Recent Results on Sparse Principle Component Analysis

T. Tony Cai, Zongming Ma and Yihong Wu

***Abstract*—Principal component analysis (PCA) is one of the most commonly used statistical procedures for dimension reduction. This paper presents some recent results on the minimax estimation of principal subspaces in high dimensions. Under mild technical conditions, we characterize the minimax risk for estimating the principal subspace under the quadratic loss within absolute constant factors.**

I. Introduction

Principal component analysis (PCA) is one of the most commonly used techniques in multivariate analysis for dimension reduction and feature extraction, and is particularly well suited for the settings where the data is high-dimensional but the signal has a low-dimensional structure. PCA has a wide array of applications, ranging from image recognition to data compression to clustering. In the conventional setting where the dimension of the data is relatively small compared with the sample size, the principal eigenvectors of the covariance matrix is typically estimated by the leading eigenvectors of the sample covariance matrix which are consistent when the dimension p is fixed and the sample size n increases [1]. However, in the high-dimensional setting where p can be much larger than n, this approach leads to very poor estimates. At various levels of rigor and generality, a series of papers showed that the sample principal eigenvectors are no longer consistent estimates of their population counterparts. For example, [2] and [3] showed that if $p/n \to \gamma \in (0,1)$ as $n \to \infty$, and the largest eigenvalue $\lambda_1 \le \sqrt{\gamma}$ and is of unit multiplicity, then the leading sample principal eigenvector $\widehat{\mathbf{v}}_1$ is asymptotically almost surely orthogonal to the leading population eigenvector $\mathbf{v}_1$, i.e., $|\mathbf{v}_1'\widehat{\mathbf{v}}_1| \to 0$ almost surely. Thus, in this case, $\widehat{\mathbf{v}}_1$ is not useful at all as an estimate of $\mathbf{v}_1$. Even when $\lambda_1 > \sqrt{\gamma}$, the angle between $\mathbf{v}_1$ and $\widehat{\mathbf{v}}_1$ still does not converge to zero unless $\lambda_1 \to \infty$. In addition to being inconsistent, sample principal eigenvectors have nonzero loadings in all the coordinates. This renders their interpretation difficult when the dimension p is large.

In view of the above negative results in the high-dimensional setting, a natural approach to principal component analysis in high dimensions is to impose certain structural constraint on the leading eigenvectors. One of the most popular assumptions is that the leading eigenvectors have a certain type of sparsity. In this case, the problem is commonly referred to as *sparse PCA* in the literature. The sparsity constraint reduces the effective number of parameters and facilitates interpretation.

Various regularized estimators of the leading eigenvectors have been proposed in the literature. Theoretical analysis has so far mainly focused on the rank-one case, i.e., estimating the leading principal eigenvector $\mathbf{v}_1$. In this case, [4] showed that the classical PCA performed on a selected subset of variables with the largest sample variances leads to a consistent estimator of $\mathbf{v}_1$ if the ordered coefficients of $\mathbf{v}_1$ have rapid decay. [5] and [6] proposed other consistent estimators when $\mathbf{v}_1$ has a bounded number of nonzero coefficients. [7] studied the rates of convergence of estimation under various sparsity assumptions on $\mathbf{v}_1$, and [8] further considers the minimax rates with missing data. [9] investigated the variable selection property of the methods by [4] and [10] when $\mathbf{v}_1$ has k nonzero entries of the same magnitude. [11] considered minimax detection when $\mathbf{v}_1$ has a bounded number of non-zeros.

More recently, for estimating a *fixed* number $r \ge 1$ of leading eigenvectors as $n, p \to \infty$, [12] studied minimax rates of convergence and adaptive estimation of the individual leading eigenvectors when the ordered coefficients of each eigenvector have rapid decay. When $r > 1$ and some of the leading eigenvalues have multiplicity great than one, the individual leading eigenvectors can be unidentifiable. On the other hand, the principal subspace spanned by them is always uniquely defined. [13] proposed a new method for estimating the principal subspace and derived rates of convergence of the estimator under similar conditions to those in [12].

The focus of this paper is to provide a non-asymptotic characterization of the minimax risk for principal subspace estimation within universal constants, without the boundedness assumption on the rank r. The full details can be found in the full manuscript [14].

II. Statistical model

Suppose we observe the $n \times p$ data matrix

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{Z}. \tag{1}$$

Here $\mathbf{U}$ is the $n \times r$ random effects matrix with iid $N(0,1)$ entries, $\mathbf{D} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})$ with $\lambda_1 \ge \cdots \ge \lambda_r > 0$, $\mathbf{V}$ is $p \times r$ orthonormal, and $\mathbf{Z}$ has iid $N(0, \sigma^2)$ entries which are independent of $\mathbf{U}$. Equivalently, one can think of $\mathbf{X}$ as an $n \times p$ matrix with rows independently drawn from the distribution $N(0, \mathbf{\Sigma})$, where the covariance matrix $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' + \mathbf{I}_p. \tag{2}$$

T. Cai and Z. Ma are with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: tcai@wharton.upenn.edu, zongming@wharton.upenn.edu. Y. Wu is with Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Email: yihongwu@illinois.edu.

978-1-4673-3146-3/13/$31.00 ©2013IEEE

Here $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ is $p \times r$ with orthonormal columns. The r largest eigenvalues of $\mathbf{\Sigma}$ are $\lambda_i + 1$, $i = 1, ..., r$, and the rest are all equal to one. The r leading eigenvectors of $\mathbf{\Sigma}$ are given by the columns of $\mathbf{V}$. Since the spectrum of $\mathbf{\Sigma}$ has r spikes, the covariance structure (2) is commonly known as the *spiked covariance matrix model* [15] in the literature.

The objective of PCA is to estimate the principal subspace $\mathrm{span}(\mathbf{V})$ based on the observation $\mathbf{X}$. Note that the principal subspace is uniquely identified with the associated projection matrix $\mathbf{V}\mathbf{V}'$. In addition, any estimator could be regarded as the subspace spanned by the columns of a matrix $\widehat{\mathbf{V}}$ with orthonormal columns, hence uniquely identified with its projection matrix $\widehat{\mathbf{V}}\widehat{\mathbf{V}}'$. Thus, estimating $\mathrm{span}(\mathbf{V})$ is equivalent to estimating $\mathbf{V}\mathbf{V}'$. Let $\|\cdot\|_{\mathrm{F}}$ denote the Frobenius norm. In this paper we consider optimal estimation of $\mathrm{span}(\mathbf{V})$ under the loss function

$$L(\mathbf{V}, \widehat{\mathbf{V}}) = \|\mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}'\|_{\mathrm{F}}^2, \tag{3}$$

which is a commonly used metric to gauge the distance between linear subspaces. It also coincides with twice the sum of the squared sines of the principal angles between the respective linear span.

The difficulty of estimating $\mathrm{span}(\mathbf{V})$ depends on the joint sparsity of the columns of $\mathbf{V}$. Let $O(p, r) = \{\mathbf{V} \in \mathbb{R}^{p\times r} : \mathbf{V}'\mathbf{V} = \mathbf{I}_r\}$ denote the collection of $p \times r$ matrices with orthonormal columns. We consider the following parameter spaces for $\mathbf{\Sigma}$ where $\mathbf{V}$ is row-sparse:

$$\Theta_q(s, p, r, \lambda) = \{\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' + \mathbf{I}_p : \lambda \le \lambda_r \le \cdots \le \lambda_1 \le \kappa\lambda, \mathbf{V} \in O(p, r),\ |\mathrm{supp}(\mathbf{V})| \le s\}, \tag{4}$$

where $\kappa > 1$ is a fixed constant and $\mathrm{supp}(\mathbf{V})$ denotes the row support of $\mathbf{V}$. More general sparsity model can be defined via the weak-ℓ_q norm of $\mathbf{V}$ and treated analogously. For conciseness we focus on the exact sparse case in this paper and refer the approximately sparse case to [14]. Row sparsity is also known as group sparsity, which is useful for high-dimensional regression, see, e.g., [16].

III. Minimax Rates for Principal Subspace Estimation

For two sequences of positive numbers a_n and b_n, we write $a_n \gtrsim b_n$ when $a_n \ge cb_n$ for some absolute constant $c > 0$, and $a_n \lesssim b_n$ when $b_n \gtrsim a_n$. Finally, we write $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

The minimax rate for principle subspace estimation with respect to the loss function (3) is given by the following theorem.

Theorem 1. *Suppose we observe data* $\mathbf{X}$ *as in* (1). *Let* $\lambda \gtrsim \sqrt{\frac{\log n}{n}}$, $s - r \gtrsim s \wedge \log \frac{ep}{s}$ *and* $n \gtrsim s \log \frac{ep}{s} \vee \log \lambda$. *The minimax risk for estimating the principal subspace* $\mathrm{span}(\mathbf{V})$ *under the loss* (3) *satisfies*

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta_0(s,p,r,\lambda)} \mathsf{E}L(\mathbf{V}, \widehat{\mathbf{V}}) \asymp \frac{\lambda+1}{n\lambda^2}\left(r(s-r) + s\log\frac{ep}{s}\right) \tag{5}$$

as long as the right-hand side of (5) does not exceed some absolute constant. Otherwise, there exists no consistent estimator.

The minimax rate in (5) depends optimally on all parameters, namely, s, p, r, n and λ. Thus Theorem 1 provides a non-asymptotic characterization of the difficulty of the principal subspace estimation problem in terms of the minimax rates over a wide range of parameter values. In particular, it is interesting to note that the minimax rate for estimating the r leading singular vectors depend on the r only through $r(s-r)$, the dimension of the Grassmannian manifold $G(s, r)$. Therefore the dependence on r is *not* monotonic, with the worst case happening at $r = \frac{s}{2}$. However, it should be noted that in order for Theorem 1 to hold, it is *necessary* to have r strictly bounded away from s. Otherwise, in the degenerate case of $r = s$, the only uncertainty is in the support of $\mathbf{V}$. The minimax rate is indeed much faster than (5), because in this regime the support can be estimated accurately.

A key step in establishing the optimal rates of convergence is the derivation of rate-sharp minimax lower bounds. It is highly non-trivial to obtain a lower bound which depends optimally on all parameters, in particular the eigenvalues and the rank. Our main technical tool for the lower bounds is based on local metric entropy [17], [18], [19], instead of the usual methods based on explicit constructions of packing sets together with Fano's Lemma used, for example, in [3], [12], [7]. Although the method is abstract in nature, the advantage is that it only relies on the analytical behavior of the metric entropy of the parameter space, thus allowing us to sidestep constructing an explicit packing, which can be a challenging task due to the need of fulfilling both the orthogonality and the sparsity constraints. See [14, Section 2.1].

IV. Optimal Estimation via Aggregation

We now show the achievability of the minimax rate given in Theorem 1. The optimal estimator of $\mathbf{V}$ is constructed using sample splitting and aggregation. This procedure is theoretically interesting but computationally intensive. A data-driven and computationally efficient estimator in constructed in [14] by reducing the sparse PCA problem to linear regression under group sparsity, which achieves the optimal rate adaptively under stronger conditions.

We first note that the loss function (3) satisfies

$$L(\mathbf{V}, \widehat{\mathbf{V}}) = 2r - 2\|\widehat{\mathbf{V}}'\mathbf{V}\|_{\mathrm{F}}^2 = 2\|(\mathbf{I} - \mathbf{V}\mathbf{V}')\widehat{\mathbf{V}}\widehat{\mathbf{V}}'\|_{\mathrm{F}}^2. \tag{6}$$

Moreover, the loss function is invariant under orthogonal complement, i.e., $L(\mathbf{V}, \widehat{\mathbf{V}}) = L(\mathbf{V}^\perp, \widehat{\mathbf{V}}^\perp)$, where $[\mathbf{V}, \mathbf{V}^\perp], [\widehat{\mathbf{V}}, \widehat{\mathbf{V}}^\perp]$ are orthogonal matrices. Therefore the loss (6) admits the following upper bound

$$L(\mathbf{V}, \widehat{\mathbf{V}}) \le 2(r \wedge (p - r)). \tag{7}$$

For notational simplicity we assume that the sample size is $2n$ and we split the sample equally according to $\mathbf{X} = \begin{bmatrix}\mathbf{X}_{(1)}\\ \mathbf{X}_{(2)}\end{bmatrix}$, where $\mathbf{X}_{(i)} = \mathbf{U}_{(i)}\mathbf{D}\mathbf{V}' + \mathbf{Z}_{(i)}, i = 1, 2$. Denote by $\mathbf{S}_{(i)} = \frac{1}{n}\mathbf{X}_{(i)}'\mathbf{X}_{(i)}$ the corresponding sample covariance matrix. The

main idea is to construct a family of estimators $\{\widehat{\mathbf{V}}_B\}$ using the first sample, indexed by the row support $B \subset [p]$, where $\widehat{\mathbf{V}}_B$ is the optimal estimator one would use if one knew beforehand that $\mathrm{supp}(\mathbf{V}) = B$. Then we aggregate these estimators by selection using the second sample. Aggregation methods have been widely used and well studied in statistics literature (see, e.g., Nemirovski [20]). To the best of our knowledge, this is the first application of the aggregation approach to sparse PCA which yields optimality results.

For each $B \subset [p]$ such that $|B| = s$, we define $\widehat{\mathbf{V}}_B \in O(p, r)$ as the r leading singular vectors of $\mathbf{J}_B \mathbf{S}_{(1)} \mathbf{J}_B$, where $\mathbf{J}_B$ is the diagonal matrix given by

$$(\mathbf{J}_B)_{ii} = \mathbf{1}_{\{i \in B\}}. \tag{8}$$

Given the collection of the $\widehat{\mathbf{V}}_B$'s, we set

$$B^* = \operatorname*{argmax}_{\substack{B \subset [p] \\ |B| = s}} \mathsf{Tr}(\widehat{\mathbf{V}}_B' \mathbf{S}_{(2)} \widehat{\mathbf{V}}_B) \tag{9}$$

and define the aggregated estimator by

$$\widehat{\mathbf{V}}_* = \mathbf{V}_{B^*}. \tag{10}$$

It is natural to use the same sample covariance matrix to construct the $\widehat{\mathbf{V}}_B$'s and to select B^*. The main advantage of sample splitting is to decouple the selection of the support and the computation of the estimator. Thus, conditioning on the first sample, we can treat the candidate estimators as if they are deterministic, which greatly simplifies the analysis. Sample splitting is commonly used in aggregation based estimation, where a sequence of estimators are constructed from the first sample and the second sample is used to aggregate these candidates to produce a final estimator.

The estimator (10) requires knowledge of the sparsity s and the rank r. Moreover, it can be computationally intensive for large values of p since in principle one needs to enumerate all $\binom{p}{s}$ possible support sets in order to obtain B^*. Nonetheless, this estimator achieves the minimax rate in Theorem 1. The proof can be found in [14, Section 6.2]. In the special case of $r = 1$, a combinatorial procedure similar to (9)–(10) has been proposed in [7].

REFERENCES

[1] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. John Wiley and Sons, 2003.

[2] J. Baik and J. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, pp. 1382–1408, 2006.

[3] D. Paul, "Asymptotics of sample eigenstruture for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[4] I. 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, pp. 682–693, 2009.

[5] D. Shen, H. Shen, and J. Marron, "Consistency of sparse PCA in high dimension, low sample size contexts," *arXiv preprint arXiv:1104.4289*, 2011.

[6] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *arXiv manuscript*, vol. arXiv:1112.2679v1, 2011.

[7] V. Q. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS'12)*, 2012. [Online]. Available: http://arxiv.org/abs/1202.0786

[8] K. Lounici, "Sparse principal component analysis with missing observations," *arXiv preprint arXiv:1205.7060*, 2012.

[9] A. Amini and M. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Annals of Statistics*, vol. 37, no. 5B, pp. 2877–2921, 2009.

[10] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, pp. 434–448, 2007.

[11] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *arXiv preprint arXiv:1202.5070*, 2012.

[12] A. Birnbaum, I. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *arXiv preprint arXiv:1203.0967*, 2012.

[13] Z. Ma, "Sparse principal component analysis and iterative thresholding," *The Annals of Statistics*, vol. 41, no. 2, pp. 772–801, 2013.

[14] T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," 2012, preprint. [Online]. Available: http://arxiv.org/abs/1211.1309

[15] I. Johnstone, "On the distribution of the largest eigenvalue in principal component analysis," *The Annals of Statistics*, vol. 29, pp. 295–327, 2001.

[16] K. Lounici, M. Pontil, S. Van De Geer, and A. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *Annals of Statistics*, vol. 39, no. 4, pp. 2164–2204, 2011.

[17] L. Le Cam, "Convergence of estimates under dimensionality restrictions," *Annals of Statistics*, vol. 1, no. 1, pp. 38 – 53, 1973.

[18] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," *Z. Wahrscheinlichkeit.*, vol. 65, no. 2, pp. 181–237, 1983.

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

[20] A. Nemirovski, "Topics in non-parametric statistics," in *Ecole d'Eté de Probabilités de Saint-Flour 1998 volume XXVIII of Lecture Notes in Mathematics*, P. Bernard, Ed. New York: Springer, 2000, p. 1738.