INVITED PAPER

A Selective Overview of Sparse Principal Component Analysis

This paper provides a selective overview of methodological and theoretical developments of sparse PCA that produce principal components that are sparse, i.e., have only a few nonzero entries.

By Hui Zou and Lingzhou Xue

ABSTRACT | Principal component analysis (PCA) is a widely used technique for dimension reduction, data processing, and feature extraction. The three tasks are particularly useful and important in high-dimensional data analysis and statistical learning. However, the regular PCA encounters great fundamental challenges under high dimensionality and may produce "wrong" results. As a remedy, sparse PCA (SPCA) has been proposed and studied. SPCA is shown to offer a "right" solution under high dimensions. In this paper, we review methodological and theoretical developments of SPCA, as well as its applications in scientific studies.

KEYWORDS | Covariance matrices; mathematical programming; principal component analysis (PCA); statistical learning.

I. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) was invented by Pearson [45]. As a dimension reduction and feature extraction method, PCA has numerous applications in statistical learning, such as handwritten zip code classification [26], human face recognition [24], eigengenes analysis [1], gene shaving [25], and so on. It would not be exaggerating to say that PCA is one of the most widely used and most important multivariate statistical techniques.

This review article focuses on the high-dimensional extension of the regular PCA, which is often called sparse PCA (SPCA). There are several popular SPCA methods in the literature, which will be reviewed in Section II. Their formulations are different but related, because the regular PCA has several equivalent definitions from different viewing angles. These definitions are equivalent without sparsity constraints, and differ with sparsity constraints. To be self-contained, we briefly discuss the several views of PCA in the following.

From a dimension reduction perspective, PCA can be described as a set of orthogonal linear transformations of the original variables such that the transformed variables maintain the information contained in the original variables as much as possible. Specifically, let $\boldsymbol{X}$ be an $n \times p$ data matrix, where n and p are the number of observations and the number of variables, respectively. For ease of presentation, assume the column means of $\boldsymbol{X}$ are all 0. The first principal component is defined as $Z_1 = \sum_{j=1}^{p} \alpha_{1j} X_j$ where $\alpha_1 = (\alpha_{11}, \dots, \alpha_{1p})^T$ is chosen to maximize the variance of Z_1, i.e.,

$$\alpha_1 = \arg\max_{\alpha} \alpha^T \hat{\boldsymbol{\Sigma}} \alpha \quad \text{subject to } \|\alpha_1\| = 1 \tag{1}$$

with $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{X}^T \boldsymbol{X})/n$. The rest principal components can be defined sequentially as follows:

$$\alpha_{k+1} = \arg\max_{\alpha} \alpha^T \hat{\boldsymbol{\Sigma}} \alpha \tag{2}$$

subject to

$$\|\alpha\| = 1 \quad \text{and} \quad \alpha^T \alpha_l = 0, \ \forall 1 \le l \le k. \tag{3}$$

Manuscript received January 30, 2018; revised May 28, 2018; accepted June 8, 2018. Date of current version August 2, 2018. The work of H. Zou was supported in part by the National Science Foundation (NSF) under Grant DMS-1505111. The work of L. Xue was supported by the National Science Foundation (NSF) under Grant DMS-1505256. *(Corresponding author: Hui Zou.)*

H. Zou is with the Department of Statistics, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: zouxx019@umn.edu).

L. Xue is with the Pennsylvania State University, State College, PA 16801 USA (e-mail: lzxue@psu.edu).

Digital Object Identifier 10.1109/JPROC.2018.2846588

This definition implies that the first K loading vectors are the first K eigenvectors of $\hat{\mathbf{\Sigma}}$.

The eigendecomposition formulation of PCA also relates PCA to the singular value decomposition (SVD) of $\boldsymbol{X}$. Let the SVD of $\boldsymbol{X}$ be

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $d_1, \ldots, d_p$ in a descending order, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times p$ and $p \times p$ orthonormal matrices, respectively. Because the columns of $\boldsymbol{V}$ are the eigenvectors of $\widehat{\mathbf{\Sigma}}$, $\boldsymbol{V}$ is the loading matrix of the principal components. By $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}$, we see that $Z_k = U_k d_k$ where U_k is the kth column of $\boldsymbol{U}$. Note that SVD can be interpreted as the best low-rank approximation to the data matrix.

PCA has another geometric interpretation, as the closest linear manifold approximation of the observed data. This definition actually matches the construction of PCA considered in [45]. Let $\boldsymbol{x}_i$ be the ith row of $\boldsymbol{X}$. Consider the first k principal components jointly $\boldsymbol{V}_k = [V_1|\cdots|V_k]$. By definition, $\boldsymbol{V}_k$ is a $p \times k$ orthonormal matrix. Project each observation to the linear space spanned by $\{V_1, \ldots, V_k\}$. The projection operator is $\boldsymbol{P}_k = \boldsymbol{V}_k\boldsymbol{V}_k^T$ and the projected data are $\boldsymbol{P}_k X_i$, $1 \leq i \leq n$. One way to define the best projection is by minimizing the total ℓ_2 approximation error

$$\min_{\boldsymbol{V}_k} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{V}_k\boldsymbol{V}_k^T\boldsymbol{x}_i\|^2. \tag{4}$$

It is easy to show that the solution is exactly the first k principal components.

In applications, variables can have different scales and units. Practitioners often standardize each variable such that its marginal sample variance is one. When this practice is applied to PCA, the resulting covariance matrix of standardized variables is the sample correlation matrix of the raw variables. Note that the eigenvalues and eigenvectors of the correlation matrix can be different from those of the covariance matrix.

II. METHODS FOR SPARSE PRINCIPAL COMPONENTS

Each principal component is a linear combination of all p variables, which makes it difficult to interpret the derived principal components as new features. Rotation techniques are commonly used to help practitioners to interpret principal components [30]. Vines [51] considered simple principal components by restricting the loadings to take values from a small set of allowable integers such as 0, 1, and -1. This restriction may be useful for certain applications but not all. Simple thresholding is an *ad hoc* way to achieve sparse loadings by setting the loadings with absolute values smaller than a threshold to zero. Although the simple thresholding is frequently used in practice, it can be potentially misleading in various respects [9]. Sparse variants of PCA aim to achieve a good balance between variance explained (dimension reduction) and sparse loadings (interpretability).

Sparse learning is ubiquitous in high-dimensional data analysis. Prior to the sparse principal component problem, an important question is how to select variables in high-dimensional regression. For the regression problem, the lasso proposed in [50] is a very promising technique for simultaneous variable selection and prediction. The lasso regression is an ℓ_1 penalized least squares method. The use of ℓ_1 penalization yields a sparse solution and also permits efficient computations.

A. SCoTLASS

Inspired by the lasso regression, Jolliffe *et al.* [31] proposed a procedure called SCoTLASS to obtain sparse loadings by directly imposing an ℓ_1 constraint on the loading vector.

SCoTLASS extends the standard PCA by taking the variance maximization perspective of the PCA. It successively maximizes the variance

$$a_k^T \hat{\mathbf{\Sigma}} a_k \tag{5}$$

subject to

$$a_k^T a_k = 1 \quad \text{and (for} \quad k \geq 2) \quad a_h^T a_k = 0, \qquad h < k \tag{6}$$

and the extra ℓ_1 constraints

$$\sum_{j=1}^{p} |a_{kj}| \leq t \tag{7}$$

for some tuning parameter t. However, SCoTLASS is high computational cost which makes it an impractical solution for high-dimensional data analysis. It motivated researchers to consider more efficient proposals for sparse principal components. A related but more efficient approach is the generalized power method presented in Section II-E.

B. Sparse Principal Component Analysis

After SCoTLASSO, the first computational efficient SPCA algorithm for high-dimensional data was introduced in [61]. Their method is named sparse principal component analysis (SPCA). Before reviewing the technical details, let us consider the application of SPCA to the pitprops data [28], a classical example showing the difficulty of interpreting principal components. The pitprops data have 180 observations and 13 measured variables. In [61], the first six ordinary principal components and the first six sparse principal components are computed. Here we only cite the results of the first three principal components in Table 1. Compared with the standard PCA, SPCA generated very sparse loading structures without losing much variance.

In the original lasso paper, Tibshirani used a quadratic programming solver to compute the lasso regression estimator, which is not very efficient for high-dimensional

Table 1 Compare PCA and SPCA on the Pitprops Data. Empty Cells Mean Zero Loadings. The Variance of SPCA Is Expected to Be Smaller Than That of PCA, by the Definition of PCA. The Differences in Variance Are Small

	PCA			SPCA		
	PC1	PC2	PC3	PC1	PC2	PC3
topdiam	-.404	.218	-.207	-.477		
length	-.406	.186	-.235	-.476		
moist	-.124	.541	.141		.785	
testsg	-.173	.456	.352		.620	
ovensg	-.057	-.170	.481	.177		.640
ringtop	-.284	-.014	.475			.589
ringbut	-.400	-.190	.253	-.250		.492
bowmax	-.294	-.189	-.243	-.344	-.021	
bowdist	-.357	.017	-.208	-.416		
whorls	-.379	-.248	-.119	-.400		
clear	.011	.205	-.070			
knots	.115	.343	.092		.013	
diaknot	.113	.309	-.326			-.015
variance	32.4	18.3	14.4	28.0	14.0	13.3

data. Efron *et al.* [17] derived the first efficient algorithm named LARS for computing the entire solution path of the lasso regression model with high-dimensional data. Motivated by LARS, Zou *et al.* [61] proposed to tackle the sparse principal component problem from a regression formulation. The resulting algorithm is SPCA.

SPCA extends the linear manifold approximation view of the PCA to derive sparse loadings. Recall that the first principal component can be defined as

$$\alpha_1 = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \alpha\alpha^T \boldsymbol{x}_i\|^2 \quad \text{subject to} \quad \|\alpha\|^2 = 1. \tag{8}$$

We reformulate (8) as

$$\underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \alpha\beta^T \boldsymbol{x}_i\|^2 \quad \text{subject to} \quad \|\alpha\|^2 = 1 \text{ and } \alpha = \beta. \tag{9}$$

The following theorem says that we can drop the equality constraint in (9) and still recover the first loading vector exactly.

Theorem 1 [61]: For any $\lambda_0 > 0$, let

$$(\hat\alpha, \hat\beta) = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \alpha\beta^T \boldsymbol{x}_i\|^2 + \lambda_0\|\beta\|^2 \quad \text{subject to} \quad \|\alpha\|^2 = 1. \tag{10}$$

Then, $\hat\beta \propto V_1$.

In Theorem 1, the extra ℓ_2 term $\lambda_0\|\beta\|^2$ is not needed when $p < n$. When $p > n$, any $\lambda_0 > 0$ should be used and it does not affect the normalized β_1. By dropping the equality constraint $\alpha = \beta$, we can use an alternating minimization algorithm to optimize the criterion in (10) because α and β are separated variables. With a fixed α, the optimization problem over β is a regression problem.

Based on Theorem 1, we can impose a sparse penalty on β to gain zero loading because the normalizing step does not change the support of β. The SPCA for the first principal component is defined as

$$(\hat\alpha, \hat\beta) = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \alpha\beta^T \boldsymbol{x}_i\|^2 + \lambda_0\|\beta\|^2 + \lambda_1\|\beta\|_1 \quad \text{subject to} \quad \|\alpha\|^2 = 1 \tag{11}$$

and the output loading vector is $\hat V_1 = \hat\beta/\|\hat\beta\|$. For $n > p$, we can let $\lambda_0 = 0$ and solving β with a fixed α is a lasso regression problem, which can be done efficiently. When $n < p$, we need to use a positive λ_0 (e.g., $\lambda_0 = 10^{-3}$), solving β with a fixed α is an elastic net regression problem [60], which can be done efficiently as well.

Theorem 1 can be generalized to handle the first k principal components simultaneously, as stated in the next theorem.

Theorem 2 [61]: Suppose we are considering the first k principal components. Let $\boldsymbol{A}_{p\times k} = [\alpha_1, \ldots, \alpha_k]$ and $\boldsymbol{B}_{p\times k} = [\beta_1, \ldots, \beta_k]$. For any $\lambda_0 > 0$, let

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \underset{\boldsymbol{A},\boldsymbol{B}}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^T \boldsymbol{x}_i\|^2 + \lambda_0 \sum_{j=1}^{k} \|\beta_j\|^2 \quad \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = I_{k\times k}. \tag{12}$$

Then, $\hat\beta_j \propto V_j$ for $j = 1, 2, \ldots, k$.

Then, the SPCA criterion for the first k sparse principal components is defined as

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \underset{\boldsymbol{A},\boldsymbol{B}}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^T \boldsymbol{x}_i\|^2 + \lambda_0 \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j}\|\beta_j\|_1 \quad \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = I_{k\times k} \tag{13}$$

where different $\lambda_{1,j}$s are allowed for penalizing the loadings of different principal components.

Zou *et al.* [61] proposed an alternating algorithm to minimize the SPCA criterion (13).

For each j, let $Y_j^* = \boldsymbol{X}\alpha_j$. It is shown that $\hat{\boldsymbol{B}} = [\hat\beta_1, \ldots, \hat\beta_k]$ and each $\hat\beta_j$ is obtained via

$$\hat\beta_j = \underset{\beta_j}{\arg\min} \|Y_j^* - \boldsymbol{X}\beta_j\|^2 + \lambda_0\|\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1. \tag{14}$$

One can use either the LARS-EN algorithm [60] or the cyclic coordinate descent algorithm [21] to solve (14). Both algorithms are efficient for high-dimensional data.

If $\boldsymbol{B}$ is fixed, the optimization problem of $\boldsymbol{A}$ is

$$\underset{\boldsymbol{A}}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^T \boldsymbol{x}_i\|^2 = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|^2$$

subject to $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_{k\times k}$. This is called a reduced rank Procrustes rotation problem in [61] because when $k = p$, it is the Procrustes rotation problem [41]. Zou *et al.* [61] derived an explicit solution to the reduced rank Procrustes rotation problem. We compute the SVD

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{15}$$

and set $\hat{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{V}^T$.

The SPCA algorithm iterates between the elastic net regression step and the SVD step till convergence. The output is the normalized $\boldsymbol{B}$ matrix $\hat{V}_j = \hat{\beta}_j/\|\hat{\beta}_j\|, 1 \le j \le k$.

Zou *et al.* [61] derived another SPCA criterion to further speed up the computation efficiency. The derivation is based on the observation that Theorem 2 is valid for all $\lambda_0 > 0$. It turns out that a thrifty solution emerges if λ_0 is taken to be a large constant.

Theorem 3 [61]: Let $\widehat{V}_j(\lambda_0) = (\hat{\beta}_j/\|\hat{\beta}_j\|)$ $(j = 1, \dots, k)$ be the sparse loadings defined in (13). Let $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$ be the solution of the optimization problem

$$\begin{aligned}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \ & \underset{\boldsymbol{A},\boldsymbol{B}}{\arg\min} -2\mathrm{Tr}\left(\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\right) + \sum_{j=1}^{k}\|\beta_j\|^2 \\ & + \sum_{j=1}^{k}\lambda_{1,j}\|\beta_j\|_1 \\ & \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_{k\times k}.\end{aligned} \tag{16}$$

When $\lambda_0 \to \infty$, $\widehat{V}_j(\lambda_0) \to (\hat{\beta}_j/\|\hat{\beta}_j\|)$.

Solving (16) can also be done via an alternating minimization algorithm. Given $\boldsymbol{A}$, we have that for each j

$$\hat{\beta}_j = \underset{\beta_j}{\arg\min} -2\alpha_j^T(\boldsymbol{X}^T\boldsymbol{X})\beta_j + \|\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1 \tag{17}$$

and the solution is given by

$$\hat{\beta}_j = S\left(\boldsymbol{X}^T\boldsymbol{X}\alpha_j, \frac{\lambda_{1,j}}{2}\right)$$

where $S(Z,\gamma)$ is the soft-thresholding operator on a vector $Z = (z_1, \dots, z_p)$ with thresholding parameter γ and

$$S(Z,\gamma)_j = (|z_j| - \gamma)_+ \operatorname{sgn}(z_j), \quad 1 \le j \le p.$$

Given $\boldsymbol{B}$, the solution of $\boldsymbol{A}$ is again $\hat{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{V}^T$ where $\boldsymbol{U}$, $\boldsymbol{V}$ are from the SVD of $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{B}$: $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$.

C. A Semidefinite Programming Approach

We introduce some necessary notation first. We use $\mathrm{Card}(M)$ to denote the number of nonzero element of M, where M can be a vector of a matrix. The notation $|M|$ means that we replace each element of M with its absolute value. Let $\mathbf{1}_p$ be the p-vector of 1.

Consider the first k-sparse principal component with at most k nonzero loadings. A natural definition of the optimal k-sparse loading vector is

$$\begin{aligned}&\underset{\alpha}{\arg\max}\ \alpha^T \ \hat{\boldsymbol{\Sigma}}\alpha \\ &\text{subject to} \quad \|\alpha\| = 1, \quad \mathrm{Card}(\alpha) \le k.\end{aligned} \tag{18}$$

When $k = p$, then the above definition gives the loadings of the first principal component. However, (18) is nonconvex and computationally difficult, especially when p is large. Convex relaxation is a standard technique used in operational research to handle difficult nonconvex problems. d'Aspremont *et al.* [15] developed a convex relation of (18), which is expressed as a semidefinite programming problem.

Let $\boldsymbol{P} = \alpha\alpha^T$. We write $\alpha^T\hat{\Sigma}\alpha = \mathrm{Tr}(\hat{\Sigma}\boldsymbol{P})$. The norm-1 constraint on α leads to a linear equality constraint on $\boldsymbol{P}$: $\mathrm{Tr}\boldsymbol{P} = 1$. Moreover, the cardinality constraint $\|\alpha\|_0 \le k$ implies $\mathrm{Card}(\boldsymbol{P}) \le k^2$. Hence, we consider the following optimization problem of $\boldsymbol{P}$:

$$\begin{aligned}&\underset{\boldsymbol{P}}{\arg\max} \quad \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{P}) \\ &\text{subject to} \ \ \mathrm{Tr}\boldsymbol{P} = 1, \quad \mathrm{Card}(\boldsymbol{P}) \le k^2 \\ &\qquad\qquad\quad \boldsymbol{P} \succeq 0 \quad \text{and} \quad \mathrm{rank}(\boldsymbol{P}) = 1.\end{aligned} \tag{19}$$

The above formulation in (19) is still nonconvex and difficult to handle due to the cardinality constraint and the rank-1 constraint. By definition, $\boldsymbol{P}$ is symmetric and $\boldsymbol{P}^2 = \boldsymbol{P}$. Observe that

$$\|\boldsymbol{P}\|_F^2 = \mathrm{Tr}(\boldsymbol{P}^T\boldsymbol{P}) = \mathrm{Tr}(\boldsymbol{P}) = 1.$$

By Cauchy–Schwartz

$$\mathbf{1}_p^T|\boldsymbol{P}|\mathbf{1}_p \le \sqrt{\mathrm{Card}(\boldsymbol{P})\|\boldsymbol{P}\|_\mathrm{F}^2} \le k.$$

Therefore, d'Aspremont *et al.* [15] suggested to relax the cardinality constraint in (19) to a linear inequality constraint $\mathbf{1}_p^T|\boldsymbol{P}|\mathbf{1}_p \le k$. Furthermore, they dropped the rank-1 constraint and ended up with the DSPCA formulation

$$\begin{aligned}&\underset{\boldsymbol{P}}{\arg\max} \quad \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{P}) \\ &\text{subject to} \ \ \mathrm{Tr}\boldsymbol{P} = 1 \\ &\qquad\qquad\quad \mathbf{1}_p^T|\boldsymbol{P}|\mathbf{1}_p \le k \\ &\qquad\qquad\quad \boldsymbol{P} \succeq 0.\end{aligned} \tag{20}$$

The above is recognized as a semidefinite programming problem and can be solved by software such as SDPT3.

DSPCA only solves for $\boldsymbol{P}$ but not α. To compute the loading vector α, d'Aspremont *et al.* [15] recommended truncating $\boldsymbol{P}$ and retaining only the dominant (sparse) eigenvector of $\boldsymbol{P}$. For the second sparse principal component, it is suggested to replace $\hat{\boldsymbol{\Sigma}}$ with $\hat{\boldsymbol{\Sigma}} - (\alpha^T\hat{\boldsymbol{\Sigma}}\alpha)\alpha\alpha^T$ in (20). The same procedure can be repeated to compute the rest sparse principal components.

For larger problems, d'Aspremont *et al.* [15] discussed a Nesterov's smooth minimization technique to handle DSPCA. The computation complexity of the algorithm is $O(p^4\sqrt{\log(p)}/\epsilon)$, where ϵ is the numerical accuracy of the solution. d'Aspremont *et al.* [14] discussed a greedy algorithm to speed up the computation. An alternating direction method of multipliers was proposed in [39].

DSPCA formulation generated many interests in the operational research and machine learning communities. Some follow-up works include [38], [53], and [13], among others.

D. Iterative Thresholding Methods

PCA can be done via the SVD of the data matrix. Thus, it is natural to consider a SPCA algorithm based on the SVD of $\boldsymbol{X}$. This idea was explored in [48] and [56].

Let the SVD of $\boldsymbol{X}$ be $\boldsymbol{X} = \boldsymbol{UDV}^T$. Consider the first principal component. We know the loading vector is V_1, the first column of $\boldsymbol{V}$. It is a well-known result that SVD of $\boldsymbol{X}$ is related to the best lower rank approximation of $\boldsymbol{X}$ [16]. Specifically, let $\tilde{U}$ be a norm-1 n-vector and $\tilde{V}$ be a p-vector. Consider $\tilde{U}\tilde{V}^T$ as a rank-1 approximation of $\boldsymbol{X}$. The best rank-1 approximation is defined as

$$\begin{aligned}&\min_{\tilde{U},\tilde{V}} \|\boldsymbol{X} - \tilde{U} \quad \tilde{V}^T\|_{\mathrm{F}}^2\\ &\text{subject to} \qquad \|\tilde{U}\| = 1\end{aligned} \tag{21}$$

and the solution is $\tilde{U} = U_1$ and $\tilde{V} = d_1V_1$ where d_1 is the first singular value.

Based on (21), Shen and Huang [48] proposed the following optimization problem:

$$\begin{aligned}(\hat{U},\hat{V}) = &\arg\min_{U,V} \|\boldsymbol{X} - UV^T\|_{\mathrm{F}}^2 + \lambda\|V\|_1\\ &\text{subject to} \quad \|U\| = 1\end{aligned} \tag{22}$$

and the sparse loading vector is normalized $\hat{V}$, $(\hat{V}/\|\hat{V}\|)$. An alternating minimization algorithm is used to solve (22). Note that given V, the optimal U is $U = \boldsymbol{X}V/\|\boldsymbol{X}V\|$. Given U, the optimal V is

$$\arg\min_{V} -2\mathrm{Tr}(\boldsymbol{X}^TUV^T) + \|V\|^2 + \lambda\|V\|_1$$

and the solution is given by the soft-thresholding operator

$$V = S\left(\boldsymbol{X}^TU, \frac{\lambda}{2}\right).$$

Thus, Shen and Huang's method is an iterative thresholding algorithm.

Note that the above procedure is similar in spirit to the SPCA algorithm in (16). The big difference is that SPCA solves k components simultaneously, but Shen and Huang's method only deals with one component at a time.

Shen and Huang [48] proposed to sequentially compute the rest sparse principal components. Suppose that we have computed the first k (U,V) pairs, then let $\boldsymbol{X}_{(k+1)} = \boldsymbol{X} - \sum_l^k U_lV_l^T$ and then the iterative thresholding algorithm is applied to $\boldsymbol{X}_{(k+1)}$ to get $(U_{(k+1)}, V_{(k+1)})$. The normalized $V_{(k+1)}$ is the loading vector of the $(k+1)$th sparse principal component. The λ parameter is allowed to differ for different principal components.

In the same vein, Witten *et al.* [56] proposed a penalized matrix decomposition (PMD) criterion as follows:

$$\begin{aligned}(\hat{U},\hat{V},\hat{d}) = \arg&\min_{U,V,d} \quad \|\boldsymbol{X} - dUV^T\|_{\mathrm{F}}^2\\ &\text{subject to} \quad \|U\| = 1, \|U\|_1 \le c_1\\ &\qquad\qquad\quad \|V\| = 1, \|V\|_1 \le c_2.\end{aligned} \tag{23}$$

By straightforward calculation, it can be shown that (23) is equivalent to the following optimization problem:

$$\begin{aligned}(\hat{U},\hat{V}) &= \arg\max_{U,V} \quad U^T\boldsymbol{X}V\\ &\text{subject to} \ \|U\| = 1, \|U\|_1 \le c_1\\ &\qquad\qquad\quad \|V\| = 1, \|V\|_1 \le c_2\end{aligned} \tag{24}$$

and $\hat{d} = \hat{U}^T\boldsymbol{X}\hat{V}$.

They also used an alternating minimization algorithm to compute (24). Given V, we update U by solving

$$\max_U U^T\boldsymbol{X}V \quad \text{subject to} \quad \|U\| = 1, \|U\|_1 \le c_1. \tag{25}$$

Given U, we update V by solving

$$\max_V U^T\boldsymbol{X}V \quad \text{subject to} \quad \|V\| = 1, \|V\|_1 \le c_2. \tag{26}$$

The equality constraint $\|U\| = 1, \|V\| = 1$ in (25) and (26) can be replaced with inequality constraint $\|U\| \le 1$, $\|V\| \le 1$ and the solutions remain the same. So, (25) and (26) are examples of the following convex optimization problem:

$$\hat{Z} = \arg\max_Z Z^TR \quad \text{subject to} \quad \|Z\| \le 1, \|Z\|_1 \le c. \tag{27}$$

It is easy to see that the solution to (27) is

$$\hat{Z} = \frac{S(R,\Delta_c)}{\|S(R,\Delta_c)\|}$$

where S is the soft-thresholding operator and Δ_c is selected as follows: $\Delta_c = 0$ if $\|(R/\|R\|)\|_1 \le c$, otherwise $\Delta_c > 0$ is chosen to satisfy $\|\hat{Z}\|_1 = c$.

E. A Generalized Power Method

Consider the first principal component. By the variance maximization definition, a direct formulation of ℓ_1 constrained sparse principal component is

$$\begin{aligned}\arg&\max_{\|\alpha\|=1} \quad \alpha^T\boldsymbol{X}^T\boldsymbol{X}\alpha\\ &\text{subject to} \ \|\alpha\|_1 \le t.\end{aligned} \tag{28}$$

Equivalently, we can solve

$$\arg \max_{\|\alpha\|=1} \quad \sqrt{\alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha}$$
$$\text{subject to} \quad \|\alpha\|_1 \leq t. \tag{29}$$

Journée *et al.* [32] considered the Lagrangian form of (29)

$$\arg \max_{\|\alpha\|=1} \sqrt{\alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha} - \lambda \|\alpha\|_1. \tag{30}$$

They offered a generalized power method for solving (31). Their idea takes advantage of this simple observation: let $\tilde{U} = \arg\max_{\|U\|=1} U^T Z$, then $\tilde{U} = Z/\|Z\|$ and $\tilde{U}^T Z = \|Z\|$. Thus, an equivalent formulation of (31) is

$$\begin{aligned}(U^*, \alpha^*) &= \arg \max_{U,\alpha} \quad U^T \boldsymbol{X} \alpha - \lambda \|\alpha\|_1 \\ &\quad \text{subject to } \|U\| = 1, \ \|\alpha\| = 1.\end{aligned} \tag{31}$$

Note that the formulation (31) is the Lagrangian form of the PMD formulation (24) without imposing the ℓ_1 constraint on U.

For any U, the optimal α and $\boldsymbol{X}^T U$ must share the same sign for each component. Let $z_j = |\alpha_j|$, and $Z = |\alpha|$. Then, the optimal Z^* must satisfy

$$\begin{aligned}Z^* = \arg \max_{Z} \quad & \sum_{j=1}^{p} (|\boldsymbol{X}^T U|_j - \lambda) z_j \\ \text{subject to} \quad & z_j \geq 0, \quad \sum_{j=1}^{p} z_j^2 = 1.\end{aligned} \tag{32}$$

When $|\boldsymbol{X}^T U|_j - \lambda \leq 0$, $z_j^* = 0$. By Cauchy–Schwartz, it is easy to see that the solution to (32) is

$$z_j^* = \frac{(|\boldsymbol{X}^T U|_j - \lambda)_+}{\sqrt{\sum_{j=1}^{p} (|\boldsymbol{X}^T U|_j - \lambda)_+^2}} \tag{33}$$

which yields

$$\alpha = \frac{S(\boldsymbol{X}^T U, \lambda)}{\|S(\boldsymbol{X}^T U, \lambda)\|} \tag{34}$$

where S is the soft-thresholding operator. Plugging (33) back to the objective function in (31), we obtain a new optimization criterion of U

$$U^* = \arg \max_{U:\|U\|=1} \sqrt{\sum_{j=1}^{p} (|\boldsymbol{X}^T U|_j - \lambda)_+^2}$$

or equivalently

$$U^* = \arg \max_{U:\|U\|\leq 1} \sum_{j=1}^{p} (|\boldsymbol{X}^T U|_j - \lambda)_+^2. \tag{35}$$

Once U^* is solved, we have

$$\alpha^* = \frac{S(\boldsymbol{X}^T U^*, \lambda)}{\|S(\boldsymbol{X}^T U^*, \lambda)\|}.$$

Solving U^* is an n-dimensional optimization problem, although the original formulation (31) is a p-dimensional optimization problem. When $p \gg n$, the generalized power method achieves great computational savings. Moreover, the objective function in (35) is differentiable and convex, and the constraint set is compact and convex. Journée *et al.* [32] used an efficient gradient method to compute U^* and analyzed its convergence property. They also showed that the generalized power method can be extended to handle the first k principal components jointly.

There are other proposals for constructing spare principal components such as the truncated power method in [58] and the exact and greedy algorithms in [42].

III. THEORETICAL RESULTS

Theoretical analysis of SPCA received considerable attention in the past decade. In what follows, we first discuss the inconsistency of the classical PCA in the high-dimensional setting, and then present recent theoretical developments of SPCA.

A. Inconsistency of PCA Under High Dimensions

Statistical analysis of PCA views $\hat{\boldsymbol{\Sigma}}$ as the empirical covariance matrix, and there is the population PCA on the true covariance matrix $\boldsymbol{\Sigma}$. In the conventional setting where the dimension is fixed and the sample size increases, the principal eigenvectors of the sample covariance matrix are the consistent estimates of the principal eigenvectors of the corresponding population covariance matrix [3].

However, the sample principal eigenvectors are inconsistent estimates of the corresponding population principal eigenvectors in the high-dimensional setting where the dimension is no longer fixed and may be much larger than the sample size. The inconsistency phenomenon was first observed in the unsupervised learning theory literature in physics (for example, [7] and [55]). About a decade ago, a series of papers in the statistics literature (for example, [5], [29], [43], [44], and [33]) investigated the inconsistency results of the classical PCA when estimating the leading principal eigenvectors in the high-dimensional setting. Baik and Silverstein [5], Paul [44], and Nadler [43] showed that when $\lim_{n\to\infty} p/n = \gamma \in (0, 1)$, the largest eigenvalue λ_1 is of unit multiplicity and $\lambda_1 \leq \sqrt{\gamma}$, the leading sample principal eigenvector $\hat{\boldsymbol{v}}_1$ is asymptotically orthogonal to the leading population principal eigenvector $\boldsymbol{v}_1$ almost surely, that is

$$P(\lim_{n\to\infty} |\boldsymbol{v}_1^T \hat{\boldsymbol{v}}_1| = 0) = 1.$$

Johnstone and Lu [29] considered the rank-1 case and gave the sufficient and necessary condition for the consistence estimation of the leading population principal eigenvector. Let $R(\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1) = \cos\alpha(\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1)$be the cosine of the angle between $\hat{\boldsymbol{v}}_1$ and $\boldsymbol{v}_1$, and let $\omega = \lim_{n\to\infty} \|\boldsymbol{v}_1\|^2/\sigma^2$ be the limiting signal-to-noise ratio. Johnstone and Lu [29] proved that

$$P\left(\lim_{n\to\infty} R^2(\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1) = R^2_\infty(\omega, c)\right) = 1$$

where $c = \lim_{n\to\infty} p/n$ and

$$R^2_\infty(\omega, c) = (\omega^2 - c)_+/(\omega^2 + c\omega).$$

Note that $R^2_\infty(\omega, c) < 1$ if and only if $c > 0$. Thus, $\hat{\boldsymbol{v}}_1$ is a consistent estimate of $\boldsymbol{v}_1$ if and only $c = 0$, which implies the inconsistency of the classical PCA in the high-dimensional setting. Jung and Marron [33] further studied the strong inconsistency of the leading sample principal eigenvector in the high dimension and low sample size context where the sample size is fixed and the dimension increases.

These inconsistency results call for new formulation of principal components that are consistent estimators of the population principal components under high dimensions.

B. Consistency of SPCA

In recent years, there have been a series of papers to develop the theoretical properties of the SPCA in the statistics literature. The consistency results are established for various regularized estimators of the leading eigenvectors. Under the rank-1 scenario with $n^{-1}\log(n \vee p) \to 0$ as $n \to \infty$, Johnstone and Lu [29] established a consistency result for the classical PCA performed on a selected subset of variables satisfying $\hat{\sigma}^2 \geq \sigma^2(1 + \alpha_n)$, where $\alpha_n = \alpha(n^{-1}\log(n \vee p))^{1/2}$. Specifically, Johnstone and Lu [29] proved that the estimated principal eigenvector $\hat{\boldsymbol{v}}_1^I$ obtained via the subset selection rule is consistent

$$P\left(\lim_{n\to\infty} \alpha(\hat{\boldsymbol{v}}_1^I, \boldsymbol{v}_1) = 0\right) = 1$$

when the magnitudes of ordered coefficients of $\boldsymbol{v}_1$ have rapid decay, i.e., the r-th largest magnitude of $\boldsymbol{v}_1$ is no greater than $Cr^{-1/q}$, $r = 1, 2, \ldots$, for some $0 < q < 2$ and $0 < C < \infty$. This marginal variance selection method fails when the variables have equal or almost equal variance. Nevertheless, Johnstone and Lu [29] proved the first theoretical justification for SPCA. Shen *et al.* [47] established the consistency of the SPCA in the high dimension and low sample size context. Amini and Wainwright [2] studied the support recovery property of the semidefinite programming approach of [15] under the k-sparse assumption for the leading eigenvector in the rank-1 spiked covariance model. Ma [40] proved the consistency of the iterative thresholding approach under a spiked covariance model. Lei and Vu [37] provided the general sufficient conditions for sparsistency for the Fantope projection and selection method. In a very recent paper, Jankova and van de Geer [27] proposed a debiased SPCA estimator and studied the asymptotic inference of the sparse eigenvectors.

C. Minimax Rates of Convergence

The minimax rate of estimation is another important theoretical development for the SPCA. The seminal paper by Birnbaum *et al.* [8] studied the minimax rates of convergence and adaptive estimation when the rank is a fixed number and the ordered coefficients of each principal eigenvector have rapid decay. Specifically, Birnbaum *et al.* [8] established a lower bound on the minimax risk of estimators under various models of sparsity for the population eigenvectors. Ma [40] showed that the iterative thresholding estimator attains the minimax rate of convergence over a certain Gaussian class of distributions when the rank is treated as a fixed constant. By allowing the rank increase with the sample size, Cai [10] and Vu and Lei [52] studied the minimax optimality and adaptive estimation of the principal subspace for the SPCA in the high-dimensional setting. Following [10], we assume that the $n \times p$ data matrix $\boldsymbol{X}$ is generated as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + \boldsymbol{Z}$$

where $\boldsymbol{U}$ is the $n \times k$ random effects matrix with independent identically distributed (i.i.d.) $N(0, 1)$ entries, $\boldsymbol{D} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})$ is a diagonal matrix with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k > 0$, $\boldsymbol{V}$ is an orthonormal matrix, $\boldsymbol{Z}$ is a random matrix with i.i.d. $N(0, \sigma^2)$ entries, and $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent. Denote by $\boldsymbol{\Sigma}$ the covariance matrix of $\boldsymbol{X}$. Note that $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T + \sigma^2\boldsymbol{I}_p$ and also that the estimation of span$(\boldsymbol{V})$ is equivalent to the estimation of $\boldsymbol{V}\boldsymbol{V}^T$. Now, we consider the optimal estimation of the principal subspace span$(\boldsymbol{V})$ under the commonly used loss function $L(\boldsymbol{V}, \hat{\boldsymbol{V}}) = \|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T\|_F^2$and the following parameter space for $\boldsymbol{\Sigma}$:

$$\begin{aligned}\Theta(s, p, k, \lambda) = \{&\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T + \sigma^2\boldsymbol{I}_p : \kappa\lambda \geq \lambda_1 \\ &\geq \cdots \geq \lambda_k \geq \lambda > 0, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_k, \|\boldsymbol{V}\|_w \leq s\}\end{aligned}$$

where $\kappa > 1$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, and $\|\boldsymbol{V}\|_w = \max_{j=1,\cdots,p} j\|\boldsymbol{V}_{(j)*}\|_0$ is the weak ℓ_0 radius of $\boldsymbol{V}$. Note that the union of the column supports of $\boldsymbol{V}$ is of size at most s. Cai *et al.* [10] used the local metric entropy [36], [57] to construct the lower bound, and then obtain the minimax risk bound in the high-dimensional setting as follows:

$$\begin{aligned}&\inf_{\hat{\boldsymbol{V}}} \sup_{\boldsymbol{\Sigma}\in\Theta(s,p,k,\lambda)} E[L(\boldsymbol{V}, \hat{\boldsymbol{V}})] \\ &\qquad \asymp \left[\frac{\lambda/\sigma^2 + 1}{n(\lambda/\sigma^2)^2}\left(k(s-k) + s\log\frac{ep}{s}\right)\right] \wedge 1.\end{aligned}$$

Cai *et al.* [11] studied the minimax rates under the spectral norm, which is directly related to estimating the rank of the factor model.

D. Statistical and Computational Tradeoff

It is important to point out that there is a fundamental tradeoff between statistical and computational performance. In general, there are no known computationally efficient methods to obtain the minimax rate optimal estimators for the SPCA. Several seminal papers highlight the tradeoff between computational and statistical efficiency for the SPCA, including [2], [6], [22], [35], [54], and others. Amini and Wainwright [2] proved that no algorithm can reliably recover the sparse eigenvector under the single-spike covariance model when $k \geq Cn/\log p$ for some positive constant C and all sufficiently large n. Krauthgamer *et al.* [35] further proved that the semidefinite programming approach [15] does not close the gap between computational and statistical efficiency as long as $k \geq C\sqrt{n}$ for some positive constant C and all the sufficiently large n. Berthet and Rigollet [6] considered the optimal detection of sparse principal components in high dimension

$$H_0 : \boldsymbol{x} \sim N(0, \boldsymbol{I}_p) \quad \text{versus} \quad H_1 : \boldsymbol{x} \sim N(0, \boldsymbol{I}_p + \theta \boldsymbol{v}_1 \boldsymbol{v}_1')$$

where $\boldsymbol{v}_1$ has a fixed number of nonzero components. To this end, Berthet and Rigollet [6] studied a minimax optimal test based on the k-sparse largest eigenvalue of the empirical covariance matrix. The computation of this sparse eigenvalue statistic depends on a well-known decision problem associated to finding whether a graph contains a clique of size k, whose computational complexity is proved to be NP-complete in general [34]. In the follow-up paper, under the hardness assumption of the planted clique problem [20], Wang *et al.* [54] showed that there is an effective sample size regime in which no randomized polynomial time algorithm can achieve the minimax optimal rate for new and larger classes satisfying a restricted covariance concentration condition. Recently, Gao *et al.* [22] obtained the first computational lower bounds for SPCA under the Gaussian single spiked covariance model and closed the gap in SPCA computational lower bounds left by [6] and [54].

IV. APPLICATIONS

SPCA can be used in applications where PCA is normally used. For example, the use of SPCA in clustering can lead to sparse clustering algorithms [12]. PCA is a part of the integrated omic-data analysis, where SPCA can be used to replace the regular PCA [46], [59]. We discuss a few recent applications of SPCA in medical imaging, ecology, and neuroscience, respectively.

Shape/image analysis: Sjöstrand [49] applied SPCA to landmark-based shape analysis of the CC brain structure. The author extracted 5, 20, and 50 nonzero principal components out of the total 156 components corresponding to landmarks, and he also applied the standard PCA as a benchmark. In the subsequent analysis, he used the univariate regression to study the relationship between the resulting deformations based on extracted variables and four clinical outcome variables (gender, age, walking speed, and verbal fluency). His findings confirmed the male/female mean shape differences and identified the deformation of the CC corresponding to the measure of walking speed. The results for verbal fluency were also meaningful anatomically. Sjöstrand [49] found that SPCA is useful to derive localized and interpretable patterns of variability while PCA did not provide much interpretational value.

Ecological study: Motivated by generating meaningful combinations of the explanatory variables, Gravuer *et al.* [23] applied SPCA to perform the dimension reduction before fitting the ABT model. The sparsity helps the interpretability of their model. Specifically, Gravuer *et al.* [23] used SPCA to study a range of human, biogeographic, and biological influences on the invasion of Trifolium species into New Zealand. The sparse principal components were obtained from 29 categorical and continuous variables for three invasion stages (i.e., introduction, naturalization, and spread), and studied the relationship of sparse principal components to invasion success by using aggregated boosted trees. Specifically, the authors identified eight sparse principal components on 22 variables for intentional introduction and unintentional introduction-naturalization stages, seven sparse principal components on 25 variables for naturalization of intentionally introduced species, and seven sparse principal components on 28 variables for relative spread rate. Gravuer *et al.* [23] found that SPCA simultaneously improves interpretability and maintains high explained variance.

Neuroscience study: SPCA was used in [4] to study the light-driven Ca^{2+} signals of the GCL cells given a set of standardized visual stimuli in a probabilistic clustering framework. Baden *et al.* [4] first used SPCA to extract features that are localized in time and readily interpretable from the responses to the chirp, color, and moving bar stimulus, and then used a Gaussian mixture model on the extracted feature set for clustering. The authors extracted 20 features from the mean response to the chirp, six features from the mean response to the color stimulus, eight features from the response time course, and four features from its temporal derivative. Many classically used temporal response features were identified, including ON- and OFF-responses with different kinetics or selectivity to different temporal frequencies. They also tried the standard PCA and found the results lead to inferior cluster quality.

SPCA is implemented in the R package elasticnet available from CRAN: http://cran.r-project.org/.

The Matlab implementation of SPCA is available in the toolbox SpaSM from http://www2.imm.dtu.dk/projects/spasm/.

V. CONCLUDING REMARKS

In our discussion, we have only presented the use of ℓ_1-norm for sparsity, but there are other equally suitable penalty functions to be used in the SPCA methods, including SCAD [18], [19] or ℓ_0, among others.

We now have a good understanding of the role of sparsity in PCA and ways to effectively exploit the sparsity. There are still remaining issues. A very important issue to be investigated further is automated SPCA. By “automated” we mean that there is a principled but not overly complicated procedure to set these sparse parameters in SPCA. This question is particularly challenging when we solve several sparse principal components jointly. We would also like to have more empirical results to help us understand the pros and cons of each proposed SPCA technique, which may also inspire new and better approaches. ■

Acknowledgments

The authors would like to thank Prof. T. Bouwmans, Prof. Y. Chi, and Prof. N. Vaswani for inviting them to contribute this review paper to the special issue. They are grateful for the very helpful comments from three anonymous reviewers.

REFERENCES

[1] O. Alter, P. O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling,” *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.

[2] A. A. Amini and M. J. Wainwright, “High-dimensional analysis of semidefinite relaxations for sparse principal components,” *Ann. Stat.*, vol. 37, no. 5B, pp. 2877–2921, 2009.

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 2003.

[4] T. Baden, P. Berens, K. Franke, M. Roseön, M. Bethge, and T. Euler, “The functional diversity of retinal ganglion cells in the mouse,” *Nature*, vol. 529, no. 7586, pp. 345–350, 2016.

[5] J. Baik and J. W. Silverstein, “Eigenvalues of large sample covariance matrices of spiked population models,” *J. Multivariate Anal.*, vol. 97, no. 6, pp. 1382–1408, 2006.

[6] Q. Berthet and P. Rigollet, “Optimal detection of sparse principal components in high dimension,” *Ann. Stat.*, vol. 41, no. 4, pp. 1780–1815, 2013.

[7] M. Biehl and A. Mietzner, “Statistical mechanics of unsupervised structure recognition,” *J. Phys. A, Math. Gen.*, vol. 27, no. 6, pp. 1885–1897, 1994.

[8] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul, “Minimax bounds for sparse PCA with noisy high-dimensional data,” *Ann. Stat.*, vol. 41, no. 3, pp. 1055–1084, 2013.

[9] J. Cadima and I. T. Jolliffe “Loading and correlations in the interpretation of principle compenents,” *J. Appl. Stat.*, vol. 22, no. 2, pp. 203–214, 1995.

[10] T. T. Cai, Z. Ma, and Y. Wu, “Sparse PCA: Optimal rates and adaptive estimation,” *Ann. Stat.*, vol. 41, no. 6, pp. 3074–3110, 2013.

[11] T. T. Cai, Z. Ma, and Y. Wu, “Optimal estimation and rank detection for sparse spiked covariance matrices,” *Probability Theory Related Fields*, vol. 161, nos. 3–4, pp. 781–815, 2015.

[12] G. Chen, P. F. Sullivan, and M. Kosorok, “Biclustering with heterogeneous variance,” *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 30, pp. 12253–12258, 2013.

[13] A. d’Aspremont, “Identifying small mean-reverting portfolios,” *Quantitative Finance*, vol. 11, no. 3, pp. 351–364, 2011.

[14] A. d’Aspremont, F. Bach, and L. E. Ghaoui, “Optimal solutions for sparse principal component analysis,” *J. Mach. Learn. Res.*, vol. 9, pp. 1269–1294, Jul. 2008.

[15] A. d’Aspremont, L. El Jordan, M. I. Jordan, and G. R. G. Lanckriet, “A direct formulation for sparse PCA using semidefinite programming,” *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, 2007.

[16] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936.

[17] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[18] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[19] J. Fan, L. Xue, and H. Zou, “Strong oracle optimality of folded concave penalized estimation,” *Ann. Stat.*, vol. 42, no. 3, pp. 819–849, 2014.

[20] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao, “Statistical algorithms and a lower bound for detecting planted cliques,” *J. ACM*, vol. 64, no. 2, p. 8, 2014.

[21] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, “Pathwise coordinate optimization,” *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.

[22] C. Gao, Z. Ma, and H. H. Zhou, “Sparse CCA: Adaptive estimation and computational barriers,” *Ann. Stat.*, vol. 45, no. 5, pp. 2074–2101, 2017.

[23] K. Gravuer, J. J. Sullivan, P. A. Williams, and R. P. Duncan, “Strong human association with plant invasion success for Trifolium introductions to New Zealand,” *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 17, pp. 6344–6349, 2008.

[24] P. Hancock, A. Burton, and V. Bruce, “Face processing: Human perception and principal components analysis,” *Memory Cogn.*, vol. 24, no. 1, pp. 26–40, 1996.

[25] T. Hastie *et al.*, “‘Gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns,” *Genome Biol.*, vol. 1, pp. 1–21, Aug. 2000.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.

[27] J. Jankova and S. van de Geer (2018). “De-biased sparse PCA: Inference and testing for eigenstructure of large covariance matrices.” [Online]. Available: https://arxiv.org/abs/1801.10567

[28] J. Jeffers, “Two case studies in the application of principal component,” *Appl. Stat.*, vol. 16, no. 3, pp. 225–236, 1967.

[29] I. M. Johnstone and A. Y. Lu, “On consistency and sparsity for principal components analysis in high dimensions,” *J. Amer. Stat. Assoc.*, vol. 104, no. 486, pp. 682–693, 2009.

[30] I. T. Jolliffe, “Rotation of principal components: Choice of normalization constraints,” *J. Appl. Stat.*, vol. 22, no. 1, pp. 29–35, 1995.

[31] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, “A modified principal component technique based on the lasso,” *J. Comput. Graph. Stat.*, vol. 12, no. 3, pp. 531–547, 2003.

[32] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, “Generalized power method for sparse principal component analysis,” *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Feb. 2010.

[33] S. Jung and J. S. Marron, “PCA consistency in high dimension, low sample size context,” *Ann. Stat.*, vol. 37, no. 6B, pp. 4104–4130, 2009.

[34] R. M. Karp, “Reducibility among combinatorial problems,” in *Proc. Complexity Comput. Symp.*, Yorktown Heights, NY, USA: IBM Thomas J. Watson Research Center, 1972, pp. 85–103.

[35] R. Krauthgamer, B. Nadler, and D. Vilenchik, “Do semidefinite relaxations solve sparse PCA up to the information limit?” *Ann. Stat.*, vol. 43, no. 3, pp. 1300–1322, 2015.

[36] L. LeCam, “Convergence of estimates under dimensionality restrictions,” *Ann. Stat.*, vol. 1, no. 1, pp. 38–53, 1973.

[37] J. Lei and V. Q. Vu, “Sparsistency and agnostic inference in sparse PCA,” *Ann. Stat.*, vol. 43, no. 1, pp. 299–322, 2015.

[38] Y. Lu and Z. Zhang, “An augmented Lagrangian approach for sparse principal component analysis,” *Math. Program.*, vol. 135, nos. 1–2, pp. 149–193, 2012.

[39] S. Ma, “Alternating direction method of multipliers for sparse principal component analysis,” *J. Oper. Res. Soc. China*, vol. 1, no. 2, pp. 253–274, 2013.

[40] Z. Ma, “Sparse principal component analysis and iterative thresholding,” *Ann. Stat.*, vol. 41, no. 2, pp. 772–801, 2013.

[41] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. Orlando, FL, USA: Academic, 1979.

[42] B. Moghaddam, Y. Weiss, and S. Avidan, “Spectral bounds for sparse PCA: Exact and greedy algorithms,” *Adv. Neural Inf. Process. Syst.*, 2006, pp. 915–922.

[43] B. Nadler, “Finite sample approximation results for principal component analysis: A matrix perturbation approach,” *Ann. Stat.*, vol. 36, no. 6, pp. 2791–2817, 2008.

[44] D. Paul, “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model,” *Stat. Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[45] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[46] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, “Methods of integrating data to uncover genotype–phenotype interactions,” *Nature Rev. Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[47] D. Shen, H. Shen, and J. S. Marron, “Consistency of sparse PCA in high dimension, low sample size contexts,” *J. Multivariate Anal.*, vol. 115, pp. 317–333, Mar. 2013.

[48] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” *J. Multivariate Anal.*, vol. 6, no. 99, pp. 1015–1034, 2008.

[49] K. Sjöstrand, “Sparse decomposition and modeling of anatomical shape variation,” *IEEE Trans. Med. Imag.*, vol. 26, no. 12, pp. 1625–1635, Dec. 2007.

[50] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[51] S. K. Vines, “Simple principal components,” *Appl. Stat.*, vol. 49, no. 4, pp. 441–451, 2000.

[52] V. Vu and J. Lei, “Minimax sparse principal subspace estimation in high dimensions,” *Ann. Stat.*, vol. 41, no. 6, pp. 2905–2947, 2013.

[53] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, “Fantope projection and selection: A near-optimal convex relaxation of sparse PCA,” in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2670–2678.

[54] T. Wang, Q. Berthet, and R. J. Samworth, “Statistical and computational trade-offs in estimation of sparse principal components,” *Ann. Stat.*, vol. 44, no. 5, pp. 1896–1930, 2016.

[55] T. L. H. Watkin and J.-P. Nadal, “Optimal unsupervised learning,” *J. Phys. A, Math. Gen.*, vol. 27, no. 6, pp. 1899–1915, 1994.

[56] 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, pp. 515–534, 2009.

[57] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Stat.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[58] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 899–925, 2013.

[59] C. Zang *et al.*, "High-dimensional genomic data bias correction and data integration using MANCIE," *Nature Commun.*, vol. 7, p. 11305, Apr. 2016.

[60] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B, Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[61] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

ABOUT THE AUTHORS

Hui Zou received the B.S. and M.S. degrees in physics from the University of Science and Technology of China, Hefei, China, in 1997 and 1999, respectively, the M.S. degree in statistics from Iowa State University, Ames, IA, USA, in 2001, and the Ph.D. degree in statistics from Stanford University, Stanford, CA, USA, in 2005.

His research interests include high-dimensional inference, statistical learning, and computational statistics.

Lingzhou Xue received the B.S. degree in mathematics from Peking University, Beijing, China, in 2008 and the M.S. and Ph.D. degrees in statistics from the University of Minnesota, Minneapolis, MN, USA, in 2011 and 2012, respectively.

His research interests include high-dimensional statistical inference and large-scale optimization.