*Submitted to the Annals of Statistics*

# MINIMAX SPARSE PRINCIPAL SUBSPACE ESTIMATION IN HIGH DIMENSIONS

By Vincent Q. Vu* and Jing Lei†

*The 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 prove optimal, non-asymptotic lower and upper bounds on the minimax subspace estimation error under two different, but related notions of $\ell_q$ subspace sparsity for $0 \leq q \leq 1$. Our upper bounds apply to general classes of covariance matrices, and they show that $\ell_q$ constrained estimates can achieve optimal minimax rates without restrictive spiked covariance conditions.

**1. Introduction.** Principal components analysis (PCA) was introduced in the early 20th century (Pearson, 1901; Hotelling, 1933) 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; Paul, 2007; Nadler, 2008).

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 on methodology that applies the concept of sparsity to eigenvector estimation (some examples include Jolliffe, Trendafilov and Uddin, 2003; d'Aspremont et al., 2007; Zou, Hastie and Tibshirani, 2006; Shen and Huang, 2008; Witten, Tibshirani and Hastie, 2009). Theoretical developments on

*Supported in part by NSF Postdoctoral Fellowship DMS-09-03120.

†Supported in part by NSF Grant BCS-0941518.

*AMS 2000 subject classifications:* Primary 62H25; secondary 62H12, 62C20

*Keywords and phrases:* principal components analysis, subspace estimation, sparsity, high-dimensional statistics, minimax bounds, random matrices, empirical process

sparsity and PCA include Johnstone and Lu (2009); Amini and Wainwright (2009); Shen, Shen and Marron (2011); Ma (2011); Vu and Lei (2012a); Birnbaum et al. (2012).

An open problem that has remained is whether sparse PCA methods can *optimally* estimate the subspace spanned by the leading eigenvectors, i.e. the *principal subspace* of variation. The subspace estimation problem is directly connected to dimension reduction and is important when there is more than one principal component of interest. Indeed, typical applications of PCA use the projection on to 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.

In this paper we study principal subspace estimation by sparse PCA in high-dimensions. We present non-asymptotic minimax lower and upper bounds with optimal dependence on the parameters of the problem. As an illustration, one consequence of our results is that the order of the minimax mean squared estimation error of the $d$-dimensional principal subspace is, ignoring constant factors,

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

where $\sigma^2$ is a measure of the noise-to-signal ratio and $R_q$ is a measure of the sparsity in an $\ell_q$ sense defined in Section 2. The $d + \log p$ factor is novel and it reflects two complementary aspects of the problem: $d$ for parametric estimation and $\log p$ for variable selection.

We obtain the minimax upper bound by analyzing a sparsity constrained principal subspace estimator and showing that it attains the optimal error (up to a constant factor). In comparison to most existing analyses in the literature, we show that the upper bound holds without assuming a spiked covariance model. A key technical ingredient in our analysis of the subspace estimator is a novel variational form of the Davis-Kahan $\sin\Theta$ Theorem (see Lemma 5.2) that allows us to bound the estimation error using some recent advances in empirical process theory. The minimax lower bound follows the standard Fano method framework, but involves nontrivial constructions of packing sets in the Stiefel Manifold.

Our results provide the first and optimal minimax lower bound for sparse principal subspace estimation. To our knowledge, the only other work that has considered sparse principal subspace estimation is that of Ma (2011) on the rate of convergence of an iterative thresholding estimator. However their analysis depends on assuming a spiked covariance model and even then

the rate of convergence has suboptimal dependence on the dimension of the principal subspace.

The remainder of the paper is organized as follows. In the next section, we introduce the sparse principal subspace estimation problem and formally setup 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. Sections 4 and 5 contain the major steps in proving the lower and upper bounds. The major steps in the proofs require some auxilliary lemmas whose proofs we defer to Appendices A and B. Section 6 closes the paper with discussion of our results and open problems.

**2. Subspace estimation.** Let $X_1, \dots, X_n \in \mathbb{R}^p$ be independent, identically distributed random vectors with mean $\mu$ and covariance matrix $\Sigma$. 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[1], and it corresponds to the optimization problem

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

where $\mathbb{G}_{p,d}$ is the Grassmann manifold[2] of $d$-dimensional subspaces of $\mathbb{R}^p$, $\Pi_{\mathcal{G}}$ is the projection onto $\mathcal{G}$, and $I_p$ is the $p \times p$ identity matrix. There is always at least one $d \leq p$ for which eq. (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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\Sigma$ 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 $\Sigma$ is

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

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

[1] This is essentially the viewpoint of Pearson (1901).

[2] For background on Grassmann and Stiefel manifolds, see Edelman, Arias and Smith (1998) and Chikuse (2003).

In practice, $\Sigma$ is unknown, so $\mathcal{S}$ must be estimated from the data. Standard PCA replaces eq. (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 $\Sigma$. 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, and are equal up to multiplication on the right by an orthogonal matrix. 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[3], $q \geq 0$, of a $p \times d$ matrix $A$ as

$$\|A\|_{2,q}^q := \begin{cases} \sum_{j=1}^{p}\left[\sum_{k=1}^{d} a_{jk}^2\right]^{\frac{q}{2}} & \text{if } q > 0, \text{ and} \\ \sum_{j=1}^{p} 1_{\{a_{j*}\neq 0\}} & \text{if } q = 0\,, \end{cases}$$

where $a_{j*}$ denotes the $j$th row of $A$. Note that $\|\cdot\|_{2,q}^q$ 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.

Definition (Row sparse subspaces). For $q \geq 0$ and $R_q \geq d$,

$$\mathcal{M}_q(R_q) := \begin{cases} \left\{ \operatorname{span}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{2,q}^q \leq R_q \right\} & \text{if } q > 0, \text{ and} \\ \left\{ \operatorname{span}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{2,0} \leq R_0 \right\} & \text{if } q = 0\,. \end{cases}$$

where $\operatorname{span}(U)$ denotes the span of the columns of $U$.

[3]To be precise, this is actually a pseudonorm when $q < 1$.

Roughly speaking, row sparsity asserts that there is a small subset of variables (coordinates of $\mathbb{R}^p$) that generate the principal subspace. Since $\|\cdot\|_{2,q}^q$ is coordinate-independent, *every* orthonormal basis of a row sparse $\mathcal{G}$ has the same $(2,q)$-norm. Column sparsity, on the other hand, asserts that there is *some* orthnormal basis of sparse vectors that spans the principal subspace. Define the $(*,q)$-norm, $q \geq 0$, of a $p \times d$ matrix $A$ as

$$\|A\|_{*,q}^q := \begin{cases} \max_{1\leq k\leq d} \sum_{j=1}^p |a_{jk}|^q & \text{if } q > 0, \text{ and} \\ \max_{1\leq k\leq d} \sum_{j=1}^p 1_{\{a_{jk}\neq 0\}} & \text{if } q = 0\,. \end{cases}$$

This is the maximum of the $\ell_q$ norms of the columns of $A$ and 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 $q \geq 0$ and $R_q \geq 1$,

$$\mathcal{M}_q^*(R_q) := \begin{cases} \{\, \text{span}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{*,q}^q \leq R_q, \,\} & \text{if } q > 0, \text{ and} \\ \{\, \text{span}(U) : U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{*,0} \leq R_0, \,\} & \text{if } q = 0\,. \end{cases}$$

The column sparse subspaces are the $d$-dimensional subspaces that have some orthonormal basis whose vectors are $\ell_q$ 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 $\mathcal{G} = \text{span}(U)$ and $\|U\|_{*,q}^q \leq R_q$ (or $\|U\|_{*,0} \leq R_0$ for $q = 0$).

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

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

for $i = 1, \ldots, n$, where $\|\cdot\|_{\psi_\alpha}$ is the Orlicz $\psi_\alpha$-norm[4] defined for $\alpha \geq 1$ as

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

This ensures that the distribution of the $X_i$'s is sub-Gaussian. 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 *noise-to-signal ratio*

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

[4]See van der Vaart and Wellner (1996, Chapter 2) for more information on the Orlicz $\psi_\alpha$-norm.

It turns out that $\sigma^2$ 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 $X_1, \dots, X_n$ that satisfy eq. (2.4), eq. (2.5), 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 eq. (2.4), eq. (2.5), and $\mathcal{S} \in \mathcal{M}_q^*(R_q)$. Throughout this paper, we consider estimating $\mathcal{S}$ over $\mathcal{P}_q(\sigma^2, R_q)$ and $\mathcal{P}_q^*(\sigma^2, R_q)$.

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.[5] For a subspace $\mathcal{E} \in \mathbb{G}_{p,d}$ and its orthogonal projection $E$, we write $E^\perp$ to denote the orthogonal projection onto $\mathcal{E}^\perp$ 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 projections $E$ 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 $k = 1, \dots, d$ and the *angle operator* between $\mathcal{E}$ and $\mathcal{F}$ is the $d \times d$ matrix

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

In this paper we will consider the following 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, for example), and can be connected to the familiar Frobenius (squared error) distance between projection matrices by the following following well-known fact from matrix perturbation theory.

[5] We refer the reader to Bhatia (1997, Chapter VII.1) and Stewart and Sun (1990) for additional background on canonical angles.

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 projections $E$ 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,

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

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

$$\sin(U_1, U_2) := \sin\Theta\big(\operatorname{span}(U_1), \operatorname{span}(U_2)\big)$$

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 projections corresponding to $\mathcal{E}, \mathcal{F} \in \mathbb{G}_{p,d}$ The following proposition, proved in the Appendix, 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*

$$\frac{1}{2}\inf_{Q\in\mathbb{V}_{d,d}}\|V_1 - V_2Q\|_F^2 \le \|\sin(V_1,V_2)\|_F^2 \le \inf_{Q\in\mathbb{V}_{d,d}}\|V_1 - V_2Q\|_F^2\,.$$

In otherwords, the distance between two subspaces is equivalent to the distance between their orthonormal bases, up to some rotation.

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

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

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

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

$$\begin{aligned} &\text{maximize} && \langle S_n, UU^T\rangle \\ &\text{subject to} && U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{2,q}^q \le R_q\,. \end{aligned} \tag{2.8}$$

If $\hat{V}$ is a solution of eq. (2.8). Then $\operatorname{span}(\hat{V})$ is a solution of eq. (2.7). The feasible set of both problems is nonempty when $R_q \ge d$ and the sparsity constraint is active only when $R_q \le d^{q/2}p^{1-q/2}$. When $q = 1$, the estimator defined by eq. (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. This estimator appears to be computationally intractable, because it involves a convex *maximization* problem.

In the column sparse case, 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.

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

$$\begin{aligned} &\text{maximize} && \langle S_n, UU^T\rangle \\ &\text{subject to} && U \in \mathbb{V}_{p,d} \text{ and } \|U\|_{*,q}^q \le R_q \end{aligned} \tag{2.10}$$

Like the row sparse estimator, the column sparse principal subspace estimator does not appear to be computationally tractable either.

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

3.1. *Minimax lower bounds.* 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

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

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

Condition 1 is necessary for the existence of a consistent estimator (see Theorems 4.1 and 4.2). 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 non-asymptotic bounds are non-trivial. Similarly, the upper bound on $R_q - d$ is only violated in trivial cases.

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

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

Here, as well as in the entire paper, $c$ denotes universal, positive constant, not necessarily the same at each occurrance. 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 Section 4.

The non-asymptotic lower bound for $0 < q < 2$ has a more complicated dependence on $(n, p, d, R_q, \sigma^2)$ because of the interaction between $\ell_q$ and $\ell_2$ norms. Therefore, our main lower bound result for $0 < q < 2$ will focus on values of $(n, p, d, R_q, \sigma^2)$ that correspond to the high-dimensional and sparse regime. (We will state more general lower bound results in Section 4.)

Let

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

The interpretation for these two quantities is natural. First, $T$ measures the relative sparsity of the problem. It ranges between 0 and 1, though the "sparse" regime generally corresponds to $T \ll 1$. The second quantity, $\gamma$ corresponds to the classic mean squared error (MSE) of standard PCA. The problem is low-dimensional if $\gamma$ is small and there is not much sparsity. We impose the following condition to preclude this case.

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

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

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

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 $\ell_p$ and $\ell_q$ balls (see Donoho and Johnstone, 1994, for example).

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

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

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. (2012).

By modifying the proof of Theorem 3.1 and Theorem 3.2 we can obtain 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 $T$ for the column sparse case is

$$T_* := \frac{d(R_q - 1)}{(p-d)^{1-\frac{q}{2}}}, \tag{3.2}$$

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

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

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

THEOREM 3.3 (Column sparse lower bound). *Let $q \in [0, 2)$. If Conditions 4 and 5 hold, then*

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

For column sparse subspaces, the lower bound is dominated by the variable selection error, because column sparsity is defined in terms of the maximal $\ell_0$ 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 $d$. We prove Theorem 3.3 in Section 4.

3.2. *Minimax upper bounds.* 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 3.4 (Row sparse upper bound in expectation). *Let $\widehat{\mathcal{S}}$ be any solution of eq. (2.7). If $6\sqrt{R_0(d + \log p)} \le \sqrt{n}$, then*

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

Although eq. (2.7) may not have a unique global minimum, Theorem 3.4 shows that *any* global minimum will be within a certain radius of the principal subspace $\mathcal{S}$. The proof of Theorem 3.4, given in Section 5.2, 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 3.5 below. We note that the rate given by Theorem 3.4 is off by a $\sqrt{\lambda_1/\lambda_{d+1}}$ factor that is due to the specific approach taken to control an empirical processes in our proof of Theorem 3.4.

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

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

The regularity conditions are

$$\epsilon_n \le 1\,, \tag{3.3}$$

$$c_1\sqrt{\frac{d}{n}}\log n\lambda_1 + c_3\epsilon_n(\log n)^{5/2}\lambda_{d+1} < \frac{1}{2}(\lambda_d - \lambda_{d+1})\,, \tag{3.4}$$

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

$$c_3\epsilon_n^2(\log n)^{5/2}\lambda_{d+1} \le \sqrt{\lambda_1\lambda_{d+1}}^{2-q}(\lambda_d - \lambda_{d+1})^{-(1-q)}\,, \tag{3.6}$$

where $c_1$ and $c_3$ are positive constants given.

THEOREM 3.5 (Row sparse upper bound in probability). *Let $q \in [0,1]$ and $\widehat{\mathcal{S}}$ be any solution of eq. (2.7). If $(X_1,\dots,X_n) \sim \mathbb{P} \in \mathcal{P}_q(\sigma^2, R_q)$ and eqs. (3.3) to (3.6) hold, then*

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

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

Theorem 3.5 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. The upper bound matches our lower bounds (Theorem 3.1 and Theorem 3.2) for the entire tuple $(n, p, d, R_q, \sigma^2)$ up to a constant if

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

for some constant $c < 1$. The proof of Theorem 3.5 is in Section 5.2. By observing that $\mathcal{M}_q^*(R_q) \subseteq \mathcal{M}_q(dR_q)$, we can reuse the proof of Theorem 3.5 to derive the following upper bound for the column sparse class.

COROLLARY 3.1 (Column sparse upper bound). *Let $q \in [0,1]$ and $\widehat{\mathcal{S}}$ be any solution of eq. (2.9). If $(X_1,\dots,X_n) \sim \mathbb{P} \in \mathcal{P}_q^*(\sigma^2, R_q)$ and eqs. (3.3) to (3.6) hold with $R_q$ replaced by $dR_q$, then*

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

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

Corollary 3.1 is slightly weaker than the corresponding result for the row sparse class. It matches the lower bound in Theorem 3.3 up to a constant if

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

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

**4. Lower bound proofs.** Theorems 3.1 to 3.3 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 3.3.

THEOREM 4.1 (Row sparse variable selection). *Let $q \in [0,2)$ and $(p,d,R_q)$ satisfy*

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

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

$$\begin{aligned}&\sup_{\mathcal{P}_q(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \\ &\ge c\left\{(R_q-d)\left[\frac{\sigma^2}{n}\left(1-\log\left(T/\gamma^{\frac{q}{2}}\right)\right)\right]^{1-\frac{q}{2}} \wedge 1\right\}^{\frac{1}{2}}.\end{aligned} \tag{4.1}$$

*Otherwise,*

$$\sup_{\mathcal{P}_q(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \ge c\left\{\frac{(p-d)\sigma^2}{n} \wedge 1\right\}^{\frac{1}{2}}. \tag{4.2}$$

The case $q=0$ is particularly simple, because $T < \gamma^{\frac{q}{2}} = 1$ holds trivially. In that case, Theorem 4.1 asserts that

$$\begin{aligned}&\sup_{\mathcal{P}_0(R_0,\sigma^2)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \\ &\ge c\left\{(R_0-d)\frac{\sigma^2}{n}\left(1+\log\frac{p-d}{R_q-d}\right) \wedge 1\right\}^{\frac{1}{2}}.\end{aligned} \tag{4.3}$$

When $q \in (0,2)$ the transition between the $T < \gamma^{\frac{q}{2}}$ and $T \ge \gamma^{\frac{q}{2}}$ regimes involves lower order ($\log\log$) terms that can be seen in eq. (4.15). Under

Condition 3, eq. (4.1) can be simplified to

$$\begin{aligned}&\sup_{\mathcal{P}_0(R_0,\sigma^2)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \\ &\geq c\left\{(R_q-d)\frac{\sigma^2}{n}\left(1+(1-a)\log\frac{(p-d)^{1-\frac{q}{2}}}{R_q-d}\right)\wedge 1\right\}^{\frac{1}{2}-\frac{q}{2}}.\end{aligned} \tag{4.4}$$

THEOREM 4.2 (Row sparse parameter estimation). *Let $q \in [0,2)$ and $(p,d,R_q)$ satisfy*

$$2 \leq d \text{ and } 2d \leq R_q - d \leq (p-d)^{1-\frac{q}{2}},$$

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

$$\sup_{\mathcal{P}_q(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \geq c\left\{(R_q-d)\left(\frac{d\sigma^2}{n}\right)^{1-\frac{q}{2}}\wedge d\right\}^{\frac{1}{2}}. \tag{4.5}$$

*Otherwise,*

$$\sup_{\mathcal{P}_q(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \geq c\left\{\frac{d(p-d)\sigma^2}{n}\wedge d\right\}^{\frac{1}{2}}. \tag{4.6}$$

This result with Equation (4.3) implies Theorem 3.1, and with Equation (4.4) it implies Theorem 3.2.

THEOREM 4.3 (Column sparse estimation). *Let $q \in [0,2)$ and $(p,d,R_q)$ satisfy*

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

*and recall the definition of $T_*$ in eq. (3.2). There exists a universal constant $c>0$ such that every estimator $\widehat{\mathcal{S}}$ satisfies the following. If $T_* < \gamma^{\frac{q}{2}}$, then*

$$\begin{aligned}&\sup_{\mathcal{P}_q^*(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \\ &\geq c\left\{d(R_q-1)\left[\frac{\sigma^2}{n}\left(1-\log\left(T_*/\gamma^{\frac{q}{2}}\right)\right)\right]^{1-\frac{q}{2}}\wedge d\right\}^{\frac{1}{2}}.\end{aligned} \tag{4.7}$$

*Otherwise,*

$$\sup_{\mathcal{P}_q^*(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \geq c\left\{\frac{(p-d)\sigma^2}{n}\wedge d\right\}^{\frac{1}{2}} \tag{4.8}$$

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 4.1 to 4.3.

4.1. *Lower bounds for principal subspace estimation via Fano's 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 4.1 (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 $\Theta$ and suppose that for all $i \neq j$*

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

*and, the Kullback-Leibler (KL) divergence*

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

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

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

The calculations required for applying Lemma 4.1 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

$$\Sigma(A) = bAA^T + I_p = (1+b)AA^T + (I_p - AA^T) \,, \tag{4.9}$$

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

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

We can choose $b$ so that $(1+b)/b^2 = \sigma^2$. The KL divergence between these multivriate Normal distributions has a simple, exact expression given in the following lemma. The proof is a straightforward and contained in the appendix.

LEMMA 4.2 (KL divergence). *For $i = 1, 2$, let $A_i \in \mathbb{V}_{p,d}$, $b \geq 0$,*

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

*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 4.2 is equivalent to the subspace distance. In applying Lemma 4.1, 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 4.3 (Local Stiefel embedding). *Let $1 \leq k \leq d < p$ and the function $A_\epsilon : \mathbb{V}_{p-d,k} \mapsto \mathbb{V}_{p,d}$ be defined in block form as*

$$A_\epsilon(J) = \begin{bmatrix} (1-\epsilon^2)^{1/2} I_k & 0 \\ 0 & I_{d-k} \\ \epsilon J & 0 \end{bmatrix} \tag{4.10}$$

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

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

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

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

By using Lemma 4.3 in conjunction with Lemmas 4.1 and 4.2, we have the following generic method for lower bounding the minimax risk of estimating the principal subspace of a covariance matrix.

LEMMA 4.4. *Let $\epsilon \in [0, 1]$ and $\{J_1, \ldots, J_N\} \subseteq \mathbb{V}_{p-d,k}$ for $1 \leq k \leq d < p$. For each $i = 1, \ldots, N$, let $\mathbb{P}_i$ be the $n$-fold product of the $\mathcal{N}(0, \Sigma(A_\epsilon(J_i)))$ probability measure, where $\Sigma(\cdot)$ is defined in eq. (4.9) and $A_\epsilon(\cdot)$ is defined in eq. (4.10). If*

$$\min_{i \neq j} \|J_i - J_j\|_F \geq \delta_N\,,$$

*then every estimator* $\widehat{\mathcal{A}}$ *of* $\mathcal{A}_i := \operatorname{span}(A_\epsilon(J_i))$ *satisfies*

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

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

4.2. *Proofs of the main lower bounds.*

Proof of Theorem 4.1. The following lemma, derived from (Massart, 2007, Lemma 4.10), allows us to analyze the variable selection aspect.

Lemma 4.5 (Hypercube construction). *Let* $m$ *be an integer satisfying* $e \leq m$ *and let* $s \in [1, m]$. *There exists a subset* $\{J_1, \ldots, J_N\} \subseteq \mathbb{V}_{m,1}$ *satisfying the following properties:*

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

Proposition 4.1. *If* $J \in \mathbb{V}_{m,d}$ *and* $q \in (0, 2]$, *then* $\|J\|_{2,q}^q \leq d^{\frac{q}{2}} \|J\|_{(2,0)}^{1-\frac{q}{2}}$.

Let $\rho \in (0, 1]$ and $\{J_1, \ldots, J_N\} \subseteq \mathbb{V}_{m,1}$ be the subset given by Lemma 4.5 with $m = p - d$ and $s = \max\{1, (p-d)\rho\}$. Then

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

Applying Lemma 4.4, with $k = 1$, $\delta_N = 1/2$, and $b$ chosen so that $(1+b)/b^2 = \sigma^2$, yields

$$\begin{aligned}\max_i \mathbb{E}_i \|\sin\Theta(\widehat{\mathcal{A}}, \mathcal{A}_i)\|_F &\geq \frac{\epsilon}{4\sqrt{2}}\left[1 - \frac{4n\epsilon^2/\sigma^2}{(1/30)(p-d)\rho(1-\log\rho)} - \frac{\log 2}{\log(p-d)}\right] \\ &= \frac{\epsilon}{4\sqrt{2}}\left[1 - \frac{120\epsilon^2}{\gamma\rho(1-\log\rho)} - \frac{\log 2}{\log(p-d)}\right] \\ &\geq \frac{\epsilon}{4\sqrt{2}}\left[\frac{1}{2} - \frac{120\epsilon^2}{\gamma\rho(1-\log\rho)}\right], \qquad (4.11)\end{aligned}$$

for every estimator $\widehat{\mathcal{A}}$ and all $\epsilon \in [0, 1/\sqrt{2}]$, because $p-d \geq 4$ by assumption. Since $J_i \in \mathbb{V}_{p-d,1}$, Proposition 4.1 implies

$$\|A_\epsilon(J_i)\|_{2,q} \leq \begin{cases} d+s\,, & \text{if } q=0\,, \text{ and} \\ \left(d+\epsilon^q s^{\frac{2-q}{2}}\right)^{1/q}, & \text{if } 0<q<2\,. \end{cases} \tag{4.12}$$

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

$$d+\epsilon^q s^{\frac{2-q}{2}} \leq R_q \iff \epsilon^{2q} \leq \frac{(R_q-d)^2}{s^{2-q}} = \frac{(R_q-d)^2}{\max\{1,(p-d)\rho\}^{2-q}}\,.$$

Thus, eq. (4.12) implies that the constraint

$$\epsilon^{2q} \leq \min\left\{(T/\rho)^2\rho^q\,,\,(R_q-d)^2\right\} \tag{4.13}$$

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

$$\epsilon^2 = \frac{1}{480}\gamma\rho(1-\log\rho) \wedge \frac{1}{2}\,.$$

If we can choose $\rho \in (0,1]$ such that eq. (4.13) is satisfied, then by eq. (4.11),

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

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

$$\rho = \begin{cases} T\left[\gamma(1-\log\rho)\right]^{-\frac{q}{2}}, & \text{if } T < \gamma^{\frac{q}{2}}\,, \text{ and} \\ 1\,, & \text{otherwise.} \end{cases} \tag{4.14}$$

We will verify that $\epsilon$ and $\rho$ satisfy eq. (4.13). The assumption that $1 \leq R_q - d$ guarantees that $\epsilon^{2q} \leq (R_q-d)^2$, because $\epsilon^{2q} \leq 1$. If $T < \gamma^{\frac{q}{2}}$, then

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

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

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

Thus, eq. (4.13) holds and so

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

Now we substitute eq. (4.14) and the definitions of $\gamma$ and $T$ into the above inequality to get the following lower bounds. If $T < \gamma^{\frac{q}{2}}$, then

$$\begin{aligned}\gamma\rho(1-\log\rho) &= T\gamma^{1-\frac{q}{2}}\Big\{1-\log\rho\Big\}^{1-\frac{q}{2}} \\ &= T\gamma^{1-\frac{q}{2}}\left\{1-\log\left(T/\gamma^{\frac{q}{2}}\right)+\frac{q}{2}\log(1-\log\rho)\right\}^{1-\frac{q}{2}} \\ &\geq T\gamma^{1-\frac{q}{2}}\left\{1-\log\left(T/\gamma^{\frac{q}{2}}\right)\right\}^{1-\frac{q}{2}}\end{aligned} \tag{4.15}$$

and so

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

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

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

Proof of Theorem 4.2. 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 4.6 (see Pajor (1998)). *Let $k$ and $s$ be integers satisfying $1 \leq k \leq s-k$, and let $\delta > 0$ There exists a subset $\{J_1,\ldots,J_N\} \subseteq \mathbb{V}_{s,k}$ 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 4.4 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,

$$\|J_i - J_j\|_F \geq \|\sin\Theta(J_i, J_j)\|_F \geq \sqrt{k}\delta\,. \tag{4.16}$$

Let $\rho \in (0,1]$ and $s = \max\{2d, \lfloor (p-d)\rho \rfloor\}$. Invoke Lemma 4.6 with $k = d$ and $\delta = c_2/e$, where $c_2 > 0$ is the constant given by Lemma 4.6. Let $\{J_1, \ldots, J_N\} \subseteq \mathbb{V}_{p-d,d}$ be the subset given by Lemma 4.6 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 eq. (4.16),

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

for all $i \neq j$. The rest of this proof mirrors that of Theorem 4.1. Let $\epsilon \in [0, 1/\sqrt{2}]$ and apply Lemma 4.4 to get

$$\begin{aligned}\max_i \mathbb{E}\|\sin\Theta(\widehat{\mathcal{A}}, \mathcal{A}_i)\|_F &\geq \frac{c_2\sqrt{d}\epsilon}{2\sqrt{2}e}\left[1 - \frac{4nd\epsilon^2/\sigma^2}{(d/4)(p-d)\rho} - \frac{\log 2}{d^2}\right] \\ &\geq c_1\sqrt{d}\epsilon\left[\frac{1}{2} - \frac{16\epsilon^2}{\gamma\rho}\right], \end{aligned} \tag{4.17}$$

where $\gamma$ is defined in eq. (3.1) and we used the assumption that $d \geq 2$. Since $J_i \in \mathbb{V}_{p-d,d}$, Proposition 4.1 implies

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

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

$$d + d^{\frac{q}{2}}\epsilon^q s^{\frac{2-q}{2}} \leq R_q \iff d^q\epsilon^{2q} \leq \frac{(R_q - d)^2}{s^{2-q}} = \frac{(R_q-d)^2}{\max\{2d, (p-d)\rho\}^{2-q}}\,.$$

So $\epsilon$ and $\rho$ must satisfy the constraint

$$d^q\epsilon^{2q} \leq \min\left\{(T/\rho)^2\rho^q\,, \frac{(R_q-d)^2}{(2d)^{2-q}}\right\} \tag{4.18}$$

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

$$\epsilon^2 = \frac{1}{64}\gamma\rho \wedge \frac{1}{2} \tag{4.19}$$

and

$$\rho = \begin{cases} T(d\gamma)^{-\frac{q}{2}} & \text{if } T < (d\gamma)^{\frac{q}{2}}, \text{ and} \\ 1, & \text{otherwise.} \end{cases} \tag{4.20}$$

Since $\epsilon^2 \le 1/2$,

$$d^q\epsilon^{2q} \le \frac{(R_q - d)^2}{(2d)^{2-q}} \iff 2^q\epsilon^{2q} \le \frac{(R_q - d)^2}{4d^2} \impliedby 2d \le R_q - d\,,$$

where the right-hand side is an assumption of the lemma. That verifies one of the inequalities in eq. (4.18). If $T < (d\gamma)^{\frac{q}{2}}$, then

$$(T/\rho)^2\rho^q = (d\gamma\rho)^q\rho^q \ge d^q\epsilon^{2q}\,.$$

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

$$(T/\rho)^2\rho^q = T^2 \ge (d\gamma)^q \ge d^q\epsilon^{2q}\,.$$

Thus, eq. (4.18) holds and by eq. (4.17),

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

Finally, we substitute the definition of $\gamma$ and eq. (4.20) into the above inequality to get the following lower bounds. If $T < (d\gamma)^{\frac{q}{2}}$, then

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

If $T \ge (d\gamma)^{\frac{q}{2}}$, then

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

PROOF OF THEOREM 4.3. The proof is a modification of the proof of Theorem 4.1. 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 4.1, 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 $\lfloor (p-d)/d \rfloor \times d$, i.e. 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 $k$th 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 4.5 with $m = \lfloor (p-d)/d \rfloor$, we obtain a subset $\{J_1, \ldots, J_M\} \subseteq \mathbb{V}_{m,1}$ such that

1. $||J_i||_0 \leq s$ for all $i$,
2. $||J_i - J_j||_2^2 \geq 1/4$ for all $i \neq j$, and
3. $\log M \geq \max\{cs(1+\log(m/s)), \log m\}$, 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 $\{J_1, \ldots, J_M\}^d$, 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; Varshamov, 1957). The $k$th coordinate of each code specifies which element of $\{J_1, \ldots, J_M\}$ to place in the $k$th 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. $||H||_{*,0} \leq s$ for all $H \in \mathcal{H}^s$.
2. $||H_1 - H_2||_2^2 \geq d/8$ for all $H_1, H_2 \in \mathcal{H}^s$ such that $H_1 \neq H_2$.
3. $\log N := \log|\mathcal{H}^s| \geq \max\{cds(1+\log(m/s)), \log m\}$, where $c > 0$ is an absolute constant.

Note that the lower bound of $\log m$ in the 3rd item arises by considering the packing set whose $N$ elements consist of matrices whose columns in $B_1, \ldots, B_d$ are all equal to some $J_i$ for $i = 1, \ldots, M$. This ensures that $\log N \geq \log M \geq \log m$. From here, the proof is a straightforward modifica-

tion of proof of Theorem 4.1 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 $\gamma$ in eq. (3.2). Apply lemma 4.4 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

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

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

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

The constraint

$$d^q\epsilon^{2q} \leq \min\left\{(T_*/\rho)^2\rho^q\,,\, d^q(R_q-1)^2\right\}$$

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

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

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

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

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

$$\sup_{\mathcal{P}_q^*(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F \geq c_2\left\{\gamma\rho(1-\log\rho)\wedge d\right\}^{\frac{1}{2}},$$

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

$$\begin{aligned}&\sup_{\mathcal{P}_q^*(\sigma^2,R_q)} \mathbb{E}\|\sin\Theta(\widehat{\mathcal{S}},\mathcal{S})\|_F\\ &\quad\geq c_3\left\{d(R_q-1)\left[\frac{\sigma^2}{n}\left(1-\log\left(T_*/\gamma^{\frac{q}{2}}\right)\right)\right]^{1-\frac{q}{2}}\wedge d\right\}^{\frac{1}{2}}.\end{aligned}$$

If $T_* \geq \gamma^{\frac{q}{2}}$, then

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

## 5. Upper bound proofs.

5.1. *A variational approach to the perturbation of spectral subspaces.* The following result allows us to bound the curvature of the matrix functional $F \mapsto \langle A, F\rangle$ at its point of maximum on the Grassmann manifold.

LEMMA 5.1 (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 $d \leq p$. Let $\mathcal{E}$ be the $d$-dimensional subspace spanned by the eigenvectors of $A$ corresponding to its $d$ largest eigenvalue, and let $E$ denote its orthogonal projection. If $\mathcal{F}$ is a $d$-dimensional subspace of $\mathbb{R}^p$ and $F$ is its orthogonal projection, then*

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

Using this lemma we have the following alternative to the traditional matrix perturbation approach to bounding subspace distances using the Davis-Kahan $\sin\Theta$ Theorem and Weyl's Inequality.

LEMMA 5.2 (Variational $\sin\Theta$). *In addition to the hypotheses of Lemma 5.1, if $F$ satisfies*

$$\langle B, E\rangle - g(E) \leq \langle B, F\rangle - g(F) \tag{5.1}$$

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

$$\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F^2 \leq \frac{\langle B-A, F-E\rangle - [g(F)-g(E)]}{\lambda_d(A)-\lambda_{d+1}(A)}. \tag{5.2}$$

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

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

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. Both Lemma 5.1 and Lemma 5.2 are proved in the appendix.

5.2. *Proofs of the main upper bounds.* $\Sigma$ and $S_n$ are both invariant under translations of $\mu$. Since our estimators only depend on $X_1, \ldots, X_n$ 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 negligble (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. We apply Lemma 5.2 taking $A = \Sigma$, $B = S_n$, $E = VV^T$, and $F = \hat{V}\hat{V}^T$, where $\hat{V}$ is a solution of eq. (2.8). Since $VV^T$ and $\hat{V}\hat{V}^T$ are feasible and $\hat{V}\hat{V}^T$ is optimal,

$$\langle S_n, VV^T\rangle \le \langle S_n, \hat{V}\hat{V}^T\rangle .$$

Thus,

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

and

$$\hat{\epsilon}^2 \le \frac{\sqrt{2}}{\lambda_d - \lambda_{d+1}}\left\langle S_n - \Sigma, \frac{\hat{V}\hat{V}^T - VV^T}{\|\hat{V}\hat{V}^T - VV^T\|_F}\right\rangle \hat{\epsilon}\,, \tag{5.3}$$

because $\|\hat{V}\hat{V}^T - VV^T\|_F^2 = 2\hat{\epsilon}^2$ by eq. (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 $\Delta$ 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

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

Equation (5.3) implies

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

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

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

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 $\delta$-net argument (see Propositions B.1 and B.2), we have, when $p>5$,

$$\big\|\|\mathcal{Z}\|_{2,\infty}\big\|_{\psi_2} \le 4.15\sqrt{d+\log p}\,. \tag{5.4}$$

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 $6\sqrt{R_0(d+\log p)}\le\sqrt{n}$. □

Proof of Theorem 3.5. Again, we start from Lemma 5.2, which gives

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

To get the correct dependence on $\lambda_i$ 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

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

For any projection matrix $\Pi$ we write $\Pi^\perp := I-\Pi$, the projection onto the orthogonal complement. By Proposition A.1 we have

$$\langle W, \hat\Pi-\Pi\rangle = \langle W, \Pi\hat\Pi^\perp\Pi\rangle + \langle W, \Pi^\perp\hat\Pi\Pi\rangle + \langle W, \Pi^\perp\hat\Pi\Pi^\perp\rangle \tag{5.5}$$

$$=: T_1+T_2+T_3 \tag{5.6}$$

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

*Controlling* $T_1$.

$$\begin{aligned} T_1 =& \langle W, \Pi\hat{\Pi}^{\perp}\Pi\rangle = \langle \Pi W\Pi, \Pi\hat{\Pi}^{\perp}\Pi\rangle \\ \le& \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^{\perp}\Pi\|_* = \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^{\perp}\hat{\Pi}^{\perp}\Pi\|_* \\ =& \|\Pi W\Pi\|_2 \|\Pi\hat{\Pi}^{\perp}\|_F^2 \le \|\Pi W\Pi\|_2 \hat{\epsilon}^2, \end{aligned} \tag{5.7}$$

where $\|\cdot\|_*$ is the nuclear norm ($\ell_1$ norm of the singular values) and $\|\cdot\|_2$ is the spectral norm (or operator norm). By Lemma B.5, we have (recall that we assume $\|Z\|_{\psi_2} \le 1$ and $\epsilon_n \le 1$ for simplicity),

$$\big\|\|\Pi W\Pi\|_2\big\|_{\psi_1} \le c_1\lambda_1\sqrt{d/n}, \tag{5.8}$$

where $c_1$ is a universal constant. Define

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

Then, when $n \ge 2$ we have

$$\mathbb{P}(\Omega_1) \le \mathbb{P}\left(\|\Pi W\Pi\|_2 \ge c_1\lambda_1 \log n\sqrt{d/n}\right) \le (n-1)^{-1}. \tag{5.9}$$

*Controlling* $T_2$.

$$\begin{aligned} T_2 &= \langle W, \Pi^{\perp}\hat{\Pi}\Pi\rangle = \langle \Pi^{\perp}W\Pi, \Pi^{\perp}\hat{\Pi}\rangle \\ &\le \|\Pi^{\perp}W\Pi\|_{2,\infty}\|\Pi^{\perp}\hat{\Pi}\|_{2,1}. \end{aligned} \tag{5.10}$$

To bound $\|\Pi^{\perp}\hat{\Pi}\|_{2,1}$, let the rows of $\Pi^{\perp}\hat{\Pi}$ be denoted by $\phi_1, \ldots, \phi_p$ and $t > 0$. Using a standard argument of bounding $\ell_1$ norm by the $\ell_q$ and $\ell_2$ norms (e.g., Raskutti, Wainwright and Yu, 2011, Lemma 5), we have for all $t > 0$, $0 < q \le 1$,

$$\begin{aligned} \|\Pi^{\perp}\hat{\Pi}\|_{2,1} &= \sum_{i=1}^{p}\|\phi_i\|_2 \\ &\le \left[\sum_{i=1}^{p}\|\phi_i\|_2^q\right]^{1/2}\left[\sum_{i=1}^{p}\|\phi_i\|_2^2\right]^{1/2} t^{-q/2} + \left[\sum_{i=1}^{p}\|\phi_i\|_2^q\right] t^{1-q} \\ &= \|\Pi^{\perp}\hat{\Pi}\|_{2,q}^{q/2}\|\Pi^{\perp}\hat{\Pi}\|_F t^{-q/2} + \|\Pi^{\perp}\hat{\Pi}\|_{2,q}^{q} t^{1-q} \\ &\le \sqrt{2}R_q^{1/2}t^{-q/2}\hat{\epsilon} + 2R_q t^{1-q}, \end{aligned} \tag{5.11}$$

where the last step uses the fact that

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

Combining eqs. (5.10) and (5.11) we obtain, for all $t > 0$, $0 < q < 1$,

$$T_2 \le \|\Pi^{\perp}W\Pi\|_{2,\infty}\left(\sqrt{2}R_q^{1/2}t^{-q/2}\hat{\epsilon} + 2R_q t^{1-q}\right)\,. \tag{5.12}$$

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

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

Taking $t = t_{2,2}$ in eq. (5.12) and using the tail bound result in Lemma B.1, we have

$$\begin{aligned}\mathbb{P}(\Omega_2) \le& \mathbb{P}(\|\Pi^{\perp}W\Pi\|_{2,\infty} \ge t_{2,1}) \\ \le& 2p5^d\exp\left(-\frac{t_{2,1}^2/8}{2\lambda_1\lambda_{d+1}/n + t_{2,1}\sqrt{\lambda_1\lambda_{d+1}}/n}\right) \\ \le& p^{-1}.\end{aligned} \tag{5.13}$$

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

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

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

$$\mathbb{P}\left(\phi(R_q,\epsilon) \ge c_3 x\lambda_{d+1}\left(\epsilon_n\epsilon^2 + \epsilon_n^2\epsilon + \epsilon_n^4\right)\right) \le 2\exp(-x^{2/5}).$$

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

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

Define function $g(\epsilon) = \epsilon_n\epsilon^2 + \epsilon_n^2\epsilon + \epsilon_n^4$. Then for all $\epsilon \geq 0$, we have $g(\epsilon) \geq \epsilon_n^4 \geq 4d^3/n^2$. On the other hand, if $\epsilon = \|\sin(U,V)\|_F$, then $\epsilon^2 \leq 2d$ and hence $g(\epsilon) \leq g(\sqrt{2d}) = 2d + \sqrt{2d} + 1$. Let $\mu = \epsilon_n^4$ 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.

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

Define

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

Then we have proved that

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

*Put things together.* Now recall the conditions in Equations (3.3) to (3.6). On $\Omega_1^c \bigcap \Omega_2^c \bigcap \Omega_3^c$, we have, from eq. (5.5) that

$$\begin{aligned}
(\lambda_d - \lambda_{d+1})\hat{\epsilon}^2 \leq &\left(c_1\sqrt{\frac{d}{n}}\log n\lambda_1 + c_3\epsilon_n(\log n)^{5/2}\lambda_{d+1}\right)\hat{\epsilon}^2 \\
&+ 21\sqrt{\lambda_1\lambda_{d+1}}^{1-q/2}(\lambda_d - \lambda_{d+1})^{q/2}\epsilon_n\hat{\epsilon} \\
&+ 21\sqrt{\lambda_1\lambda_{d+1}}^{2-q}(\lambda_d - \lambda_{d+1})^{-(1-q)}\epsilon_n^2, \\
\Longrightarrow\ \frac{1}{2}(\lambda_d - \lambda_{d-1})\hat{\epsilon}^2 \leq &21\sqrt{\lambda_1\lambda_{d+1}}^{1-q/2}(\lambda_d - \lambda_{d+1})^{q/2}\epsilon_n\hat{\epsilon} \\
&+ 21\sqrt{\lambda_1\lambda_{d+1}}^{2-q}(\lambda_d - \lambda_{d+1})^{-(1-q)}\epsilon_n^2, \\
\Longrightarrow\ \hat{\epsilon} \leq &9\left(\frac{\sqrt{\lambda_1\lambda_{d+1}}}{\lambda_d - \lambda_{d+1}}\right)^{1-q/2}\epsilon_n.
\end{aligned}$$

□

**6. Discussion.** We have derived non-asymptotic minimax upper and lower bounds for principal subspace estimation over two classes of sparse subspaces. In the row sparse case, our upper and lower bounds match up to constants and are optimal in $(n, p, d, R_q, \sigma^2)$ in the sparse, high-dimensional regime. In the column sparse case, our upper and lower bounds match up to constants and are optimal in $(n, p, R_q, \sigma^2)$. We conjecture that the $d + \log p$ term that appears in the column sparse upper bound (Corollary 3.1) can be improved to $1 + \log p$, and thus match the lower bound (Theorem 3.3). It appears to us that the primary obstacle is tightening our analysis of a cross-product term ($T_2$ in eq. (5.6)) that roughly corresponds to the cross-covariance of the principal subspace and its orthocomplement. This is an interesting technical challenge, but after all, deriving non-asymptotic bounds that are optimal in all five parameters $(n, p, d, R_q, \sigma^2)$ seems too ambitious.

Interestingly, in the case $d = 1$ (where row and column sparsity coincide), the form of the minimax optimal error for the principal subspace estimation problem parallels that for the coefficient vector in the sparse linear model (see Raskutti, Wainwright and Yu, 2011) with the noise-to-signal ratio $\sigma^2$ playing the same role in the error as the noise variance in the linear model. For $d > 1$, we suspect that this parallel relationship will continue to hold with the multivariate (or multiple response) sparse linear model under appropriate sparsity conditions. However, minimax rates have yet to be established for that problem.

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 $\ell_1$ constrained PCA, the optimization problem remains challenging as there are no known algorithms to efficiently compute a global maximum. Finally, although the minimax optimal estimators that we propose do not require knowledge of the noise-to-signal ratio $\sigma^2$, they do require knowledge of (or an upperbound on) the sparsity $R_q$. It is not hard to modify our techniques to produce an estimator that gives up adaptivity to $\sigma^2$ in exchange for adaptivity to $R_q$. One could do this by using penalized versions of our estimators with a penalty factor proportional to $\sigma^2$. An extension along this line has already been considered by Lounici (2012) for the $d = 1$ case. A more interesting question is whether or not there exists fully adaptive principal subspace estimators. In other words, under what conditions can one find an estimator that achieves the minimax optimal error without requiring knowledge of either $\sigma^2$ or $R_q$?

## APPENDIX A: ADDITIONAL PROOFS

PROOF OF PROPOSITION 2.2. Let $\gamma_i$ be the cosine of the $i$th 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. □

### A.1. Proofs related to the lower bounds.

PROOF OF LEMMA 4.2. Write $\Sigma_i = \Sigma(A_i)$ for $i = 1, 2$. Since $\Sigma_1$ and $\Sigma_2$ are nonsingular and have the same determinant,

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

Now

$$\Sigma_2^{-1} = (1+b)^{-1}A_2A_2^T + (I_p - A_2A_2^T)$$

and

$$\Sigma_1 - \Sigma_2 = b(A_1A_1^T - A_2A_2^T)\,.$$

Thus,

$$\begin{aligned} &\operatorname{Tr}\left(\Sigma_2^{-1}(\Sigma_1-\Sigma_2)\right) \\ &= \frac{b}{1+b}\left\{(1+b)\langle I_p - A_2A_2^T, A_1A_1^T\rangle - \langle A_2A_2^T, A_2A_2^T - A_1A_1^T\rangle\right\} \\ &= \frac{b-1}{b}\{b\langle I_p - A_2A_2^T, A_1A_1^T\rangle - \langle I_p, A_2A_2^T - A_2A_2^TA_1A_1^T\rangle\} \\ &= \frac{b}{1+b}\left\{(1+b)\langle I_p - A_2A_2^T, A_1A_1^T\rangle - \langle A_2A_2^T, I_p - A_1A_1^T\rangle\right\} \\ &= \frac{b^2}{1+b}\|\sin(A_1,A_2)\|_F^2\,, \end{aligned}$$

by Proposition 2.1. □

PROOF OF LEMMA 4.3. By Proposition 2.1 and the definition of $A_\epsilon(\cdot)$,

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

The upper bound follows from Proposition 2.2:

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

PROOF OF LEMMA 4.5. 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. $\log|\Omega_m^{s_0}| \geq c s_0 \log(m/s_0)$, where $c > 0.251$.

Let

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

Clearly, $\{J_1, \ldots, J_N\} \subseteq \mathbb{V}_{m,1}$ 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 $\{J_1, \ldots, J_N\}$ satisfies

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

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

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

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

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

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 $\{J_1, \ldots, J_N\}$ taken to be the $N = m$ vectors $\{(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)\} \subseteq \{0, 1\}^m$. That yields, in combination with eq. (A.1),

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

### A.2. Proofs related to the upper bounds.

Proof of Lemma 5.1. 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

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

Since orthogonal projections are idempotent,

$$\begin{aligned}
\lambda_d\langle E, F^\perp\rangle - \lambda_{d+1}\langle E^\perp, F\rangle &= \lambda_d\langle EF^\perp, EF^\perp\rangle - \lambda_{d+1}\langle E^\perp F, E^\perp F\rangle \\
&= \lambda_d\|EF^\perp\|_2^2 - \lambda_{d+1}\|E^\perp F\|_F^2\,.
\end{aligned}$$

Now apply Proposition 2.1 to conclude that

$$\lambda_d\|EF^\perp\|_2^2 - \lambda_{d+1}\|E^\perp F\|_F^2 = (\lambda_d - \lambda_{d+1})\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F^2\,. \qquad \square$$

PROOF OF LEMMA 5.2. Equation (5.1) is equivalent to

$$0 \le \langle B, F - E\rangle - [g(F) - g(E)]\,.$$

Then by Lemma 5.1,

$$\begin{aligned}[\lambda_d(A) - \lambda_{d+1}(A)]\,\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F^2 &\le -\langle A, F - E\rangle \\ &\le \langle B - A, F - E\rangle - [g(F) - g(E)]\,. \quad \square\end{aligned}$$

PROPOSITION A.1. *If $W$ is symmetric, and $E$ and $F$ are orthogonal projections, then*

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

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 B: EMPIRICAL PROCESS RELATED PROOFS

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

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

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

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

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

*for a universal constant $c > 0$.*

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

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

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

PROPOSITION B.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

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

and the triangle inequality, we have that

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

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

PROOF OF LEMMA B.1. Let $N_\delta$ be a minimal $\delta$-net in $\mathbb{S}_2^{d-1}$ for some $\delta \in (0,1)$ to be chosen later. By Proposition B.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 $j$th column of $I_{p\times p}$. Taking $\delta = 1/2$, by Proposition B.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 B.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 B.2) implies that for all $t > 0$ and every $u \in \mathcal{N}_\delta$

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

### B.2. The quadratic terms.

LEMMA B.4. *Let $\epsilon \geq 0$, $q \in (0,1]$, and*

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

*There exists a constant $c > 0$ such that*

$$\mathbb{E}\phi(R_q,\epsilon) \leq c\|Z_1\|_{\psi_2}^2\lambda_{d+1}\left\{\epsilon\frac{E(R_q,\epsilon)}{\sqrt{n}} + \frac{E^2(R_q,\epsilon)}{n}\right\},$$

*where*

$$E(R_q,\epsilon) = \mathbb{E}\sup\{\langle \mathcal{Z}, U\rangle \,:\, U \in \mathbb{V}_{p,d}\,, \|U\|_{2,q}^q \leq 2R_q\,, \|U\|_F \leq \epsilon\}$$

*and $\mathcal{Z}$ is a $(p-d)\times d$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Moreover, we have, for another numerical constant $c'$,*

$$\frac{E(R_q,\epsilon)}{\sqrt{n}} \leq c'(R_q^{1/2}t^{1-q/2}\epsilon + R_q t^{2-q}) \tag{B.1}$$

*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,

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

and so

$$\mathbb{E}(R_q,\epsilon) \leq \mathbb{E}\|\mathcal{Z}\|_{2,\infty}\sup\{\|U\|_{2,1} \,:\, U \in \mathbb{V}_{p,d}\,, \|U\|_{2,q}^q \leq 2R_q\,, \|U\|_F \leq \epsilon\}\,.$$

By eq. (5.4) and the fact that the Orlicz $\psi_2$-norm bounds the expectation,

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

Now $\|U\|_{2,1}$ is just the $\ell_1$ norm of the vector of row-wise norms of $U$. So we use a standard argument to bound the $\ell_1$ norm in terms of the $\ell_2$ and $\ell_q$ 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,

$$\sup\{\|U\|_{2,1} : U \in \mathbb{V}_{p,d}\,, \|U\|_{2,q}^q \le 2R_q\,, \|U\|_F \le \epsilon\} \le R_q^{1/2}t^{-q/2} + R_q t^{1-q}\,.$$

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

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

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

PROOF. Let $\mathcal{N}_\delta$ be a minimal $\delta$-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 B.1, for all $t > 0$ and every $u \in \mathcal{N}_\delta$

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

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

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

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

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

Thus,

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

## 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.

## REFERENCES

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics* **37** 2877–2921.

Bhatia, R. (1997). *Matrix Analysis*. Springer-Verlag.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2012). Minimax bounds for sparse PCA with noisy high-dimensional data. to appear.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Annals of Statistics* **38** 3696–3723.

Chikuse, Y. (2003). *Statistics on Special Manifolds*. Springer.

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 Review* **49** 434–448.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over $l_p$-balls for $l_q$-error. *Probability Theory and Related Fields* **99** 277–303.

Edelman, A., Arias, T. A. and Smith, S. T. (1998). The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications* **20** 303–353.

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. *Journal of the American Statistical Association* **104** 682–693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12** 531-547.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.

Lounici, K. (2012). Sparse principal component analysis with missing observations.

Ma, Z. (2011). Sparse principal component analysis and iterative thresholding.

Massart, P. (2007). *Concentration Inequalities and Model Selection*. Springer-Verlag.

Mendelson, S. (2010). Empirical Processes with a Bounded $\psi_1$ Diameter. *Geometric and Functional Analysis* **20** 988–1027.

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

PAJOR, A. (1998). Metric Entropy of the Grassmann Manifold. In *Convex Geometric Analysis. MSRI Publications* **34** 181–188.

PAUL, D. (2007). Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. *Statistica Sinica* **17** 1617–1642.

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 $\ell_q$-balls. *IEEE Transactions on Information Theory*. to appear.

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* **99** 1015-1034.

SHEN, D., SHEN, H. and MARRON, J. S. (2011). Consistency of sparse PCA in high dimension, low sample size contexts.

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

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag.

VARSHAMOV, R. R. (1957). Estimate of the number of signals in error correcting codes. *Dokl. Acad. Nauk SSSR* **117** 739–741.

VU, V. Q. and LEI, J. (2012a). Minimax Rates of Estimation for Sparse PCA in High Dimensions. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

VU, V. Q. and LEI, J. (2012b). Squared-norm empirical process in Banach space. Manuscript available at http://vince.vu/papers/squared-norm-process.pdf.

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: Research Papers in Probability and Statistics* 423–435.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics* **15** 265–286.

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

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