The Annals of Statistics
2013, Vol. 41, No. 6, 3074–3110
DOI: 10.1214/13-AOS1178

SPARSE PCA: OPTIMAL RATES AND ADAPTIVE ESTIMATION

BY T. TONY CAI[1], ZONGMING MA[2] AND YIHONG WU[3]

University of Pennsylvania, University of Pennsylvania and University of Illinois at Urbana-Champaign

Principal component analysis (PCA) is one of the most commonly used statistical procedures with a wide range of applications. This paper considers both minimax and adaptive estimation of the principal subspace in the high dimensional setting. Under mild technical conditions, we first establish the optimal rates of convergence for estimating the principal subspace which are sharp with respect to all the parameters, thus providing a complete characterization of the difficulty of the estimation problem in term of the convergence rate. The lower bound is obtained by calculating the local metric entropy and an application of Fano's lemma. The rate optimal estimator is constructed using aggregation, which, however, might not be computationally feasible.

We then introduce an adaptive procedure for estimating the principal subspace which is fully data driven and can be computed efficiently. It is shown that the estimator attains the optimal rates of convergence simultaneously over a large collection of the parameter spaces. A key idea in our construction is a reduction scheme which reduces the sparse PCA problem to a high-dimensional multivariate regression problem. This method is potentially also useful for other related problems.

1. Introduction. Due to dramatic advances in science and technology, high-dimensional data are now routinely collected in a wide range of fields including genomics, signal processing, risk management and portfolio allo-

Received November 2012; revised May 2013.
[1]Supported in part by NSF FRG Grant DMS-08-54973, NSF Grant DMS-12-08982 and NIH Grant R01 CA 127334-05.
[2]Supported in part by the Dean's Research Fund of the Wharton School.
[3]Supported in part by NSF FRG Grant DMS-08-54973 when he was a postdoctoral fellow at the University of Pennsylvania.
AMS 2000 subject classifications. Primary 62H12; secondary 62H25, 62C20.
Key words and phrases. Adaptive estimation, aggregation, covariance matrix, eigenvector, group sparsity, low-rank matrix, minimax lower bound, optimal rate of convergence, principal component analysis, thresholding.

cation. In many applications, the signal of interest lies in a subspace of much lower dimension and the between-sample variation is determined by a small number of factors. For example, in spectroscopy, the variation of the infrared and ultraviolet spectra is driven by the concentration levels of a small number of chemical components in the system [53]. In financial econometrics, it is commonly believed that the variation in asset returns is driven by a small number of common factors combined with random noise [16].

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 [3]. 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 [4, 9, 23, 26, 30, 39, 43] showed that the sample principal eigenvectors are no longer consistent estimates of their population counterparts. For example, Baik and Silverstein [4] and Paul [43] 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 $\hat{\mathbf{v}}_1$ is asymptotically almost surely orthogonal to the leading population eigenvector $\mathbf{v}_1$, that is, $|\mathbf{v}_1' \hat{\mathbf{v}}_1| \to 0$ almost surely. Thus, in this case, $\hat{\mathbf{v}}_1$ 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 $\hat{\mathbf{v}}_1$ 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.

1.1. *Sparse PCA.* 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. See, for example, [18, 27, 28, 48, 52, 56, 60]. Theoretical analysis has so far mainly focused on the rank-one case, that is, estimating the leading principal eigenvector $\mathbf{v}_1$. In this case, Johnstone and Lu [26] 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. Shen, Shen and Marron [47] and Yuan and Zhang [59] proposed other consistent estimators when $\mathbf{v}_1$ has a bounded number of nonzero coefficients. Vu and Lei [54] studied the rates of convergence of estimation under various sparsity assumptions on $\mathbf{v}_1$, and Lounici [35] further considers the minimax rates with missing data. Amini and Wainwright [2] investigated the variable selection property of the methods by [26] and [18] when $\mathbf{v}_1$ has k nonzero entries all of the same magnitude. Berthet and Rigollet [5] considered minimax detection when $\mathbf{v}_1$ has a bounded number of nonzeros.

More recently, for estimating a fixed number $r \geq 1$ of leading eigenvectors as $n, p \to \infty$, Birnbaum et al. [9] 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. Ma [37] proposed a new method for estimating the principal subspace and derived rates of convergence of the estimator under similar conditions to those in [9].

1.2. *Estimation of principal subspace.* In this paper, we focus on the estimation of the principal subspace. Both minimax and adaptive estimation are considered. Throughout the paper, let $\mathbf{X}$ be an $n \times p$ data matrix generated as

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

Here $\mathbf{U}$ is the $n \times r$ random effects matrix with i.i.d. $N(0,1)$ entries, $\mathbf{D} = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$, $\mathbf{V}$ is $p \times r$ orthonormal and $\mathbf{Z}$ has i.i.d. $N(0, \sigma^2)$ 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

$$\mathbf{\Sigma} = \operatorname{Cov}(\mathbf{X}_{i*}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' + \sigma^2\mathbf{I}_p. \tag{2}$$

Here $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ 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 $\lambda_i + \sigma^2$, $i = 1, \ldots, r$, and the rest are all equal to σ^2. 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* [24] in the literature.

The goal of the present paper is to estimate the principal subspace $\operatorname{span}(\mathbf{V})$ 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 $\operatorname{span}(\mathbf{V})$ is equivalent to estimating $\mathbf{V}\mathbf{V}'$. Let $\|\cdot\|_{\mathrm{F}}$ denote the Frobenius norm. In this paper we consider optimal and adaptive estimation of $\operatorname{span}(\mathbf{V})$ 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 $\operatorname{span}(\mathbf{V})$ depends on the joint sparsity of the columns of $\mathbf{V}$. Let $\|\mathbf{V}_{j*}\|$ denote the Euclidean norm of the jth row of $\mathbf{V}$. Order the row norms in decreasing order as $\|\mathbf{V}_{(1)*}\| \geq \cdots \geq \|\mathbf{V}_{(p)*}\|$. We define the *weak ℓ_q radius* of $\mathbf{V}$ as

$$\|\mathbf{V}\|_{q,w} \triangleq \max_{j\in[p]} j\|\mathbf{V}_{(j)*}\|^q \tag{4}$$

and let

$$O(p,r) = \{\mathbf{V} \in \mathbb{R}^{p\times r} : \mathbf{V}'\mathbf{V} = \mathbf{I}_r\} \tag{5}$$

denote the collection of $p \times r$ matrices with orthonormal columns. We consider the following parameter spaces for $\mathbf{\Sigma}$ where the weak ℓ_q radius of $\mathbf{V}$ is constrained:

$$\begin{aligned}\Theta_q(s,p,r,\lambda) = \{\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' + \mathbf{I}_p : 0 < \lambda \leq \lambda_r \leq \cdots \leq \lambda_1 \leq \kappa\lambda,\\ \mathbf{V} \in O(p,r), \|\mathbf{V}\|_{q,w} \leq s\},\end{aligned} \tag{6}$$

where $q \in [0,2)$ and $\kappa > 1$ is a fixed constant. Note that in the rank-one case, our structural assumption coincides with [26], (9), or [37], (3.5). In the special case of $q = 0$, the union of the column supports of $\mathbf{V}$ is of size at most s. Weak ℓ_q-ball is a commonly used model for sparsity. See, for example, Abramovich et al. [1] for wavelet estimation and Cai and Zhou [15] for sparse covariance matrix estimation. Group sparsity is also useful for high-dimensional regression, see, for example, Lounici et al. [36].

Let $q \in [0,2)$ and $s > 0$. Denote the weak-ℓ_q ball on $O(p,r)$ by

$$\mathcal{G}_q(s,p) = \{\mathbf{V} \in O(p,r) : \|\mathbf{V}\|_{q,w} \leq s\}, \tag{7}$$

which is the parameter space of $\mathbf{V}$. In order for $\mathcal{G}_q(s,p)$ to be nontrivial, that is, neither empty nor the whole $O(p,r)$, the weak-ℓ_q radius must satisfy (see Section 7.1 in the supplementary material [12] for a proof)

$$\frac{2-q}{2} r \leq s \leq r^{q/2} p^{(2-q)/2}. \tag{8}$$

In particular, if $q = 0$, then we have $1 \leq r \leq s \leq p$. Throughout the paper, we assume that (8) holds.

1.3. *Optimal rates of convergence.* Combining the upper and lower bound results developed in Section 2, we establish the following minimax rates of convergence for estimating the principal subspace $\operatorname{span}(\mathbf{V})$ under the loss (3). We focus here on the exact sparse case of $q = 0$; the optimal rates for the general case of $q \in (0, 2)$ are given in Section 2. For two sequences of positive numbers a_n and b_n, we write $a_n \gtrsim b_n$ when $a_n \geq cb_n$ 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.

THEOREM 1. *Suppose we observe data* $\mathbf{X}$ *as in (1). Let* $\frac{\lambda}{\sigma^2} \gtrsim \sqrt{\frac{\log n}{n}}$, $s - r \gtrsim s \wedge \log \frac{ep}{s}$ *and* $n \gtrsim s \log \frac{ep}{s} \vee \log \frac{\lambda}{\sigma^2}$. *The minimax risk for estimating the principal subspace* $\operatorname{span}(\mathbf{V})$ *under the loss (3) satisfies*

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma} \in \Theta_0(s,p,r,\lambda)} \mathsf{E} \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \asymp \frac{\lambda/\sigma^2 + 1}{n(\lambda/\sigma^2)^2} \left(r(s-r) + s \log \frac{ep}{s} \right) \tag{9}$$

as long as the right-hand side of (9) does not exceed some absolute constant. Otherwise, there exists no consistent estimator.

The rate of convergence in (9) depends optimally on all the parameters s, p, r, n and λ. The result thus provides a precise characterization of the difficulty of the principal subspace estimation problem in terms of the minimax rates over a wide range of parameter values.

A key step in establishing the optimal rates of convergence is the derivation of rate-sharp minimax lower bounds. It is highly nontrivial 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 [7, 34, 58], instead of the usual methods based on explicit constructions of packing sets together with Fano's lemma used, for example, in [9, 43, 54]. 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 weak-ℓ_q ball constraints.

We then construct an explicit estimator using an aggregation scheme, which is shown to attain the same rates of convergence as those of the minimax lower bounds. The matching lower and upper bounds together establish the optimal rates of convergence. This aggregation method can potentially be useful for other high-dimensional sparse PCA problems as well. Aggregation methods have been widely used and well studied in statistics literature. See, for example, Juditsky and Nemirovski [29], Yang [57], Nemirovski [41] and Rigollet and Tsybakov [45]. To the best of our knowledge, this is the first application of the aggregation approach to sparse PCA which yields optimality results.

1.4. *Adaptive estimation.* The rate-optimal aggregation estimator depends on the model parameters that are usually unknown in practice and is unfortunately not computationally feasible when p is large. We then propose an adaptive estimation procedure that is fully data driven and easily implementable. The estimator is shown to attain the optimal rate of convergence simultaneously over a large collection of the parameter spaces defined in (6).

The proposed method is based on a reduction scheme. By a conditioning argument, the original sparse PCA problem is reduced to a high-dimensional regression problem with orthogonal design and group sparsity on the regression coefficients. Then, we apply the model selection penalty idea from [8] to construct the final estimator.

A key step in the reduction scheme is the construction of two new samples in the form of (1), which share the same realization of the random effects $\mathbf{U}$ but have independent copies of the noise matrix $\mathbf{Z}$. This construction works because a common realization of $\mathbf{U}$ is critical in guaranteeing a sufficient signal-to-noise ratio in the resulting regression problem. In contrast, splitting the original sample into two halves fails to achieve this goal. On the other hand, the independence of the noise components ensures that the regression problem has white noise structure. The adaptivity and minimax optimality of the subspace estimator depend heavily on those of the regression coefficient estimator. Thus, as a byproduct of the analysis, we also show that our estimator for regression coefficients is adaptively rate optimal under group sparsity. To the best of our knowledge, the specific estimator and its adaptive optimality is also new in the literature.

1.5. *Other related work.* The present paper is related to a fast growing literature on estimating sparse covariance/precision matrices as well as low-rank matrices. Significant advances have been made on optimal estimation of the whole covariance or precision matrix. Many regularization methods, including banding, tapering, thresholding and penalization, have been proposed. In particular, Cai, Zhang and Zhou [14] established the optimal rate of convergence for estimating a class of bandable covariance matrices under the spectral norm. Cai and Yuan [13] proposed a block thresholding procedure which is shown to be adaptively rate-optimal over a wide range of collections of bandable covariance matrices. Bickel and Levina [6] introduced a thresholding procedure and obtained rates of convergence for sparse covariance matrix estimation. Cai and Zhou [15] established the minimax rates of convergence for estimating sparse covariance matrices under a range of matrix norms including the spectral norm. Cai, Liu and Zhou [10] obtained the optimal rate of convergence for estimating the sparse precision matrices.

Our work is also related to another active area of research, namely, the recovery of low-rank matrices based on noisy observations. Negahban and Wainwright [40] studied (near) low-rank matrix recovery by M-estimators under restricted strong convexity based on the penalized nuclear norm min-

imization over matrices. Koltchinskii, Lounici and Tsybakov [32] considered estimation of low-rank matrices based on a trace regression model which includes matrix completion as a special case. A nuclear norm penalized estimator was proposed and a general sharp oracle inequality was established. See also Recht, Fazel and Parrilo [44] and Rohde and Tsybakov [46].

1.6. *Organization of the paper.* The rest of the paper is organized as follows. After introducing basic notation, Section 2 establishes the minimax rates of convergence for estimating the principal subspace by obtaining matching minimax lower and upper bounds. An aggregation estimator is constructed and shown to be rate optimal. Section 3 introduces an adaptive estimation procedure for the principal subspace which is fully data driven and easily computable. It is shown that this estimator attains the optimal rates of convergence simultaneously over a large collection of parameter spaces. Connections to other related problems are discussed in Section 5. The proofs of the main results and key technical lemmas are given in Section 6 and some additional technical arguments are contained in the supplementary material [12].

2. Minimax rates for principal subspace estimation. We establish in this section the minimax rates of convergence for estimating the principal subspace in two steps. First, minimax lower bounds are obtained for the estimation problem under the loss (3). Then an aggregation estimator is introduced and is shown to attain the same rates as given in the lower bounds, under mild conditions on the parameters. The matching lower and upper bounds thus establish the minimax rates of convergence.

We begin by introducing some basic notation. Throughout the paper, for any matrix $\mathbf{X} = (x_{ij})$ and any vector $\mathbf{u}$, denote by $\|\mathbf{X}\|$ the spectral norm, $\|\mathbf{X}\|_{\mathrm{F}}$ the Frobenius norm and $\|\mathbf{u}\|$ the vector ℓ_2 norm. Moreover, the ith row of $\mathbf{X}$ is denoted by $\mathbf{X}_{i*}$ and the jth column by $\mathbf{X}_{*j}$. Let $\operatorname{supp}(\mathbf{X}) = \{i : \mathbf{X}_{i*} \neq 0\}$ denote the row support of $\mathbf{X}$. For a positive integer p, $[p]$ denotes the index set $\{1, 2, \ldots, p\}$. For two subsets I and J of indices, denote by $\mathbf{X}_{IJ}$ the $|I| \times |J|$ submatrices formed by x_{ij} with $(i, j) \in I \times J$. Let $\mathbf{X}_{I*} = \mathbf{X}_{I[p]}$ and $\mathbf{X}_{*J} = \mathbf{X}_{[n]J}$. For any square matrix $\mathbf{A} = (a_{ij})$, we let $\operatorname{Tr}(\mathbf{A}) = \sum_i a_{ii}$ be its trace. Define the inner product of matrices $\mathbf{B}$ and $\mathbf{C}$ of the same size by $\langle \mathbf{B}, \mathbf{C} \rangle = \operatorname{Tr}(\mathbf{B}'\mathbf{C})$. For any matrix $\mathbf{A}$, we use $\sigma_i(\mathbf{A})$ to denote its ith largest singular value. When $\mathbf{A}$ is positive semi-definite, $\sigma_i(\mathbf{A})$ is also the ith largest eigenvalue of $\mathbf{A}$. Let $\operatorname{span}(\mathbf{A})$ denote the linear subspace spanned by the columns of $\mathbf{A}$. For any real numbers a and b, set $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any set A, $|A|$ denotes its cardinality. Let $\mathbb{S}^{p-1}$ denote the unit sphere in $\mathbb{R}^p$. Let $G(k, r)$ denote the Grassmannian manifold consisting of all r-dimensional linear subspace of $\mathbb{R}^k$. Let $O(p)$ denote the collection of all $p \times p$ orthogonal matrices. Throughout the paper, we use c

and C to denote generic absolute positive constants, though the actual value may vary at different occasions. For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ when $a_n \geq cb_n$ for some absolute constant c, and $a_n \lesssim b_n$ when $a_n \leq Cb_n$ for some absolute constant C. Finally, we write $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

2.1. *Lower bounds.* We first establish the minimax lower bounds which are instrumental in obtaining the optimal rates of convergence. In view of the upper bounds to be given in Section 2.2 by an aggregation procedure, these lower bounds are minimax rate optimal under mild conditions.

Before proceeding to the precise statements, we introduce the following notation: let

$$h(\lambda) = \frac{\lambda^2}{\lambda + 1}, \tag{10}$$

$$\Psi(k, p, r, n, \lambda) = \frac{1}{nh(\lambda)}\left(rk + k\log\frac{\mathrm{e}p}{k}\right) \tag{11}$$

and

$$\Psi_0(k, p, r, n, \lambda) = \frac{1}{nh(\lambda)}\left(r(k-r) + k\log\frac{\mathrm{e}p}{k}\right). \tag{12}$$

Define the *effective dimension* by

$$k_q^*(s, p, r, n, \lambda) \triangleq \lceil x_q(s, p, r, n, \lambda)\rceil, \tag{13}$$

where $\lceil a\rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$, and

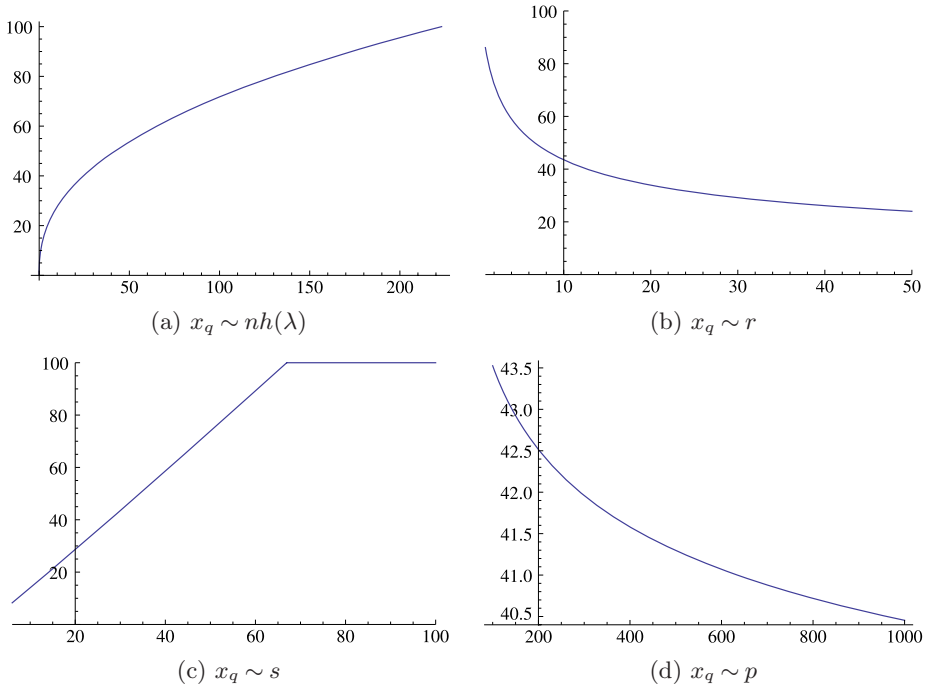
$$x_q(s, p, r, n, \lambda) \triangleq \max\left\{0 \leq x \leq p : x \leq s\left(\frac{nh(\lambda)}{r + \log(\mathrm{e}p/x)}\right)^{q/2}\right\}. \tag{14}$$

REMARK 1 (Effective dimension). The effective dimension k_q^* is a proxy which captures the massiveness of the parameter set for the principle subspace under the weak-ℓ_q constraint. In addition, the minimax estimation rate turns out to be a strictly increasing function of k_q^*. In the exact sparse case, it is evident from (13) that $k_0^* = s$. Therefore in this case, the effective dimension coincides with the row sparsity of $\mathbf{V}$. For $q \in (0, 2)$, the effective dimension satisfies the following properties (proved in Section 7.2 in the supplementary material [12]):

(1) $k_q^* \geq 1$.

(2) $k_q^* = p$ if and only if $s \geq p(\frac{r+1}{nh(\lambda)})^{q/2}$, in which case the effective dimension coincides with the ambient dimension.

(3) $s \mapsto k_q^*$ is increasing. Moreover, there exists a function τ_q, such that $k_q^*(as, p, r, n, \lambda) \leq k_q^*(s, p, r, n, \lambda)\tau_q(a)$ for any $a \geq 1$.



(a) $x_q \sim nh(\lambda)$ (b) $x_q \sim r$

(c) $x_q \sim s$ (d) $x_q \sim p$

FIG. 1. *Plots of x_q against individual parameters [default values: $p = 100$, $s = 30$, $r = 10$, $nh(\lambda) = 30$, $q = 0.8$]. The effective dimension is $k_q^* = \lceil x_q \rceil$.*

(4) $k_q^* \gtrsim s$ if and only if the assumption (16) holds.

See Figure 1 for a graphical illustration on the dependence of the effective dimension k_q^* on various parameters.

Without loss of generality, we assume unit noise variance ($\sigma^2 = 1$) from now on. All results hold for a general σ by replacing λ with λ/σ^2. We consider the lower bounds separately in two cases: $0 < q < 2$ and $q = 0$.

THEOREM 2 (Lower bound: $0 < q < 2$). *Let $p \in \mathbb{N}$, $r \in [p]$ and k_q^* be defined in (13). Let the observed matrix $\mathbf{X}$ be generated by model (1) with $\sigma = 1$. Assume that*

$$r \le \frac{s}{2} \wedge (p + 1 - k_q^*) \tag{15}$$

and that

$$nh(\lambda) \ge C_0^{2/q} \left(r + \log \frac{ep}{s} \right) \tag{16}$$

for some sufficiently large absolute constant C_0. Then there exists a constant c depending only on q and an absolute constant c_0, such that the minimax

risk for estimating $\mathbf{V}$ *over the parameter space* $\Theta = \Theta_q(s,p,r,\lambda)$ *satisfies*

$$\inf_{\widehat{\mathbf{V}}} \sup_{\boldsymbol{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \geq c\Psi(k_q^*, p, r, n, \lambda) \wedge c_0, \tag{17}$$

where Ψ *is defined in (11).*

For the case of $q = 0$ we have the following lower bound:

THEOREM 3 (Lower bound: $q = 0$). *Let* p, s, r *be integers such that* $1 \leq r \leq s \leq p$. *Let the observed matrix* $\mathbf{X}$ *be generated by model (1) with* $\sigma = 1$. *Then the minimax risk for estimating* $\mathbf{V}$ *over the parameter space* $\Theta = \Theta_0(s,p,r,\lambda)$ *satisfies*

$$\begin{aligned}&\inf_{\widehat{\mathbf{V}}} \sup_{\boldsymbol{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \\ &\quad\gtrsim \left[\frac{1}{nh(\lambda)}\left(r(s-r) + (s-r)\log\frac{\mathrm{e}(p-r)}{s-r}\right)\right] \wedge 1.\end{aligned} \tag{18}$$

2.2. *Optimal estimation via aggregation.* We now show that the lower bounds given in Section 2.1 are indeed rate optimal under mild technical conditions. The optimal estimator of $\mathbf{V}$ is constructed using sample splitting and aggregation. The estimator is theoretically interesting but computationally intensive. We will construct a data-driven and easily implementable estimator in Section 3 under stronger conditions.

We first note that the loss function (3) satisfies

$$L(\mathbf{V}, \widehat{\mathbf{V}}) = 2r - 2\|\widehat{\mathbf{V}}'\mathbf{V}\|_{\mathrm{F}}^2 = 2\|(\mathbf{I} - \mathbf{V}\mathbf{V}')\widehat{\mathbf{V}}\widehat{\mathbf{V}}'\|_{\mathrm{F}}^2. \tag{19}$$

Moreover, the loss function is invariant under orthogonal complement, that is, $L(\mathbf{V}, \widehat{\mathbf{V}}) = L(\mathbf{V}^\perp, \widehat{\mathbf{V}}^\perp)$, where $[\mathbf{V}, \mathbf{V}^\perp], [\widehat{\mathbf{V}}, \widehat{\mathbf{V}}^\perp]$ are orthogonal matrices. Therefore the loss (19) admits the following upper bound:

$$L(\mathbf{V}, \widehat{\mathbf{V}}) \leq 2(r \wedge (p-r)). \tag{20}$$

For notational simplicity we assume that the sample size is $2n$ and we split the sample equally according to $\mathbf{X} = \left[\begin{smallmatrix}\mathbf{X}_{(1)}\\ \mathbf{X}_{(2)}\end{smallmatrix}\right]$, 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 $\operatorname{supp}(\mathbf{V}) = B$. Then we aggregate these estimators by selection using the second sample.

Recall the effective dimension k_q^* defined in (13). For each $B \subset [p]$ such that $|B| = k_q^*$, 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

$$(\mathbf{J}_B)_{ii} = \mathbf{1}_{\{i\in B\}}. \tag{21}$$

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

$$B^* = \operatorname*{argmax}_{\substack{B\subset[p]\\ |B|=k_q^*}} \operatorname{Tr}(\widehat{\mathbf{V}}_B' \mathbf{S}_{(2)} \widehat{\mathbf{V}}_B) \tag{22}$$

and define the aggregated estimator by

$$\widehat{\mathbf{V}}_* = \mathbf{V}_{B^*}. \tag{23}$$

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 facilitates the analysis. Sample splitting is commonly used in aggregation based estimation, where a sequence of estimators is constructed from the first sample and the second sample is used to aggregate these candidates to produce a final estimator.

Estimator (23) requires knowledge of the value of q, the weak-ℓ_q radius s, the rank r and the spike size λ. Moreover, it can be computationally intensive for large values of p since in principle one needs to enumerate all $\binom{p}{k_q^*}$ possible column supports in order to obtain B^*. Nonetheless, the next theorem establishes its minimax rate optimality:

THEOREM 4. *Let $q \in [0,2)$. Let k_q^* be defined in (13). Let $\widehat{\mathbf{V}}_*$ be the aggregated estimator defined in (23). Assume that*

$$\lambda \geq C_0 \sqrt{\frac{\log n}{n}}, \tag{24}$$

$$nh(\lambda) \geq C_0 k_q^* \left(r + \log \frac{\mathrm{e}p}{k_q^*} \right) \tag{25}$$

and

$$n \geq C_0 \left(k_q^* \log \frac{\mathrm{e}p}{k_q^*} \vee \log \lambda \right) \tag{26}$$

for some sufficiently large constant C_0. Then there exists a constant C depending only on κ and q such that for $\Theta = \Theta_q(s,r,p,\lambda)$,

$$\sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E} \|\widehat{\mathbf{V}}_* \widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \leq 2(r \wedge (p-r)) \wedge C\Psi(k_q^*, p, r, n, \lambda), \tag{27}$$

where $\Psi(k,p,r,n,\lambda)$ and k_q^* are defined in (11) and (13), respectively. Moreover, if $q=0$, then Ψ in (27) can be replaced by Ψ_0 defined in (12) with $k_0^* = s$, and condition (25) can be dropped.

When $q \in (0,2)$, under the conditions of Theorems 2 and 4, the lower and upper bounds together yield the minimax rates of convergence $\Psi(k_q^*, p, r, n, \lambda)$ given in (11) with the optimal dependence on all the parameters, in particular the eigenvalues and the rank. When $q = 0$, the lower and upper bounds match under less restrictive conditions, which will be discussed in more detail in Remark 2 below.

2.3. *Comments.* We conclude this section with a few important remarks.

REMARK 2 (Minimax rates in the exact sparse case). Comparing the lower and upper bounds for $q = 0$ in Theorems 2 and 4, we see a sufficient condition for the minimax rate to match (and hence coincide with Ψ_0) is

$$s - r \gtrsim s \wedge \log \frac{ep}{s}. \tag{28}$$

To see this, suppose that $s - r \gtrsim s$. Then there exists $c \in (0,1)$, such that $r \le (1-c)s \le (1-c)p$. Then $(s-r)\log\frac{e(p-r)}{s-r} \ge cs \log\frac{ecp}{s} \gtrsim s\log\frac{ep}{s}$ and the lower bound in (18) agrees with Ψ_0. Now suppose that $s - r \lesssim s$ and $s - r \gtrsim \log\frac{ep}{s}$. Then $r \asymp s$ and $r(s-r) + s\log\frac{ep}{s} \asymp r(s-r) \asymp s(s-r)$. Since $r \mapsto (s-r)\log\frac{e(p-r)}{s-r}$ is decreasing on $[0,s]$, we have $r(s-r) + (s-r)\log\frac{e(p-r)}{s-r} \asymp s(s-r)$. Hence the lower bound in (18) also agrees with Ψ_0.

It is interesting to note that under the condition (28), the minimax rate for estimating the r leading singular vectors depend on the r only through $r(s-r)$, which is 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 the minimax rate to coincide with Ψ_0, it is *necessary* to have r strictly bounded away from s, for example, in the regime of (28). When $r = s$, the lower bound in Theorem 3 becomes zero. In this degenerate case, the only uncertainty is in the support of $\mathbf{V}$. The minimax rate is indeed much faster than Ψ_0, because in this regime the support can be estimated accurately. See Section 7.3 in the supplementary material [12].

REMARK 3. For $q \in (0,2)$, the minimax rate Ψ depends on the effective dimension k_q^* which is defined implicitly through equations (13)–(14). It is possible to obtain an explicit formula of the minimax rate in some regime. For example, if $s \ge p^{1-\epsilon}(\frac{r+\log p}{nh(\lambda)})^{q/2}$ for some constant $\epsilon \in (0,1)$, then the effective dimension satisfies $k_q^* \le p^{1-\epsilon}$. Moreover, we have $k_q^* \asymp s(\frac{nh(\lambda)}{r+\log p})^{q/2}$. Hence the minimax rate is given by

$$\Psi(k_q^*, p, r, n, \lambda) \asymp s\left(\frac{r+\log p}{nh(\lambda)}\right)^{1-q/2}.$$

An interesting side product of the proofs of Theorems 3 and 4 is the following nonasymptotic minimax rate for the regular PCA problem without

structural assumptions on the principle subspaces. It is a classical result (see, e.g., [21, 49]) that when $p \le n$, the sample covariance matrix is not exact minimax optimal for estimating the whole covariance matrix under certain losses (e.g., the Stein loss). As shown in the next theorem, in the unstructured case, it turns out that the sample version of the principle subspace is minimax *rate* optimal even in high dimensions. For more details see Theorems 8 and 9 in Sections 6.1 and 6.2.

THEOREM 5. *Let* $\Theta = \Theta_0(p, p, r, \lambda)$. *Let* $n \ge C_0(r + \log \lambda)$ *and* $\lambda \ge C_0\sqrt{\log(n)/n}$ *for some sufficiently large constant* C_0. *Then for all* $r \in [p]$,

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma} \in \Theta} \mathsf{E} \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \asymp r \wedge (p-r) \wedge \frac{r(p-r)}{nh(\lambda)}, \tag{29}$$

which can be attained by $\widehat{\mathbf{V}}$ *consisting of the* r *leading eigenvectors of the sample covariance matrix* $\mathbf{S}$.

Theorem 5 implies that, without structural assumptions on the principle subspace $\mathbf{V}$, consistent estimators exist if and only $\frac{nh(\lambda)}{r(p-r)} \to \infty$. Moreover, unless $nh(\lambda)$ exceeds a constant factor of p, even the optimal estimator is within a constant factor of $r \wedge (p-r)$, the upper bound of the loss function.

In the structured case, we can also investigate when regular PCA is rate optimal. It is intuitive to expect that regular PCA is strictly suboptimal if the principal eigenvectors are highly sparse, since the procedure ignores the structure of the problem. Indeed, Theorem 5, together with Theorems 2–4, reveals the precise regime where regular PCA is minimax rate optimal: under the conditions of Theorem 9, regular PCA achieves minimax rate if and only if the effective dimension $k_q^* \asymp p$. In view of definition (13), this is equivalent to that the weak-ℓ_q radius satisfies $s \gtrsim p(\frac{r}{nh(\lambda)})^{q/2}$. In the exact sparse case ($q = 0$), this condition reduces to that the sparsity $s \asymp p$.

In the special case of $r = 1$, a similar combinatorial procedure to (22)–(23) has been proposed in [54]. Using Mendelson's results on empirical processes [38], this procedure requires no sample splitting but can only be shown to attain a convergence rate that is suboptimal in λ [54], Theorem 2.2: with $\lambda \to \infty$ and all the other parameters fixed, the upper bound in [54] does not vanish. In contrast, the optimal rate Ψ decays at the rate $\lambda^{-(1-q/2)}$ when $k_q^* < p$ and λ^{-1} when $k_q^* = p$. Comparing with the analysis in [54], the proof of Theorem 4 is more elementary. By exploring the structure of the difference between the sample covariance matrix and the true covariance matrix, we obtained an upper bound that is optimal in all parameters.

3. Adaptive estimation. The aggregation estimator constructed in Section 2.2 has been shown to be rate optimal. However, it depends on the unknown parameters and is computationally infeasible when p is large. We

construct in this section an adaptive estimation procedure for principal subspaces which is fully data driven and easily computable. Furthermore, it is shown that the estimator attains the optimal rate of convergence simultaneously over a large collection of the parameter spaces defined in (6).

A key idea in our construction is a reduction scheme which reduces the sparse PCA problem to a high-dimensional multivariate regression problem. This method is potentially applicable to other sparsity patterns of the leading eigenvectors. We first introduce the general reduction scheme in Section 3.1 which transforms the principal subspace estimation problem to a high-dimensional multivariate regression problem. The specialization of this general method under weak-ℓ_q constraint will be detailed in Section 3.2.

3.1. *A general reduction scheme.* The general reduction scheme involves four steps, which are introduced in order below. The procedure used in step 2 for initial estimation will be specified in Section 3.2 for weak-ℓ_q constrained parameter spaces. For ease of exposition, we regard the rank r as given in the statement below. Data-based choice of r will be discussed at the end of Section 3.2.

Step 1: *Sample generation.* Given the data matrix $\mathbf{X}$ in (1) with $\sigma = 1$, we generate an $n \times p$ random matrix $\tilde{\mathbf{Z}}$ with i.i.d. $N(0,1)$ entries which are independent of $\mathbf{U}$ and $\mathbf{Z}$, and form two samples $\mathbf{X}^i = \mathbf{X} + (-1)^i\tilde{\mathbf{Z}}$, $i = 0, 1$. Let $\mathbf{Z}^i = \mathbf{Z} + (-1)^i\tilde{\mathbf{Z}}$ for $i = 0, 1$, then $\mathbf{Z}^0$ and $\mathbf{Z}^1$ are independent, and their entries are i.i.d. $N(0, 2)$ distributed. Then the two samples $\mathbf{X}^0$ and $\mathbf{X}^1$ can be equivalently written as

$$\mathbf{X}^i = \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{Z}^i, \qquad i = 0, 1. \tag{30}$$

Let $\mathbf{S}^i = \frac{1}{n}(\mathbf{X}^i)'\mathbf{X}^i$, $i = 0, 1$, be the sample covariance matrices for the two samples.

Step 2: *Initial estimation.* We use the sample $\mathbf{X}^0$ to compute an initial estimator $\mathbf{V}^0$. A specific procedure for computing the initial estimator $\mathbf{V}^0$ will be given in Section 3.2.

Step 3: *Reduction to regression.* Form

$$(\mathbf{X}^1)'\mathbf{X}^0\mathbf{V}^0 = \mathbf{V}\mathbf{A} + (\mathbf{Z}^1)'\mathbf{B}, \tag{31}$$

where $\mathbf{B} = \mathbf{X}^0\mathbf{V}^0$ and $\mathbf{A} = \mathbf{D}\mathbf{U}'\mathbf{B}$. We now "whiten" the matrix in (31) as follows. Note that $\mathbf{B} = \mathbf{X}^0\mathbf{V}^0$ can be explicitly computed after step 2. Let its singular value decomposition be $\mathbf{B} = \mathbf{L}\mathbf{C}\mathbf{R}'$, where $\mathbf{L} \in \mathbb{R}^{n\times r}, \mathbf{C} \in \mathbb{R}^{r\times r}$ and $\mathbf{R} \in \mathbb{R}^{p\times r}$. Post-multiply both sides of (31) by $\frac{1}{\sqrt{2}}\mathbf{R}\mathbf{C}^{-1}$ to obtain

$$\mathbf{Y} = \mathbf{\Theta} + \mathbf{E}, \tag{32}$$

where $\mathbf{Y} = \frac{1}{\sqrt{2}}(\mathbf{X}^1)'\mathbf{X}^0\mathbf{V}^0\mathbf{R}\mathbf{C}^{-1}$, $\mathbf{\Theta} = \frac{1}{\sqrt{2}}\mathbf{V}\mathbf{A}\mathbf{R}\mathbf{C}^{-1}$ and $\mathbf{E} = \frac{1}{\sqrt{2}}(\mathbf{Z}^1)'\mathbf{L}$. We shall treat (32) as a regression problem, where $\mathbf{Y}$ is the observed matrix, $\mathbf{\Theta}$ is the signal matrix of interest and $\mathbf{E}$ is the additive noise matrix. Equivalently,

we think of $\mathbf{\Theta}$ as the coefficient matrix, and the design matrix is $\mathbf{X} = \mathbf{I}_p$. The reason why this is plausible will be detailed in Section 3.2.

Given $\mathbf{Y}$, we propose the following method for computing $\widehat{\mathbf{\Theta}}$. Define

$$t_k = r + \sqrt{2r\beta \log \frac{ep}{k}} + \beta \log \frac{ep}{k}. \tag{33}$$

Fix an arbitrary $\delta \in (0,1)$. With slight abuse of notation, define

$$\operatorname{pen}(\mathbf{\Theta}) = \operatorname{pen}(|\operatorname{supp}(\mathbf{\Theta})|), \qquad \text{where } \operatorname{pen}(k) = (1+\delta)^2 \sum_{i=1}^{k} t_i. \tag{34}$$

Then the estimator for $\mathbf{\Theta}$ is defined as

$$\widehat{\mathbf{\Theta}} = \operatorname*{argmin}_{\mathbf{\Theta} \in \mathbb{R}^{p \times r}} \|\mathbf{Y} - \mathbf{\Theta}\|_{\mathrm{F}}^2 + \operatorname{pen}(\mathbf{\Theta}). \tag{35}$$

Such a penalized least squares approach has been widely used in orthogonal regression with various choices of the penalty functions. See, for example, Birgé and Massart [8] and Abramovich et al. [1].

REMARK 4. The penalized least squares estimator $\widehat{\mathbf{\Theta}}$ in (35) can be easily computed. Recall (32) and write the ith row of the matrix $\mathbf{Y}$ as $\mathbf{y}_i$ and so $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_p]'$. Let $\mathbf{y}_{(i)}$ denote the row in $\mathbf{Y}$ with the ith largest ℓ_2 norm, that is, $\|\mathbf{y}_{(1)}\| \geq \|\mathbf{y}_{(2)}\| \geq \cdots \geq \|\mathbf{y}_{(p)}\|$. Define

$$\hat{k} = \operatorname*{argmin}_{k \in [p]} \left\{ (1+\delta)^2 \sum_{i=1}^{k} t_i + \sum_{i=k+1}^{p} \|\mathbf{y}_{(i)}\|^2 \right\}. \tag{36}$$

In case of multiple minimizers, $\hat{k}$ is chosen to be the smallest one. It is also clear that $\hat{k}$ is easy to compute. With $\hat{k}$, the estimator $\widehat{\mathbf{\Theta}}$ is given by $\widehat{\mathbf{\Theta}} = [\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_p]'$ where

$$\hat{\boldsymbol{\theta}}_i = \mathbf{y}_i \cdot \mathbf{1}_{\{\|\mathbf{y}_i\|^2 > (1+\delta)^2 t_{\hat{k}}\}}.$$

Note that $\hat{k}$ can be equivalently defined as $\operatorname{argmin}_{k \in [p]} \sum_{i=1}^{k} [(1+\delta)^2 t_i - \|\mathbf{y}_{(i)}\|^2]$. Therefore $\|\mathbf{y}_{(\hat{k})}\|^2 > (1+\delta)^2 t_{\hat{k}}$ and $\|\mathbf{y}_{(\hat{k}+1)}\|^2 \leq (1+\delta)^2 t_{\hat{k}+1}$. Since t_k is strictly decreasing in k, we obtain that $\|\mathbf{y}_{(1)}\|^2 \geq \cdots \geq \|\mathbf{y}_{(\hat{k})}\|^2 > (1+\delta)^2 t_{\hat{k}} \geq \|\mathbf{y}_{(\hat{k}+1)}\|^2 \geq \cdots$. Thus, $|\operatorname{supp}(\widehat{\mathbf{\Theta}})| = \hat{k}$.

Step 4: *Final estimation*. Last but not least, we obtain the estimator $\widehat{\mathbf{V}}$ for $\mathbf{V}$ by orthonormalizing the columns of $\widehat{\mathbf{\Theta}}$. The orthonormalization can be completed by the Gram–Schmidt procedure or QR factorization. The estimated subspace is $\operatorname{span}(\widehat{\mathbf{V}}) = \operatorname{span}(\widehat{\mathbf{\Theta}})$.

An important feature of the above reduction scheme is that the two samples $\mathbf{X}^0$ and $\mathbf{X}^1$ share the *same* realization of random factors $\mathbf{U}$ and their only difference is in the noise matrices $\mathbf{Z}^0$ and $\mathbf{Z}^1$. This is critical for main-

taining the right level of signal-to-noise ratio in the regression problem (32). In contrast, splitting the original sample into two halves as in Section 2.2 does not achieve this goal here. Since our analysis relies on the independence of $\mathbf{Z}^0$ and $\mathbf{Z}^1$, the normality of the noise is crucial to this scheme.

3.2. *Sparse PCA and regression with group sparsity.* We now apply the general reduction scheme to the principal subspace estimation problem with the parameter spaces defined in (6). In what follows, we first introduce and study a specific estimator for the initial estimation step. Then we derive properties of the proposed estimator for the regression with group sparsity problem. Furthermore, we show that the general reduction scheme paired with the two specific estimators leads to a final estimator which adaptively achieves the optimal rates of estimation over a large collection of the parameter spaces of interest. For clarity of exposition, we regard the rank r as given when introducing the estimators. Data-driven choice of r is discussed at the end of this subsection.

Initial estimation. Let $p_n \triangleq p \vee n$. We construct the initial estimator $\mathbf{V}^0$ via the diagonal thresholding method [26] as follows:

(1) Define the set of features

$$J = \{j : s^0_{jj} \geq 2(1 + \alpha\sqrt{\log p_n / n})\}, \tag{37}$$

where $\{s^0_{jj}\}_{j=1}^p$ are the diagonal elements of $\mathbf{S}^0 = \frac{1}{n}(\mathbf{X}^0)'\mathbf{X}^0$, and $\alpha > 0$ is a tuning parameter.

(2) Compute the first r eigenvectors $\{\hat{\mathbf{v}}_1^J, \ldots, \hat{\mathbf{v}}_r^J\}$ of the submatrix $\mathbf{S}^0_{JJ}$.

(3) Define $\mathbf{V}^0 \in O(p, r)$, where

$$\mathbf{V}^0_{J*} = [\hat{\mathbf{v}}_1^J, \ldots, \hat{\mathbf{v}}_r^J], \qquad \mathbf{V}^0_{J^c *} = \mathbf{0}. \tag{38}$$

The following result, proved in Section 7.5 in the supplementary material [12], gives sufficient conditions on the model parameters and the choice of α to guarantee that the initial estimator $\mathbf{V}^0$ is reasonably close to $\mathbf{V}$, which suffices for the initialization of our scheme.

PROPOSITION 1. *Suppose that* $\log n \geq M_0 \log \lambda$ *for some constant* $M_0 > 0$. *Suppose that*

$$n(\lambda^2 \wedge 1) \geq C_0(r + \log p)^2 / \log p \tag{39}$$

and

$$s^2 \left(\frac{\log(p \vee n)}{n\lambda^2} \right)^{1-q/2} < \kappa^{-2}(2-q)^q / C_0 \tag{40}$$

for a sufficiently large constant $C_0 > 0$. *If* $\mathbf{V}^0$ *is defined in (38) with a sufficiently large* $\alpha \geq \sqrt{10(1 + 1/M_0)}$ *in (37), then uniformly over* $\Theta = \Theta_q(s, p,$

r, λ), we have

$$|\mathrm{supp}(\mathbf{V}^0)| \le k_q^* \quad \textit{and} \quad \sigma_r(\mathbf{V}'\mathbf{V}^0) \ge 1/2 \tag{41}$$

hold with probability at least $1 - C/[nh(\lambda)]$, where k_q^* is defined in (13).

We note that condition (40) is critical in establishing the second claim in (41), which ensures that $\mathbf{V}^0$ is a reasonable estimator of $\mathbf{V}$. Such a condition is needed for diagonal thresholding to work even when $r = 1$. See, for example, condition C3 in [42], page 95. Theorem 4.1 of [9] showed that diagonal thresholding could be suboptimal even under a stronger condition than (40).

REMARK 5. When M_0 in Proposition 1 is unknown, we replace it by

$$\widehat{M_0} = \log n / \log(\sigma_1(\mathbf{S}^0) - 2), \tag{42}$$

where $\sigma_1(\mathbf{S}^0)$ is the largest eigenvalue of $\mathbf{S}^0$. This estimate works because $\sigma_1(\mathbf{S}^0) - 2$ is an over-estimate of λ with high probability [39, 43], since the noise variance here is two. The estimator (42) allows us to choose α in (37) without explicit knowledge of M_0.

Orthogonal regression with group sparsity. We first explain why we can treat (32) as a regression problem. When we condition on the values of $\mathbf{U}$ and $\mathbf{Z}^0$, the matrix $\mathbf{X}^0$ becomes deterministic. Thus, as deterministic functions of $\mathbf{X}^0$, the matrices $\mathbf{V}^0, \mathbf{B}, \mathbf{L}, \mathbf{C}$ and $\mathbf{R}$ are also deterministic. Furthermore, $\mathbf{A}$ and hence $\mathbf{\Theta}$, as deterministic functions of $\mathbf{U}$ and $\mathbf{B}$, are also deterministic. On the other hand, $\mathbf{Z}^1$ is independent of both $\mathbf{U}$ and $\mathbf{Z}^0$ and hence is independent of $\mathbf{X}^0$, $\mathbf{B}$ and $\mathbf{L}$. Thus, the conditional distribution of $\mathbf{Z}^1$ on $(\mathbf{U}, \mathbf{Z}^0)$ always has i.i.d. $N(0, 2)$ entries, and so the conditional distribution of $\mathbf{E}$ has i.i.d. standard normal entries. Therefore, when we condition on the values of $\mathbf{U}$ and $\mathbf{Z}^0$, problem (32) indeed reduces to a standard multivariate regression problem with *orthogonal design* and *white noise*.

When the sparsity of $\mathbf{V}$ is specified as in (6), we need to consider the following parameter space for $\mathbf{\Theta}$:

$$\mathcal{F}_q(s', p) = \{\mathbf{\Theta} : \|\mathbf{\Theta}\|_{q,w} \le s'\} \tag{43}$$

with $q \in [0, 2)$. The parameter s' is typically different from s in (6), as it also depends on the other model parameters as well as the realization of $\mathbf{U}$ and $\mathbf{Z}^0$. However, this will not cause any difficulty in practice, because the estimator proposed in (35) and the associated theorem below remain valid for all values of $s' > 0$. In the literature of high-dimensional regression, (43) is usually referred to as the group sparsity constraint on the regression coefficients $\mathbf{\Theta}$.

For the estimator $\widehat{\mathbf{\Theta}}$ in (35), we have following upper bound on its risk. By the lower bounds in [36] for $q = 0$, the rates in Theorem 6 are optimal.

THEOREM 6. *Consider the regression problem*

$$\mathbf{Y} = \mathbf{\Theta} + \mathbf{E},$$

where $\mathbf{\Theta} \in \mathbb{R}^{p\times r}$ *is deterministic and* $\mathbf{E}$ *has i.i.d.* $N(0,1)$ *entries. Let the parameter space* $\mathcal{F}_q(s',p)$ *be defined in (43) for some* $q \in [0,2)$ *and* $s' > 0$. *If* $\beta > 2$ *in (33) and* $\delta \in (0,1)$ *in (34), then there is a positive constant* C *that depends only on* q, β *and* δ, *such that the estimator in (35) satisfies*

$$\sup_{\mathbf{\Theta}\in\mathcal{F}_q(s',p)} \mathsf{E}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\mathrm{F}}^2 \le Ck'\left(r + \log\frac{ep}{k'}\right),$$

where

$$k' \triangleq \min\{k \in [p] : t_k^{q/2} k \ge s'\} \tag{44}$$

for t_k *defined in (33), and if the set in (44) is empty, we set* $k' = p$.

Adaptation. With the above preparation, we are now ready to show that if we start with a proper initial estimator $\mathbf{V}^0$ [such as that in (38)] and estimate $\mathbf{\Theta}$ by (35), then the estimator $\widehat{\mathbf{V}}$ resulting from orthonormalizing the columns of $\widehat{\mathbf{\Theta}}$ achieves the optimal rates of convergence. We state the theorem in a slightly more general format. In particular, it holds for the initial estimator in (38) under the conditions of Proposition 1.

THEOREM 7 (Adaptation). *Let* $\lambda \ge C_0$ *for some sufficiently large constant* C_0. *Let* $\Theta = \Theta_q(s,p,r,\lambda)$ *satisfy the conditions in Theorem 4. Suppose that there exists an initial estimator* $\mathbf{V}^0$ *which satisfies (41) with probability at least* $1 - C'/(nh(\lambda))$. *Then the estimator* $\widehat{\mathbf{V}}$ *obtained by orthonormalizing* $\widehat{\mathbf{\Theta}}$ *in (35) with* $\beta > 2$ *in (33) and* $\delta \in (0,1)$ *in (34) satisfies*

$$\sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \le 2(r \wedge (p-r)) \wedge C\Psi(k_q^*, p, r, n, \lambda),$$

where k_q^* *is defined in (13), and* $C > 0$ *is a constant depending only on* q, β *and* δ.

We note that the assumption $\lambda > C_0$ is imposed to ensure that the "whitening" procedure in step 3 of the reduction scheme can be performed.

It is interesting to compare the statement of Theorem 7 to the minimax lower bound in Theorems 2–3 as well as the performance of the combinatorial aggregation estimator $\widehat{\mathbf{V}}^*$ established in Theorem 4. For any parameter space $\Theta = \Theta_q(s,p,r,\lambda)$ such that the conditions of Proposition 1 hold, we could use the $\mathbf{V}^0$ in (38), and the resulting $\widehat{\mathbf{V}}$ is guaranteed to achieve the optimal rates of convergence on Θ, which matches the performance of the aggregation estimator for any $q > 0$. Moreover, in this case both $\mathbf{V}^0$ and $\widehat{\mathbf{V}}$ can be efficiently computed. Hence $\widehat{\mathbf{V}}$ can be used in practice while $\widehat{\mathbf{V}}^*$ is computationally intensive. However, in the exact sparse case of $q = 0$, the

upper bound in Theorem 7 depends on the rank r linearly through sr, while the true minimax rate in Theorem 3 depends on r quadratically through $r(s-r)$, which is smaller than rs if $s-r$ is small. The suboptimality of $\widehat{\mathbf{V}}$ in this specific regime is partially due to the fact that our reduction scheme transforms the problem into a regression problem without taking account of the orthogonality structure of the parameter space.

REMARK 6. Theorem 7 shows that any estimator $\mathbf{V}^0$ satisfying (41) can be used to produce an adaptive estimator. So the task of constructing adaptive optimal estimators is reduced to constructing a "reasonable" estimator.

Consistent estimator of r. Last but not least, we discuss how to construct a consistent estimator of r based on data. To this end, recall the definition of the set J in (37), and the matrix $\mathbf{S}^0_{JJ}$. We propose to estimate r by

$$\hat{r} = \max\{l : \sigma_l(\mathbf{S}^0_{JJ}) > 2(1+\delta_{|J|})\}, \tag{45}$$

where for any $m > 0$ and M_0 in the conditions of Proposition 1, we define

$$\delta_m = 2(\sqrt{m/n} + t_m) + (\sqrt{m/n} + t_m)^2$$

with $t_m^2 = \frac{2}{n}((m+1)\log(ep) + (1+2/M_0)\log n)$. Here, we regard M_0 as known. Otherwise, we could always replace it with the estimator (42) proposed in Remark 5. Note that the estimator (45) could be easily integrated with the diagonal thresholding method for computing $\mathbf{V}^0$. In particular, $\hat{r}$ can be computed after we select the set J in (37).

For this estimator, we have the following result.

PROPOSITION 2. *Under the condition of Proposition 1, $\hat{r} = r$ holds with probability at least $1 - C[nh(\lambda)]^{-1}$.*

Under the conditions of Proposition 1 and Theorem 7, Proposition 2 implies that the conclusion in Theorem 7 still holds if we replace r by $\hat{r}$.

4. Numerical experiments. In this section, we report simulation results comparing the adaptive method proposed in Section 3 with the iterative thresholding method proposed in Ma [37].

In all the results reported here, the sample size $n = 1000$ and the ambient dimension $p = 2000$. We focus on the case of exact sparsity, that is, $q = 0$. The sparsity parameter s takes value in $\{40, 80, 120, 160, 200\}$, and the rank r takes value in $\{1, 5, 10, 20\}$. For each (s, r) combination, the $\mathbf{V}$ matrix is obtained from orthonormalizing an $p \times r$ matrix $\mathbf{M}$ where $\mathbf{M}_{i*}$ have i.i.d. $N(0, i^4)$ entries for $i = 1, \ldots, s$ and $\mathbf{M}_{i*} = 0$ for all $i > s$. We set the variances of different rows to be different so that the ordered norms of the nonzero

TABLE 1
Average loss $\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2$ *over* 50 *repetitions for each* (s, r) *combinations*

		s				
r	**Method**	**40**	**80**	**120**	**160**	**200**
1	RegSPCA	0.0236	0.0660	0.0892	0.1074	0.1754
	ITSPCA	0.0117	0.0366	0.0483	0.0619	0.0712
5	RegSPCA	0.0348	0.0718	0.1134	0.1470	0.1992
	ITSPCA	0.0520	0.1209	0.1848	0.2368	0.3042
10	RegSPCA	0.0544	0.1247	0.1777	0.2394	0.3052
	ITSPCA	0.0914	0.2284	0.3535	0.4866	0.6313
20	RegSPCA	0.0640	0.1826	0.2904	0.4030	0.5083
	ITSPCA	0.1185	0.3740	0.6449	0.9045	1.1715

rows in $\mathbf{V}$ also exhibit fast decay. When $r = 1$, the spike size $\lambda_1 = 20$. When $r > 1$, the λ_i's take r equispaced values such that $\lambda_r = 10$ and $\lambda_1 = 20$.

When implementing the method in Section 3, we take $\alpha = 3$ in (37), $\beta = 2.1$ in (33) and $\delta = 0.05$ in (34) in all the simulations reported here. In addition, we made a slight modification to the proposed method to obtain better numerical results. We first run the method to obtain an estimator, denoted by $\widehat{\mathbf{V}}_1$. Then we switch the roles of $\mathbf{X}^0$ and $\mathbf{X}^1$ and run the proposed procedure again to obtain a second estimator $\widehat{\mathbf{V}}_2$. Finally, we use the r leading eigenvectors of $\widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1' + \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2'$ as the columns of the final estimator $\widehat{\mathbf{V}}$. By Theorem 10 in Section 7.11 in the supplementary material [12], we have

$$\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}\|_{\mathrm{F}} \le \|\widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1' + \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2' - 2\mathbf{V}\mathbf{V}'\|_{\mathrm{F}} \le \sum_{i=1,2} \|\widehat{\mathbf{V}}_i\widehat{\mathbf{V}}_i' - \mathbf{V}\mathbf{V}\|_{\mathrm{F}}.$$

Here, the first inequality holds because $\sigma_r(\widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1' + \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2') \ge \sigma_r(\widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1') = 1$ and $\sigma_{r+1}(2\mathbf{V}\mathbf{V}') = 0$, while the second is by the triangle inequality. By the last display, the theoretical results in Section 3, which apply to both $\widehat{\mathbf{V}}_1$ and $\widehat{\mathbf{V}}_2$, also apply to the final estimator $\widehat{\mathbf{V}}$. When implementing the iterative thresholding method in Ma [37], we set all tuning parameters at their recommended values.

Table 1 summarizes the average squared Frobenius losses of the proposed method (RegSPCA) and the iterative thresholding method (ITSPCA) over 50 repetitions for each (s, r) combination. Table 1 shows that for all values of the sparsity parameter, RegSPCA outperformed ITSPCA when $r = 5, 10$ or 20, while ITSPCA led to smaller average losses when $r = 1$. This demonstrates the competitiveness of RegSPCA in the group sparse setting considered in the present paper. On the other hand, we note that ITSPCA was not designed specifically for handling the group sparsity structure which is the case when $r > 1$, and hence its underperformance is not unexpected.

5. Discussions. We have focused in the present paper on the estimation of the principal subspace span($\mathbf{V}$) under the loss (3). The minimax rates of convergence are established and a computationally efficient adaptive estimator is constructed.

Both the current paper and Ma [37] consider the problem of sparse subspace estimation under the spiked model, but they differ in several important ways. First, in addition to the sparsity constraint on the leading eigenvectors, the current paper requires them to share support. This extra assumption is motivated by real data applications. For instance, if the observed vectors are the leading Fourier coefficients of random functions with a common covariance kernel, then we expect the leading eigenvectors to have large coefficients only at low frequency coordinates so that the resulting leading eigen-functions in the time domain are smooth. Second, Ma [37] focused on the error upper bounds of an adaptive estimator with the subspace rank r assumed to be a fixed constant. Whether the dependence of the bounds on r is optimal was not studied. The current paper conducts an investigation on the dependence of the minimax rates on key model parameters, including r which can grow with n and p. Last but not least, we have focused exclusively on the subspace span($\mathbf{V}$) which is natural when the spikes are of the same order, while Ma [37] considered estimating subspaces spanned by the first few rather than all columns of $\mathbf{V}$. The optimal rates of the latter estimation problem is of most interest when the spikes scale at different rates with n and p, which we leave as an interesting problem for future research.

A problem closely related to principal subspace estimation is the estimation of the whole covariance matrix $\mathbf{\Sigma}$ under the same structural assumption (6). Both minimax estimation and adaptive estimation are of significant interest. Results on minimax rates under the spectral norm loss $L(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2$ can be found in [11].

It is interesting to extend the aggregation method in Section 2.2 to other settings beyond sparsity or weak ℓ_q constraints. In the exact sparse case ($q = 0$), note that the rate-optimal estimator in (27) is constructed by choosing the best estimator from a collection of estimators, each of which is designed for a specific sparsity pattern. Theorem 4 can now be interpreted as an oracle inequality for the average risk, which is within a constant factor of the oracle risk $\frac{r(k-r)}{nh(\lambda)}$ plus the excess risk $\frac{1}{nh(\lambda)} \log \binom{p}{k}$. One immediate generalization of Theorem 4 is that we can also construct aggregated estimators if it is known that the true principle subspace belongs to a collection of N subspaces. Then the excess risk does not exceed $\frac{1}{nh(\lambda)} \log N$.

It should be noted that our analysis in this paper relies on the normality of the model, which allows us to express the sample in the form of (1). In particular the adaptive procedure requires the independence of $\mathbf{Z}^0$ and $\mathbf{Z}^1$, which is a consequence of the normality of the noise. It is unclear whether

the same results hold for all noise distributions with sub-Gaussian tails. It is an interesting problem to study the robustness of the adaptive procedure and to extend the results to other noise distributions.

6. Proofs. In this section we prove Theorems 3, 4 and 7. The proofs of the other results, together with those of the key lemmas and some additional technical arguments, are given in the supplementary material [12].

6.1. *Proof of Theorem 3.* We first give a lower bound on the oracle risk where we know beforehand the row support of $\mathbf{V}$. This corresponds to a k-dimensional unstructured PCA problem, where the goal is to estimate the r leading singular vectors of the covariance matrix. In view of the upper bound in Theorem 9, the rates are minimax optimal.

THEOREM 8 (Oracle risk: lower bound). *Let $\Theta = \Theta_0(k, k, r, \lambda)$. Then*

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma} \in \Theta} \mathsf{E} \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \gtrsim r \wedge (k-r) \wedge \frac{r(k-r)}{nh(\lambda)}. \tag{46}$$

To prove Theorem 8, we use a minimax lower bound due to Yang and Barron [58], Section 7, via *local* metric entropy, which in turn relies on an argument by Birgé [7]. For completeness, we state the result in Proposition 3 and provide a short proof in Section 7.8 in the supplementary material [12]. The method of local metric entropy in an $\frac{1}{\sqrt{n}}$-neighborhood dates back to Le Cam [34]. The advantage of this method is that it only relies on the analytical behavior of the metric entropy of the parameter space, thus allowing us to sidestep constructing explicit packing set in the parameter space.

PROPOSITION 3. *Let (Θ, ρ) be a totally bounded metric space and $\{P_\theta : \theta \in \Theta\}$ a collection of probability measures. For any $E \subset \Theta$, denote by $\mathcal{N}(E, \epsilon)$ the ϵ-covering number of E, that is, the minimal number of balls of radius ϵ whose union contains E. Denote by $\mathcal{M}(E, \epsilon)$ the ϵ-packing number of E, that is, the maximal number of points in E whose pairwise distance is at least ϵ. Put*

$$A \triangleq \sup_{\theta \neq \theta'} \frac{D(P_\theta || P_{\theta'})}{\rho^2(\theta, \theta')}. \tag{47}$$

If there exist $0 < c_0 < c_1 < \infty$ and $d \geq 1$ such that

$$\left(\frac{c_0}{\epsilon}\right)^d \leq \mathcal{N}(\Theta, \epsilon) \leq \left(\frac{c_1}{\epsilon}\right)^d \tag{48}$$

for all $0 < \epsilon < \epsilon_0$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathsf{E}_\theta[\rho^2(\hat{\theta}(X), \theta)] \geq \frac{c_0^2}{840 c_1^2} \left(\frac{d}{A} \wedge \epsilon_0^2\right). \tag{49}$$

We also need the following result regarding the metric entropy of the Grassmannian manifold $G(k,r)$ due to Szarek [50].

LEMMA 1. *For any $\mathbf{V} \in O(k,r)$, identifying the subspace* $\operatorname{span}(\mathbf{V})$ *with its projection matrix $\mathbf{V}\mathbf{V}'$, define the metric on $G(k,r)$ by $\rho(\mathbf{V}\mathbf{V}', \mathbf{U}\mathbf{U}') = \|\mathbf{V}\mathbf{V}' - \mathbf{U}\mathbf{U}'\|_{\mathrm{F}}$. Then for any $\epsilon \in (0, \sqrt{2(r \wedge (k-r))}]$,*

$$\left(\frac{c_0}{\epsilon}\right)^{r(k-r)} \le \mathcal{N}(G(k,r), \epsilon) \le \left(\frac{c_1}{\epsilon}\right)^{r(k-r)}, \tag{50}$$

where c_0, c_1 are absolute constants. Moreover, for any $\mathbf{V} \in O(k,r)$ and any $\alpha \in (0,1)$,

$$\mathcal{M}(B(\mathbf{V}, \epsilon), \alpha\epsilon) \ge \left(\frac{c_0}{\alpha c_1}\right)^{r(k-r)}. \tag{51}$$

PROOF. Note that $\rho(\mathbf{V}\mathbf{V}', \mathbf{U}\mathbf{U}') = \sqrt{2}\|(\mathbf{I} - \mathbf{V}\mathbf{V}')\mathbf{U}\mathbf{U}'\|_{\mathrm{F}}$, in view of (19). This metric is unitarily invariant; see ρ'_α in [50], Remark 5, page 175. Applying [50], Proposition 8, page 169, with $\alpha(\cdot) = \|\cdot\|$ gives (50). By the proof of equation (158) in the supplementary material [12] for any $\epsilon \in (0, \sqrt{2(r \wedge (k-r))}]$ and any $\alpha \in (0,1)$, there exists $\mathbf{V}^* \in O(k,r)$ such that $\mathcal{M}(B(\mathbf{V}^*, \epsilon), \alpha\epsilon) \ge (\frac{c_0}{\alpha c_1})^{r(k-r)}$. Now for any $\mathbf{V} \in O(k,r)$, there exists $\mathbf{T} \in O(p)$, such that $\mathbf{V} = \mathbf{T}\mathbf{V}^*$. Then (51) holds since the metric d is unitarily invariant. □

PROOF OF THEOREM 8. For the purpose of lower bound, we consider the special case of $\lambda_1 = \cdots = \lambda_r = \lambda$, that is, $\mathbf{\Sigma} = \lambda\mathbf{V}\mathbf{V}' + \mathbf{I}_k$. Note that the Kullback–Leibler divergence between normal distributions is given by $D(N(0, \mathbf{\Sigma}_1) || N(0, \mathbf{\Sigma}_0)) = \frac{1}{2}(\operatorname{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_1 - \mathbf{I}_k) - \log\det \mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_1)$. Then for any $\mathbf{U}, \mathbf{V} \in O(k,r)$, we have

$$\begin{aligned} &D(N(0, \lambda\mathbf{V}\mathbf{V}' + \mathbf{I}_k)^n || N(0, \lambda\mathbf{U}\mathbf{U}' + \mathbf{I}_k)^n) \\ &\quad = \frac{n}{2}\operatorname{Tr}\left(-\frac{\lambda}{\lambda+1}\mathbf{V}\mathbf{V}' + \lambda\mathbf{U}\mathbf{U}' - \frac{\lambda^2}{\lambda+1}\mathbf{V}\mathbf{V}'\mathbf{U}\mathbf{U}'\right) \\ &\quad = \frac{n\lambda^2}{2(\lambda+1)}(r - \|\mathbf{U}'\mathbf{V}\|_{\mathrm{F}}^2) = \frac{nh(\lambda)}{2}\|\mathbf{V}\mathbf{V}' - \mathbf{U}\mathbf{U}'\|_{\mathrm{F}}^2, \end{aligned} \tag{52}$$

where the first and second inequalities are by the matrix inversion lemma and the fact that $\operatorname{Tr}(\mathbf{V}\mathbf{V}') = \operatorname{Tr}(\mathbf{V}'\mathbf{V}) = r$, respectively. In view of (47), we have $A = nh(\lambda)/2$. Applying Proposition 3 with $\epsilon_0 = \sqrt{2(r \wedge (k-r))}$ yields the desired (46). □

PROOF OF THEOREM 3. Let $\Theta = \Theta_0(s,p,r,\lambda)$. By definition (13), k_0^* coincides with s. In view of the fact that $(a \wedge b) + (c \wedge d) \ge (a \wedge c)(b+d)$, it

is sufficient to prove the following inequalities separately:

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \gtrsim r \wedge (s-r) \wedge \frac{r(s-r)}{nh(\lambda)}, \tag{53}$$

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \gtrsim 1 \wedge \frac{s-r}{nh(\lambda)} \log \frac{\mathrm{e}(p-r)}{s-r}. \tag{54}$$

Inequality (53) follows from an oracle argument: consider the following sub-collection:

$$\left\{ \mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{0} \end{bmatrix} : \mathbf{V}_1 \in O(s,r) \right\}.$$

Split the data matrix according to $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ consists of the first s columns. Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_s)$. Then the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d. according to $\mathcal{N}(0, \mathbf{V}_1\mathbf{\Lambda}\mathbf{V}_1' + \mathbf{I}_s)$ and $N(0, \mathbf{I}_{p-s})$, respectively. Therefore a sufficient statistic for estimating $\mathbf{V}$ is $\mathbf{X}_1$. This reduces the problem to an s-dimensional unconstrained PCA problem. Applying the lower bound in Theorem 8 yields (53).

Inequality (54) follows from existing results on rank-one estimation (e.g., [9, 54]). To make the argument rigorous, we focus on the special case where $\{\mathbf{v}_2, \ldots, \mathbf{v}_r\}$ are fixed to be standard basis. Denote the following sub-collection:

$$\left\{ \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r-1} \end{bmatrix} : \mathbf{v}_1 \in \mathbb{S}^{p-r}, |\operatorname{supp}(\mathbf{v}_1)| \le s-r+1 \right\}, \tag{55}$$

which is well defined since $s \le p$ by definition. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ denotes the first $p-r+1$ columns of $\mathbf{X}$. Then $\mathbf{X}_1$ and $\mathbf{X}_2$ consists of n independent samples from $N(0, \mathbf{I}_{p-r+1} + \lambda \mathbf{v}_1\mathbf{v}_1')$ and $N(0, \mathbf{I}_{r-1})$, respectively. Restricted on the subset (55), the minimax estimation error of $\mathbf{V}$ is equal to the minimax estimation error of $\mathbf{v}_1$ based on $\mathbf{X}_1$. This is equivalent to replacing the ambient dimension p by $p-r+1$ and estimating only the leading singular vector $\mathbf{v}_1$, which is $(s-r+1)$-sparse, under the loss $\|\mathbf{v}_1\mathbf{v}_1' - \hat{\mathbf{v}}_1\hat{\mathbf{v}}_1'\|_{\mathrm{F}}^2$. Applying the minimax lower bound from [54], Theorem 2.1[4] (see also [9], Theorem 2), we have

$$\begin{aligned} \inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 &\ge \inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|^2 \\ &\gtrsim 1 \wedge \frac{(s-r)}{nh(\lambda)} \log \frac{\mathrm{e}(p-r)}{s-r}, \end{aligned} \tag{56}$$

completing the proof of Theorem 3. □

[4]Note that [54], Theorem 2.1, for $q=0$ only applies to the regime where $s-r \le (p-r)/\mathrm{e}$. This does not affect the *rate* of the lower bound (56) because the minimax rate is a nondecreasing function of the sparsity s. Therefore if $s-r > (p-r)/\mathrm{e}$, we can use the lower bound for $s-r = (p-r)/\mathrm{e}$ to obtain (56), since $s-r \le p-r$ by definition.

6.2. *Proof of Theorem 4.* We first state a few technical lemmas (proved in Section 7.10 in the supplementary material [12]) and an oracle upper bound (proved in Section 7.9 in the supplementary material [12]), which, in view of the lower bound in Theorem 8, gives the optimal rates of the regular PCA problem. Some of the proofs are relegated to the supplementary material [12].

LEMMA 2. *Let $a,b,c>0$. Then $ax^2 \le bx+c$ implies that $x^2 \le \frac{b^2}{a^2}+\frac{2c}{a}$.*

PROOF. Since $|x-\frac{b}{2a}| \le \frac{\sqrt{b^2+4ac}}{2a}$, we have $x^2 \le \frac{b^2+b^2+4ac}{2a^2}$. □

LEMMA 3. *Let $\mathbf{\Sigma} = \mathbf{I}_p + \mathbf{VDV}'$. For any $\mathbf{T} \in O(p,r)$, we have*

$$\frac{\lambda_r}{2}\|\mathbf{VV}'-\mathbf{TT}'\|_{\mathrm{F}}^2 \le \langle \mathbf{\Sigma}, \mathbf{VV}'-\mathbf{TT}'\rangle \le \frac{\lambda_1}{2}\|\mathbf{VV}'-\mathbf{TT}'\|_{\mathrm{F}}^2. \tag{57}$$

LEMMA 4. *Let $\mathbf{K} \in \mathbb{R}^{p\times p}$ be symmetric such that $\operatorname{Tr}(\mathbf{K})=0$ and $\|\mathbf{K}\|_{\mathrm{F}}=1$. Let $\mathbf{Z}$ be $n\times p$ consisting of independent standard normal entries. Then for any $t>0$, we have*

$$\mathsf{P}\left(\frac{1}{\sqrt{n}}|\langle \mathbf{Z}'\mathbf{Z}, \mathbf{K}\rangle| \ge 2t + \frac{2t^2}{\sqrt{n}}\right) \le 2\exp(-t^2). \tag{58}$$

LEMMA 5. *Let $X_1,\dots,X_N$ be i.i.d. such that*

$$\mathsf{P}(|X_1| \ge at+bt^2) \le c\exp(-t^2), \tag{59}$$

where $a,b,c>0$. Then

$$\mathsf{E}\max_{i\in[N]}|X_i|^2 \le (2a^2+8b^2)\log(ecN) + 2b^2\log^2(cN). \tag{60}$$

LEMMA 6. *Let $\mathbf{E}$ be a symmetric positive definite matrix. Let $\mathbf{F}$ be a symmetric matrix. Then $|\langle \mathbf{E},\mathbf{F}\rangle| \le \|\mathbf{F}\|\operatorname{Tr}(\mathbf{E})$.*

PROOF. This is a special case of von Neumann's trace inequality. □

LEMMA 7. *Let $\mathbf{\Theta} \in \mathcal{F}_q(s,p)$ and $k\in[p]$, where $\mathcal{F}_q(s,p)$ is defined in (43). Let $\|\mathbf{\Theta}_{(i)*}\|$ denote its ith largest row norm. Then*

$$\sum_{i>k}\|\mathbf{\Theta}_{(i)*}\|^2 \le \frac{q}{2-q}k(s/k)^{2/q}. \tag{61}$$

PROOF. By the definition of $\mathcal{F}_q(s,p)$ in (43), we have

$$\sum_{i>k}\|\mathbf{\Theta}_{(i)*}\|^2 \le s^{q/2}\sum_{i>k} i^{-2/q} \le s^{q/2}\int_k^\infty x^{-2/q}\,\mathrm{d}x = \frac{q}{2-q}k(s/k)^{2/q}.$$ □

THEOREM 9 (Oracle risk: upper bound). *Let $p=k$ and $r\in[k]$. Let $n \ge C_0(r+\log\lambda)$ and $\lambda \ge C_0\sqrt{\log(n)/n}$ for some sufficiently large constant C_0.*

Let $\widehat{\mathbf{V}} \in O(k,r)$ *be formed by the* r *leading singular vectors of the sample covariance matrix* $\mathbf{S}$. *Let* $\Theta = \Theta_0(k,k,r,\lambda,\kappa)$. *Then*

$$\sup_{\mathbf{\Sigma}\in\Theta} \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \lesssim r \wedge (k-r) \wedge \frac{r(k-r)}{nh(\lambda)}. \tag{62}$$

Proof of Theorem 4. Before delving into the details, we give an outline of the proof as follows:

(1) We find a good sparse approximation of the true singular vectors which lies in the weak-ℓ_q ball defined by (43).

(2) We decompose the risk into a summation of three terms, namely the *approximation error*, *oracle risk* and *excess risk*, the first two of which are upper bounded in Lemma 7 and Theorem 9, respectively.

(3) The excess risk is controlled by a careful concentration-of-measure analysis, which forms the core of the proof.

We also remark that by (8), (13) and condition (25), we have

$$k_q^* \geq r. \tag{63}$$

To see this, first note that $k_0^* \geq r$ by (8) directly. When $q \in (0,2)$, if $k_q^* = p$, then $k_q^* \geq r$. Otherwise, we have

$$k_q^* \geq s\left(\frac{nh(\lambda)}{r + \log(ep/k_q^*)}\right)^{q/2} \geq s(C_0 k_q^*)^{q/2} \geq C_0^{q/2} s \geq r.$$

Here the first inequality comes from (13), the second is due to condition (25), the third holds since $k_q^* \geq 1$ and the last holds for sufficiently large C_0 in view of (8).

Step 1: *Sparse approximation* . Fix $\mathbf{V} \in O(p,r) \cap \mathcal{F}_q(s,p)$. We assume that $q > 0$. Note that this step is superfluous if $q = 0$ since $\mathbf{V}$ is already sparse. Let $k = k_q^*$ be defined in (13). Let $\mathcal{B}(k) = \{B \subset [p] : |B| = k\}$. Let $A \in \mathcal{B}(k)$ denote the collection of row indices of $\mathbf{V}$ corresponding to the k largest row norm. Put

$$\tilde{\mathbf{\Sigma}} = \mathbf{J}_A \mathbf{\Sigma} \mathbf{J}_A + \mathbf{J}_{A^{\mathrm{c}}} = \mathbf{J}_A \mathbf{V} \mathbf{\Lambda} \mathbf{V}' \mathbf{J}_A + \mathbf{I}_p, \tag{64}$$

where $\mathbf{J}_A$ is the diagonal matrix defined in (21). Denote the SVD of $\mathbf{J}_A \mathbf{V} \mathbf{\Lambda} \mathbf{V}' \mathbf{J}_A$ by $\tilde{\mathbf{V}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{V}}'$, where $\tilde{\mathbf{\Lambda}} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r, 0, \ldots, 0)$ and $\tilde{\mathbf{V}} \in O(p,r) \cap \mathcal{F}_0(s,p)$, since $\operatorname{supp}(\tilde{\mathbf{V}}) = A$. Now we claim that $\tilde{\mathbf{V}}$ is in fact the r leading singular vectors of $\tilde{\mathbf{\Sigma}}$. To this end, note that the singular values of $\tilde{\mathbf{\Sigma}}$ are $\{1 + \tilde{\lambda}_1, \ldots, 1 + \tilde{\lambda}_r, 1\}$. In view of (64), it is sufficient to show that the rth largest singular value of $\tilde{\mathbf{\Sigma}}$ is separated from one, that is, $\sigma_r(\tilde{\mathbf{\Sigma}}) > 1$. By Weyl's theorem ([22], Theorem 4.3.1),

$$\sigma_r(\tilde{\mathbf{\Sigma}}) \geq \sigma_r(\mathbf{\Sigma}) - \|\mathbf{\Sigma} - \tilde{\mathbf{\Sigma}}\| \geq 1 + \lambda_r - \|\mathbf{\Sigma} - \tilde{\mathbf{\Sigma}}\|_{\mathrm{F}}.$$

Put $\mathbf{U} = \mathbf{J}_A \mathbf{V}$. Then

$$\begin{aligned}
\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{F}} &= \|\mathbf{V}\mathbf{\Lambda}\mathbf{V}' - \mathbf{U}\mathbf{\Lambda}\mathbf{U}'\|_{\mathrm{F}} \\
&\le \|(\mathbf{V}-\mathbf{U})\mathbf{\Lambda}\mathbf{V}'\|_{\mathrm{F}} + \|\mathbf{U}\mathbf{\Lambda}(\mathbf{V}-\mathbf{U})'\|_{\mathrm{F}} \le 2\lambda_1 \|\mathbf{V}-\mathbf{U}\|_{\mathrm{F}} \\
&\le 2\lambda_1 \sqrt{\frac{q}{2-q} k (s/k)^{2/q}} \qquad (65) \\
&\le 2\lambda_1 \sqrt{\frac{q}{2-q} \Psi(k,p,r,n,\lambda)} \qquad (66) \\
&\le \frac{\lambda_r}{2}, \qquad (67)
\end{aligned}$$

where (65) follows from applying Lemma 7, (66) follows from the choice of $k = k_q^*$ in (13), and (67) is implied by the assumption (25). Therefore

$$\sigma_r(\tilde{\mathbf{\Sigma}}) \ge 1 + \frac{\lambda_r}{2}, \qquad (68)$$

which implies that $\tilde{\mathbf{V}}$ indeed corresponds to the r leading singular vectors of $\tilde{\mathbf{\Sigma}}$. Hence we obtain the SVD of (64) as $\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}' + \mathbf{I}_p$. Using Theorem 10 in the supplementary material [12] we show that $\tilde{\mathbf{V}}$ provides a good sparse approximation of $\mathbf{V}$,

$$\|\mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}'\|_{\mathrm{F}}^2 \le \frac{2\|\mathbf{\Sigma} - \tilde{\mathbf{\Sigma}}\|_{\mathrm{F}}^2}{(\sigma_r(\tilde{\mathbf{\Sigma}}) - 1)^2} \le \frac{32 q \kappa^2}{2-q} \Psi(k,p,r,n,\lambda), \qquad (69)$$

where the last inequality follows from (65) and (68). If $q = 0$, then we define $\tilde{\mathbf{V}} = \mathbf{V}$.

Step 2: *Risk decomposition.* By definition of the maximizer B^* in (22), $\langle \mathbf{S}_{(2)}, \mathbf{V}_A\mathbf{V}_A' - \mathbf{V}_*\mathbf{V}_*' \rangle \le 0$. In view of Lemma 3, we have

$$\begin{aligned}
&\frac{\lambda_r}{2}\|\widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \\
&\quad \le \langle \mathbf{\Sigma}, \mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' \rangle \\
&\quad = \langle \mathbf{\Sigma}, \mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}' \rangle + \langle \mathbf{\Sigma}, \tilde{\mathbf{V}}\tilde{\mathbf{V}}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' \rangle + \langle \mathbf{\Sigma}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' \rangle \\
&\quad \le \langle \mathbf{\Sigma}, \mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}' \rangle + \langle \mathbf{\Sigma}, \tilde{\mathbf{V}}\tilde{\mathbf{V}}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' \rangle + \langle \mathbf{\Sigma} - \mathbf{S}_{(2)}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' \rangle \\
&\quad = \langle \mathbf{\Sigma}, \mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}' \rangle + \langle \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}}\tilde{\mathbf{V}}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' \rangle + \langle \mathbf{\Sigma} - \mathbf{S}_{(2)}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' \rangle \qquad (70) \\
&\quad \le \frac{\lambda_1}{2} \underbrace{\|\mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}'\|_{\mathrm{F}}^2}_{\text{approximation error}} + \frac{\lambda_1}{2} \underbrace{\|\tilde{\mathbf{V}}\tilde{\mathbf{V}}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A'\|_{\mathrm{F}}^2}_{\text{oracle risk}} \\
&\qquad + \underbrace{\langle \mathbf{\Sigma} - \mathbf{S}_{(2)}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' \rangle}_{\text{excess risk}}, \qquad (71)
\end{aligned}$$

where (70) follows from that $\operatorname{supp}(\tilde{\mathbf{V}}) = \operatorname{supp}(\widehat{\mathbf{V}}_A) = A$, and (71) follows from Lemma 3.

Note that the expected oracle risk is upper bounded by Theorem 9 because the conditions of Theorem 4 imply those of Theorem 9. The sparse approximation error can be upper bounded by (69). Moreover, in the exact sparse case ($q = 0$), we have $\tilde{\mathbf{V}} = \mathbf{V}$ and the approximation error is zero.

Step 3: *Excess risk*. The hard part is to control the third term (the worst-case fluctuation) in (71). To this end, we decompose the sample covariance matrix as

$$\mathbf{S}_{(2)} = \frac{1}{n}\mathbf{X}'_{(2)}\mathbf{X}_{(2)} = \frac{1}{n}(\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)} + \mathbf{Z}'_{(2)})(\mathbf{U}_{(2)}\mathbf{D}\mathbf{V}' + \mathbf{Z}_{(2)}).$$

Then

$$\mathbf{\Sigma} - \mathbf{S}_{(2)} = \mathbf{G} + \mathbf{H}, \tag{72}$$

where

$$\mathbf{G} \triangleq \mathbf{V}\mathbf{D}\left(\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right)\mathbf{D}\mathbf{V}', \tag{73}$$

$$\mathbf{H} \triangleq \mathbf{I}_p - \frac{1}{n}\mathbf{Z}'_{(2)}\mathbf{Z}_{(2)} - \frac{1}{n}\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\mathbf{Z}_{(2)} - \frac{1}{n}\mathbf{Z}'_{(2)}\mathbf{U}_{(2)}\mathbf{D}\mathbf{V}'. \tag{74}$$

We first deal the inner product with $\mathbf{G}$: write $\langle \mathbf{G}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_*\rangle = \langle \mathbf{G}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A - \mathbf{V}\mathbf{V}'\rangle - \langle \mathbf{G}, \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_* - \mathbf{V}\mathbf{V}'\rangle$. Note that

$$\begin{aligned}
\langle \mathbf{G}, \mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A\rangle &= \left\langle \mathbf{D}\left(\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right)\mathbf{D}, \mathbf{V}'(\mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A)\mathbf{V}\right\rangle \\
&= \left\langle \mathbf{D}\left(\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right)\mathbf{D}, \mathbf{I}_r - \mathbf{V}'\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A\mathbf{V}\right\rangle \\
&\le \left\|\mathbf{D}\left(\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right)\mathbf{D}\right\| \operatorname{Tr}(\mathbf{I}_r - \mathbf{V}'\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A\mathbf{V}) \qquad (75) \\
&\le \frac{\lambda_1}{2}\left\|\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right\| \|\mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A\|_{\mathrm{F}}^2, \qquad (76)
\end{aligned}$$

where (76) is due to (19) and (75) is a consequence of Lemma 6, in view of the fact that $\mathbf{I}_r - \mathbf{V}'\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A\mathbf{V}$ is symmetric positive semi-definite while $\mathbf{D}(\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r)\mathbf{D}$ is symmetric. Similarly, we have

$$\langle \mathbf{G}, \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_* - \mathbf{V}\mathbf{V}'\rangle \le \frac{\lambda_1}{2}\left\|\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right\| \|\mathbf{V}\mathbf{V}' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_*\|_{\mathrm{F}}^2. \tag{77}$$

Combining (76) and (77), we arrive at

$$|\langle \mathbf{G}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_*\rangle| \le 2\lambda_1\left\|\frac{1}{n}\mathbf{U}'_{(2)}\mathbf{U}_{(2)} - \mathbf{I}_r\right\| \|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}'_A - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}'_*\|_{\mathrm{F}}^2. \tag{78}$$

Next we control the inner product with $\mathbf{H}$: recall that $A = \operatorname{supp}(\tilde{\mathbf{V}})$ is fixed. We define a collection of $p \times p$ symmetric matrices indexed by $B \in \mathcal{B}(k)$ as follows:

$$\mathbf{K}_B \triangleq \|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_B\widehat{\mathbf{V}}_B'\|_{\mathrm{F}}^{-1}(\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_B\widehat{\mathbf{V}}_B'), \tag{79}$$

which has *zero trace* and unit Frobenius norm. Recall that $\widehat{\mathbf{V}}_* = \widehat{\mathbf{V}}_{B^*}$. Then

$$\begin{aligned}\langle \mathbf{H}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\rangle &= \|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\|_{\mathrm{F}}\langle \mathbf{H}, \mathbf{K}_{B^*}\rangle \\ &\le \|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\|_{\mathrm{F}} \underbrace{\max_{B\in\mathcal{B}(k)} |\langle \mathbf{H}, \mathbf{K}_B\rangle|}_{\triangleq T}\end{aligned} \tag{80}$$

Assembling (72), (78) and (80), we can upper bound the excess risk by

$$\begin{aligned}&\langle \boldsymbol{\Sigma} - \mathbf{S}_{(2)}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\rangle \\ &= \langle \mathbf{G}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\rangle + \langle \mathbf{H}, \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\rangle \\ &\le 2\lambda_1\left\|\frac{1}{n}\mathbf{U}_{(2)}'\mathbf{U}_{(2)} - \mathbf{I}_r\right\|\|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\|_{\mathrm{F}}^2 + T\|\widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A' - \widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*'\|_{\mathrm{F}}.\end{aligned} \tag{81}$$

Now we combine the risk decomposition (71) with the upper bounds above to control the risk of our aggregated estimator $\widehat{\mathbf{V}}_*$: to simplify notation, denote

$$\begin{aligned}\delta &= \|\widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}, \qquad \Delta = \|\mathbf{V}\mathbf{V}' - \tilde{\mathbf{V}}\tilde{\mathbf{V}}'\|_{\mathrm{F}}, \\ R &= \|\tilde{\mathbf{V}}\tilde{\mathbf{V}}' - \widehat{\mathbf{V}}_A\widehat{\mathbf{V}}_A'\|_{\mathrm{F}}, \qquad M = \left\|\frac{1}{n}\mathbf{U}_{(2)}'\mathbf{U}_{(2)} - \mathbf{I}_r\right\|.\end{aligned}$$

Assembling (71) and (81), we have

$$\left(\frac{\lambda_r}{2} - 6\lambda_1 M\right)\delta^2 \le T\delta + (\Delta^2 + R^2)\left(\frac{\lambda_1}{2} + 6\lambda_1 M\right) + T(R + \Delta). \tag{82}$$

Introduce the event $E = \{M \le \frac{1}{24\kappa}\}$. By assumption (26), $r \le c''n$ for a sufficiently small constant c''. Then there exists a constant $c' > 0$ only depending on κ, such that $\frac{1}{24\kappa} \ge 2(\sqrt{\frac{r}{n}} + t) + (\sqrt{\frac{r}{n}} + t)^2$, where $t = \sqrt{\frac{\log(c'nh(\lambda))}{n}}$. Applying Proposition 4 in the supplementary material [12] yields

$$\mathsf{P}(E^c) \le \frac{1}{c'nh(\lambda)}. \tag{83}$$

Conditioning on the event E and using Lemma 2, we have

$$\delta^2 \le \frac{32T^2}{\lambda_r^2} + \frac{3\lambda_1(\Delta^2 + R^2) + 4T(R + \Delta)}{\lambda_r}. \tag{84}$$

Recall from (19) that the loss function is upper bounded by $r \wedge (p-r)$. Taking expectation on both sides of (84), and using (83) together with the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\mathsf{E}\|\widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \\
&\le \frac{32\mathsf{E}T^2}{\lambda_r^2} + 3\kappa(\Delta^2 + \mathsf{E}R^2) + \frac{4\mathsf{E}[T(R+\Delta)]}{\lambda_r} + r\mathsf{P}(E^c) \qquad (85)\\
&\le \frac{20\mathsf{E}T^2}{\lambda_r^2} + (3\kappa+8)(\Delta^2 + \mathsf{E}R^2) + \frac{r}{c'nh(\lambda)}. \qquad (86)
\end{aligned}$$

In view of the oracle upper bound in Theorem 9, we have

$$\mathsf{E}R^2 \le C\left(r \wedge (k-r) \wedge \frac{(k-r)r}{nh(\lambda)}\right). \qquad (87)$$

By (69), if $q>0$, the approximation is upper bounded by

$$\Delta^2 \le \frac{32q\kappa^2}{2-q}\Psi(k,p,r,n,\lambda). \qquad (88)$$

If $q=0$, then $\Delta=0$. To control the right-hand side of (86), it boils down to upper bound $\mathsf{E}T^2$. In the sequel we shall prove that

$$\mathsf{E}T^2 \le C(1+\lambda_1)\frac{k}{n}\log\frac{ep}{k} \qquad (89)$$

for some absolutely constant C. Plugging (87), (88) and (89) into (86), we arrive at

$$\begin{aligned}
&\mathsf{E}\|\widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \\
&\le \frac{C}{h(\lambda)}\frac{k}{n}\log\frac{ep}{k} + \frac{32q\kappa^2}{2-q}\Psi(k,p,r,n,\lambda) + r \wedge \frac{(k-r)r}{nh(\lambda)} + \frac{r}{c'nh(\lambda)} \qquad (90)\\
&\le C'\Psi(k,p,r,n,\lambda), \qquad (91)
\end{aligned}$$

where the constant C' only depends on κ. In the special case of $q=0$, the approximation error is $\Delta=0$, which implies that the second term in (90) is zero. Hence we have the following stronger result:

$$\begin{aligned}
\mathsf{E}\|\widehat{\mathbf{V}}_*\widehat{\mathbf{V}}_*' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 &\le \frac{C}{h(\lambda)}\frac{k}{n}\log\frac{ep}{k} + r \wedge \frac{(k-r)r}{nh(\lambda)} + \frac{r}{c'nh(\lambda)} \\
&\le C'\Psi_0(s,p,r,n,\lambda), \qquad (92)
\end{aligned}$$

where Ψ_0 is defined in (12). Then (91) and (92) imply the statement of the theorem for $q>0$ and $q=0$, respectively.

To finish the proof of the theorem, it remains to establish (89). To this end, recall that $\mathbf{K}_B$ is symmetric and $\operatorname{Tr}(\mathbf{K}_B) = 0$. By the definitions of T and $\mathbf{H}$ in (80) and (74), respectively, we have

$$T \le T_1 + 2T_2, \tag{93}$$

where we define

$$T_1 \triangleq \frac{1}{n} \max_{B \in \mathcal{B}(k)} |\langle \mathbf{Z}'_{(2)} \mathbf{Z}_{(2)}, \mathbf{K}_B \rangle| \tag{94}$$

$$T_2 \triangleq \frac{1}{n} \max_{B \in \mathcal{B}(k)} |\langle \mathbf{V}\mathbf{D}\mathbf{U}'_{(2)} \mathbf{Z}_{(2)}, \mathbf{K}_B \rangle| = \frac{1}{n} \max_{B \in \mathcal{B}(k)} |\langle \mathbf{Z}'_{(2)} \mathbf{U}_{(2)} \mathbf{D}\mathbf{V}', \mathbf{K}_B \rangle|. \tag{95}$$

We shall prove that

$$\mathsf{E} T_1^2 \le \frac{24k}{n} \log \frac{ep}{k} + \frac{32k^2}{n^2} \log^2 \frac{ep}{k} + \frac{62}{n}. \tag{96}$$

$$\mathsf{E} T_2^2 \le \lambda_1 \left(\frac{40k}{n} \log \frac{ep}{k} + \frac{24k^2}{n^2} \log^2 \frac{ep}{k} + \frac{103}{n} + \frac{17k}{n^2} \right). \tag{97}$$

Assembling (93) with (96)–(95) and using the fact that $(a+b)^2 \le 2(a^2+b^2)$, we arrive at

$$\begin{aligned} \mathsf{E} T^2 &\le \mathsf{E} T_1^2 + 8 \mathsf{E} T_2^2 \\ &\le 1500(1+\lambda_1) \left(\frac{k}{n} \log \frac{ep}{k} + \frac{k^2}{n^2} \log^2 \frac{ep}{k} \right) \qquad (98) \\ &\le 3000(1+\lambda_1) \frac{k}{n} \log \frac{ep}{k}, \qquad (99) \end{aligned}$$

where we used $\frac{k}{n} \log \frac{p}{k} \le 1$ implied by the assumption (26).

It then remains to establish (96)–(97). Note that the collection $\{\mathbf{K}_B : B \in \mathcal{B}(k)\}$ belongs to the σ-algebra generated by the first sample $\mathbf{X}_{(1)}$, which is independent of $(\mathbf{Z}_{(2)}, \mathbf{U}_{(2)})$. By conditioning on $\mathbf{X}_{(1)}$, we can treat $\{\mathbf{K}_B : B \in \mathcal{B}(k)\}$ as fixed matrices. □

Proof of (96). For each fixed $B \in \mathcal{B}(k)$, $\mathbf{K}_B \perp\!\!\!\perp \mathbf{Z}_{(2)}$. Applying Lemma 4, we have

$$\mathsf{P}\left(\frac{1}{\sqrt{n}} |\langle \mathbf{Z}'\mathbf{Z}, \mathbf{K}_B \rangle| \ge 2t + \frac{2t^2}{\sqrt{n}} \right) \le 2\exp(-t^2).$$

Applying Lemma 5 with $N = |\mathcal{B}(k)| = \binom{p}{k} \le (\frac{ep}{k})^k$, $a = 2, b = \frac{2}{\sqrt{n}}$ and $c = 2$, we have

$$\mathsf{E} T_1^2 \le \frac{1}{n} \left(8 \log(2eN) + \frac{8}{n} (\log^2(2N) + 2\log(2eN)) \right) \tag{100}$$

$$= \frac{24}{n} \log(2eN) + \frac{8}{n^2} \log^2(2N), \tag{101}$$

which implies (96). □

PROOF OF (97). Fix $B \in \mathcal{B}(k)$. Since $\mathbf{U}_{(2)} \perp\!\!\!\perp \mathbf{Z}_{(2)}$, conditioned on the realization of $\mathbf{U}_{(2)}$, $\langle \mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\mathbf{Z}_{(2)}, \mathbf{K}_B\rangle = \langle \mathbf{K}_B\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}, \mathbf{Z}'_{(2)}\rangle$ is distributed according to $N(0, \|\mathbf{K}_B\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\|_{\mathrm{F}}^2)$. Therefore

$$\langle \mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\mathbf{Z}_{(2)}, \mathbf{K}_B\rangle \overset{\mathrm{(d)}}{=} \|\mathbf{K}_B\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\|_{\mathrm{F}} W$$

for some $W \sim N(0,1)$ independent of $\mathbf{U}_{(2)}$.

Using the fact that $\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}} \le \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|$, we have

$$\|\mathbf{K}_B\mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\|_{\mathrm{F}} \le \|\mathbf{K}_B\|_{\mathrm{F}}\|\mathbf{V}\|\|\mathbf{D}\|\|\mathbf{U}'_{(2)}\| \le \sqrt{\lambda_1}\|\mathbf{U}_{(2)}\|.$$

Consequently, $\langle \mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\mathbf{Z}_{(2)}, \mathbf{K}_B\rangle$ is stochastically dominated by $\sqrt{\lambda_1}\|\mathbf{U}_{(2)}\||W|$. Since $\mathbf{U}_{(2)}$ is an $n \times r$ standard Gaussian matrix, Lemma 10 in the supplementary material [12] yields

$$\mathsf{P}(\|\mathbf{U}_{(2)}\| \ge \sqrt{n} + \sqrt{r} + t) \le \exp\left(-\frac{t^2}{2}\right), \qquad t > 0. \tag{102}$$

Applying the union bound yields

$$\begin{aligned}
&\mathsf{P}(\|\mathbf{U}_{(2)}\||W| \ge \sqrt{2}(\sqrt{n}+\sqrt{r})t + 2t^2) \\
&\quad \le \mathsf{P}((\|\mathbf{U}_{(2)}\| - \sqrt{n} - \sqrt{r})|W| \ge 2t^2) + \mathsf{P}(|W| \ge \sqrt{2}t) \\
&\quad \le \mathsf{P}(\|\mathbf{U}_{(2)}\| \ge \sqrt{n} + \sqrt{r} + \sqrt{2}t) + 2\mathsf{P}(|W| \ge \sqrt{2}t) \\
&\quad \le 3\exp(-t^2),
\end{aligned}$$

which the last inequality follows from (102) and the Chernoff bound $\mathsf{P}(W \ge \sqrt{2}t) \le \frac{1}{2}\exp(-t)$. Therefore,

$$\mathsf{P}\left(\frac{\langle \mathbf{V}\mathbf{D}\mathbf{U}'_{(2)}\mathbf{Z}_{(2)}, \mathbf{K}_B\rangle}{\sqrt{\lambda_1}} \ge \sqrt{2}(\sqrt{n}+\sqrt{r})t + 2t^2\right) \le 3\exp(-t^2).$$

Applying Lemma 5 with $N = \binom{p}{k}$ yields

$$\mathsf{E}T_2^2 \le \frac{4\lambda_1}{n^2}((8 + (\sqrt{n}+\sqrt{r})^2)\log(3eN) + 2\log^2(3N)),$$

which, in view of $r \le k$, implies the desired (97). □

6.3. *Proof of Theorem 7.* We prove the theorem in three steps. First, we verify that the "whitening" procedure in step 3 of the reduction scheme can be performed. Next, we investigate the signal-to-noise ratio of the regression problem conditional on the values of $\mathbf{U}$ and $\mathbf{Z}^0$. Finally, we derive the desired rates by using Theorem 6 and Wedin's sin-theta theorem [55].

(1°) As a first step, we verify that the "whitening" step is indeed possible, which requires that $\sigma_r(\mathbf{B}) > 0$. To this end, let $J = \operatorname{supp}(\mathbf{V}^0)$. Since $\mathbf{B} = \mathbf{UDV}'\mathbf{V}^0 + \mathbf{Z}^0\mathbf{V}^0$, we have

$$\begin{aligned}\sigma_r(\mathbf{B}) &\geq \sigma_r(\mathbf{UDV}'\mathbf{V}^0) - \sigma_1(\mathbf{Z}^0\mathbf{V}^0)\\ &\geq \sigma_r(\mathbf{U})\sigma_r(\mathbf{D})\sigma_r(\mathbf{V}'\mathbf{V}^0) - \sigma_1(\mathbf{Z}^0_J).\end{aligned} \tag{103}$$

By our assumption on $\mathbf{V}^0$, condition (41) is satisfied with probability at least $1 - C/[nh(\lambda)]$. By Lemma 10 in the supplementary material [12] and the union bound,

$$\begin{aligned}\sigma_r(\mathbf{U}) &\geq \sqrt{n}\left(1 - \sqrt{\frac{r}{n}} - \sqrt{\frac{2\log[nh(\lambda)]}{n}}\right),\\ \sigma_r(\mathbf{V}'\mathbf{V}^0) &\geq \frac{1}{2}, \qquad |J| \leq k_q^*\end{aligned} \tag{104}$$

holds with probability at least $1 - C/[nh(\lambda)]$. Note that assumption (26) implies that $n \geq C_0 r$ and that $n \geq C_0 \log[nh(\lambda)]$. Thus, for sufficiently large C_0 in (26), the first inequality in (104) leads to $\sigma_r(\mathbf{U}) \geq \frac{2}{3}\sqrt{n}$. Together with $\sigma_r(\mathbf{D}) = \sqrt{\lambda_r}$, the first term in (103) is thus lower bounded by $\frac{1}{3}\sqrt{n\lambda_r}$, and hence

$$\sigma_r(\mathbf{B}) \geq \frac{1}{3}\sqrt{n\lambda_r} - \sigma_1(\mathbf{Z}^0_J) \tag{105}$$

with probability at least $1 - C/[nh(\lambda)]$.

Turning to the second term in (103), we first note that it is upper bounded by $\max_{I\subset[p],|I|=k_q^*} \|\mathbf{Z}^0_I\|$ conditioned on the event that $|J| \leq k_q^*$. Note that for any $t > 0$, we have

$$\begin{aligned}&\mathsf{P}\Big\{\max_{I\subset[p],|I|=k_q^*} \|\mathbf{Z}^0_I\| > \sqrt{n} + \sqrt{k_q^*} + t\Big\}\\ &\quad\leq \sum_{I\subset[p],|I|=k_q^*} \mathsf{P}\{\|\mathbf{Z}^0_I\| > \sqrt{n} + \sqrt{k_q^*} + t\} \leq \binom{p}{k_q^*}\exp(-t^2/2)\\ &\quad\leq \left(\frac{ep}{k_q^*}\right)^{k_q^*}\exp(-t^2/2) = \exp\left(-\frac{t^2}{2} + k_q^*\log\left(\frac{ep}{k_q^*}\right)\right).\end{aligned}$$

Set $t = t^* = \sqrt{2k_q^*\log(ep/k_q^*)} + \sqrt{2\log[nh(\lambda)]}$. The rightmost side of the last display is then bounded by $C/[nh(\lambda)]$. Thus, by (104) and the union bound,

$$\sigma_1(\mathbf{Z}^0_J) \leq \sqrt{n} + \sqrt{k_q^*} + t^* \leq 2\sqrt{n} \tag{106}$$

with probability at least $1 - C/[nh(\lambda)]$, where the last inequality holds because the assumption (26) implies that $k_q^* \leq n/4$ and $t^* \leq n/2$ as long as C_0 is sufficiently large.

Under the assumption that $\lambda_r \geq C_0$ for some sufficiently large $C_0 > 36$, (105) and (106) lead to $\sigma_r(\mathbf{B}) \geq c\sqrt{n\lambda_r} > 0$ with probability at least $1 - C/[nh(\lambda)]$. This completes the first step in the proof.

(2°) Let $\bar{\mathbf{A}} = \frac{1}{\sqrt{2}}\mathbf{A}\mathbf{R}\mathbf{C}^{-1} = \frac{1}{\sqrt{2}}\mathbf{D}\mathbf{U}'\mathbf{B}\mathbf{R}\mathbf{C}^{-1} = \frac{1}{\sqrt{2}}\mathbf{D}\mathbf{U}'\mathbf{L}$. Then $\boldsymbol{\Theta} = \mathbf{V}\bar{\mathbf{A}}$ in (32). In the second step, we show that there exist two constants $C_2 > C_1 > 0$ depending only on κ, such that with probability at least $1 - C/[nh(\lambda)]$,

$$C_1\sqrt{n\lambda} \leq \sigma_r(\bar{\mathbf{A}}) \leq \sigma_1(\bar{\mathbf{A}}) \leq C_2\sqrt{n\lambda}. \tag{107}$$

To this end, note that (104) and assumption (26) imply

$$\sigma_r(\bar{\mathbf{A}}) \geq \frac{1}{\sqrt{2}}\sigma_r(\mathbf{D})\sigma_r(\mathbf{U}) \geq \sqrt{\frac{n\lambda_r}{2}}\left(1 - \sqrt{\frac{r}{n}} - \sqrt{\frac{2\log[nh(\lambda)]}{n}}\right) \geq C_1\sqrt{n\lambda}$$

holds with probability at least $1 - C/[nh(\lambda)]$. Under the same assumption, Lemma 10 in the supplementary material [12] implies

$$\sigma_1(\bar{\mathbf{A}}) \leq \frac{1}{\sqrt{2}}\sigma_1(\mathbf{D})\sigma_1(\mathbf{U}) \leq \sqrt{\frac{n\lambda_1}{2}}\left(1 + \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log[nh(\lambda)]}{n}}\right) \leq C_2\sqrt{n\lambda}.$$

Thus (107) is established.

(3°) Next we show that, conditioned on the event that (107) holds, the signal matrix $\boldsymbol{\Theta}$ lies in $\mathcal{F}_q(s', p)$ where

$$s' \leq s\sigma_1^q(\bar{\mathbf{A}}) \leq Cs(n\lambda)^{q/2} \leq Cs(nh(\lambda))^{q/2}, \tag{108}$$

where the middle inequality is due to (107), the last inequality follows from the assumption that $\lambda \geq C_0$ and the first inequality is due to $\|\boldsymbol{\Theta}\|_{q,w} \leq \|\mathbf{V}\|_{q,w}\|\bar{\mathbf{A}}\|^q$, which is a consequence of equation (110) in Section 7.1 of the supplementary material [12].

Let k' be defined in (44). We show that whenever (108) holds, we have

$$k' \leq C'k_q^*, \tag{109}$$

where k_q^* is the effective dimension defined in (13), and the constant C' depends only on q. To see this, note that $k_q^* \geq 1$ by Remark 1. Then (109) holds trivially if $k' = 1$. Next assume that $k' \geq 2$. By definition, $t_{k'-1}^{q/2}(k' - 1) \leq s'$. Note that $\beta > 1$ and $t_k \geq r + \log\frac{ep}{k}$. By (108), we have $(k' - 1)(r + \log\frac{ep}{k'-1})^{q/2} \leq Cs(nh(\lambda))^{q/2}$. Hence $k' - 1 \leq k_q^*(Cs, p, r, n, \lambda) \leq \tau_q(C)k_q^*(s, p, r, n, \lambda)$, where the last inequality follows from the third property of k_q^* in Remark 1. This proves the desired (109).

Let E denote the event that both (104) and (107) hold. Then

$$\begin{aligned}\mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 &= \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2\mathbf{1}_{\{E\}} + \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2\mathbf{1}_{\{E^c\}} \\ &\leq \mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2\mathbf{1}_{\{E\}} + \frac{Cr}{nh(\lambda)}.\end{aligned}$$

Here, the last inequality holds because the loss function is upper bounded by r and $\mathsf{P}(E^c) \le C/[nh(\lambda)]$.

To further bound the first term on the rightmost hand side, we note that E is completely determined by $\mathbf{U}$ and $\mathbf{Z}^0$. Hence, it is nonrandom conditioned on $\mathbf{U}$ and $\mathbf{Z}^0$. Thus

$$\begin{aligned}\mathsf{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'\|_{\mathrm{F}}^2 \mathbf{1}_{\{E\}} &\le 2\mathsf{E}\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2}{\sigma_r^2(\bar{\mathbf{A}})}\mathbf{1}_{\{E\}} \le \frac{C}{n\lambda}\mathsf{E}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2 \mathbf{1}_{\{E\}} \\ &= \frac{C}{n\lambda}\mathsf{E}[\mathsf{E}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2 \mathbf{1}_{\{E\}} | \mathbf{U}, \mathbf{Z}^0]\mathbf{1}_{\{E\}}] \\ &\le \frac{C}{n\lambda}\mathsf{E}\left[k'\left(r + \log\frac{\mathrm{e}p}{k'}\right)\mathbf{1}_{\{E\}}\right] \\ &\le \frac{Ck_q^*}{n\lambda}\left(r + \log\frac{\mathrm{e}p}{k_q^*}\right).\end{aligned}$$

Here, the first inequality comes from Wedin's sin-theta theorem for SVD [55]. The second inequality comes from (107). The second-to-last inequality comes from Theorem 6. The last inequality holds because on the event E, $k' \le Ck_q^*$ in view of (109), and $k \mapsto k(r + \log(\mathrm{e}p/k))$ is increasing. We complete the proof by noting that $1/\lambda \le C/h(\lambda)$ holds since $\lambda > C_0$. The upper bound $2(r \wedge (p-r))$ holds in view of (20).

SUPPLEMENTARY MATERIAL

Supplement to "Sparse PCA: Optimal rates and adaptive estimation" (DOI: 10.1214/13-AOS1178SUPP; .pdf). We provide proofs for all the remaining theoretical results in the paper. The proofs rely on results in [17, 19, 20, 25, 31, 33] and [51].

REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

[2] Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[3] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[4] Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

[5] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

[6] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[7] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129

[8] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268. MR1848946

[9] Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. Preprint. Available at arXiv:1203.0967.

[10] Cai, T. T., Liu, W. and Zhou, H. H. (2012). Optimal estimation of large sparse precision matrices. Technical Report, Univ. Pennsylvania, PA.

[11] Cai, T. T., Ma, Z. and Wu, Y. (2013). Optimal estimation and rank detection for sparse spiked covariance matrices. Preprint. Available at arXiv:1305.3235.

[12] Cai, T. T., Ma, Z. and Wu, Y. (2013). Supplement to "Sparse PCA: Optimal rates and adaptive estimation." DOI:10.1214/13-AOS1178SUPP.

[13] Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

[14] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[15] Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

[16] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean–variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050

[17] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

[18] d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448 (electronic). MR2353806

[19] Davidson, K. R. and Szarek, S. J. (2001). *Handbook of the Geometry of Banach Spaces* 317–366. North-Holland, Amsterdam. MR1863696

[20] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

[21] Eaton, M. L. (1970). Some problems in covariance estimation. Technical Report 49, Dept. Statistics, Univ. Stanford, Stanford, CA.

[22] Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

[23] Hoyle, D. C. and Rattray, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E (3)* **69** 026124.

[24] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[25] Johnstone, I. M. (2001). Thresholding for weighted chi-squared. *Statist. Sinica* **11** 691–704.

[26] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

[27] Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

[28] Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.* **11** 517–553. MR2600619

[29] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. MR1792783

[30] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454

[31] Kolmogorov, A. N. and Tihomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk* **14** 3–86. MR0112032

[32] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[33] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

[34] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[35] Lounici, K. (2013). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI* 327–356. Springer, Basel.

[36] Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

[37] Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

[38] Mendelson, S. (2010). Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.* **20** 988–1027. MR2729283

[39] Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

[40] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

[41] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

[42] Paul, D. (2005). Nonparametric estimation of principal components. Ph.D. thesis, Univ. Stanford, Stanford, CA. MR2707156

[43] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[44] Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543

[45] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[46] Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

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

[48] Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

[49] Stein, C. (1956). Some problems in multivariate analysis, part i. Technical Report 6, Dept. Statistics, Univ. Stanford.
[50] Szarek, S. J. (1982). Nets of Grassmann manifold and orthogonal group. In *Proceedings of Research Workshop on Banach Space Theory (Iowa City, Iowa, 1981)* 169–185. Univ. Iowa, Iowa City, IA. MR0724113
[51] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359
[52] Ulfarsson, M. O. and Solo, V. (2008). Sparse variable PCA using geodesic steepest descent. *IEEE Trans. Signal Process.* **56** 5823–5832. MR2518261
[53] Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, FL.
[54] Vu, V. Q. and Lei, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. In *The Fifteenth International Conference on Artificial Intelligence and Statistics* (*AISTATS'12*), available at http://arxiv.org/abs/1202.0786.
[55] Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *Nordisk Tidskr. Informationsbehandling (BIT)* **12** 99–111. MR0309968
[56] Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.
[57] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** 135–161. MR1790617
[58] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500
[59] Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.* **14** 899–925. MR3063614
[60] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

T. T. Cai
Z. Ma
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania 19104
USA
E-mail: tcai@wharton.upenn.edu
zongming@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~tcai
http://www-stat.wharton.upenn.edu/~zongming

Y. Wu
Department of Electrical
and Computer Engineering
University of Illinois
Urbana-Champaign
Urbana, Illinois 61801
USA
E-mail: yihongwu@illinois.edu
URL: http://www.ifp.illinois.edu/~yihongwu