The Annals of Statistics
2013, Vol. 41, No. 6, 2905–2947
DOI: 10.1214/13-AOS1151

MINIMAX SPARSE PRINCIPAL SUBSPACE ESTIMATION IN HIGH DIMENSIONS

BY VINCENT Q. VU[1] AND JING LEI[2]

Ohio State University and Carnegie Mellon University

We study sparse principal components analysis in high dimensions, where p (the number of variables) can be much larger than n (the number of observations), and analyze the problem of estimating the subspace spanned by the principal eigenvectors of the population covariance matrix. We introduce two complementary notions of ℓ_q subspace sparsity: row sparsity and column sparsity. We prove nonasymptotic lower and upper bounds on the minimax subspace estimation error for $0 \leq q \leq 1$. The bounds are optimal for row sparse subspaces and nearly optimal for column sparse subspaces, they apply to general classes of covariance matrices, and they show that ℓ_q constrained estimates can achieve optimal minimax rates without restrictive spiked covariance conditions. Interestingly, the form of the rates matches known results for sparse regression when the effective noise variance is defined appropriately. Our proof employs a novel variational $\sin\Theta$ theorem that may be useful in other regularized spectral estimation problems.

1. Introduction. Principal components analysis (PCA) was introduced in the early 20th century [Hotelling (1933), Pearson (1901)] and is arguably the most well known and widely used technique for dimension reduction. It is part of the mainstream statistical repertoire and is routinely used in numerous and diverse areas of application. However, contemporary applications often involve much higher-dimensional data than envisioned by the early developers of PCA. In such high-dimensional situations, where the number of variables p is of the same order or much larger than the number of observations n, serious difficulties emerge: standard PCA can produce inconsistent estimates of the principal directions of variation and lead to unreliable conclusions [Johnstone and Lu (2009), Nadler (2008), Paul (2007)].

The principal directions of variation correspond to the eigenvectors of the covariance matrix, and in high-dimensions consistent estimation of the eigenvectors is generally not possible without additional assumptions about the covariance matrix or its eigenstructure. Much of the recent development in PCA has focused

Received February 2013; revised June 2013.

[1]Supported in part by NSF Postdoctoral Fellowship DMS-09-03120.

[2]Supported in part by NSF Grant BCS-0941518.

MSC2010 subject classifications. Primary 62H25; secondary 62H12, 62C20.

Key words and phrases. Principal components analysis, subspace estimation, sparsity, high-dimensional statistics, minimax bounds, random matrices.

on methodology that applies the concept of sparsity to the estimation of individual eigenvectors [examples include d'Aspremont et al. (2007), Jolliffe, Trendafilov and Uddin (2003), Journée et al. (2010), Shen and Huang (2008), Witten, Tibshirani and Hastie (2009), Zou, Hastie and Tibshirani (2006)]. Theoretical developments on sparsity and PCA include consistency [Johnstone and Lu (2009), Shen, Shen and Marron (2013)], variable selection properties [Amini and Wainwright (2009)], rates of convergence and minimaxity [Vu and Lei (2012a)], but have primarily been limited to results about estimation of the leading eigenvector. Very recently, Birnbaum et al. (2013) established minimax lower bounds for the estimation of individual eigenvectors. However, an open problem that has remained is whether sparse PCA methods can *optimally* estimate the subspace spanned by the leading eigenvectors, that is, the *principal subspace* of variation.

The subspace estimation problem is directly connected to dimension reduction and is important when there may be more than one principal component of interest. Indeed, typical applications of PCA use the projection onto the principal subspace to facilitate exploration and inference of important features of the data. In that case, the assumption that there are distinct principal directions of variation is mathematically convenient but unnatural: it avoids the problem of unidentifiability of eigenvectors by imposing an artifactual choice of principal axes. Dimension reduction by PCA should emphasize subspaces rather than eigenvectors.

An important conceptual issue in applying sparsity to principal subspace estimation is that, unlike the case of sparse vectors, it is not obvious how to formally define what is meant by a *sparse principal subspace*. In this article, we present two complementary notions of sparsity based on ℓ_q (pseudo-) norms: row sparsity and column sparsity. Roughly, a subspace is row sparse if every one of its orthonormal bases consists of sparse vectors. In the $q = 0$ case, this intuitively means that a row sparse subspace is generated by a small subset of variables, independent of the choice of basis. A column sparse subspace, on the other hand, is one which has some orthonormal basis consisting of sparse vectors. This means that the choice of basis is crucial; the existence of a sparse basis is an implicit assumption behind the frequent use of rotation techniques by practitioners to help interpret principal components.

In this paper, we study sparse principal subspace estimation in high-dimensions. We present nonasymptotic minimax lower and upper bounds for estimation of both row sparse and column sparse principal subspaces. Our upper bounds are constructive and apply to a wide class of distributions and covariance matrices. In the row sparse case they are optimal up to constant factors, while in the column sparse case they are nearly optimal. As an illustration, one consequence of our results is that the order of the minimax mean squared estimation error of a row sparse d-dimensional principal subspace (for $d \ll p$) is

$$R_q\left(\frac{\sigma^2}{n}(d + \log p)\right)^{1-q/2}, \qquad 0 \le q \le 1,$$

where σ^2 is the effective noise variance (a function of the eigenvalues of population covariance matrix) and R_q is a measure of the sparsity in an ℓ_q sense defined in Section 2. Our analysis allows σ, R_q, and d to change with n and p. When $q = 0$, the rate has a very intuitive explanation. There are R_0 variables active in generating the principal subspace. For each active variable, we must estimate the corresponding d coordinates of the basis vectors. Since we do not know in advance which variables are active, we incur an additional cost of $\log p$ for variable selection.

To our knowledge, the only other work that has considered sparse principal subspace estimation is that of Ma (2013). He proposed a sparse principal subspace estimator based on iterative thresholding, and derived its rate of convergence under a spiked covariance model (where the covariance matrix is assumed to be a rank-d perturbation of the identity) similar to that in Birnbaum et al. (2013). He showed that it nearly achieves the optimal rate when estimating a single eigenvector, but was not able to track its dependence on the dimension of the principal subspace.

We obtain the minimax upper bounds by analyzing a sparsity constrained principal subspace estimator and showing that it attains the optimal error (up to constant factors). In comparison to most existing works in the literature, we show that the upper bounds hold without assuming a spiked covariance model. This spiked covariance assumption seems to be necessary for two reasons. The first is that it simplifies analyses and enables the exploitation of special properties of the multivariate Gaussian distribution. The second is that it excludes the possibility of the variables having equal variances. Estimators proposed by Paul (2007), Johnstone and Lu (2009), and Ma (2013) require an initial estimate based on *diagonal thresholding*—screening out variables with small sample variances. Such an initial estimate will not work when the variables have equal variances or have been standardized. The spiked covariance model excludes that case and, in particular, does not allow PCA on correlation matrices.

A key technical ingredient in our analysis of the subspace estimator is a novel variational form of the Davis–Kahan $\sin\Theta$ theorem (see Corollary 4.1) that may be useful in other regularized spectral estimation problems. It allows us to bound the estimation error using some recent advanced results in empirical process theory, without Gaussian or spiked covariance assumptions. The minimax lower bounds follow the standard Fano method framework [e.g., Yu (1997)], but their proofs involve nontrivial constructions of packing sets in the Stiefel manifold. We develop a generic technique that allows us to convert global packing sets without orthogonality constraints into local packing sets in the Stiefel manifold, followed by a careful combinatorial analysis on the cardinality of the resulting matrix class.

The remainder of the paper is organized as follows. In the next section, we introduce the sparse principal subspace estimation problem and formally describe our minimax framework and estimator. In Section 3, we present our main conditions and results, and provide a brief discussion about their consequences and intuition. Section 4 outlines the key ideas and main steps of the proof. Section 5 concludes the paper with discussion of related problems and practical concerns. Appendices A, B contain the details in proving the lower and upper bounds. The major steps in the proofs require some auxiliary lemmas whose proofs we defer to Appendices C, D.

2. Subspace estimation. Let $X_1, \dots, X_n \in \mathbb{R}^p$ be independent, identically distributed random vectors with mean μ and covariance matrix Σ. To reduce the dimension of the X_i's from p down to d, PCA looks for d mutually uncorrelated, linear combinations of the p coordinates of X_i that have maximal variance. Geometrically, this is equivalent to finding a d-dimensional linear subspace that is closest to the centered random vector $X_i - \mu$ in a mean squared sense, and it corresponds to the optimization problem

$$\begin{aligned} &\text{minimize} \quad \mathbb{E}\|(I_p - \Pi_{\mathcal{G}})(X_i - \mu)\|_2^2 \\ &\text{subject to} \quad \mathcal{G} \in \mathbb{G}_{p,d}, \end{aligned} \tag{2.1}$$

where $\mathbb{G}_{p,d}$ is the Grassmann manifold of d-dimensional subspaces of $\mathbb{R}^p$, $\Pi_{\mathcal{G}}$ is the orthogonal projector of $\mathcal{G}$, and I_p is the $p \times p$ identity matrix. [For background on Grassmann and Stiefel manifolds, see Chikuse (2003), Edelman, Arias and Smith (1999).] There is always at least one $d \le p$ for which (2.1) has a unique solution. That solution can be determined by the spectral decomposition

$$\Sigma = \sum_{j=1}^{p} \lambda_j v_j v_j^T, \tag{2.2}$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ are the eigenvalues of Σ and $v_1, \dots, v_p \in \mathbb{R}^p$, orthonormal, are the associated eigenvectors. If $\lambda_d > \lambda_{d+1}$, then the d-dimensional *principal subspace* of Σ is

$$\mathcal{S} = \operatorname{span}\{v_1, \dots, v_d\}, \tag{2.3}$$

and the orthogonal projector of $\mathcal{S}$ is given by $\Pi_{\mathcal{S}} = VV^T$, where V is the $p \times d$ matrix with columns $v_1, \dots, v_d$.

In practice, Σ is unknown, so $\mathcal{S}$ must be estimated from the data. Standard PCA replaces (2.1) with an empirical version. This leads to the spectral decomposition of the sample covariance matrix

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T,$$

where $\bar{X}$ is the sample mean, and estimating $\mathcal{S}$ by the span of the leading d eigenvectors of S_n. In high-dimensions however, the eigenvectors of S_n can be inconsistent estimators of the eigenvectors of Σ. Additional structural constraints are necessary for consistent estimation of $\mathcal{S}$.

2.1. *Subspace sparsity.* The notion of sparsity is appealing and has been used successfully in the context of estimating vector valued parameters such as the leading eigenvector in PCA. Extending this notion to subspaces requires care because sparsity is inherently a coordinate-dependent concept while subspaces are coordinate-independent. For a given d-dimensional subspace $\mathcal{G} \in \mathbb{G}_{p,d}$, the set of orthonormal matrices whose columns span $\mathcal{G}$ is a subset of the Stiefel manifold $\mathbb{V}_{p,d}$ of $p \times d$ orthonormal matrices. We will consider two complementary notions of subspace sparsity defined in terms of those orthonormal matrices: *row sparsity* and *column sparsity*.

Define the $(2,q)$-norm, $q \in [0,\infty]$, of a $p \times d$ matrix A as the usual ℓ_q norm of the vector of row-wise ℓ_2 norms of A:

$$\|A\|_{2,q} := \big\| \big(\|a_{1*}\|_2 \quad \cdots \quad \|a_{p*}\|_2 \big) \big\|_q,$$

where a_{j*} denotes the jth row of A. (To be precise, this is actually a pseudonorm when $q < 1$.) Note that $\|\cdot\|_{2,q}$ is coordinate-independent, because $\|AO\|_{2,q} = \|A\|_{2,q}$ for any orthogonal matrix $O \in \mathbb{R}^{d\times d}$. We define the *row sparse subspaces* using this norm. Let $\operatorname{col}(U)$ denotes the span of the columns of U.

DEFINITION (Row sparse subspaces). For $0 \le q < 2$ and $d \le R_q \le d^{q/2} \times p^{1-q/2}$,

$$\mathcal{M}_q(R_q) := \begin{cases} \{\operatorname{col}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{2,q}^q \le R_q\}, & \text{if } 0 < q < 2 \quad \text{and} \\ \{\operatorname{col}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{2,0} \le R_0\}, & \text{if } q = 0. \end{cases}$$

The constraints on R_q arise from the fact that the vector of row-wise ℓ_2 norms of a $p \times d$ orthonormal matrix belongs to a sphere of radius d. Roughly speaking, row sparsity asserts that there is a small subset of variables (coordinates of $\mathbb{R}^p$) that generate the principal subspace. Since $\|\cdot\|_{2,q}$ is coordinate-independent, *every* orthonormal basis of a $\mathcal{G} \in \mathcal{M}_q(R_q)$ has the same $(2,q)$-norm.

Another related notion of subspace sparsity is *column sparsity*, which asserts that there is *some* orthonormal basis of sparse vectors that spans the principal subspace. Define the $(*,q)$-norm, $q \in [0,\infty]$, of a $p \times d$ matrix A as the maximal ℓ_q norm of its columns:

$$\|A\|_{*,q} := \max_{1\le j\le d} \|a_{*j}\|_q,$$

where a_{*j} denotes the jth column of A. This is not coordinate-independent. We define the column sparse subspaces to be those that have some orthonormal basis with small $(*,q)$-norm.

DEFINITION (Column sparse subspaces). For $0 \le q < 2$ and $1 \le R_q \le p^{1-q/2}$,

$$\mathcal{M}_q^*(R_q) := \begin{cases} \{\operatorname{col}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{*,q}^q \le R_q\}, & \text{if } 0 < q < 2 \quad \text{and} \\ \{\operatorname{col}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{*,0} \le R_0\}, & \text{if } q = 0. \end{cases}$$

The column sparse subspaces are the d-dimensional subspaces that have some orthonormal basis whose vectors are ℓ_q sparse in the usual sense. Unlike row sparsity, the orthonormal bases of a column sparse $\mathcal{G}$ do not all have the same $(*, q)$-norm, but if $\mathcal{G} \in \mathcal{M}_q^*(R_q)$, then there exists some $U \in \mathbb{V}_{p,d}$ such that $\mathcal{G} = \mathrm{col}(U)$ and $\|U\|_{*,q}^q \le R_q$ (or $\|U\|_{*,q} \le R_q$ for $q = 0$).

2.2. *Parameter space.* We assume that there exist i.i.d. random vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$, with $\mathbb{E}Z_1 = 0$ and $\mathrm{Var}(Z_1) = I_p$, such that

$$X_i = \mu + \Sigma^{1/2} Z_i \quad \text{and} \quad \|Z_i\|_{\psi_2} \le 1 \tag{2.4}$$

for $i = 1, \ldots, n$, where $\|\cdot\|_{\psi_\alpha}$ is the Orlicz ψ_α-norm [e.g., van der Vaart and Wellner (1996), Chapter 2] defined for $\alpha \ge 1$ as

$$\|Z\|_{\psi_\alpha} := \sup_{b:\|b\|_2 \le 1} \inf\left\{C > 0 : \mathbb{E}\exp\left|\frac{\langle Z, b\rangle}{C}\right|^\alpha \le 2\right\}.$$

This ensures that all one-dimensional marginals of X_i have sub-Gaussian tails. We also assume that the eigengap $\lambda_d - \lambda_{d+1} > 0$ so that the principal subspace $\mathcal{S}$ is well defined. Intuitively, $\mathcal{S}$ is harder to estimate when the eigengap is small. This is made precise by the *effective noise variance*

$$\sigma_d^2(\lambda_1, \ldots, \lambda_p) := \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2}. \tag{2.5}$$

It turns out that this is a key quantity in the estimation of $\mathcal{S}$, and that it is analogous to the noise variance in linear regression. Let

$$\mathcal{P}_q(\sigma^2, R_q)$$

denote the class of distributions on $X_1, \ldots, X_n$ that satisfy (2.4), $\sigma_d^2 \le \sigma^2$, and $\mathcal{S} \in \mathcal{M}_q(R_q)$. Similarly, let

$$\mathcal{P}_q^*(\sigma^2, R_q)$$

denote the class of distributions that satisfy (2.4), $\sigma_d^2 \le \sigma^2$, and $\mathcal{S} \in \mathcal{M}_q^*(R_q)$.

2.3. *Subspace distance.* A notion of distance between subspaces is necessary to measure the performance of a principal subspace estimator. The *canonical angles* between subspaces generalize the notion of angles between lines and can be used to define subspace distances. There are several equivalent ways to describe canonical angles, but for our purposes it will be easiest to describe them in terms of projection matrices. See Bhatia [(1997), Chapter VII.1] and Stewart and Sun (1990) for additional background on canonical angles. For a subspace $\mathcal{E} \in \mathbb{G}_{p,d}$ and its orthogonal projector E, we write $E^\perp$ to denote the orthogonal projector of $\mathcal{E}^\perp$ and recall that $E^\perp = I_p - E$.

DEFINITION. Let $\mathcal{E}$ and $\mathcal{F}$ be d-dimensional subspaces of $\mathbb{R}^p$ with orthogonal projectors E and F. Denote the singular values of $EF^{\perp}$ by $s_1 \geq s_2 \geq \cdots$. The *canonical angles* between $\mathcal{E}$ and $\mathcal{F}$ are the numbers

$$\theta_k(\mathcal{E}, \mathcal{F}) = \arcsin(s_k)$$

for $k = 1, \ldots, d$ and the *angle operator* between $\mathcal{E}$ and $\mathcal{F}$ is the $d \times d$ matrix

$$\Theta(\mathcal{E}, \mathcal{F}) = \operatorname{diag}(\theta_1, \ldots, \theta_d).$$

In this paper we will consider the distance between subspaces $\mathcal{E}, \mathcal{F} \in \mathbb{G}_{p,d}$

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm. This distance is indeed a metric on $\mathbb{G}_{p,d}$ [see Stewart and Sun (1990), e.g.], and can be connected to the familiar Frobenius (squared error) distance between projection matrices by the following fact from matrix perturbation theory.

PROPOSITION 2.1 [See Stewart and Sun (1990), Theorem I.5.5]. *Let $\mathcal{E}$ and $\mathcal{F}$ be d-dimensional subspaces of $\mathbb{R}^p$ with orthogonal projectors E and F. Then*:

1. *The singular values of $EF^{\perp}$ are*

$$s_1, s_2, \ldots, s_d, 0, \ldots, 0.$$

2. *The singular values of $E - F$ are*

$$s_1, s_1, s_2, s_2, \ldots, s_d, s_d, 0, \ldots, 0.$$

In other words, $EF^{\perp}$ has at most d nonzero singular values and the nonzero singular values of $E - F$ are the nonzero singular values of $EF^{\perp}$, each counted twice.

Thus,

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F^2 = \|EF^{\perp}\|_F^2 = \tfrac{1}{2}\|E - F\|_F^2 = \|E^{\perp}F\|_F^2. \tag{2.6}$$

We will frequently use these identities. For simplicity, we will overload notation and write

$$\sin(U_1, U_2) := \sin\Theta\bigl(\operatorname{col}(U_1), \operatorname{col}(U_2)\bigr)$$

for $U_1, U_2 \in \mathbb{V}_{p,d}$. We also use a similar convention for $\sin(E, F)$, where E, F are the orthogonal projectors corresponding to $\mathcal{E}, \mathcal{F} \in \mathbb{G}_{p,d}$ The following proposition, proved in Appendix C, relates the subspace distance to the ordinary Euclidean distance between orthonormal matrices.

PROPOSITION 2.2. *If $V_1, V_2 \in \mathbb{V}_{p,d}$, then*

$$\frac{1}{2}\inf_{Q \in \mathbb{V}_{d,d}} \|V_1 - V_2 Q\|_F^2 \leq \|\sin(V_1, V_2)\|_F^2 \leq \inf_{Q \in \mathbb{V}_{d,d}} \|V_1 - V_2 Q\|_F^2.$$

In other words, the distance between two subspaces is equivalent to the minimal distance between their orthonormal bases.

2.4. *Sparse subspace estimators.* Here we introduce an estimator that achieves the optimal (up to a constant factor) minimax error for row sparse subspace estimation. To estimate a row sparse subspace, it is natural to consider the empirical minimization problem corresponding to (2.1) with an additional sparsity constraint corresponding to $\mathcal{M}_q(R_q)$.

We define the row sparse principal subspace estimator to be a solution of the following constrained optimization problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\|(I_p - \Pi_{\mathcal{G}})(X_i - \bar{X})\|_2^2 \\ &\text{subject to} \quad \mathcal{G} \in \mathcal{M}_q(R_q). \end{aligned} \tag{2.7}$$

For our analysis, it is more convenient to work on the Stiefel manifold. Let $\langle A, B\rangle := \operatorname{trace}(A^T B)$ for matrices A, B of compatible dimension. It is straightforward to show that following optimization problem is equivalent to (2.7):

$$\begin{aligned} &\text{maximize} \quad \langle S_n, UU^T\rangle \\ &\text{subject to} \quad U \in \mathbb{V}_{p,d} \\ &\qquad\qquad \|U\|_{2,q}^q \le R_q \qquad (\text{or } \|U\|_{2,0} \le R_0 \text{ if } q = 0). \end{aligned} \tag{2.8}$$

If $\hat{V}$ is a global maximizer of (2.8), then $\operatorname{col}(\hat{V})$ is a solution of (2.7). When $q = 1$, the estimator defined by (2.8) is essentially a generalization to subspaces of the Lasso-type sparse PCA estimator proposed by Jolliffe, Trendafilov and Uddin (2003). A similar idea has also been used by Chen, Zou and Cook (2010) in the context of sufficient dimension reduction. The constraint set in (2.8) is clearly nonconvex, however this is unimportant, because the objective function is convex and we know that the maximum of a convex function over a set D is unaltered if we replace D by its convex hull. Thus, (2.8) is equivalent to a convex *maximization* problem. Finding a global maximum of convex maximization problems is computationally challenging and efficient algorithms remain to be developed. Nevertheless, in the most popular case $q = 1$, some algorithms have been proposed with promising empirical performance [Shen and Huang (2008), Witten, Tibshirani and Hastie (2009)].

We define the column sparse principal subspace estimator analogously to the row sparse principal subspace estimator, using the column sparse subspaces $\mathcal{M}_q^*(R_q)$ instead of the row sparse ones. This leads to the following equivalent Grassmann and Stiefel manifold optimization problems:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\|(I_p - \Pi_{\mathcal{G}})(X_i - \bar{X})\|_2^2 \\ &\text{subject to} \quad \mathcal{G} \in \mathcal{M}_q^*(R_q) \end{aligned} \tag{2.9}$$

and

$$\text{(2.10)}\qquad \begin{aligned} &\text{maximize} && \langle S_n, UU^T \rangle \\ &\text{subject to} && U \in \mathbb{V}_{p,d} \\ & && \|U\|_{*,q}^q \le R_q \qquad (\text{or } \|U\|_{*,0} \le R_0 \text{ if } q = 0) \end{aligned}$$

3. Main results. In this section, we present our main results on the minimax lower and upper bounds on sparse principal subspace estimation over the row sparse and column sparse classes.

3.1. *Row sparse lower bound.* To highlight the key results with minimal assumptions, we will first consider the simplest case where $q = 0$. Consider the following two conditions.

CONDITION 1. There is a constant $M > 0$ such that

$$(R_q - d)\left[\frac{\sigma^2}{n}\left(d + \log \frac{(p-d)^{1-q/2}}{R_q - d}\right)\right]^{1-q/2} \le M.$$

CONDITION 2. $4 \le p - d$ and $2d \le R_q - d \le (p-d)^{1-q/2}$.

Condition 1 is necessary for the existence of a consistent estimator (see Theorems A.1 and A.2). Without Condition 1, the statements of our results would be complicated by multiple cases to deal with the fact that the subspace distance is bounded above by $\sqrt{d}$. The lower bounds on $p - d$ and $R_q - d$ are minor technical conditions that ensure our nonasymptotic bounds are nontrivial. Similarly, the upper bound on $R_q - d$ is only violated in trivial cases (detailed discussion given below).

THEOREM 3.1 (Row sparse lower bound, $q = 0$). *If Conditions 1 and 2 hold, then*

$$\inf_{\hat{\mathcal{S}}} \sup_{\mathcal{P}_0(\sigma^2, R_0)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \ge c(R_0 - d)\frac{\sigma^2}{n}\left[d + \log \frac{p-d}{R_0 - d}\right].$$

Here, as well as in the entire paper, c denotes a universal, positive constant, not necessarily the same at each occurrence. This lower bound result reflects two separate aspects of the estimation problem: *variable selection* and *parameter estimation after variable selection*. Variable selection refers to finding the variables that generate the principal subspace, while estimation refers to estimating the subspace after selecting the variables. For each variable, we accumulate two types of errors: one proportional to d that reflects the coordinates of the variable in the

d-dimensional subspace, and one proportional to $\log[(p-d)/(R_0-d)]$ that reflects the cost of searching for the R_0 active variables. We prove Theorem 3.1 in Appendix A.

The nonasymptotic lower bound for $0<q<2$ has a more complicated dependence on (n, p, d, R_q, σ^2) because of the interaction between ℓ_q and ℓ_2 norms. Therefore, our main lower bound result for $0<q<2$ will focus on combinations of (n, p, d, R_q, σ^2) that correspond to the high-dimensional and sparse regime. (We state more general lower bound results in Appendix A.) Let

$$T := \frac{R_q - d}{(p-d)^{1-q/2}} \quad \text{and} \quad \gamma := \frac{(p-d)\sigma^2}{n}. \tag{3.1}$$

The interpretation for these two quantities is natural. First, T measures the relative sparsity of the problem. Roughly speaking, it ranges between 0 and 1 when the sparsity constraint in (2.8) is active, though the "sparse" regime generally corresponds to $T \ll 1$. The second quantity, γ corresponds to the classic mean squared error (MSE) of standard PCA. The problem is low-dimensional if γ is small compared to T. We impose the following condition to preclude this case.

CONDITION 3. There is a constant $a<1$ such that $T^a \le \gamma^{q/2}$.

This condition lower bounds the classic MSE in terms of the sparsity and is mild in high-dimensional situations. When $a = q/2$, for example, Condition 3 reduces to

$$R_q - d \le \frac{\sigma^2}{n}(p-d)^{2-q/2}.$$

We also note that this assumption becomes milder for larger values of a and it is related to conditions in other minimax inference problem involving ℓ_p and ℓ_q balls [see Donoho and Johnstone (1994), e.g.].

THEOREM 3.2 (Row sparse lower bound, $0<q<2$). *Let $q \in (0,2)$. If Conditions 1 to 3 hold, then*

$$\inf_{\hat{\mathcal{S}}} \sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \ge c(R_q - d)\left\{\frac{\sigma^2}{n}\left[d + \log\frac{(p-d)^{1-q/2}}{R_q - d}\right]\right\}^{1-q/2}.$$

This result generalizes Theorem 3.1 and reflects the same combination of variable selection and parameter estimation. When Condition 3 does not hold, the problem is outside of the sparse, high-dimensional regime. As we show in the proof, there is actually a "phase transition regime" between the high-dimensional sparse and the classic dense regimes for which sharp minimax rate remains unknown. A similar phenomenon has been observed in Birnbaum et al. (2013).

3.2. *Row sparse upper bound.* Our upper bound results are obtained by analyzing the estimators given in Section 2.4. The case where $q = 0$ is the clearest, and we begin by stating a weaker, but simpler minimax upper bound for the row sparse class.

THEOREM 3.3 (Row sparse upper bound, $q = 0$). *Let $\hat{\mathcal{S}}$ be any global maximizer of (2.7). If $6\sqrt{R_0(d + \log p)} \le \sqrt{n}$, then*

$$\sup_{\mathcal{P}_0(\sigma^2, R_0)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le cR_0 \frac{\lambda_1}{\lambda_{d+1}} \frac{\sigma^2(d + \log p)}{n}.$$

Although (2.7) may not have a unique global optimum, Theorem 3.3 shows that *any* global optimum will be within a certain radius of the principal subspace $\mathcal{S}$. The proof of Theorem 3.3, given in Section 4.2, is relatively simple but still nontrivial. It also serves as a prototype for the much more involved proof of our main upper bound result stated in Theorem 3.4 below. We note that the rate given by Theorem 3.3 is off by a λ_1/λ_{d+1} factor that is due to the specific approach taken to control an empirical process in our proof of Theorem 3.3.

To state the main upper bound result with optimal dependence on (n, p, d, R_q, σ^2), we first describe some regularity conditions. Let

$$\varepsilon_n := \sqrt{2} R_q^{1/2} \left(\frac{d + \log p}{n} \right)^{1/2 - q/4}.$$

The regularity conditions are

$$\varepsilon_n \le 1, \tag{3.2}$$

$$c_1 \sqrt{\frac{d}{n}} \log n \lambda_1 + c_3 \varepsilon_n (\log n)^{5/2} \lambda_{d+1} < \frac{1}{2}(\lambda_d - \lambda_{d+1}), \tag{3.3}$$

$$c_3 \varepsilon_n (\log n)^{5/2} \lambda_{d+1} \le \sqrt{\lambda_1 \lambda_{d+1}}^{1-q/2} (\lambda_d - \lambda_{d+1})^{q/2} \tag{3.4}$$

and

$$c_3 \varepsilon_n^2 (\log n)^{5/2} \lambda_{d+1} \le \sqrt{\lambda_1 \lambda_{d+1}}^{2-q} (\lambda_d - \lambda_{d+1})^{-(1-q)}, \tag{3.5}$$

where c_1 and c_3 are positive constants involved in the empirical process arguments. Equations (3.2) to (3.5) require that ε_n, the minimax rate of estimation (except the factor involving λ), to be small enough, compared to empirical process constants and some polynomials of λ. Such conditions are mild in the high dimensional, sparse regime, since to some extent, they are qualitatively similar and analogous to Conditions 1 to 3 required by the lower bound.

REMARK 1. Conditions (3.2) to (3.5) are general enough to allow R_q, d and λ_j ($j = 1, d, d + 1$) to scale with n. For example, consider the case $q = 0$, and let

$d = n^a$, $R_0 = n^b$, $p = n^c$, $\lambda_1 = n^{r_1}$, $\lambda_d = n^{r_2}$, $\lambda_{d+1} = n^{r_3}$, where $0 < a < b < c$, and $r_1 \ge r_2 > r_3$. Note that the r_j's can be negative. Then it is straightforward to verify that conditions (3.2) to (3.5) hold for large values of n whenever $a + b < 1$ and $r_1 < r_2 + (1 - a)/2$. Condition (3.2) implies that d cannot grow faster than $\sqrt{n}$.

THEOREM 3.4 (Row sparse upper bound in probability). *Let $q \in [0, 1]$ and $\hat{\mathcal{S}}$ be any solution of* (2.7). *If $(X_1, \dots, X_n) \sim \mathbb{P} \in \mathcal{P}_q(\sigma^2, R_q)$ and* (3.2) *to* (3.5) *hold, then*

$$\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le c R_q \left(\frac{\sigma^2(d + \log p)}{n}\right)^{1-q/2}$$

with probability at least $1 - 4/(n-1) - 6\log n/n - p^{-1}$.

Theorem 3.4 is presented in terms of a probability bound instead of an expectation bound. This stems from technical aspects of our proof that involve bounding the supremum of an empirical process over a set of random diameter. For $q \in [0, 1]$, the upper bound matches our lower bounds (Theorems 3.1 and 3.2) for the entire tuple (n, p, d, R_q, σ^2) up to a constant if

$$R_q^{2/(2-q)} \le p^c \tag{3.6}$$

for some constant $c < 1$. To see this, combining this additional condition and Condition 2, the term $\log \frac{p^{1-q/2}}{R_q}$ in the lower bound given in Theorem 3.2 is within a constant factor of $\log p$ in the upper bound given in Theorem 3.4. It is straightforward to check that the other terms in lower and upper bounds agree up to constants with obvious correspondence. Moreover, we note that the additional condition (3.6) is only slightly stronger than the last inequality in Condition 2. The proof of Theorem 3.4 is in Appendix B.1.

Using the probability upper bound result and the fact that $\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le d$, one can derive an upper bound in expectation.

COROLLARY 3.1. *Under the same condition as in Theorem* 3.4, *we have for some constant c,*

$$\mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le c\left\{R_q\left[\frac{\sigma^2(d + \log p)}{n}\right]^{1-q/2} + d\left(\frac{\log n}{n} + \frac{1}{p}\right)\right\}.$$

REMARK 2. The expectation upper bound has an additional $d(\log n/n + 1/p)$ term that can be further reduced by refining the argument (see Remark 3 below). It is not obvious if one can completely avoid such a term. But in many situations it is dominated by the first term. Again, we invoke the scaling considered in Remark 1. When $q = 0$, the first term is of order $n^{a+b+(r_1+r_3)/2-r_2-1}$, and the additional term is $n^{a-1}\log n + n^{a-c}$, which is asymptotically negligible if $b > r_2 - (r_1 + r_3)/2 + (1 - c)_+$.

REMARK 3. Given any $r > 0$, it is easy to modify the proof of Theorem 3.4 [as well as conditions (3.2) to (3.5)] such that the results of Theorem 3.4 and Corollary 3.1 hold with c replaced by some constant $c(r)$, and the probability bound becomes $1 - 4/(n^r - 1) - 6 \log n/n^r - 1/p^r$.

3.3. *Column sparse lower bound.* By modifying the proofs of Theorems 3.1 and 3.2, we can obtain lower bound results for the column sparse case that are parallel to the row sparse case. For brevity, we present the $q = 0$ and $q > 0$ cases together. The analog of T, the degree of sparsity, for the column sparse case is

$$T_* := \frac{d(R_q - 1)}{(p - d)^{1-q/2}}, \tag{3.7}$$

and the analogs of Conditions 2 and 3 are the following.

CONDITION 4. $4d \le p - d$ and $d \le d(R_q - 1) \le (p - d)^{1-q/2}$.

CONDITION 5. There is a constant $a < 1$ such that $T_*^a \le \gamma^{q/2}$.

THEOREM 3.5 (Column sparse lower bound). *Let $q \in [0, 2)$. If Conditions 4 and 5 hold, then*

$$\inf_{\hat{\mathcal{S}}} \sup_{\mathcal{P}_q^*(\sigma^2, R_q)} \mathbb{E}\|\sin(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \ge c d(R_q - 1)\left\{\frac{\sigma^2}{n}\left[1 + \log \frac{(p - d)^{1-q/2}}{d(R_q - 1)}\right]\right\}^{1-q/2}.$$

For column sparse subspaces, the lower bound is dominated by the variable selection error, because column sparsity is defined in terms of the maximal ℓ_0 norms of the vectors in an orthonormal basis and R_0 variables must be selected for each of the d vectors. So the variable selection error is inflated by a factor of d, and hence becomes the dominating term in the total estimation error. We prove Theorem 3.5 in Appendix A.

3.4. *Column sparse upper bound.* A specific challenge in analyzing the column sparse principal subspace problem (2.10) is to bound the supremum of the empirical process

$$\langle S_n - \Sigma, UU^T - VV^T \rangle$$

indexed by all $U \in \mathcal{U}(p, d, R_q, \varepsilon)$ where

$$\mathcal{U}(p, d, R_q, \varepsilon) \equiv \{U : \mathbb{V}_{p,d}, \|U\|_{*,q}^q \le R_q, \|UU^T - VV^T\|_F \le \varepsilon\}.$$

Unlike the row sparse matrices, the matrices UU^T and VV^T are no longer column sparse with the same radius R_q.

By observing that $\mathcal{M}_q^*(R_q) \subseteq \mathcal{M}_q(dR_q)$, we can reuse the proof of Theorem 3.4 to derive the following upper bound for the column sparse class.

COROLLARY 3.2 (Column sparse upper bound). *Let $q \in [0, 1]$ and $\hat{\mathcal{S}}$ be any solution of* (2.9). *If $(X_1, \dots, X_n) \sim \mathbb{P} \in \mathcal{P}_q^*(\sigma^2, R_q)$ and* (3.2) *to* (3.5) *hold with R_q replaced by dR_q, then*

$$\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le c d R_q \left(\frac{\sigma^2 (d + \log p)}{n} \right)^{1-q/2}$$

with probability at least $1 - 4/(n-1) - 6 \log n / n - p^{-1}$.

Corollary 3.2 is slightly weaker than the corresponding result for the row sparse class. It matches the lower bound in Theorem 3.5 up to a constant if

$$\big(d(R_q - 1)\big)^{2/(2-q)} \le p^c$$

for some constant $c < 1$, and $d < C \log p$ for some other constant C.

3.5. *A conjecture for the column sparse case.* Note that Theorem 3.5 and Corollary 3.2 only match when $d \le C \log p$. For larger values of d, we believe that the lower bound in Theorem 3.5 is optimal and the upper bound can be improved.

CONJECTURE (Minimax error bound for column sparse case). Under the same conditions as in Corollary 3.2, there exists an estimator $\hat{\mathcal{S}}$ such that

$$\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le c d R_q \left(\frac{\sigma^2 (1 + \log p)}{n} \right)^{1-q/2}$$

with high probability. As a result, the optimal minimax lower and upper bounds for this case shall be

$$\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \asymp d R_q \left(\frac{\sigma^2 \log p}{n} \right)^{1-q/2}.$$

One reason for the conjecture is based on the following intuition. Suppose that $\lambda_1 > \lambda_2 > \cdots > \lambda_d > \lambda_{d+1}$ (there is enough gap between the leading eigenvalues) one can recover the individual leading eigenvectors with an error rate whose dependence on (n, R_q, p) is the same as in the lower bound [cf. Birnbaum et al. (2013), Vu and Lei (2012a)]. As a result, the estimator $\hat{V} = (\hat{v}_1, \hat{v}_2, \dots, \hat{v}_d)$ shall give the desired upper bound. On the other hand, it remains open to us whether the estimator in (2.10) can achieve this rate for d much larger than $\log p$.

4. Sketch of proofs. For simplicity, we focus on the row sparse case with $q = 0$, assuming also the high dimensional and sparse regime. For more general cases, see Theorems A.1 and A.2 in Appendix A.

4.1. *The lower bound*. Our proof of the lower bound features a combination of the general framework of the Fano method and a careful combinatorial analysis of packing sets of various classes of sparse matrices. The particular challenge is to construct a rich packing set of the parameter space $\mathcal{P}_q(\sigma^2, R_q)$. We will consider centered p-dimensional Gaussian distributions with covariance matrix Σ given by

$$\Sigma(A) = bAA^T + I_p, \tag{4.1}$$

where $A \in \mathbb{V}_{p,d}$ is constructed from the "local Stiefel embedding" as given below. Let $1 \le k \le d < p$ and the function $A_\varepsilon : \mathbb{V}_{p-d,k} \mapsto \mathbb{V}_{p,d}$ be defined in block form as

$$A_\varepsilon(J) = \begin{bmatrix} (1-\varepsilon^2)^{1/2} I_k & 0 \\ 0 & I_{d-k} \\ \varepsilon J & 0 \end{bmatrix} \tag{4.2}$$

for $0 \le \varepsilon \le 1$. We have the following generic method for lower bounding the minimax risk of estimating the principal subspace of a covariance matrix. It is proved in Appendix A as a consequence of Lemmas A.1 to A.3.

LEMMA 4.1 (Fano method with Stiefel embedding). *Let* $\varepsilon \in [0, 1]$ *and* $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{p-d,k}$ *for* $1 \le k \le d < p$. *For each* $i = 1, \dots, N$, *let* $\mathbb{P}_i$ *be the* n-*fold product of the* $\mathcal{N}(0, \Sigma(A_\varepsilon(J_i)))$ *probability measure, where* $\Sigma(\cdot)$ *is defined in* (4.1) *and* $A_\varepsilon(\cdot)$ *is defined in* (4.2). *If*

$$\min_{i \ne j} \|J_i - J_j\|_F \ge \delta_N,$$

then every estimator $\hat{\mathcal{A}}$ *of* $\mathcal{A}_i := \operatorname{col}(A_\varepsilon(J_i))$ *satisfies*

$$\max_i \mathbb{E}_i \|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F \ge \frac{\delta_N \varepsilon \sqrt{1-\varepsilon^2}}{2} \left[1 - \frac{4nk\varepsilon^2/\sigma^2 + \log 2}{\log N}\right],$$

where $\sigma^2 = (1+b)/b^2$.

Note that if $\|J\|_{2,0} \le R_0 - d$, then $\|A_\varepsilon(J)\|_{2,0} \le R$. Thus Lemma 4.1 with appropriate choices of J_i can yield minimax lower bounds over p-dimensional Gaussian distributions whose principal subspace is R_0 row sparse.

The remainder of the proof consists of two applications of Lemma 4.1 that correspond to the two terms in Theorem 3.1. In the first part, we use a variation of the Gilbert–Varshamov bound (Lemma A.5) to construct a packing set in $\mathbb{V}_{p-d,1}$ consisting of $(R_0 - d)$-sparse vectors. Then we apply Lemma 4.1 with

$$k = 1, \qquad \delta_N = 1/4, \qquad \varepsilon^2 \asymp \frac{\sigma^2 R_0 \log p}{n}.$$

This yields a minimax lower bound that reflects the variable selection complexity. In the second part, we leverage existing results on the metric entropy of the

Grassmann manifold (Lemma A.6) to construct a packing set of $\mathbb{V}_{R_0-d,d}$. Then we apply Lemma 4.1 with

$$k = d, \qquad \delta_N = c_0\sqrt{d}/e, \qquad \varepsilon^2 \asymp \frac{\sigma^2 R_0}{n}.$$

This yields a minimax lower bound that reflects the complexity of post-selection estimation. Putting these two results together, we have for a subset of Gaussian distributions $G \subseteq \mathcal{P}_0(\sigma^2, R_q)$ the minimax lower bound:

$$\max_G \mathbb{E}\|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F^2 \geq cR_0\frac{\sigma^2}{n}(\log p \wedge d) \geq (c/2)R_0\frac{\sigma^2}{n}(d + \log p).$$

4.2. *The upper bound.* The upper bound proof requires a careful analysis of the behavior of the empirical maximizer of the PCA problem under sparsity constraints. The first key ingredient is to provide a lower bound of the curvature of the objective function at its global maxima. Traditional results of this kind, such as Davis–Kahan sinΘ theorem and Weyl's inequality, are not sufficient for our purpose.

The following lemma, despite its elementary form, has not been seen in the literature (to our knowledge). It gives us the right tool to bound the curvature of the matrix functional $F \mapsto \langle A, F\rangle$ at its point of maximum on the Grassmann manifold.

LEMMA 4.2 (Curvature lemma). *Let A be a $p \times p$ positive semidefinite matrix and suppose that its eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_p(A)$ satisfy $\lambda_d(A) > \lambda_{d+1}(A)$ for $d < p$. Let $\mathcal{E}$ be the d-dimensional subspace spanned by the eigenvectors of A corresponding to its d largest eigenvalues*, *and let E denote its orthogonal projector. If $\mathcal{F}$ is a d-dimensional subspace of $\mathbb{R}^p$ and F is its orthogonal projector, then*

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F^2 \leq \frac{\langle A, E - F\rangle}{\lambda_d(A) - \lambda_{d+1}(A)}.$$

Lemma 4.2 is proved in Appendix C.2. An immediate corollary is the following alternative to the traditional matrix perturbation approach to bounding subspace distances using the Davis–Kahan sinΘ theorem and Weyl's inequality.

COROLLARY 4.1 (Variational sinΘ). *In addition to the hypotheses of Lemma 4.2, if B is a symmetric matrix and F satisfies*

$$\langle B, E\rangle - g(E) \leq \langle B, F\rangle - g(F) \tag{4.3}$$

for some function $g : \mathbb{R}^{p\times p} \mapsto \mathbb{R}$, then

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F^2 \leq \frac{\langle B - A, F - E\rangle - [g(F) - g(E)]}{\lambda_d(A) - \lambda_{d+1}(A)}. \tag{4.4}$$

The corollary is different from the Davis–Kahan $\sin\Theta$ theorem because the orthogonal projector F does not have to correspond to a subspace spanned by eigenvectors of B. F only has to satisfy

$$\langle B, E\rangle - g(E) \le \langle B, F\rangle - g(F).$$

This condition is suited ideally for analyzing solutions of regularized and/or constrained maximization problems where E and F are feasible, but F is optimal. In the simplest case, where $g \equiv 0$, combining (4.4) with the Cauchy–Schwarz inequality and (2.6) recovers a form of the Davis–Kahan $\sin\Theta$ theorem in the Frobenius norm:

$$\frac{1}{\sqrt{2}}\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F \le \frac{\|B - A\|_F}{\lambda_d(A) - \lambda_{d+1}(A)}.$$

In the upper bound proof, let $V \in \mathbb{V}_{p,d}$ be the true parameter, and $\hat{V}$ be a solution of (2.8). Then we have

$$\langle S_n, \hat{V}\hat{V}^T - VV^T\rangle \ge 0.$$

Applying Corollary 4.1 with $B = S_n$, $A = \Sigma$, $E = VV^T$, $F = \hat{V}\hat{V}^T$, and $g \equiv 0$, we have

$$\|\sin\Theta(V, \hat{V})\|_F^2 \le \frac{\langle S_n - \Sigma, \hat{V}\hat{V}^T - VV^T\rangle}{\lambda_d(\Sigma) - \lambda_{d+1}(\Sigma)}. \tag{4.5}$$

Obtaining a sharp upper bound for $\langle S - \Sigma, \hat{V}\hat{V}^T - VV^T\rangle$ is nontrivial. First, one needs to control $\sup_{F\in\mathcal{F}}\langle S - \Sigma, F\rangle$ for some class $\mathcal{F}$ of sparse and symmetric matrices. This requires some results on quadratic form empirical process. Second, in order to obtain better bounds, we need to take advantage of the fact that $\hat{V}\hat{V}^T - VV^T$ is probably small. Thus, we need to use a peeling argument to deal with the case where $\mathcal{F}$ has a random (but probably) small diameter. These details are given in Appendices B.1 and D. Here we present a short proof of Theorem 3.3 to illustrate the idea.

PROOF OF THEOREM 3.3. By (4.5), we have

$$\hat{\varepsilon}^2 := \|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le \frac{\langle S_n - \Sigma, \hat{V}\hat{V}^T - VV^T\rangle}{\lambda_d - \lambda_{d+1}}$$

and

$$\hat{\varepsilon}^2 \le \frac{\sqrt{2}}{\lambda_d - \lambda_{d+1}}\left\langle S_n - \Sigma \frac{\hat{V}\hat{V}^T - VV^T}{\|\hat{V}\hat{V}^T - VV^T\|_F}\right\rangle\hat{\varepsilon}, \tag{4.6}$$

because $\|\hat{V}\hat{V}^T - VV^T\|_F^2 = 2\hat{\varepsilon}^2$ by (2.6). Let

$$\Delta = \frac{\hat{V}\hat{V}^T - VV^T}{\|\hat{V}\hat{V}^T - VV^T\|_F}.$$

Then $\|\Delta\|_{2,0} \le 2R_0$, $\|\Delta\|_F = 1$, and Δ has at most d positive eigenvalues and at most d negative eigenvalues (see Proposition 2.1). Therefore, we can write $\Delta = AA^T - BB^T$ where $\|A\|_{2,0} \le 2R_0$, $\|A\|_F \le 1$, $A \in \mathbb{R}^{p\times d}$, and the same holds for B. Let

$$\mathcal{U}(R_0) = \{U \in \mathbb{R}^{p\times d} : \|U\|_{2,0} \le 2R_0 \text{ and } \|U\|_F \le 1\}.$$

Equation (4.6) implies

$$\mathbb{E}\hat{\varepsilon} \le \frac{2\sqrt{2}}{\lambda_d - \lambda_{d+1}} \mathbb{E} \sup_{U\in\mathcal{U}(R_0)} |\langle S_n - \Sigma, UU^T\rangle|.$$

The empirical process $\langle S_n - \Sigma, UU^T\rangle$ indexed by U is a generalized quadratic form, and a sharp bound of its supremum involves some recent advances in empirical process theory due to Mendelson (2010) and extensions of his results. By Corollary 4.1 of Vu and Lei (2012b), we have

$$\begin{aligned} &\mathbb{E} \sup_{U\in\mathcal{U}(R_0)} |\langle S_n - \Sigma, UU^T\rangle| \\ &\quad \le c\lambda_1 \left\{ \frac{\mathbb{E}\sup_{U\in\mathcal{U}(R_0)}\langle \mathcal{Z}, U\rangle}{\sqrt{n}} + \left(\frac{\mathbb{E}\sup_{U\in\mathcal{U}(R_0)}\langle \mathcal{Z}, U\rangle}{\sqrt{n}}\right)^2 \right\}, \end{aligned}$$

where $\mathcal{Z}$ is a $p \times d$ matrix of i.i.d. standard Gaussian variables. To control $\mathbb{E}\sup_{U\in\mathcal{U}}\langle \mathcal{Z}, U\rangle$, note that

$$\langle \mathcal{Z}, U\rangle \le \|\mathcal{Z}\|_{2,\infty}\|U\|_{2,1} \le \|\mathcal{Z}\|_{2,\infty}\sqrt{2R_0},$$

because $U \in \mathcal{U}(R_0)$. Using a standard δ-net argument (see Propositions D.1 and D.2), we have, when $p > 5$,

$$\big\| \|\mathcal{Z}\|_{2,\infty} \big\|_{\psi_2} \le 4.15\sqrt{d + \log p} \tag{4.7}$$

and hence

$$\mathbb{E} \sup_{U\in\mathcal{U}} \langle \mathcal{Z}, U\rangle \le 6\sqrt{R_0(d + \log p)}.$$

The proof is complete since we assume that $6\sqrt{R_0(d + \log p)} \le \sqrt{n}$. □

5. Discussion. There is a natural correspondence between the sparse principal subspace optimization problem (2.7) and some optimization problems considered in the sparse regression literature. We have also found that there is a correspondence between minimax results for sparse regression and those that we presented in this article. In spite of these connections, results on computation for sparse principal subspaces (and sparse PCA) are far less developed than for sparse regression. In this final section, we will discuss the connections with sparse regression, both optimization and minimax theory, and then conclude with some open problems for sparse principal subspaces.

5.1. *Connections with sparse regression.* Letting $\tilde{X}_i = X_i - \bar{X}$ denote a centered observation, we can write (2.7) in the $d = 1$ case as an equivalent penalized regression problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\|\tilde{X}_i - uu^T\tilde{X}_i\|_2^2 + \tau_q\|u\|_q^q \\ &\text{subject to} \quad u \in \mathbb{R}^p \quad \text{and} \quad \|u\|_2 = 1 \end{aligned}$$

for $0 < q \le 1$ and similarly for $q = 0$. The penalty parameter $\tau_q \ge 0$ plays a similar role as R_q. When $q = 1$ this is equivalent to a penalized form of the sparse PCA estimator considered in Jolliffe, Trendafilov and Uddin (2003) and it also bears similarity to the estimator considered by Shen and Huang (2008). It is also similar to the famous ℓ_1-penalized optimization often used in high-dimensional regression [Tibshirani (1996)]. In the subspace case $d > 1$, one can write an analogous penalized multivariate regression problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\|\tilde{X}_i - UU^T\tilde{X}_i\|_2^2 + \tau_q\|U\|_{2,q}^q \\ &\text{subject to} \quad U \in \mathbb{V}_{p,d} \end{aligned}$$

for $0 < q \le 1$ and similarly for $q = 0$. When $q = 1$, this corresponds to a "group Lasso" penalty where entries in the same row of U are penalized simultaneously [Yuan and Lin (2006), Zhao, Rocha and Yu (2009)]. The idea being that as τ_q varies, a variable should enter/exit all d coordinates simultaneously. In the column sparse case, when $q = 1$ the analogous penalized multivariate regression problem has a penalty which encourages each column of U to be sparse, but does not require that the pattern of sparsity to be the same across columns.

The analogy between row sparse principal subspace estimation and sparse regression goes beyond the optimization problems formulated above—it is also reflected in terms of the minimax rate. In the sparse regression problem, we assume an i.i.d. sample $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ for $1 \le i \le n$ satisfying

$$Y_i = \beta^T X_i + \varepsilon_i,$$

where ε_i is mean zero, independently of X_i, and $\beta \in \mathbb{R}^p$ is the regression coefficient vector. Raskutti, Wainwright and Yu (2011) showed (with some additional conditions on the distribution of X_i) that if $\|\beta\|_q^q \le R_q$ and $\operatorname{Var}\varepsilon_i \le \sigma^2$, then the minimax rate of estimating β in ℓ_2 norm is (ignoring constants)

$$\sqrt{R_q}\left(\frac{\sigma^2\log p}{n}\right)^{1/2-q/4}.$$

The estimator that achieves this error rate is obtained by solving the ℓ_q constrained least square problem. The $d > 1$ case corresponds to the multivariate regression model, where $Y_i \in \mathbb{R}^d$, $\beta \in \mathbb{R}^{p\times d}$, and $\operatorname{Var}\varepsilon_i = \sigma^2 I_d$. The results of Negahban et al. (2012), with straightforward modifications, imply that if $\|\beta\|_{2,q}^q \le R_q$, then a penalized least squares estimator can achieve the ℓ_2 error rate

$$\sqrt{R_q}\left(\frac{\sigma^2(d+\log p)}{n}\right)^{1/2-q/4},$$

agreeing with our minimax lower and upper bounds for the row sparse principal subspace problem.

5.2. *Practical concerns.* The nature of this work is theoretical and it leaves open many challenges for methodology and practice. The minimax optimal estimators that we present appear to be computationally intractable because they involve convex *maximization* rather than convex *minimization* problems. Even in the case $q = 1$, which corresponds to a subspace extension of ℓ_1 constrained PCA, the optimization problem remains challenging as there are no known algorithms to efficiently compute a global maximum.

Although the minimax optimal estimators that we propose do not require knowledge of the noise-to-signal ratio σ^2, they do require knowledge of (or an upper bound on) the sparsity R_q. It is not hard to modify our techniques to produce an estimator that gives up adaptivity to σ^2 in exchange for adaptivity to R_q. One could do this by using penalized versions of our estimators with a penalty factor proportional to σ^2. An extension along this line has already been considered by Lounici (2013) for the $d = 1$ case. A more interesting question is whether or not there exist fully adaptive principal subspace estimators.

Under what conditions can one find an estimator that achieves the minimax optimal error without requiring knowledge of either σ^2 or R_q? Works by Paul (2007) and Ma (2013) on refinements of diagonal thresholding for the spiked covariance model seems promising on this front, but as we mentioned in the Introduction, the spiked covariance model is restrictive and necessarily excludes the common practice of standardizing variables. Is it possible to be adaptive outside the spiked covariance model? One possible approach can be described in the following three steps. (1) use a conservative choice of R_q (say, p^a, for some $0 < a < 1$); (2) estimate σ^2 using eigenvalues obtained from the sparsity constrained principal subspace estimator; and (3) use a sparsity penalized principal subspace estimator with σ^2 replaced by its estimate. We will pursue this idea in further detail in future work.

APPENDIX A: LOWER BOUND PROOFS

Theorems 3.1, 3.2 and 3.5 are consequences of three more general results stated below. An essential part of the strategy of our proof is to analyze the *variable selection* and *estimation* aspects of the problem separately. We will consider two types of subsets of the parameter space that capture the essential difficulty of each aspect: one where the subspaces vary over different subsets of variables, and another where the subspaces vary over a fixed subset of variables. The first two results give lower bounds for each aspect in the row sparse case. Theorems 3.1 and 3.2 follow easily from them. The third result directly addresses the proof of Theorem 3.5.

THEOREM A.1 (Row sparse variable selection). *Let $q \in [0,2)$ and (p, d, R_q) satisfy*

$$4 \le p - d \quad \text{and} \quad 1 \le R_q - d \le (p-d)^{1-q/2}.$$

There exists a universal constant $c > 0$ such that every estimator $\hat{\mathcal{S}}$ satisfies the following. If $T < \gamma^{q/2}$, then

$$\begin{aligned}(A.1)\qquad &\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ &\quad \ge c\left\{(R_q - d)\left[\frac{\sigma^2}{n}\left(1 - \log(T/\gamma^{q/2})\right)\right]^{1-q/2} \wedge 1\right\}^{1/2}.\end{aligned}$$

Otherwise,

$$(A.2)\qquad \sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \ge c\left\{\frac{(p-d)\sigma^2}{n} \wedge 1\right\}^{1/2}.$$

The case $q = 0$ is particularly simple, because $T < \gamma^{q/2} = 1$ holds trivially. In that case, Theorem A.1 asserts that

$$\begin{aligned}(A.3)\qquad &\sup_{\mathcal{P}_0(R_0, \sigma^2)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ &\quad \ge c\left\{(R_0 - d)\frac{\sigma^2}{n}\left(1 + \log\frac{p-d}{R_q - d}\right) \wedge 1\right\}^{1/2}.\end{aligned}$$

When $q \in (0,2)$ the transition between the $T < \gamma^{q/2}$ and $T \ge \gamma^{q/2}$ regimes involves lower order (log log) terms that can be seen in (A.15). Under Condition 3, (A.1) can be simplified to

$$\begin{aligned}(A.4)\qquad &\sup_{\mathcal{P}_0(R_0, \sigma^2)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ &\quad \ge c\left\{(R_q - d)\frac{\sigma^2}{n}\left(1 + (1-a)\log\frac{(p-d)^{1-q/2}}{R_q - d}\right) \wedge 1\right\}^{1/2-q/2}.\end{aligned}$$

THEOREM A.2 (Row sparse parameter estimation). *Let $q \in [0,2)$ and (p, d, R_q) satisfy*

$$2 \le d \quad \text{and} \quad 2d \le R_q - d \le (p-d)^{1-q/2},$$

and let T and γ be defined as in (3.1). There exists a universal constant $c > 0$ such that every estimator $\hat{\mathcal{S}}$ satisfies the following. If $T < (d\gamma)^{q/2}$, then

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq c\left\{(R_q - d)\left(\frac{d\sigma^2}{n}\right)^{1-q/2} \wedge d\right\}^{1/2}. \tag{A.5}$$

Otherwise,

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq c\left\{\frac{d(p-d)\sigma^2}{n} \wedge d\right\}^{1/2}. \tag{A.6}$$

This result with (A.3) implies Theorem 3.1, and with (A.4) it implies Theorem 3.2.

THEOREM A.3 (Column sparse estimation). *Let $q \in [0, 2)$ and (p, d, R_q) satisfy*

$$4 \leq (p-d)/d \quad \textit{and} \quad d \leq d(R_q - 1) \leq (p-d)^{1-q/2},$$

and recall the definition of T_* in (3.7). There exists a universal constant $c > 0$ such that every estimator $\hat{\mathcal{S}}$ satisfies the following. If $T_* < \gamma^{q/2}$, then

$$\begin{aligned} &\sup_{\mathcal{P}^*_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ &\quad \geq c\left\{d(R_q - 1)\left[\frac{\sigma^2}{n}\left(1 - \log\left(T_*/\gamma^{q/2}\right)\right)\right]^{1-q/2} \wedge d\right\}^{1/2}. \end{aligned} \tag{A.7}$$

Otherwise,

$$\sup_{\mathcal{P}^*_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq c\left\{\frac{(p-d)\sigma^2}{n} \wedge d\right\}^{1/2}. \tag{A.8}$$

In the next section we setup a general technique, using Fano's inequality and Stiefel manifold embeddings, for obtaining minimax lower bounds in principal subspace estimation problems. Then we move on to proving Theorems A.1 and A.3.

A.1. Lower bounds for principal subspace estimation via Fano method. Our main tool for proving minimax lower bounds is the generalized Fano method. We quote the following version from Yu (1997), Lemma 3.

LEMMA A.1 (Generalized Fano method). *Let $N \geq 1$ be an integer and $\{\theta_1, \ldots, \theta_N\} \subset \Theta$ index a collection of probability measures $\mathbb{P}_{\theta_i}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Let d be a pseudometric on Θ and suppose that for all $i \neq j$*

$$d(\theta_i, \theta_j) \geq \alpha_N$$

and, the Kullback–Leibler (KL) divergence

$$D(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta_j}) \le \beta_N.$$

Then every $\mathcal{A}$-measurable estimator $\hat{\theta}$ satisfies

$$\max_i \mathbb{E}_{\theta_i} d(\hat{\theta}, \theta_i) \ge \frac{\alpha_N}{2}\left[1 - \frac{\beta_N + \log 2}{\log N}\right].$$

The calculations required for applying Lemma A.1 are tractable when $\{\mathbb{P}_{\theta_i}\}$ is a collection of multivariate Normal distributions. Let $A \in \mathbb{V}_{p,d}$ and consider the mean zero p-variate Normal distribution with covariance matrix

$$\Sigma(A) = bAA^T + I_p = (1+b)AA^T + (I_p - AA^T), \tag{A.9}$$

where $b > 0$. The noise-to-signal ratio of the principal d-dimensional subspace of these covariance matrices is

$$\sigma^2 = \frac{1+b}{b^2}$$

and can choose b to achieve any $\sigma^2 > 0$. The KL divergence between these multivariate Normal distributions has a simple, exact expression given in the following lemma. The proof is straightforward and contained in Appendix C.1.

LEMMA A.2 (KL divergence). *For $i = 1, 2$, let $A_i \in \mathbb{V}_{p,d}$, $b \ge 0$,*

$$\Sigma(A_i) = (1+b)A_i A_i^T + (I_p - A_i A_i^T),$$

and $\mathbb{P}_i$ be the n-fold product of the $\mathcal{N}(0, \Sigma(A_i))$ probability measure. Then

$$D(\mathbb{P}_1 \| \mathbb{P}_2) = \frac{nb^2}{1+b} \|\sin(A_1, A_2)\|_F^2.$$

The KL divergence between the probability measures in Lemma A.2 is equivalent to the subspace distance. In applying Lemma A.1, we will need to find packing sets in $\mathbb{V}_{p,d}$ that satisfy the sparsity constraints of the model *and* have small diameter according to the subspace Frobenius distance. The next lemma, proved in the Appendix, provides a general method for constructing such local packing sets.

LEMMA A.3 (Local Stiefel embedding). *Let $1 \le k \le d < p$ and the function $A_\varepsilon : \mathbb{V}_{p-d,k} \mapsto \mathbb{V}_{p,d}$ be defined in block form as*

$$A_\varepsilon(J) = \begin{bmatrix} (1-\varepsilon^2)^{1/2} I_k & 0 \\ 0 & I_{d-k} \\ \varepsilon J & 0 \end{bmatrix} \tag{A.10}$$

for $0 \le \varepsilon \le 1$. If $J_1, J_2 \in \mathbb{V}_{p-d,k}$, then

$$\varepsilon^2(1-\varepsilon^2)\|J_1 - J_2\|_F^2 \le \|\sin(A_\varepsilon(J_1), A_\varepsilon(J_2))\|_F^2 \le \varepsilon^2 \|J_1 - J_2\|_F^2.$$

This lemma allows us to convert global $O(1)$-separated packing sets in $\mathbb{V}_{p-d,k}$ into $O(\varepsilon)$-separated packing sets in $\mathbb{V}_{p,d}$ that are localized within a $O(\varepsilon)$-diameter. Note that

$$\|J_i - J_j\|_F \le \|J_i\|_F + \|J_j\|_F \le 2\sqrt{k}.$$

By using Lemma A.3 in conjunction with Lemmas A.1 and A.2, we have the following generic method for lower bounding the minimax risk of estimating the principal subspace of a covariance matrix.

LEMMA A.4. *Let $\varepsilon \in [0,1]$ and $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{p-d,k}$ for $1 \le k \le d < p$. For each $i = 1, \dots, N$, let $\mathbb{P}_i$ be the n-fold product of the $\mathcal{N}(0, \Sigma(A_\varepsilon(J_i)))$ probability measure, where $\Sigma(\cdot)$ is defined in (A.9) and $A_\varepsilon(\cdot)$ is defined in (A.10). If*

$$\min_{i \ne j} \|J_i - J_j\|_F \ge \delta_N,$$

then every estimator $\hat{\mathcal{A}}$ of $\mathcal{A}_i := \operatorname{col}(A_\varepsilon(J_i))$ satisfies

$$\max_i \mathbb{E}_i \|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F \ge \frac{\delta_N \varepsilon \sqrt{1-\varepsilon^2}}{2}\left[1 - \frac{4nk\varepsilon^2/\sigma^2 + \log 2}{\log N}\right],$$

where $\sigma^2 = (1+b)/b^2$.

A.2. Proofs of the main lower bounds.

PROOF OF THEOREM A.1. The following lemma, derived from Massart [(2007), Lemma 4.10], allows us to analyze the variable selection aspect.

LEMMA A.5 (Hypercube construction). *Let m be an integer satisfying $e \le m$ and let $s \in [1, m]$. There exists a subset $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{m,1}$ satisfying the following properties:*

1. $\|J_i\|_{2,0} \le s$ *for all i,*
2. $\|J_i - J_j\|_2^2 \ge 1/4$ *for all $i \ne j$, and*
3. $\log N \ge \max\{cs[1 + \log(m/s)], \log(m)\}$, *where $c > 1/30$ is an absolute constant.*

PROPOSITION A.1. *If $J \in \mathbb{V}_{m,d}$ and $q \in (0, 2]$, then $\|J\|_{2,q}^q \le d^{q/2} \|J\|_{2,0}^{1-q/2}$.*

Let $\rho \in (0,1]$ and $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{m,1}$ be the subset given by Lemma A.5 with $m = p - d$ and $s = \max\{1, (p-d)\rho\}$. Then

$$\begin{aligned}\log N &\ge \max\{cs(1 + \log[(p-d)/s]), \log(p-d)\}\\ &\ge \max\{(1/30)(p-d)\rho(1 - \log\rho), \log(p-d)\}.\end{aligned}$$

Applying Lemma A.4, with $k = 1$, $\delta_N = 1/2$, and b chosen so that $(1+b)/b^2 = \sigma^2$, yields

$$\begin{aligned}
&\max_i \mathbb{E}_i \|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F \\
&\quad\geq \frac{\varepsilon}{4\sqrt{2}}\left[1 - \frac{4n\varepsilon^2/\sigma^2}{(1/30)(p-d)\rho(1-\log\rho)} - \frac{\log 2}{\log(p-d)}\right] \\
&\quad= \frac{\varepsilon}{4\sqrt{2}}\left[1 - \frac{120\varepsilon^2}{\gamma\rho(1-\log\rho)} - \frac{\log 2}{\log(p-d)}\right] \\
&\quad\geq \frac{\varepsilon}{4\sqrt{2}}\left[\frac{1}{2} - \frac{120\varepsilon^2}{\gamma\rho(1-\log\rho)}\right]
\end{aligned} \tag{A.11}$$

for every estimator $\hat{\mathcal{A}}$ and all $\varepsilon \in [0, 1/\sqrt{2}]$, because $p - d \geq 4$ by assumption. Since $J_i \in \mathbb{V}_{p-d,1}$, Proposition A.1 implies

$$\|A_\varepsilon(J_i)\|_{2,q} \leq \begin{cases} d+s, & \text{if } q = 0 \quad \text{and} \\ \left(d + \varepsilon^q s^{(2-q)/2}\right)^{1/q}, & \text{if } 0 < q < 2. \end{cases} \tag{A.12}$$

For every $q \in [0, 2)$

$$d + \varepsilon^q s^{(2-q)/2} \leq R_q \quad \Longleftrightarrow \quad \varepsilon^{2q} \leq \frac{(R_q - d)^2}{s^{2-q}} = \frac{(R_q - d)^2}{\max\{1, (p-d)\rho\}^{2-q}}.$$

Thus, (A.12) implies that the constraint

$$\varepsilon^{2q} \leq \min\{(T/\rho)^2 \rho^q, (R_q - d)^2\} \tag{A.13}$$

is sufficient for $\mathcal{A}_i \in \mathcal{M}_q(R_q)$ and hence $\mathbb{P}_i \in \mathcal{P}_q(\sigma^2, R_q)$. Now fix

$$\varepsilon^2 = \tfrac{1}{480}\gamma\rho(1 - \log\rho) \wedge \tfrac{1}{2}.$$

If we can choose $\rho \in (0, 1]$ such that (A.13) is satisfied, then by (A.11),

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq \max_i \mathbb{E}_i \|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F \geq \frac{\varepsilon}{16\sqrt{2}}.$$

Choose $\rho \in (0, 1]$ to be the unique solution of the equation

$$\rho = \begin{cases} T[\gamma(1 - \log\rho)]^{-q/2}, & \text{if } T < \gamma^{q/2} \quad \text{and} \\ 1, & \text{otherwise.} \end{cases} \tag{A.14}$$

We will verify that ε and ρ satisfy (A.13). The assumption that $1 \leq R_q - d$ guarantees that $\varepsilon^{2q} \leq (R_q - d)^2$, because $\varepsilon^{2q} \leq 1$. If $T < \gamma^{q/2}$, then

$$(T/\rho)^2 \rho^q = [\gamma\rho(1 - \log\rho)]^q \geq \varepsilon^{2q}.$$

If $T \geq \gamma^{q/2}$, then $\rho = 1$ and

$$(T/\rho)^2 \rho^q = T^2 \geq \gamma^q \geq \varepsilon^{2q}.$$

Thus, (A.13) holds and so

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq \frac{\varepsilon}{16\sqrt{2}} \geq \frac{1}{496}[\gamma\rho(1-\log\rho)]^{1/2} \wedge \frac{1}{32}.$$

Now we substitute (A.14) and the definitions of γ and T into the above inequality to get the following lower bounds. If $T < \gamma^{q/2}$, then

$$\begin{aligned}\gamma\rho(1-\log\rho) &= T\gamma^{1-q/2}\{1-\log\rho\}^{1-q/2} \\ &= T\gamma^{1-q/2}\left\{1-\log(T/\gamma^{q/2}) + \frac{q}{2}\log(1-\log\rho)\right\}^{1-q/2} \\ &\geq T\gamma^{1-q/2}\{1-\log(T/\gamma^{q/2})\}^{1-q/2}\end{aligned} \tag{A.15}$$

and so

$$\begin{aligned}&\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ &\qquad \geq c_0\left\{(R_q - d)\left[\frac{\sigma^2}{n}(1-\log(T/\gamma^{q/2}))\right]^{1-q/2} \wedge 1\right\}^{1/2}.\end{aligned}$$

If $T \geq \gamma^{q/2}$, then $\gamma\rho(1-\log\rho) = \gamma$ and

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \geq c_0(\gamma \wedge 1)^{1/2} = c_0\left\{\frac{(p-d)\sigma^2}{n} \wedge 1\right\}^{1/2}. \qquad \square$$

PROOF OF THEOREM A.2. For a fixed subset of s variables, the challenge in estimating the principal subspace of these variables is captured by the richness of packing sets in the Stiefel manifold $\mathbb{V}_{s,d}$. A packing set in the Stiefel manifold can be constructed from a packing set in the Grassman manifold by choosing a single element of the Stiefel manifold as a representative for each element of the packing set in the Grassmann manifold. This is well defined, because the subspace distance is invariant to the choice of basis. The following lemma specializes known results Pajor [(1998), Proposition 8] for packing sets in the Grassman manifold.

LEMMA A.6 [See Pajor (1998)]. *Let k and s be integers satisfying $1 \leq k \leq s - k$, and let $\delta > 0$ There exists a subset $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{s,k}$ satisfying the following properties*:

1. $\|\sin(J_i, J_j)\|_F \geq \sqrt{k}\delta$ *for all* $i \neq j$, *and*
2. $\log N \geq k(s-k)\log(c_2/\delta)$, *where* $c_2 > 0$ *is an absolute constant.*

To apply this result to Lemma A.4 we will use Proposition 2.2 to convert the lower bound on the subspace distance into a lower bound on the Frobenius distance between orthonormal matrices. Thus,

$$\|J_i - J_j\|_F \geq \|\sin\Theta(J_i, J_j)\|_F \geq \sqrt{k}\delta. \tag{A.16}$$

Let $\rho \in (0, 1]$ and $s = \max\{2d, \lfloor (p-d)\rho \rfloor\}$. Invoke Lemma A.6 with $k = d$ and $\delta = c_2/e$, where $c_2 > 0$ is the constant given by Lemma A.6. Let $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{p-d,d}$ be the subset given by Lemma A.6 after augmenting with rows of zeroes if necessary. Then

$$\log N \geq d(s-d) \geq \max\{d(s/2), d^2\} \geq \max\{(d/4)(p-d)\rho, d^2\}$$

and by (A.16),

$$\|J_i - J_j\|_F^2 \geq d(c_2/e)^2$$

for all $i \neq j$. The rest of this proof mirrors that of Theorem A.1. Let $\varepsilon \in [0, 1/\sqrt{2}]$ and apply Lemma A.4 to get

$$\begin{aligned} \max_i \mathbb{E}\|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F &\geq \frac{c_2\sqrt{d}\varepsilon}{2\sqrt{2}e}\left[1 - \frac{4nd\varepsilon^2/\sigma^2}{(d/4)(p-d)\rho} - \frac{\log 2}{d^2}\right] \\ &\geq c_1\sqrt{d}\varepsilon\left[\frac{1}{2} - \frac{16\varepsilon^2}{\gamma\rho}\right], \end{aligned} \tag{A.17}$$

where γ is defined in (3.1) and we used the assumption that $d \geq 2$. Since $J_i \in \mathbb{V}_{p-d,d}$, Proposition A.1 implies

$$\|A_\varepsilon(J_i)\|_{2,q} \leq \begin{cases} d + s, & \text{if } q = 0 \quad \text{and} \\ (d + d^{q/2}\varepsilon^q s^{(2-q)/2})^{1/q}, & \text{if } 0 < q < 2. \end{cases}$$

For every $q \in (0, 2]$

$$\begin{aligned} d + d^{q/2}\varepsilon^q s^{(2-q)/2} \leq R_q \quad &\iff \\ d^q\varepsilon^{2q} \leq \frac{(R_q - d)^2}{s^{2-q}} &= \frac{(R_q - d)^2}{\max\{2d, (p-d)\rho\}^{2-q}}. \end{aligned}$$

So ε and ρ must satisfy the constraint

$$d^q\varepsilon^{2q} \leq \min\left\{(T/\rho)^2\rho^q, \frac{(R_q - d)^2}{(2d)^{2-q}}\right\} \tag{A.18}$$

to ensure that $\mathbb{P}_i \in \mathcal{P}_q(\sigma^2, R_q)$. Fix

$$\varepsilon^2 = \tfrac{1}{64}\gamma\rho \wedge \tfrac{1}{2} \tag{A.19}$$

and

$$\rho = \begin{cases} T(d\gamma)^{-q/2}, & \text{if } T < (d\gamma)^{q/2} \quad \text{and} \\ 1, & \text{otherwise.} \end{cases} \tag{A.20}$$

Since $\varepsilon^2 \leq 1/2$,

$$d^q\varepsilon^{2q} \leq \frac{(R_q - d)^2}{(2d)^{2-q}} \quad \iff \quad 2^q\varepsilon^{2q} \leq \frac{(R_q - d)^2}{4d^2} \quad \Longleftarrow \quad 2d \leq R_q - d,$$

where the right-hand side is an assumption of the lemma. That verifies one of the inequalities in (A.18). If $T < (d\gamma)^{q/2}$, then

$$(T/\rho)^2 \rho^q = (d\gamma\rho)^q \rho^q \geq d^q \varepsilon^{2q}.$$

If $T \geq (d\gamma)^{q/2}$, then $\rho = 1$ and

$$(T/\rho)^2 \rho^q = T^2 \geq (d\gamma)^q \geq d^q \varepsilon^{2q}.$$

Thus, (A.18) holds and by (A.17),

$$\begin{aligned}
\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F &\geq \max_i \mathbb{E}_i \|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F \\
&\geq c_1 \sqrt{d}\varepsilon \left[\frac{1}{2} - \frac{16\varepsilon^2}{\gamma^{(2-q)/q}\rho}\right] \\
&\geq \frac{c_1}{4}\sqrt{d}\varepsilon \\
&\geq c_0 (d\gamma\rho \wedge d)^{1/2}.
\end{aligned}$$

Finally, we substitute the definition of γ and (A.20) into the above inequality to get the following lower bounds. If $T < (d\gamma)^{q/2}$, then

$$\begin{aligned}
\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin\Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F &\geq c_0 \{T(d\gamma)^{1-q/2} \wedge d\}^{1/2} \\
&= c_0 \left\{(R_q - d)\left(\frac{d\sigma^2}{n}\right)^{1-q/2} \wedge d\right\}^{1/2}.
\end{aligned}$$

If $T \geq (d\gamma)^{q/2}$, then

$$\sup_{\mathcal{P}_q(\sigma^2, R_q)} \mathbb{E}\|\sin(\hat{V}, V)\|_F \geq c_0 (d\gamma \wedge d)^{1/2} = c_0 \left\{\frac{d(p-d)\sigma^2}{n} \wedge d\right\}^{1/2}. \qquad \square$$

PROOF OF THEOREM A.3. The proof is a modification of the proof of Theorem A.1. The difficulty of the problem is captured by the difficulty of variable selection within each column of V. Instead of using a single hypercube construction as in the proof of Theorem A.1, we apply a hypercube construction on each of the d columns. We do this by dividing the $(p-d) \times d$ matrix into d submatrices of size $\lfloor (p-d)/d \rfloor \times d$, that is, constructing matrices of the form

$$\begin{bmatrix} B_1^T & B_2^T & \cdots & B_d^T & 0 & \cdots \end{bmatrix}^T$$

and confining the hypercube construction to the kth column of each $\lfloor (p-d)/d \rfloor \times d$ matrix B_k, $k = 1, \ldots, d$. This ensures that the resulting $(p-d) \times d$ matrix has orthonormal columns with disjoint supports.

Let $\rho \in (0, 1]$ and $s \in \max\{1, \lfloor (p-d)/d \rfloor \rho\}$. Applying Lemma A.5 with $m = \lfloor (p-d)/d \rfloor$, we obtain a subset $\{J_1, \ldots, J_M\} \subseteq \mathbb{V}_{m,1}$ such that:

1. $\|J_i\|_0 \le s$ for all i,
2. $\|J_i - J_j\|_2^2 \ge 1/4$ for all $i \ne j$, and
3. $\log M \ge \max\{cs(1 + \log(m/s)), \log m\}$, where $c > 1/30$ is an absolute constant.

Next we will combine the elements of this packing set in $\mathbb{V}_{m,1}$ to form a packing set in $\mathbb{V}_{p-d,d}$. A naive approach takes the d-fold product $\{J_1, \dots, J_M\}^d$, however this results in too small a packing distance because two elements of this product set may differ in only one column.

We can increase the packing distance by requiring a substantial number of columns to be different between any two elements of our packing set without much sacrifice in the size of the final packing set. This is achieved by applying an additional combinatorial round with the Gilbert–Varshamov bound on M-ary codes of length d with minimum Hamming distance $d/2$ [Gilbert (1952), Varšamov (1957)]. The kth coordinate of each code specifies which element of $\{J_1, \dots, J_M\}$ to place in the kth column of B_k, and so any two elements of the resulting packing set will differ in at least $d/2$ columns. Denote the resulting subset of $\mathbb{V}_{p-d,d}$ by $\mathcal{H}^s$. We have:

1. $\|H\|_{*,0} \le s$ for all $H \in \mathcal{H}^s$.
2. $\|H_1 - H_2\|_2^2 \ge d/8$ for all $H_1, H_2 \in \mathcal{H}^s$ such that $H_1 \ne H_2$.
3. $\log N := \log |\mathcal{H}^s| \ge \max\{cds(1+\log(m/s)), \log m\}$, where $c > 0$ is an absolute constant.

Note that the lower bound of $\log m$ in the third item arises by considering the packing set whose N elements consist of matrices whose columns in $B_1, \dots, B_d$ are all equal to some J_i for $i = 1, \dots, M$. This ensures that $\log N \ge \log M \ge \log m$. From here, the proof is a straightforward modification of proof of Theorem A.1 with the substitution of $p - d$ by $(p - d)/d$. For brevity we will only outline the major steps.

Recall the definitions of T_* and γ in (3.7). Apply Lemma A.4 with the subset $\mathcal{H}^s$, $k = d$, $\delta_N = \sqrt{d}/\sqrt{8}$, and b chosen so that $(1 + b)/b^2 = \sigma^2$. Then

$$\begin{aligned}\max_i \mathbb{E}\|\sin\Theta(\hat{\mathcal{A}}, \mathcal{A}_i)\|_F &\ge c_0\sqrt{d}\varepsilon\left[1 - \frac{4n\varepsilon^2/\sigma^2}{cm\rho(1 - \log\rho)} - \frac{\log 2}{\log m}\right] \\ &\ge c_0\sqrt{d}\varepsilon\left[\frac{1}{4} - \frac{(8/c)d\varepsilon^2}{\gamma\rho(1 - \log\rho)}\right]\end{aligned}$$

by the assumption that $(p - d)/d \ge 4$, and

$$\|A_i\|_{*,q} \le \begin{cases} 1 + s, & \text{if } q = 0 \quad \text{and} \\ \left(1 + \varepsilon^q s^{(2-q)/2}\right)^{1/q}, & \text{if } 0 < q < 2. \end{cases}$$

The constraint

$$d^q\varepsilon^{2q} \le \min\{(T_*/\rho)^2\rho^q, d^q(R_q - 1)^2\}$$

ensures that $\mathbb{P}_i \in \mathcal{P}_q^*(\sigma^2, R_q)$. It is satisfied by choosing ε so that

$$d\varepsilon^2 = c_1 \gamma \rho (1 - \log \rho) \wedge \tfrac{1}{2},$$

where $c_1 > 0$ is a sufficiently small constant, the assumption that $d < d(R_q - 1)$, and letting ρ be the unique solution of the equation

$$\rho = \begin{cases} T_*[\gamma(1 - \log \rho)]^{-q/2}, & \text{if } T_* < \gamma^{q/2} \quad \text{and} \\ 1, & \text{otherwise.} \end{cases}$$

We conclude that every estimator $\hat{V}$ satisfies

$$\sup_{\mathcal{P}_q^*(\sigma^2, R_q)} \mathbb{E}\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \ge c_2 \{\gamma \rho (1 - \log \rho) \wedge d\}^{1/2},$$

and we have the following explicit lower bounds. If $T_* < \gamma^{q/2}$, then

$$\sup_{\mathcal{P}_q^*(\sigma^2, R_q)} \mathbb{E}\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \\ \ge c_3 \left\{ d(R_q - 1) \left[\frac{\sigma^2}{n} \bigl(1 - \log(T_*/\gamma^{q/2})\bigr) \right]^{1-q/2} \wedge d \right\}^{1/2}.$$

If $T_* \ge \gamma^{q/2}$, then

$$\sup_{\mathcal{P}_q^*(\sigma^2, R_q)} \mathbb{E}\|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F \ge c_3 \left\{ \frac{(p-d)\sigma^2}{n} \wedge d \right\}^{1/2}. \qquad \square$$

APPENDIX B: UPPER BOUND PROOFS

B.1. Proofs of the main upper bounds. Σ and S_n are both invariant under translations of μ. Since our estimators only depend on $X_1, \dots, X_n$ only through S_n, we will assume without loss of generality that $\mu = 0$ for the remainder of the paper. The sample covariance matrix can be written as

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - \bar{X} \bar{X}^T.$$

It can be show that $\bar{X}\bar{X}^T$ is a higher order term that is negligible [see the proofs in Vu and Lei (2012a), for an example of such arguments]. Therefore, we will ignore this term and focus on the dominating $\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$ term in our proofs below.

PROOF OF THEOREM 3.4. Again, we start from Corollary 4.1, which gives

$$\hat{\varepsilon}^2 := \|\sin \Theta(\hat{\mathcal{S}}, \mathcal{S})\|_F^2 \le \frac{\langle S_n - \Sigma, \hat{V}\hat{V}^T - VV^T \rangle}{\lambda_d - \lambda_{d+1}}.$$

To get the correct dependence on λ_i and for general values of q, we need a more refined analysis to control the random variable $\langle S_n - \Sigma, \hat{V}\hat{V}^T - VV^T\rangle$. Let

$$W := S_n - \Sigma, \qquad \Pi := VV^T \quad \text{and} \quad \hat{\Pi} := \hat{V}\hat{V}^T.$$

Recall that for an orthogonal projector Π we write $\Pi^\perp := I - \Pi$. By Proposition C.1 we have

$$\langle W, \hat{\Pi} - \Pi\rangle = -\langle W, \Pi\hat{\Pi}^\perp\Pi\rangle + 2\langle W, \Pi^\perp\hat{\Pi}\Pi\rangle + \langle W, \Pi^\perp\hat{\Pi}\Pi^\perp\rangle \tag{B.1}$$

$$=: -T_1 + 2T_2 + T_3. \tag{B.2}$$

We will control T_1 (the upper-quadratic term), T_2 (the cross-product term), and T_3 (the lower-quadratic term) separately.

Controlling T_1.

$$\begin{aligned} |T_1| &= |\langle W, \Pi\hat{\Pi}^\perp\Pi\rangle| = |\langle \Pi W\Pi, \Pi\hat{\Pi}^\perp\Pi\rangle| \\ &\le \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^\perp\Pi\|_* = \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^\perp\hat{\Pi}^\perp\Pi\|_* \\ &= \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^\perp\|_F^2 \le \|\Pi W\Pi\|_2\hat{\varepsilon}^2, \end{aligned} \tag{B.3}$$

where $\|\cdot\|_*$ is the nuclear norm (ℓ_1 norm of the singular values) and $\|\cdot\|_2$ is the spectral norm (or operator norm). By Lemma D.5, we have (recall that we assume $\|Z\|_{\psi_2} \le 1$ and $\varepsilon_n \le 1$ for simplicity)

$$\big\| \|\Pi W\Pi\|_2 \big\|_{\psi_1} \le c_1\lambda_1\sqrt{d/n}, \tag{B.4}$$

where c_1 is a universal constant. Define

$$\Omega_1 = \left\{ |T_1| \ge c_1\sqrt{\frac{d}{n}}\log n\lambda_1\hat{\varepsilon}^2 \right\}.$$

Then, when $n \ge 2$ we have

$$\mathbb{P}(\Omega_1) \le \mathbb{P}\big(\|\Pi W\Pi\|_2 \ge c_1\lambda_1\log n\sqrt{d/n}\big) \le (n-1)^{-1}. \tag{B.5}$$

Controlling T_2.

$$\begin{aligned} T_2 &= \langle W, \Pi^\perp\hat{\Pi}\Pi\rangle = \langle \Pi^\perp W\Pi, \Pi^\perp\hat{\Pi}\rangle \\ &\le \|\Pi^\perp W\Pi\|_{2,\infty}\|\Pi^\perp\hat{\Pi}\|_{2,1}. \end{aligned} \tag{B.6}$$

To bound $\|\Pi^\perp\hat{\Pi}\|_{2,1}$, let the rows of $\Pi^\perp\hat{\Pi}$ be denoted by $\phi_1, \dots, \phi_p$ and $t > 0$. Using a standard argument of bounding ℓ_1 norm by the ℓ_q and ℓ_2 norms [e.g., Raskutti, Wainwright and Yu (2011), Lemma 5], we have for all $t > 0$, $0 < q \leq 1$,

$$
\begin{aligned}
\|\Pi^{\perp}\hat{\Pi}\|_{2,1} &= \sum_{i=1}^{p} \|\phi_i\|_2 \\
&\leq \left[\sum_{i=1}^{p} \|\phi_i\|_2^q\right]^{1/2} \left[\sum_{i=1}^{p} \|\phi_i\|_2^2\right]^{1/2} t^{-q/2} + \left[\sum_{i=1}^{p} \|\phi_i\|_2^q\right] t^{1-q} \\
&= \|\Pi^{\perp}\hat{\Pi}\|_{2,q}^{q/2} \|\Pi^{\perp}\hat{\Pi}\|_F t^{-q/2} + \|\Pi^{\perp}\hat{\Pi}\|_{2,q}^q t^{1-q} \\
&\leq \sqrt{2} R_q^{1/2} t^{-q/2} \hat{\varepsilon} + 2R_q t^{1-q},
\end{aligned} \tag{B.7}
$$

where the last step uses the fact that

$$
\begin{aligned}
\|\Pi^{\perp}\hat{\Pi}\|_{2,q}^q &= \|\Pi^{\perp}\hat{V}\|_{2,q}^q = \|\hat{V} - \Pi\hat{V}\|_{2,q}^q \leq \|\hat{V}\|_{2,q}^q + \|VV^T\hat{V}\|_{2,q}^q \\
&\leq \|\hat{V}\|_{2,q}^q + \|V\|_{2,q}^q \leq 2R_q.
\end{aligned}
$$

Combining (B.6) and (B.7) we obtain, for all $t > 0$, $0 < q < 1$,

$$
T_2 \leq \|\Pi^{\perp} W \Pi\|_{2,\infty} (\sqrt{2} R_q^{1/2} t^{-q/2} \hat{\varepsilon} + 2R_q t^{1-q}). \tag{B.8}
$$

The case where $q = 0$ is simpler and omitted. Now define

$$
\begin{aligned}
\Omega_2 &:= \{T_2 \geq 20(\sqrt{\lambda_1 \lambda_{d+1}}^{1-q/2} (\lambda_d - \lambda_{d+1})^{q/2} \varepsilon_n \hat{\varepsilon} \\
&\qquad + \sqrt{\lambda_1 \lambda_{d+1}}^{2-q} (\lambda_d - \lambda_{d+1})^{-(1-q)} \varepsilon_n^2)\} \\
&= \{T_2 \geq t_{2,1}(\sqrt{2R_q} t_{2,2}^{-q/2} \hat{\varepsilon} + 2R_q t_{2,2}^{1-q})\}, \\
t_{2,1} &= 20\sqrt{\lambda_1 \lambda_{d+1}} \sqrt{\frac{d + \log p}{n}}, \\
t_{2,2} &= \frac{\sqrt{\lambda_1 \lambda_{d+1}}}{\lambda_d - \lambda_{d+1}} \sqrt{\frac{d + \log p}{n}}.
\end{aligned}
$$

Taking $t = t_{2,2}$ in (B.8) and using the tail bound result in Lemma D.1, we have

$$
\begin{aligned}
\mathbb{P}(\Omega_2) &\leq \mathbb{P}(\|\Pi^{\perp} W \Pi\|_{2,\infty} \geq t_{2,1}) \\
&\leq 2p5^d \exp\left(-\frac{t_{2,1}^2/8}{2\lambda_1 \lambda_{d+1}/n + t_{2,1}\sqrt{\lambda_1 \lambda_{d+1}}/n}\right) \\
&\leq p^{-1}.
\end{aligned} \tag{B.9}
$$

Controlling T_3. The bound on T_3 involves a quadratic form empirical process over a random set. Let $\varepsilon \geq 0$ and define

$$
\phi(R_q, \varepsilon) := \sup\{\langle W, \Pi^{\perp} U U^T \Pi^{\perp}\rangle : U \in \mathbb{V}_{p,d}, \|U\|_{2,q}^q \leq R_q, \|\Pi^{\perp} U\|_F \leq \varepsilon\}.
$$

Then by Lemma D.4, we have, with some universal constants c_3, for $x > 0$

$$\mathbb{P}(\phi(R_q, \varepsilon) \geq c_3 x \lambda_{d+1}(\varepsilon_n \varepsilon^2 + \varepsilon_n^2 \varepsilon + \varepsilon_n^4)) \leq 2\exp(-x^{2/5}).$$

Let $T_3(U) = \langle W, \Pi^\perp U U^T \Pi^\perp \rangle$, for all $U \in \mathcal{U}_p(R_q)$, where

$$\mathcal{U}_p(R_q) := \{U \in \mathbb{V}_{p,d} : \mathrm{col}(U) \in \mathcal{M}_p(R_q)\}.$$

Define function $g(\varepsilon) = \varepsilon_n \varepsilon^2 + \varepsilon_n^2 \varepsilon + \varepsilon_n^4$. Then for all $\varepsilon \geq 0$, we have $g(\varepsilon) \geq \varepsilon_n^4 \geq 4d^3/n^2$. On the other hand, if $\varepsilon = \|\sin\Theta(U, V)\|_F$, then $\varepsilon^2 \leq 2d$ and hence $g(\varepsilon) \leq g(\sqrt{2d}) = 2d + \sqrt{2d} + 1$. Let $\mu = \varepsilon_n^4$ and $J = \lceil \log_2(g(\sqrt{2d})/\mu) \rceil$. Then we have $J \leq 3\log n + 6/5$.

Note that g is strictly increasing on $[0, \sqrt{2d}]$. Then we have the following peeling argument:

$$\begin{aligned}
&\mathbb{P}[\exists U \in \mathcal{U}_p(R_q) : T_3(U) \geq 2c_3 \lambda_{d+1} (\log n)^{5/2} g(\|\sin(U, V)\|_F)] \\
&\quad \leq \mathbb{P}[\exists 1 \leq j \leq J, U \in \mathcal{U}_p(R_q) : 2^{j-1}\mu \leq g(\|\sin\Theta(U, V)\|_F) \leq 2^j \mu, \\
&\qquad\qquad T_3(U) \geq 2c_3 \lambda_{d+1} (\log n)^{5/2} g(\|\sin\Theta(U, V)\|_F)] \\
&\quad \leq \sum_{j=1}^{J} \mathbb{P}[\phi(R_q, g^{-1}(2^j \mu)) \geq c_3 \lambda_{d+1} (\log n)^{5/2} 2^j \mu] \\
&\quad \leq J 2n^{-1} \leq \frac{6\log n}{n} + \frac{3}{n}.
\end{aligned}$$

Define

$$\Omega_3 := \{\phi(R_q, \hat{\varepsilon}) \geq c_3 (\log n)^{5/2} \lambda_{d+1} (\varepsilon_n \hat{\varepsilon}^2 + \varepsilon_n^2 \hat{\varepsilon} + \varepsilon_n^4)\}.$$

Then we have proved that

$$\mathbb{P}(\Omega_3) \leq \frac{6\log n}{n} + \frac{3}{n}.$$

Putting things together. Now recall the conditions in (3.2) to (3.5). On $\Omega_1^c \cap \Omega_2^c \cap \Omega_3^c$, we have, from (B.1) that

$$\begin{aligned}
(\lambda_d - \lambda_{d+1})\hat{\varepsilon}^2 &\leq \left(c_1 \sqrt{\frac{d}{n}} \log n \lambda_1 + c_3 \varepsilon_n (\log n)^{5/2} \lambda_{d+1}\right) \hat{\varepsilon}^2 \\
&\quad + 41 \sqrt{\lambda_1 \lambda_{d+1}}^{1-q/2} (\lambda_d - \lambda_{d+1})^{q/2} \varepsilon_n \hat{\varepsilon} \\
&\quad + 41 \sqrt{\lambda_1 \lambda_{d+1}}^{2-q} (\lambda_d - \lambda_{d+1})^{-(1-q)} \varepsilon_n^2 \quad \Longrightarrow \\
\frac{1}{2}(\lambda_d - \lambda_{d-1})\hat{\varepsilon}^2 &\leq 41 \sqrt{\lambda_1 \lambda_{d+1}}^{1-q/2} (\lambda_d - \lambda_{d+1})^{q/2} \varepsilon_n \hat{\varepsilon} \\
&\quad + 41 \sqrt{\lambda_1 \lambda_{d+1}}^{2-q} (\lambda_d - \lambda_{d+1})^{-(1-q)} \varepsilon_n^2 \quad \Longrightarrow \\
\hat{\varepsilon} &\leq 83 \left(\frac{\sqrt{\lambda_1 \lambda_{d+1}}}{\lambda_d - \lambda_{d+1}}\right)^{1-q/2} \varepsilon_n. \qquad \square
\end{aligned}$$

APPENDIX C: ADDITIONAL PROOFS

PROOF OF PROPOSITION 2.2. Let γ_i be the cosine of the ith canonical angle between the subspaces spanned by V_1 and V_2. By Theorem II.4.11 of Stewart and Sun (1990),

$$\inf_{Q \in \mathbb{V}_{k,k}} \|V_1 - V_2 Q\|_F^2 = 2 \sum_i (1 - \gamma_i).$$

The inequalities

$$1 - x \le (1 - x^2) \le 2(1 - x)$$

hold for all $x \in [0, 1]$. So

$$\frac{1}{2} \inf_{Q \in \mathbb{V}_{k,k}} \|V_1 - V_2 Q\|_F^2 \le \sum_i (1 - \gamma_i^2) \le \inf_{Q \in \mathbb{V}_{k,k}} \|V_1 - V_2 Q\|_F^2.$$

Apply the trigonometric identity $\sin^2 \theta = 1 - \cos^2 \theta$ to the preceding display to conclude the proof. □

C.1. Proofs related to the lower bounds.

PROOF OF LEMMA A.2. Write $\Sigma_i = \Sigma(A_i)$ for $i = 1, 2$. Since Σ_1 and Σ_2 are nonsingular and have the same determinant,

$$\begin{aligned} D(\mathbb{P}_1 \| \mathbb{P}_2) &= n D(\mathcal{N}(0, \Sigma_1) \| \mathcal{N}(0, \Sigma_2)) \\ &= \frac{n}{2} \{\operatorname{trace}(\Sigma_2^{-1} \Sigma_1) - p - \log\det(\Sigma_2^{-1} \Sigma_1)\} \\ &= \frac{n}{2} \operatorname{trace}(\Sigma_2^{-1} (\Sigma_1 - \Sigma_2)). \end{aligned}$$

Now

$$\Sigma_2^{-1} = (1 + b)^{-1} A_2 A_2^T + (I_p - A_2 A_2^T)$$

and

$$\Sigma_1 - \Sigma_2 = b(A_1 A_1^T - A_2 A_2^T).$$

Thus,

$$\begin{aligned} &\operatorname{trace}(\Sigma_2^{-1} (\Sigma_1 - \Sigma_2)) \\ &\quad = \frac{b}{1 + b} \{(1 + b)\langle I_p - A_2 A_2^T, A_1 A_1^T\rangle - \langle A_2 A_2^T, A_2 A_2^T - A_1 A_1^T\rangle\} \\ &\quad = \frac{b - 1}{b} \{b\langle I_p - A_2 A_2^T, A_1 A_1^T\rangle - \langle I_p, A_2 A_2^T - A_2 A_2^T A_1 A_1^T\rangle\} \end{aligned}$$

$$= \frac{b}{1+b}\{(1+b)\langle I_p - A_2A_2^T, A_1A_1^T\rangle - \langle A_2A_2^T, I_p - A_1A_1^T\rangle\}$$
$$= \frac{b^2}{1+b}\|\sin(A_1, A_2)\|_F^2$$

by Proposition 2.1. □

PROOF OF LEMMA A.3. By Proposition 2.1 and the definition of $A_\varepsilon(\cdot)$,

$$\begin{aligned}\|\sin(A_\varepsilon(J_1), A_\varepsilon(J_2))\|_F^2 &= \frac{1}{2}\|[A_\varepsilon(J_1)][A_\varepsilon(J_1)]^T - [A_\varepsilon(J_2)][A_\varepsilon(J_2)]^T\|_F^2 \\ &= \varepsilon^2(1-\varepsilon^2)\|J_1 - J_2\|_F^2 + \frac{\varepsilon^4}{2}\|J_1J_1^T - J_2J_2^T\|_F^2 \\ &\geq \varepsilon^2(1-\varepsilon^2)\|J_1 - J_2\|_F^2.\end{aligned}$$

The upper bound follows from Proposition 2.2:

$$\|\sin(A_\varepsilon(J_1), A_\varepsilon(J_2))\|_F^2 \leq \|A_\varepsilon(J_1) - A_\varepsilon(J_2)\|_F^2 = \varepsilon^2\|J_1 - J_2\|_F^2. \qquad \square$$

PROOF OF LEMMA A.5. Let $s_0 = \lfloor\min(m/e, s)\rfloor$. The assumptions that $m/e \geq 1$ and $s \geq 1$ guarantee that $s_0 \geq 1$. According to Massart [(2007), Lemma 4.10] [with $\alpha = 7/8$ and $\beta = 8/(7e)$], there exists a subset $\Omega_m^{s_0} \subseteq \{0,1\}^m$ satisfying the following properties:

1. $\|\omega\|_0 = s_0$ for all $\omega \in \Omega_m^{s_0}$,
2. $\|\omega - \omega'\|_0 > s_0/4$ for all distinct pairs $\omega, \omega' \in \Omega_m^{s_0}$, and
3. $\log|\Omega_m^{s_0}| \geq cs_0\log(m/s_0)$, where $c > 0.251$.

Let

$$\{J_1, \dots, J_N\} := \{s_0^{-1/2}\omega : \omega \in \Omega_m^{s_0}\}.$$

Clearly, $\{J_1, \dots, J_N\} \subseteq \mathbb{V}_{m,1}$ and

$$\|J_i\|_{(2,0)} = \|\omega\|_0 = s_0 \leq s$$

for every i. If $i \neq j$, then

$$\|J_i - J_j\|_F^2 = s_0^{-1}\|\omega_i - \omega_j\|_0 > 1/4.$$

The cardinality of $\{J_1, \dots, J_N\}$ satisfies

$$\log N = \log|\Omega_m^{s_0}| \geq cs_0\log(m/s_0).$$

As a function of s_0, the above right-hand side is increasing on the interval $[0, m/e]$. Since $\min(m/e, s)/2 \leq s_0$ belongs to that interval

$$\begin{aligned}\log N &\geq c(\min(m/e, s)/2)\log[m/(\min(m/e, s)/2)] \\ &\geq (c/2)\min(m/e, s)\log[m/\min(m/e, s)].\end{aligned}$$

It is easy to see that

$$\min(m/e, s)\log\bigl[m/\min(m/e, s)\bigr] \geq \max\bigl\{s\log(m/s), s/e\bigr\}$$

for all $s \in [1, m]$. Thus,

$$\min(m/e, s)\log[m/\min(m/e, s) \geq (1+e)^{-1}s + (1+e)^{-1}s\log(m/s)$$

and

$$\log N \geq (c/2)(1+e)^{-1}s\bigl(1+\log(m/s)\bigr), \tag{C.1}$$

where $(c/2)(1+e)^{-1} > 1/30$. If the above right-hand side is $\leq \log m$, then we may repeat the entire argument from the beginning with $\{J_1, \dots, J_N\}$ taken to be the $N = m$ vectors $\{(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)\} \subseteq \{0, 1\}^m$. That yields, in combination with (C.1),

$$\log N \geq \max\bigl\{(1/30)s\bigl[1+\log(m/s)\bigr], \log m\bigr\}. \qquad \square$$

C.2. Proofs related to the upper bounds.

PROOF OF LEMMA 4.2. For brevity, denote the eigenvalues of A by $\lambda_d := \lambda_d(A)$. Let $A = \sum_{i=1}^p \lambda_i u_i u_i^T$ be the spectral decomposition of A so that $E = \sum_{i=1}^d u_i u_i^T$ and $E^\perp = \sum_{i=d+1}^p u_i u_i^T$. Then

$$\begin{aligned}
\langle A, E - F\rangle &= \bigl\langle A, E(I-F) - (I-E)F\bigr\rangle \\
&= \bigl\langle EA, F^\perp\bigr\rangle - \bigl\langle E^\perp A, F\bigr\rangle \\
&= \sum_{i=1}^d \lambda_i \bigl\langle u_i u_i^T, F^\perp\bigr\rangle - \sum_{i=d+1}^p \lambda_i \bigl\langle u_i u_i^T, F\bigr\rangle \\
&\geq \lambda_d \sum_{i=1}^d \bigl\langle u_i u_i^T, F^\perp\bigr\rangle - \lambda_{d+1} \sum_{i=d+1}^p \bigl\langle u_i u_i^T, F\bigr\rangle \\
&= \lambda_d\bigl\langle E, F^\perp\bigr\rangle - \lambda_{d+1}\bigl\langle E^\perp, F\bigr\rangle.
\end{aligned}$$

Since orthogonal projectors are idempotent,

$$\begin{aligned}
\lambda_d\bigl\langle E, F^\perp\bigr\rangle - \lambda_{d+1}\bigl\langle E^\perp, F\bigr\rangle &= \lambda_d\bigl\langle EF^\perp, EF^\perp\bigr\rangle - \lambda_{d+1}\bigl\langle E^\perp F, E^\perp F\bigr\rangle \\
&= \lambda_d\bigl\|EF^\perp\bigr\|_F^2 - \lambda_{d+1}\bigl\|E^\perp F\bigr\|_F^2.
\end{aligned}$$

Now apply Proposition 2.1 to conclude that

$$\lambda_d\bigl\|EF^\perp\bigr\|_F^2 - \lambda_{d+1}\bigl\|E^\perp F\bigr\|_F^2 = (\lambda_d - \lambda_{d+1})\bigl\|\sin\Theta(\mathcal{E}, \mathcal{F})\bigr\|_F^2. \qquad \square$$

PROPOSITION C.1. *If W is symmetric, and E and F are orthogonal projectors, then*

$$\langle W, F - E\rangle = \bigl\langle E^\perp W E^\perp, F\bigr\rangle - \bigl\langle EWE, F^\perp\bigr\rangle + 2\bigl\langle E^\perp W E, F\bigr\rangle. \tag{C.2}$$

PROOF. Using the expansion

$$W = E^\perp W E^\perp + EWE + EWE^\perp + E^\perp WE$$

and the symmetry of W, F and E, we can write

$$\begin{aligned}\langle W, F - E\rangle &= \langle E^\perp W E^\perp, F - E\rangle + \langle EWE, F - E\rangle \\ &\quad + 2\langle E^\perp WE, F - E\rangle \\ &= \langle E^\perp W E^\perp, E^\perp(F - E)\rangle + \langle EWE, E(F - E)\rangle \\ &\quad + 2\langle E^\perp WE, E^\perp(F - E)\rangle \\ &= \langle E^\perp W E^\perp, F\rangle + \langle EWE, E(F - E)\rangle + 2\langle E^\perp WE, F\rangle.\end{aligned}$$

Now note that

$$E(F - E) = EF - E = -EF^\perp. \qquad \square$$

APPENDIX D: EMPIRICAL PROCESS RELATED PROOFS

D.1. The cross-product term. This section is dedicated to proving the following bound on the cross-product term.

LEMMA D.1. *There exists a universal constant $c > 0$ such that*

$$\mathbb{P}(\|\Pi^\perp W \Pi\|_{2,\infty} > t) \le 2p5^d \exp\left(-\frac{t^2/8}{2\lambda_1\lambda_{d+1}/n + t\sqrt{\lambda_1\lambda_{d+1}}/n}\right).$$

The proof of Lemma D.1 builds on the following two lemmas. They are adapted from Lemmas 2.2.10 and 2.2.11 of van der Vaart and Wellner (1996).

LEMMA D.2 (Bernstein's inequality). *Let $Y_1, \dots, Y_n$ be independent random variables with zero mean. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Y_i\right| > t\right) \le 2\exp\left(-\frac{t^2/2}{2\sum_{i=1}^n \|Y_i\|_{\psi_1}^2 + t \max_{i\le n} \|Y_i\|_{\psi_1}}\right)$$

LEMMA D.3 (Maximal inequality). *Let $Y_1, \dots, Y_m$ be arbitrary random variables that satisfy the bound*

$$\mathbb{P}(|Y_i| > t) \le 2\exp\left(-\frac{t^2/2}{b + at}\right)$$

for all $t > 0$ (and i) and fixed $a, b > 0$. Then

$$\left\|\max_{1\le i\le m} Y_i\right\|_{\psi_1} \le c\left(a\log(1+m) + \sqrt{b\log(1+m)}\right)$$

for a universal constant $c > 0$.

We bound $\|\Pi^{\perp}(S_n - \Sigma)\Pi\|_{2,\infty}$ by a standard δ-net argument.

PROPOSITION D.1. *Let A be a $p \times d$ matrix, $(e_1, \dots, e_p)$ be the canonical basis of $\mathbb{R}^p$ and $\mathcal{N}_\delta$ be a δ-net of $\mathbb{S}_2^{d-1}$ for some $\delta \in [0, 1)$. Then*

$$\|A\|_{2,\infty} \le (1-\delta)^{-1} \max_{1 \le j \le p} \max_{u \in \mathcal{N}_\delta} \langle e_j, Au \rangle.$$

PROOF. By duality and compactness, there exists $u_* \in \mathbb{S}^{d-1}$ and $u \in \mathcal{N}_\delta$ such that

$$\|A\|_{2,\infty} = \max_{1 \le j \le p} \|e_j^T A\|_2 = \max_{1 \le j \le p} \langle e_j, Au_* \rangle$$

and $\|u_* - u\|_2 \le \delta$. Then by the Cauchy–Schwarz inequality,

$$\begin{aligned} \|A\|_{2,\infty} &= \max_{1 \le j \le p} \langle e_j, Au \rangle + \langle e_j, A(u_* - u) \rangle \\ &\le \max_{1 \le j \le p} \langle e_j, Au \rangle + \delta \|e_j^T A\|_2 \\ &\le \max_{1 \le j \le p} \max_{u \in \mathcal{N}_\delta} \langle e_j, Au \rangle + \delta \|A\|_{2,\infty}. \end{aligned}$$

Thus,

$$\|A\|_{2,\infty} \le (1-\delta)^{-1} \max_{1 \le j \le p} \max_{u \in \mathcal{N}_\delta} \langle e_j, Au \rangle. \qquad \square$$

The following bound on the covering number of the sphere is well known [see, e.g., Ledoux (2001), Lemma 3.18].

PROPOSITION D.2. *Let $\mathcal{N}_\delta$ be a minimal δ-net of $\mathbb{S}_2^{d-1}$ for $\delta \in (0, 1)$. Then*

$$|\mathcal{N}_\delta| \le (1 + 2/\delta)^d.$$

PROPOSITION D.3. *Let X and Y be random variables. Then*

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

PROOF. Let $A = X/\|X\|_{\psi_2}$ and $Y/\|Y\|_{\psi_2}$. Using the elementary inequality

$$|ab| \le \tfrac{1}{2}(a^2 + b^2)$$

and the triangle inequality we have that

$$\|AB\|_{\psi_1} \le \tfrac{1}{2}(\|A^2\|_{\psi_1} + \|B^2\|_{\psi_1}) = \tfrac{1}{2}(\|A\|_{\psi_2}^2 + \|B\|_{\psi_2}^2) = 1.$$

Multiplying both sides of the inequality by $\|X\|_{\psi_2} \|Y\|_{\psi_2}$ gives the desired result. $\square$

PROOF OF LEMMA D.1. Let N_δ be a minimal δ-net in $\mathbb{S}_2^{d-1}$ for some $\delta \in (0,1)$ to be chosen later. By Proposition D.1 we have

$$\|\Pi^\perp W \Pi\|_{2,\infty} \le \frac{1}{1-\delta} \max_{1\le j\le p} \max_{u\in N_\delta} \langle \Pi^\perp e_j, WVu\rangle,$$

where e_j is the jth column of $I_{p\times p}$. Taking $\delta = 1/2$, by Proposition D.2 we have $|N_\delta| \le 5^d$.

Now $\Pi^\perp \Sigma V = 0$ and so

$$\langle \Pi^\perp e_j, WVu\rangle = \frac{1}{n}\sum_{i=1}^n \langle X_i, \Pi^\perp e_j\rangle\langle X_i, Vu\rangle$$

is the sum of independent random variables with mean zero. By Proposition D.3, the summands satisfy

$$\begin{aligned}\|\langle X_i, \Pi^\perp e_j\rangle\langle X_i, Vu\rangle\|_{\psi_1} &\le \|\langle X_i, \Pi^\perp e_j\rangle\|_{\psi_2}\|\langle X_i, Vu\rangle\|_{\psi_2} \\ &= \|\langle Z_i, \Sigma^{1/2}\Pi^\perp e_j\rangle\|_{\psi_2}\|\langle Z_i, \Sigma^{1/2}Vu\rangle\|_{\psi_2} \\ &\le \|Z_1\|_{\psi_2}^2 \|\Sigma^{1/2}\Pi^\perp e_j\|_2 \|\Sigma^{1/2}Vu\|_2 \\ &\le \|Z_1\|_{\psi_2}^2 \sqrt{\lambda_1\lambda_{d+1}}.\end{aligned}$$

Recall that $\|Z\|_{\psi_2}^2 = 1$. Then Bernstein's inequality (Lemma D.2) implies that for all $t > 0$ and every $u \in \mathcal{N}_\delta$

$$\begin{aligned}\mathbb{P}(\|\Pi^\perp W\Pi\|_{2,\infty} > t) &\le \mathbb{P}\Big(\max_{1\le j\le p}\max_{u\in N_\delta}\langle \Pi^\perp e_j, WVu\rangle > t/2\Big) \\ &\le p5^d \mathbb{P}(|\langle \Pi^\perp e_j, WVu\rangle| > t/2) \\ &\le 2p5^d \exp\biggl(-\frac{t^2/8}{2\lambda_1\lambda_{d+1}/n + t\sqrt{\lambda_1\lambda_{d+1}}/n}\biggr).\end{aligned}$$ □

D.2. The quadratic terms.

LEMMA D.4. *Let $\varepsilon \ge 0$, $q \in (0,1]$, and*

$$\phi(R_q, \varepsilon) = \sup\{\langle S_n - \Sigma, \Pi^\perp U U^T \Pi^\perp\rangle : U \in \mathbb{V}_{p,d}, \|U\|_{2,q}^q \le R_q, \|\Pi^\perp U\|_F \le \varepsilon\}.$$

There exist constants $c > 0$ and c_1 such that for all $x \ge c_1$,

$$\mathbb{P}\biggl[\phi(R_q,\varepsilon) \ge cx\|Z_1\|_{\psi_2}^2 \lambda_{d+1}\biggl\{\varepsilon\frac{E(R_q,\varepsilon)}{\sqrt{n}} + \frac{E^2(R_q,\varepsilon)}{n}\biggr\}\biggr] \le 2\exp(-x^{2/5}),$$

where

$$E(R_q,\varepsilon) = \mathbb{E}\sup\{\langle \mathcal{Z}, U\rangle : U \in \mathbb{R}^{p\times d}, \|U\|_{2,q}^q \le 2R_q, \|U\|_F \le \varepsilon\}$$

and $\mathcal{Z}$ is a $p \times d$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. As a consequence, we have, for another constant c_2

$$\mathbb{E}\phi(R_q,\varepsilon) \le c_2\|Z_1\|_{\psi_2}^2\lambda_{d+1}\left\{\varepsilon\frac{E(R_q,\varepsilon)}{\sqrt{n}} + \frac{E^2(R_q,\varepsilon)}{n}\right\}.$$

Moreover, we have, for another numerical constant c',

$$\frac{E(R_q,\varepsilon)}{\sqrt{n}} \le c'\bigl(R_q^{1/2}t^{1-q/2}\varepsilon + R_q t^{2-q}\bigr) \tag{D.1}$$

with $t = \sqrt{\frac{d+\log p}{n}}$.

PROOF. The first part follows from Corollary 4.1 of Vu and Lei (2012b). It remains for us to prove the "moreover" part. By the duality of the (2, 1)- and $(2,\infty)$-norms,

$$\langle \mathcal{Z}, U\rangle \le \|\mathcal{Z}\|_{2,\infty}\|U\|_{2,1}$$

and so

$$\mathbb{E}(R_q,\varepsilon) \le \mathbb{E}\|\mathcal{Z}\|_{2,\infty}\sup\bigl\{\|U\|_{2,1} : U \in \mathbb{R}^{p\times d}, \|U\|_{2,q}^q \le 2R_q, \|U\|_F \le \varepsilon\bigr\}.$$

By (4.7) and the fact that the Orlicz ψ_2-norm bounds the expectation,

$$\mathbb{E}\|\mathcal{Z}\|_{2,\infty} \le c'\sqrt{d+\log p}.$$

Now $\|U\|_{2,1}$ is just the ℓ_1 norm of the vector of row-wise norms of U. So we use a standard argument to bound the ℓ_1 norm in terms of the ℓ_2 and ℓ_q norms for $q \in (0,1]$ [e.g., Raskutti, Wainwright and Yu (2011), Lemma 5], and find that for every $t > 0$

$$\begin{aligned}\|U\|_{2,1} &\le \|U\|_{2,q}^{q/2}\|U\|_{2,2}t^{-q/2} + \|U\|_{2,q}^q t^{1-q}\\ &= \|U\|_{2,q}^{q/2}\|U\|_F t^{-q/2} + \|U\|_{2,q}^q t^{1-q}.\end{aligned}$$

Thus,

$$\sup\bigl\{\|U\|_{2,1} : U \in \mathbb{R}^{p\times d}, \|U\|_{2,q}^q \le 2R_q, \|U\|_F \le \varepsilon\bigr\} \le R_q^{1/2}t^{-q/2} + R_q t^{1-q}.$$

Letting $t = \mathbb{E}\|\mathcal{Z}\|_{2,\infty}/\sqrt{n}$, and combining the above inequalities completes the proof. $\square$

LEMMA D.5. *There exists a constant $c > 0$ such that*

$$\bigl\|\|\Pi(S_n - \Sigma)\Pi\|_2\bigr\|_{\psi_1} \le c\|Z_1\|_{\psi_2}^2\lambda_1(\sqrt{d/n} + d/n).$$

PROOF. Let $\mathcal{N}_\delta$ be a minimal δ-net of $\mathbb{S}_2^{d-1}$ for some $\delta \in (0,1)$ to be chosen later. Then

$$\|\Pi(S_n - \Sigma)\Pi\|_2 = \|V^T(S_n - \Sigma)V\|_2 \le (1-2\delta)^{-1} \max_{u \in \mathcal{N}_\delta} |\langle Vu, (S_n - \Sigma)Vu\rangle|.$$

Using a similar argument as in the proof of Lemma D.1, for all $t > 0$ and every $u \in \mathcal{N}_\delta$

$$\mathbb{P}(|\langle Vu, (S_n - \Sigma)Vu\rangle| > t) \le 2\exp\left(-\frac{t^2/2}{2\sigma^2/n + t\sigma/n}\right),$$

where $\sigma = 2\|Z_1\|_{\psi_2}^2 \lambda_1$. Then Lemma D.3 implies that

$$\begin{aligned}\big\|\|\Pi(S_n - \Sigma)\Pi\|_2\big\|_{\psi_1} &\le (1-2\delta)^{-1} \Big\|\max_{u \in \mathcal{N}_\delta} |\langle Vu, (S_n - \Sigma)Vu\rangle|\Big\|_{\psi_1} \\ &\le (1-2\delta)^{-1} C\sigma\left(\sqrt{\frac{\log(1+|\mathcal{N}_\delta|)}{n}} + \frac{\log(1+|\mathcal{N}_\delta|)}{n}\right),\end{aligned}$$

where $C > 0$ is a constant. Choosing $\delta = 1/3$ and applying Proposition D.2 yields $|\mathcal{N}_\delta| \le 7^d$ and

$$\log(1 + |\mathcal{N}_\delta|) \le \log(8)\log(d).$$

Thus,

$$\big\|\|\Pi(S_n - \Sigma)\Pi\|_2\big\|_{\psi_1} \le 7C\sigma(\sqrt{d/n} + d/n). \qquad \square$$

Acknowledgments. The research reported in this article was completed while V. Q. Vu was visiting the Department of Statistics at Carnegie Mellon University. He thanks them for their hospitality and support. We also thank the referees for their helpful comments.

REFERENCES

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

BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865

CHIKUSE, Y. (2003). *Statistics on Special Manifolds*. *Lecture Notes in Statistics* **174**. Springer, New York. MR1960435

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

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303.

EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856

GILBERT, E. N. (1952). A comparison of signalling alphabets. *Bell System Technical Journal* **31** 504–522.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** 498–520.

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

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

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

LEDOUX, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

LOUNICI, K. (2013). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI* (C. Houdré, D. M. Mason, J. Rosiński and J. A. Wellner, eds.). *Progress in Probability* **66** 327–356. Springer, Basel.

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

MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

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

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

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

PAJOR, A. (1998). Metric entropy of the Grassmann manifold. In *Convex Geometric Analysis* (*Berkeley, CA*, 1996). *Mathematical Sciences Research Institute Publications* **34** 181–188. Cambridge Univ. Press, Cambridge. MR1665590

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

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2** 559–572.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

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. MR3004561

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

VARŠAMOV, R. R. (1957). The evaluation of signals in codes with correction of errors. *Dokl. Akad. Nauk SSSR* (*N.S*) **117** 739–741. MR0095090

VU, V. Q. and LEI, J. (2012a). Minimax rates of estimation for sparse PCA in high dimensions. In *Proceedings of the* 15*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*). *JMLR Workshop and Conference Proceedings Volume* 22.

VU, V. Q. and LEI, J. (2012b). Squared-norm empirical process in Banach space. Available at arXiv:1312.1005.

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.

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
USA
E-MAIL: vqv@stat.osu.edu

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jinglei@andrew.cmu.edu