interpretability, but also it can yield consistent estimates if sparsity is truly presented in the population for high dimensional data. Statistical analysis of sparse PCA has been received significant attention recently and how to obtain dependable estimates statistically of eigenvectors and eigenspace for PCA on high-dimensional data has been the focus of recently literatures. The main questions needed to be answered in sparse PCA is whether there has an algorithm not only asymptotically consistent but also computationally efficient. Theoretical research from statistical guarantees view of sparse PCA includes consistency [2,8,14,38,41,50,53,55], minimax risk bounds for estimating eigenvectors [40,42-43,45,61], optimal sparsity level detection [4,44,48,59] and principal subspaces estimation [5,15-16,36,9,40,51,57] have been established under various statistical models. Because most of the methods based on spiked covariance model, so we firstly given an introduction about spiked variance model and then give a high dimensional sparse PCA theoretical analysis review from above several aspects.

5.1. Spiked-Covariance Model

Given a sample $X_1,...,X_n \in R^d$ drawn from a multivariate normal distribution $X_i \sim N(0,\sum)$ with mean is 0 and the population covariance matrix is $\sum$. The problem is the consistency of the sample covariance matrix $\hat{\sum} = \frac{1}{n} X'X$ as an estimator of the population covariance matrix $\sum$ when the dimension $d, n \to \infty$. Because the sample covariance matrix is not a good approximation to the population covariance matrix when the data dimension is larger than the sample size, Johnstone [2] proposed a spiked population model in which all but a fixed finite number of population eigenvalues(the spikes) are taken to be 1 as n, d become large.

The population covariance matrix can be formulated as $\sum = \hat{\sum} + \Delta$ under the

assumption $\Delta = \sigma^2 I$ .where Δ is noise. Using spectral decomposition of Sample covariance matrix $\hat{\Sigma} = U \Lambda U^T$ is get, then the covariance matrix of X_i is computed as $\hat{\Sigma} = U \Lambda U' + \sigma^2 I = \sum_{j=1}^{r} \lambda_j^2 u_j u'_j + \sigma^2 I$, where $\lambda_1^2 \geq ... \geq \lambda_r^2 > 0$ are the eigenvalues of $\hat{\Sigma}$, $\Lambda = diag(\lambda_1, ..., \lambda_r)$. Therefore the j-th eigenvalue of Σ is obtained as: $\Sigma = diag(\lambda_1^2 + \sigma^2, \lambda_2^2 + \sigma^2, ..., \lambda_r^2 + \sigma^2, \sigma^2, ..., \sigma^2)$, if $r = 1$, It is a single-spiked model.

5.2. Statistical Properties of High-Dimensional Sparse PCA

(a) Consistency. Johnstone *et al.* [2] presented a two-step method based on variable selection by largest entries in the diagonal of the sample covariance matrix, They proved that the conventional PCA performed on a selected subset of variables with the largest sample variances leads to a consistent estimator of u_1. They also proved that if the support of $k \leq \sqrt{n / \log p}$, Their method would be succeed with high probability and if $k \geq \sqrt{n / \log p}$, their method would be failed with high probability. Shen *et al.* [41] established conditions for consistency of a sparse PCA method in [11] when $p \to \infty$ and n is fixed. Yuan [98] also derived the convergence rate of TPower methods.

Under the single spike population model, Amin *et al*. [38] considered the variable selection property and developed conditions for recovering the non-zero entries of eigenvectors by the methods of diagonal thresholding [2] and DSPCA based on semi-definite programming relaxation[8] where both sample size and dimension tend to infinity. They proved that the information-theoretic critical rate is $k \leq n / \log p$. If $k << n / \log p$, no method can succeed in variable selection. In contrast, if $k \leq \sqrt{n / \log p}$, diagonal thresholding is consistent. Continue to the work of [38], Krauthgamer *et al*. [50] proved that no computationally-efficient algorithm can recover the support if $k \geq \sqrt{n}$. They proposed covariance thresholding (CT) method and proved that if the sparsity levels $k = O(\sqrt{n})$, recovery is possible. They also showed that the rank-one condition assumed by [38] does not hold if $\sqrt{n} \leq k \leq (n / \log p)$. Deshpande [55] considered another question whether covariance thresholding be a polynomial time algorithm that is guaranteed to solve the sparse PCA problem for $\sqrt{n / \log p} \leq k \leq \sqrt{n}$.

(b) Minmax rates of convergence. Birnbaum *et al.* [42] considered the minimax rates of convergence and adaptive estimation of the individual leading vector as the ordered coefficients of each eigenvector have rapid decay. Vu *et al.* [39] studied the rates of convergence of estimation under various sparsity assumption on the leading vector. They established minimax rates for estimation under l_2 loss with l_q -penalized estimators with suitably model parameters. Wang *et al.* [59] considered the question of whether it is possible to find an estimator of u_1 that is computable in polynomial time, and it attained the minimax optimal rate of convergence u_1 . They showed that no randomized polynomial time algorithm can achieve the minimax optimal rate.

(c) Optimal sparisty levels detection. Sparse detection method wants to detect the presence of a sparse structure in high dimensional data. [2,15] suggested heuristics when the detection levels are unknown, but they are not proven to achieve the optimal detection levels. Berthet *et al.* [47,53] proved whether there exists a polynomial-time computable statistic for reliably detecting the presence of a single spike of l_0 -sparsity. They proved that no polynomial algorithm will reconstruct the support unless $k \leq \sqrt{n}$. An interesting work should be addressed here, [47-48] established computational lower bounds in sparse principal component detection by the same difficult problem — the Planted Clique problem [84].

(d) Principal subspaces estimation. Most of researches above focused on estimating the

leading eigenvector u_1, but if some leading eigenvalues are identical or close to each other, individual eigenvectors are not identifiable. Moreover, if PCA is considered as a dimension reduction technique, the low-dimensional subspace onto which we project data should be of the interest [15]. So recently most works are presented on principal subspaces estimation which focused primarily on finding principal subspaces of $\sum$ spanned by sparse leading eigenvectors. Paul and Johnstone [5,36] studied multiple-spike model and proposed an augmented sparse PCA method to estimate each of the leading eigenvectors attaining near optimal rate of convergence of their procedure in the high dimensional setting. Their work provided asymptotic lower bounds for the minimax rate of convergence over l_q balls for $q \in (0,2]$.They also analyzed the performance of an estimator based on the multistage theresholding procedure and show that it can nearly attain the optimal rate of convergence. In contrast, Vu *et al.* [39] presented a model allowing a more general class of covariance matrices. Ma *et al.* [15] presented a iterative thresholding method and proved its consistency and achieved a near optimal statistical convergence rates when estimating several individual leading vectors under the spiked covariance model with the similar condition in [42]. Cai *et al.* [43,45] attained an optimal principal subspace estimator based on a regression-type method, and the minimax rates of convergence are derived and a computationally efficient adaptive estimator is constructed. Vu *et al.* [16] proposed a new method called FPS which generalized DSPCA to estimate the principal subspace spanned by the top k leading eigenvectors. Nevertheless, [16,40]established a near-optimal Frobenius norm error bound for the FPS estimator under general conditions and showed that the obtained estimator only attained the suboptimal $s*\sqrt{\log d / n}$ statistical rate of convergence. Lei *et al.* [51] considered the variable selection consistency and agnostic inference properties of Fantope projection and selection (FPS) method which didn't need spiked-covariance model anymore. When the eigenvalues of $\sum$ are fixed, a sufficient condition for consistent variable selection using FPS is $s \le \sqrt{n / \log p}$ while being computationally tractable. Similar but different with [51], Gu [57] proposed a family of estimators for subspace of a population matrix based on the semi-definite relaxation of sparse PCA with novel regularizations which didn't rely on the spiked covariance model. One is convex sparse PCA which had oracle property and the same convergence rate as standard PCA. The second estimator is non-convex sparse PCA which can also attained faster rate than [51]. Wang [61] also proposed a two-stage sparse PCA procedure employed sparse orthogonal iteration pursuit that attained the optimal principal subspace estimator in polynomial time which converged at the rate of $1/\sqrt{t}$ within the initialization stage, and at a geometric rate within the main stage.

6. Discussions and Challenges

In our paper, a literature survey of current sparse PCA in sparse PCA has been given. Two important issues have been studies: the summarization of sparse PCA's various formulations and algorithms and the survey of the theoretical analysis for sparse PCA in previous research. Based on this review, we now take on the challenge of discussing some perspective research directions.

6.1. Performance Improvements of Algorithms (Sparse PCA)

Although we categorized the sparse PCA algorithm from the optimization formulation added by the different constraint and penalty form as l_0 or l_1-norm sparse PCA in Figure1 in our paper, but the algorithms for sparse PCA can be categorized from different aspects. In order to evaluation the performance of the algorithm, the sparse PCA algorithm can derive a new kind of category of sparse PCA algorithms as shown in Figure 2. Seen from this Figure, we noticed that most of the nowadays sparse PCA algorithms are deflation-

based method which focused on the first leading principal component, iterative deflation technique can be used to obtain the additional components from the input matrix. The weakness of most of the listed methods is that they produced sparse loadings that are not completely orthogonal and the components are correlated [16, 21]. How to improve the orthogonality of eigenvectors and decrease the correlation of PCs is an open problem in the development of sparse PCA algorithms. The first answer is the principal subspace estimation, the theoretical analysis for principal subspace estimation of sparse PCA has been paid more attention since 2013 as described in section 5 which will improve the performance of sparse PCA, but the work for estimating the principal subspace or even multiple eigenvectors simultaneously is very little [16]. To our best knowledge, only FPS [16] and ITSPCS [15] are principal subspace estimation based methods. We believe that the development of principal subspace estimation methods for sparse PCA will be developed in the next few years.

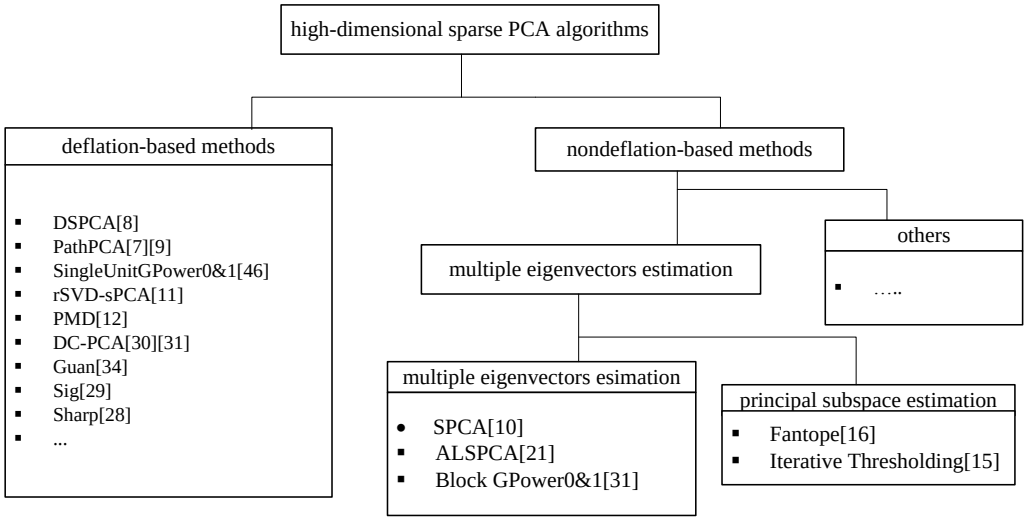

Figure 2. A New Kind of Category of Sparse PCA Algorithm Based on Deflation

Secondly, how to derive a new formulation for Sparse PCA like [21, 27]to obtain the sparse and orthogonal loading vectors, the components are uncorrelated while capturing as much variance as possible is also an important directions .

In most of algorithms of sparse PCA, the degree of sparsity is controlled via a penalization parameter in the sparse PCA algorithms, then how to tuning such parameter corresponding is another open problem. A similar problem is that the user does not know in advance if and how sparse the loadings will be, but tuning the penalty parameters in the methods is time consuming for high dimensional data[11,85], an efficient tuning algorithm of the parameters trying to avoid them is preferred.

Besides the typical sparse PCA, there also exists many research on the extensible sparse PCA which doesn't included in our paper, such as structured sparse PCA [86], Robust sparse PCA [87], sparse PCA for Rank-deficient matrix [19] or constant-rank matrix[20], a sparse logistic principal component analysis for binary data [88], sparse principal components via semi-partition clustering [89], interpretable principal components using clustering [90], principal component analysis with sparse fused loadings [85] and so on. So how to extend the typical sparse PCA suitable for special circumstance is also important and interesting problem.

6.2. Trade-Off Theoretical and Computational Sparse PCA

Despite this comprehensive literature review, although the consistency and convergence has established for sparse PCA in high dimensional data, most of existing statistical guarantees are known hold under on the spiked covariance models. However the real application is not as this, theoretical analysis of sparse PCA on the general model is an open problem. Moreover, although there are various kinds of algorithm to solve sparse PCA, but only the Thresholding methods and Semi-definite Programming based method has been statistically analyzed, Most of existing methods lack statistical guarantees. How to expand the theoretical analysis of other methods for sparse PCA is also a hard problem. From our review process, we also noticed that there remains a big gap between the computational and statistical aspects of sparse PCA. There is no tractable algorithm is known to attain the statistical optimal sparse PCA estimator provably without relying on the spiked covariance assumption. Is there a polynomial time method with strong statistical guarantees for the general model? Is there a polynomial time method with principal subspace estimation in high dimension circumstances is still need us to make a deeply exploration.

6.3. Extending the Application of Sparse PCA

In the past several years, Sparse PCA has been successfully applied in diverse areas as bioinformatics, natural language processing and machine vision. Because automatically learning the features from high dimensional data has been a major research topic in machine learning and pattern recognition, and sparse PCA can be used as an unsupervised feature extraction step which can derive new feature learning algorithm. It is an important direction to derive a new sparse PCA for special application, and the fast and simple sparse PCA must be also considered firstly to extend the practical application of sparse PCA.

Acknowledgements

This paper is partially supported by Fundamental Research Funds for the Central University (NS2015092).

References

[1] S. Jung and J. S. Marron, “PCA consistency in high dimension, low sample size context”, The Annals of Statistics, vol. 37, no. 6B, **(2009)**, pp. 4104–4130.
[2] I. M. Johnstone and A. Lu, “On Consistency and Sparsity for Principal Components Analysis in High Dimensions”, Journal of the American Statistical Association, vol. 104, no. 486, **(2009)**, pp. 682-693.
[3] I. M. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis”, Annals of Statistics, vol. 29, no. 2, **(2001)**, pp. 295-327.
[4] N. Naikal, A. Y. Yang and S. S. Sastry, “Informative feature selection for object recognition via sparse PCA”, IEEE International Conference on Computer Vision (ICCV), **(2011)**, pp. 818-825.
[5] D. Paul, “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model”, Statistica Sinica, vol. 14, no. 4, **(2007)**, pp. 1617-1642.
[6] I. Jolliffe, N. T. Trendafilov and M. Uddin (2003), “A modified principal component technique based on the LASSO”, Journal of Computational and Graphical Statistics, vol. 12, no. 13,**(2013)**, pp. 531-547.
[7] A. d'Aspremont, F. Bach and L. El Ghaoui, “Optimal Solutions for Sparse Principal Component Analysis”, Journal of Machine Learning Research, vol. 9, **(2008)**, pp.1269-1294.
[8] A. d'Aspremont, L. El-Ghaoui, M. Jordan, and G. Lanckriet, “A direct formulation for sparse PCA using semi-definite programming”, SIAM Review, vol. 49, no. 3, **(2004)**, pp. 434-448.
[9] A. d'Aspremont, B. Alexandre, R. Francis and L. El. Ghaoui, “Full regularization path for sparse principal component analysis”, In Proceedings of the 24th international conference on Machine learning, ICML '07, **(2007)**, pp. 177-184.
[10] H. Zou, T. Hastie, R. Tibshirani, “Sparse principal component analysis”, Journal of Computational and Graphical Statistics, vol. 15, no. 2, **(2006)**, pp. 265-286.
[11] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation”, Journal of Multivariate Analysis vol. 99, no. 6, **(2008)**, pp. 1015-1034.

[12] D. M. Witten, R. Tibshirani and T. Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis", Biostatistics, vol. 10, no. 3, **(2009)**, pp. 515-534.
[13] M. Journée, Y. Nesterov, P. Richtarik and R. Sepulchre, "Generalized power method for sparse PCA", The Journal of Machine Learning Research, vol. 11, **(2010)**, pp. 517-553.
[14] X. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems", The Journal of Machine Learning Research, vol. 14, no. 1, **(2013)**, pp. 899-925.
[15] Z. Ma, "Sparse principal component analysis and iterative thresholding", Annals of Statistics, no. 41, vol.2, **(2013)**, pp. 772-801.
[16] V. Q. Vu, J. Cho, J. Lei and K. Rohe, "Fantope Projection and Selection: A near-optimal convex relaxation of Sparse PCA", Advances in Neural Information Processing Systems (NIPS), **(2013)**, pp. 2670-2678.
[17] A. Farcomeni, "An exact approach to sparse principal component analysis", Computational Statistics, vol. 24, no. 4, **(2009)**, pp. 583–604.
[18] F. H. Kaiser, "The Varimax Criterion for Analytic Rotation in Factor Analysis", Psychometricka, vol. 23, no. 3, **(1958)**, pp. 187-200.
[19] M. Asteris, D. Papailiopoulos and G. N. Karystinos, "Sparse principal component of a rank-deficient matrix", Proceedings of IEEE International Symposium on Information Theory, **(2011)**, pp. 673-677.
[20] M. Asteris, D. Papailiopoulos and G. N. Karystinos, "The sparse principal component of a constant-rank matrix", IEEE Transactions on Information Theory, accepted, **(2014)**.
[21] Z. Lu and Y. Zhang, "An augmented Lagrangian approach for sparse principal component analysis", Math. Program, vol. 135, no. 1-2, **(2012)**, pp. 149-193.
[22] R. Luss and M. Teboulle, "Conditional Gradient Algorithms for Rank One Matrix Approximations with a Sparsity Constraint", SIAM Review, vol. 55, no. 1**(2013)**, pp. 65-98.
[23] S. Ma, "Alternating direction method of multipliers for sparse principal component analysis", Journal of the Operations Research Society of China, vol. 1, no.2**, (2013)**, pp. 253-274.
[24] G. M. Merolo, "Sparse principal component analysis: a least squares approximation approach", arXiv preprint arXiv: 1406.1381, **(2014)**.
[25] B. Moghaddam, Y. Weiss and S. Avidan, "Spectral bounds for sparse PCA: Exact and greedy algorithms. In Advances in Neural Information Processing Systems", **(2006)**, pp. 915-922.
[26] D. S. Papailiopoulous, A. G. Dimakis and S. Korokythakis, "Sparse PCA through low-rank approximations", arXiv preprint arXiv: 1303.0551, **(2013)**.
[27] X. Qi, R. Luo and H. Zhao, "Sparse principal component analysis by choice of norm", Journal of Multivariate Analysis, vol. 114, **(2013)**, pp. 127-160.
[28] K. Sharp and M. Rattray, "Dense message passing for sparse principal component analysis", Proceedings of 13th international conference on artificial intelligence and statistics, **(2010)**, pp. 725-732.
[29] C. Sigg and J. Buhmann, "Expectation-maximization for sparse and non-negative PCA", Proceedings of the 25th International Conference on Machine Learing (ICML), **(2008)**, pp. 960-967.
[30] B. Sriperumbudur, D. Torres and G. Lanckriet, "A D.C. programming approach to the sparse generalized eigenvalue problem", arXiv preprint arXiv: 0901.1504. **(2009)**.
[31] B. Sriperumbudur, D. Torres and G. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem", Machine Learning, vol. 85, no. 1-2, **(2011)**, pp. 3-39.
[32] Y. Wang and Q. Wu, "Sparse PCA by iterative elimination algorithm", Advances in Computational Mathematics, vol. 36, no. 1, **(2012)**, pp. 137-151.
[33] R. Zass and A. Shashua, "Nonnegative sparse PCA", Advances in Neural Information Processing Systems, **(2007)**, pp. 1561-1568.
[34] Y. Guan and J. Dy, "Sparse probabilistic principal component analysis", Proceedings of 12th International Conference on Artificial Intelligence and Statistics, **(2009)**, pp.185-192.
[35] W. Hager and J. Zhu, "Projection algorithms for non-convex minimization with application to sparse principal component analysis", arXiv preprint arXiv: 1404.4132, **(2014)**.
[36] D. Paul and I. M. Johnstone, "Augmented sparse principal component analysis for high dimensional data", manuscript **(2007).**
[37] D. Paul and I. M. Johnstone, "Augmented sparse principal component analysis for high dimensional data", arXiv preprint arXiv: 1202.1242, **(2012)**.
[38] A. Amini and M. Wainwright, "High dimensional analysis of semi-definite relaxations for sparse principal component analysis", Annals of Statistics, vol. 37, no. 5B, **(2009)**, pp. 2877-2921.
[39] V. Q. Vu and J. Lei. "Minimax rates of estimation for sparse PCA in high dimensions", arXiv preprint arXiv: 1202.0786, **(2012)**.
[40] V. Q. Vu and J. Lei, "Minimax Sparse Principal Subspace Estimation in High Dimensions", Annals of Statistics, vol. 41, no. 6, **(2013)**, pp. 2905-2947.
[41] D. Shen, H. Shen and J. S. Marron, "Consistency of Sparse PCA in High Dimension, Low Sample Size Contexts", Journal of Multivariate Analysis, vol. 115, **(2013)**, pp. 317-333.
[42] A. Birnbaum, I. M. Johnstone, B. Nadler and D. Paul, "Minimax bounds for sparse-PCA with noisy high dimensional data", Annals of Statistics, vol. 41, no. 3, **(2013)**, pp. 1055-1084.

[43] T. T. Cai, Z. M. Ma and Y. Wu, "Recent results on sparse principle component analysis", IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), **(2013)**, pp. 181-183.
[44] T. T. Cai, Z. Ma and Y. Wu, "Optimal estimation and rank detection for sparse spiked covariance matrices", Probability Theory and Related Fields, **(2013)**, pp. 1-35.
[45] T. T. Cai, Z. Ma and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation", The Annals of Statistics, vol. 41, no. 6, **(2013)**, pp. 3074-3110.
[46] T. T. Cai, Z. Ma and Y. Wu, "Complexity theoretic lower bounds for sparse principal component detection", Conference on Learning Theory, **(2013)**, pp. 1046-1066.
[47] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension", The Annals of Statistics, vol. 41, no. 4, **(2013)**, pp. 1780-1815.
[48] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection", Conference on Learning Theory, **(2013)**, pp. 1046-1066.
[49] K. Lounici, "Sparse principal component analysis with missing observations", High Dimensional Probability VI, Springer, **(2013)**, pp. 327–356.
[50] R. Krauthgamer, B. Nadler and D. Vilenchick, "Do semi-definite relaxations really solve sparse PCA up to the information limit?" to appear in the Annals of Statistics, **(2015)**.
[51] J. Lei and V. Q. Vu, "Sparsistency and agnostic inference in sparse PCA", arXiv preprint arXiv: 1401.6978, **(2014)**.
[52] M. Asteris, D. Papailiopoulos and A. G. Dimakis, "Nonnegative sparse PCA with provable guarantees", Proceedings of International Conference on Machine Learning (ICML), Beijing, China, June 22-24, **(2014)**.
[53] J. Bickel and E. Levina, "Regularized estimation of large covariance matrices", Annals of Statistics, vol. 36, **(2008)**, pp.199-227.
[54] A. d'Aspremont, F. Bach and L. El Ghaoui, "Approximation Bounds for Sparse Principal Component Analysis", Mathematical Programming Series B, vol. 148, no. 1-2, **(2014)**, pp. 89-110.
[55] Y. Deshpande and A. Montanari, "Sparse PCA via Covariance Thresholding", Advances in Neural Information Processing Systems (NIPS), **(2014)**, pp. 334-342.
[56] Y. Deshpande and A. Montanari. "Information-theoretically optimal sparse PCA", IEEE International Symposium on Information Theory, **(2014)**, pp. 2197-2201.
[57] Q. Q. Gu, Z. R. Wang and H. Liu, "Sparse PCA with oracle property", Advances in Neural Information Processing Systems **(2014)**, pp. 1529-1537.
[58] A. J. Rothman, P. J. Bickel, E. Levina and J. Zhu, "Sparse permutation invariant covariance estimation", Electronic Journal of Statistics, vol. 2, **(2008)**, pp. 494-515.
[59] T. Y. Wang and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components", arXiv preprint arXiv: 1408.5369, **(2014)**.
[60] Z. R. Wang, H. Liu and T. Zhang, "Optimal Computational and Statistical Rates of Convergence for Sparse Non-convex Learning Problems", Annals of Statistics, vol. 42, no. 6, **(2014)**, pp. 2164.
[61] Z. R. Wang, H. R. Lu and H. Liu, "Non-convex Statistical Optimization: Minimax-Optimal Sparse PCA in Polynomial Time", arXiv preprint arXiv: 1408.5352, **(2014)**.
[62] A. J. Bonner and J. Beyene, "Detecting networks of genes associated with human drug induced liver injury (DILI) concern using sparse principal components", Revised manuscript submitted to Systems Biomedicine, **(2014)**.
[63] X. Chen, "Adaptive elastic-net sparse principal component analysis for pathway association testing", Statistical applications in genetics and molecular biology, vol. 10, no. 1, **(2011)**, pp. 1-21.
[64] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe and J. P. Mesirov, "Multiclass cancer diagnosis using tumor gene expression signatures", Proceedings of the National Academy of Sciences, vol. 98, **(2001)**, pp. 15149–15154.
[65] Z. R. Wang, F. Han and H. Liu, "Sparse Principal Subspace Analysis for High Dimensional Vector Autoregressive Models", **(2013),** Submitted.
[66] Z. R. Wang, F. Han and H. Liu, "Sparse Principal Component Analysis for High Dimensional Multivariate Time Series", International Conference on Artificial Intelligence and Statistics (AISTATS)**(2013)**, pp. 48-56.
[67] Y. W. Zhang and L. El. Ghaoui, "Large-Scale Sparse Principal Component Analysis with Application to Text Data", Proceedings Advances in Neural Information Processing Systems (NIPS)**(2011)**, pp. 532-539.
[68] Y. Zhang, A. d'Aspremont and L. El Ghaoui, "Sparse PCA: Convex Relaxations, Algorithms and Applications", Preprint on ArXiv: 1011.3781, Handbook on Semi-definite, Cone and Polynomial Optimization, **(2011)**, pp. 915-940.
[69] P. Richtárik, M. Takáč, and S. D. Ahipaşaoğlu, "Alternating maximization: unifying framework for 8 sparse PCA formulations and efficient parallel codes", arXiv preprint arXiv: 1212.4137, **(2012)**.
[70] Y. L. Hsu, P. Y. Huang and D. T. Chen, "Sparse principal component analysis in cancer research", Translational Cancer Research, vol. 3, no. 3, **(2014)**, pp. 182-190.
[71] N. T. Trendafilov, "From simple structure to sparse components: a review", Computational Statistics, vol. 29, no. 3-4, **(2014)**, pp. 431-454.

[72] M. E. Tipping, C. M. Bishop, "Probabilistic principal component analysis", Journal of the Royal Statistical Society, Series B, vol. 61, no. 3, **(1999)**, pp. 611-622.
[73] L. Mackey, "Deflation methods for sparse PCA", Advances in Neural Information Processing Systems, **(2009)**, pp. 1017-1024.
[74] J. N. R. Jeffers, "Two Case Studies in the Application of Principal component Analysis", Applied statistics, vol. 16, no. 3, **(1967)**, pp. 225-236.
[75] S. K. Vienes, "Simple principal components", applied statistics, vol. 49, no. 4, **(2000)**, pp. 441-451.
[76] J. Cadima and I. T. Jolliffe, "Loadings and correlations in the interpretation of principal components", Journal of Applied Statistics, vol. 22, no. 2, **(1995)**, pp. 203–215.
[77] V. Kuleshov, "Fast algorithms for sparse principal component analysis based on Rayleigh quotient iteration", Proceedings of the 30th International Conference on Machine Learning, **(2013)**, pp. 1418-1425.
[78] N. T. Trendafilov and I. T. Jolliffe, "Projected gradient approach to the numerical solution of the SCoTLASS", Computational Statistics and Data Analysis, vol. 50, no.1, **(2006)**, pp. 242-253.
[79] Y. Nesterov and A. Nemirovskii, "Interior-Point Polynomial Algorithms in Convex Programming", Society for Industrial Mathematics 13. SIAM, Philadelphia, PA, **(1994)**.
[80] Y. Nesterov, "Smooth minimization of non-smooth functions", Mathematical Programming, vol. 103, no.1, **(2005)**, pp.127-152.
[81] S. Boyd, N. Parikh, E. Chu, B.Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", In: Foundations and Trends in Machine Learning, vol. 3, no. 1**(2010)**, pp. 1-122.
[82] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, **(2005)**, pp. 301-320.
[83] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant", Journal of the Royal Statistical Society, Series B. vol. 73, no. 5, **(2011),** pp. 753-772.
[84] V. Feldman, E. Grigorescu, L. Reyzin, S. S. Vempala and Y. Xiao, "Statistical algorithms and a lower bound for detecting planted cliques", Proceedings of the forty-fifth annual ACM Symposium on Theory of Computing, **(2013)**, pp. 655-664.
[85] F. Guo, J. Gareth, E. Levina, G. Michailidis and J. Zhu, "Principal component analysis with sparse fused loadings", Journal of Computational Graphical Statistics, vol. 19, no. 4, **(2010)**, pp. 947-962.
[86] R. Jenatton, G. Obozinski and F. Bach, "Structured sparse principal component analysis", arXiv preprint arXiv: 0909.1440, **(2009)**.
[87] C. Crou, P. Filzmose and H. Fritz, "Robust Sparse Principal Component Analysis", Technometrics, vol. 55, no. 2, **(2013)**, pp. 202–214.
[88] S. Lee, J. Z. Huang and J. Hu, "Sparse logistic principal component analysis for binary data", Annuals of Applied Statistics, vol. 4, no. 3, **(2010)**, pp. 1579-1601.
[89] D. Enki, N. T. Trendafilov, "Sparse principal components by semi-partition clustering", Computational Statistics, vol. 27, no. 4, **(2012)**, pp. 605-626.
[90] D. Enki, N. T. Trendafilov and T. Jolliffe, "A clustering approach to interpretable principal components", Journal of Applied Statistics, vol. 40, no. 3, **(2013)**, pp. 583-599.

Authors

Shen Ning-min, He was born in 1991. He received the B.S. degree in Computer Science from Jiangxi Normal University in 2013. Now he is a Master Candidate of Nanjing University of Aeronautics and Astronautics. His research interests are data mining, parallel computing and software verification.

Li Jing, She was born in 1976.She received the Ph.D. degree in Computer Science and Technology from Nanjing University in 2004. Currently, she is an associate professor of Nanjing University of Aeronautics and Astronautics. Her research interests include data mining, image processing and software verification.