The Annals of Statistics
2013, Vol. 41, No. 3, 1055–1084
DOI: 10.1214/12-AOS1014

MINIMAX BOUNDS FOR SPARSE PCA WITH NOISY HIGH-DIMENSIONAL DATA

BY AHARON BIRNBAUM, IAIN M. JOHNSTONE[1], BOAZ NADLER[2] AND DEBASHIS PAUL[3]

Hebrew University, Stanford University, Weizmann Institute of Science and University of California, Davis

We study the problem of estimating the leading eigenvectors of a high-dimensional population covariance matrix based on independent Gaussian observations. We establish a lower bound on the minimax risk of estimators under the l_2 loss, in the joint limit as dimension and sample size increase to infinity, under various models of sparsity for the population eigenvectors. The lower bound on the risk points to the existence of different regimes of sparsity of the eigenvectors. We also propose a new method for estimating the eigenvectors by a two-stage coordinate selection scheme.

1. Introduction. Principal components analysis (PCA) is widely used to reduce dimensionality of multivariate data. A traditional setting involves repeated observations from a multivariate normal distribution. Two key theoretical questions are: (i) *what is the relation between the sample and population eigenvectors*, *and* (ii) *how well can population eigenvectors be estimated under various sparsity assumptions*? When the dimension N of the observations is fixed and the sample size $n \to \infty$, the asymptotic properties of the sample eigenvalues and eigenvectors are well known [Anderson (1963), Muirhead (1982)]. This asymptotic analysis works because the sample covariance approximates the population covariance well when the sample size is large. However, it is increasingly common to encounter statistical problems where the dimensionality N is comparable to, or larger than, the sample size n. In such cases, the sample covariance matrix, in general, is not a reliable estimate of its population counterpart.

Better estimators of large covariance matrices, under various models of sparsity, have been studied recently. These include development of banding and thresholding schemes [Bickel and Levina (2008a, 2008b), Cai and Liu (2011), El Karoui (2008), Rothman, Levina and Zhu (2009)], and analysis of their rate of convergence in the spectral norm. More recently, Cai, Zhang and Zhou (2010) and Cai

Received February 2012.

[1] Supported by NSF Grants DMS-09-06812 and NIH R01 EB001988.

[2] Supported by a grant from the Israeli Science Foundation (ISF).

[3] Supported by the NSF Grants DMR 1035468 and DMS-11-06690.

MSC2010 subject classifications. Primary 62G20; secondary 62H25.

Key words and phrases. Minimax risk, high-dimensional data, principal component analysis, sparsity, spiked covariance model.

and Zhou (2012) established the minimax rate of convergence for estimation of the covariance matrix under the matrix l_1 norm and the spectral norm, and its dependence on the assumed sparsity level.

In this paper we consider a related but different problem, namely, the estimation of the leading eigenvectors of the covariance matrix. We formulate this eigenvector estimation problem under the well-studied "spiked population model" which assumes that the ordered set of eigenvalues $\mathcal{L}(\Sigma)$ of the population covariance matrix Σ satisfies

$$\mathcal{L}(\Sigma) = \{\lambda_1 + \sigma^2, \dots, \lambda_M + \sigma^2, \sigma^2, \dots, \sigma^2\} \tag{1.1}$$

for some $M \geq 1$, where $\sigma^2 > 0$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_M > 0$. This is a standard model in several scientific fields, including, for example, array signal processing [see, e.g., van Trees (2002)] where the observations are modeled as the sum of an M-dimensional random signal and an independent, isotropic noise. It also arises as a latent variable model for multivariate data, for example, in factor analysis [Jolliffe (2002), Tipping and Bishop (1999)]. The assumption that the leading M eigenvalues are distinct is made to simplify the analysis, as it ensures that the corresponding eigenvectors are identifiable up to a sign change. The assumption that all remaining eigenvalues are equal is not crucial as our analysis can be generalized to the case when these are only bounded by σ^2. Asymptotic properties of the eigenvalues and eigenvectors of the sample covariance matrix under this model, in the setting when $N/n \to c \in (0, \infty)$ as $n \to \infty$, have been studied by Lu (2002), Baik and Silverstein (2006), Nadler (2008), Onatski (2006) and Paul (2007), among others. A key conclusion is that when $N/n \to c > 0$, the eigenvectors of standard PCA are inconsistent estimators of the population eigenvectors.

Eigenvector and covariance matrix estimation are related in the following way. When the population covariance is a low rank perturbation of the identity, as in this paper, sparsity of the eigenvectors corresponding to the nonunit eigenvalues implies sparsity of the whole covariance. Consistency of an estimator of the whole covariance matrix in spectral norm implies convergence of its leading eigenvalues to their population counterparts. If the gaps between the distinct eigenvalues remain bounded away from zero, it also implies convergence of the corresponding eigen-subspaces [El Karoui (2008)]. In such cases, upper bounds for sparse covariance estimation in the spectral norm, as in Bickel and Levina (2008a) and Cai and Zhou (2012), also yield upper bounds on the rate of convergence of the corresponding eigenvectors under the l_2 loss. These works, however, did not study the following fundamental problem, considered in this paper: *How well can the leading eigenvectors be estimated*, *and namely*, *what are the minimax rates for eigenvector estimation*? Indeed, it turns out that the optimal rates for covariance matrix estimation and leading eigenvector estimation are different. Moreover, schemes based on thresholding the entries of the sample covariance matrix do not achieve the minimax rate for eigenvector estimation. The latter result is beyond the scope of this paper and will be reported in a subsequent publication by the current authors.

Several works considered various models of sparsity for the leading eigenvectors and developed improved sparse estimators. For example, Witten, Tibshirani and Hastie (2009) and Zou, Hastie and Tibshirani (2006), among others, imposed l_1-type sparsity constraints directly on the eigenvector estimates and proposed optimization procedures for obtaining them. Shen and Huang (2008) suggested a regularized low rank approach to sparse PCA. The consistency of the resulting leading eigenvectors was recently proven in Shen, Shen and Marron (2011), in a model in which the sample size n is fixed while $N \to \infty$. d'Aspremont et al. (2007) suggested a semi-definite programming (SDP) problem as a relaxation to the l_0-penalty for sparse population eigenvectors. Assuming a single spike, Amini and Wainwright (2009) studied the asymptotic properties of the resulting leading eigenvector of the covariance estimator in the joint limit as both sample size and dimension tend to infinity. Specifically, they considered a leading eigenvector with exactly $k \ll N$ nonzero entries all of the form $\{-1/\sqrt{k}, 1/\sqrt{k}\}$. For this hardest subproblem in the k-sparse l_0-ball, Amini and Wainwright (2009) derived information theoretic lower bounds for such eigenvector estimation.

In this paper, in contrast, following Johnstone and Lu (2009) (JL), we study estimation of the leading eigenvectors of Σ assuming that these are *approximately* sparse, with a bounded l_q norm. Under this model, JL developed an estimation procedure based on coordinate selection by thresholding the diagonal of the sample covariance matrix, followed by the spectral decomposition of the submatrix corresponding to the selected coordinates. JL further proved consistency of this estimator assuming dimension grows at most polynomially with sample size, but did not study its convergence rate. Since this estimation procedure is considerably simpler to implement and computationally much faster than the l_1 penalization procedures cited above, it is of interest to understand its theoretical properties. More recently, Ma (2011) developed iterative thresholding sparse PCA (ITSPCA), which is based on repeated filtering, thresholding and orthogonalization steps that result in sparse estimators of the subspaces spanned by the leading eigenvectors. He also proved consistency and derived rates of convergence under appropriate loss functions and sparsity assumptions. In a later work, Cai, Ma and Wu (2012) considered a two-stage estimation scheme for the leading population eigenvector, in which the first stage is similar to the DT scheme applied to a stochastically perturbed version of the data. The estimates of the leading eigenvectors from this step are then used to project another stochastically perturbed version of the data to obtain the final estimates of the eigenvectors through solving an orthogonal regression problem. They showed that this two-stage scheme achieves the optimal rate for estimation of eigen-subspaces under suitable sparsity conditions.

In this paper, which is partly based on the Ph.D. thesis of Paul [Paul (2005)], we study the estimation of the leading eigenvectors of Σ, all assumed to belong to appropriate l_q spaces. Our analysis thus extends the JL setting and complements the work of Amini and Wainwright (2009) in the l_0-sparsity setting. For simplicity, we assume Gaussian observations in our analysis.

The main contributions of this paper are as follows. First, we establish lower bounds on the rate of convergence of the minimax risk for any eigenvector estimator under the l_2 loss. This analysis points to three different regimes of sparsity, which we denote *dense*, *thin and sparse*, each having its own rate of convergence. We show that in the "dense" setting (as defined in Section 3), the standard PCA estimator attains the optimal rate of convergence, whereas in sparser settings it is not even consistent. Next, we show that while the JL diagonal thresholding (DT) scheme is consistent under these sparsity assumptions, it is not rate optimal in general. This motivates us to propose a new refined thresholding method (Augmented Sparse PCA, or ASPCA) that is based on a two-stage coordinate selection scheme. In the sparse setting, both our ASPCA procedure, as well as the method of Ma (2011) achieve the lower bound on the minimax risk obtained by us, and are thus rate-optimal procedures, so long as DT is consistent. For proofs see Ma (2011) and Paul and Johnstone (2007). There is a somewhat special, intermediate, "thin" region where a gap exists between the current lower bound and the upper bound on the risk. It is an open question whether the lower bound can be improved in this scenario, or a better estimator can be derived. Table 1 provides a comparison of the lower bounds and rates of convergence of various estimators.

The theoretical results also show that under comparable scenarios, the optimal rate for eigenvector estimation $O((\log N/n)^{-(1-q/2)})$ (under squared-error loss) is faster than the rate obtained for sparse covariance estimation, $O((\log N/n)^{-(1-q)})$ (under squared operator norm loss), by Bickel and Levina (2008a) and shown to be optimal by Cai and Zhou (2012).

Finally, we emphasize that to obtain good finite-sample performance for both our two-stage scheme, as well as for other thresholding methods, the exact thresholds need to be carefully tuned. This issue and the detailed theoretical analysis of the ASPCA estimator are beyond the scope of this paper, and will be presented in a future publication. After this paper was completed, we learned of Vu and Lei

TABLE 1
Comparison of lower bounds on eigenvector estimation and worst case rates of various procedures

Estimator **Lower bound**	**Dense** $O(N/n)$	**Thin** $O(n^{-(1-q/2)})$	**Sparse** $O((\log N/n)^{1-q/2})$
PCA	Rate optimal*	Inconsistent	Inconsistent
DT	Inconsistent	Inconsistent	Not rate optimal
ASPCA	Inconsistent	Inconsistent	Rate optimal†

*When $N/n \to 0$.

†So long as DT is consistent.

(2012), which cites Paul and Johnstone (2007) and contains results overlapping with some of the work of Paul and Johnstone (2007) and this paper.

The rest of the paper is organized as follows. In Section 2, we describe the model for the eigenvectors and analyze the risk of the standard PCA estimator. In Section 3, we present the lower bounds on the minimax risk of any eigenvector estimator. In Section 4, we derive a lower bound on the risk of the diagonal thresholding estimator proposed by Johnstone and Lu (2009). In Section 5, we propose a new estimator named ASPCA (augmented sparse PCA) that is a refinement of the diagonal thresholding estimator. In Section 6, we discuss the question of attainment of the risk bounds. Proofs of the results are given in Appendix A.

Throughout, $\mathbb{S}^{N-1}$ denotes the unit sphere in $\mathbb{R}^N$ centered at the origin, $\lfloor x \rfloor$ denotes the largest integer less than or equal to $x \in \mathbb{R}$.

2. Problem setup.

We suppose a triangular array model, in which for each n, the random vectors $X_i := X_i^n, i = 1, \dots, n$, each have dimension $N = N(n)$ and are independent and identically distributed on a common probability space. Throughout we assume that X_i's are i.i.d. as $N_N(\mathbf{0}, \Sigma)$, where the population matrix Σ, also depending on N, is a finite rank perturbation of (a multiple of) the identity. In other words,

$$\Sigma = \sum_{\nu=1}^{M} \lambda_\nu \theta_\nu \theta_\nu^T + \sigma^2 I, \tag{2.1}$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_M > 0$, and the vectors $\theta_1, \dots, \theta_M \in \mathbb{R}^N$ are orthonormal, which implies (1.1). θ_ν is the eigenvector of Σ corresponding to the νth largest eigenvalue, namely, $\lambda_\nu + \sigma^2$. The term "finite rank" means that M remains fixed even as $n \to \infty$. The asymptotic setting involves letting both n and N grow to infinity simultaneously. For simplicity, we assume that the λ_ν's are fixed while the parameter space for the θ_ν's varies with N.

The observations can be described in terms of the model

$$X_{ik} = \sum_{\nu=1}^{M} \sqrt{\lambda_\nu} v_{\nu i} \theta_{\nu k} + \sigma Z_{ik}, \qquad i = 1, \dots, n, k = 1, \dots, N. \tag{2.2}$$

Here, for each i, $v_{\nu i}$, Z_{ik} are i.i.d. $N(0, 1)$. Since the eigenvectors of Σ are invariant to a scale change in the original observations, it is henceforth assumed that $\sigma = 1$. Hence, $\lambda_1, \dots, \lambda_M$ in the asymptotic results should be replaced by $\lambda_1/\sigma^2, \dots, \lambda_M/\sigma^2$ when (2.1) holds with an arbitrary $\sigma > 0$. Since the main focus of this paper is estimation of eigenvectors, without loss of generality we consider the uncentered sample covariance matrix $\mathbf{S} := n^{-1}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X} = [X_1 : \dots : X_n]$.

The following condition, termed *Basic assumption*, will be used throughout the asymptotic analysis, and will be referred to as (BA).

(BA) (2.2) holds with $\sigma = 1$; $N = N(n) \to \infty$ as $n \to \infty$; $\lambda_1 > \cdots > \lambda_M > 0$ are fixed (do not vary with N); M is unknown but fixed.

2.1. *Eigenvector estimation with squared error loss.* Given data $\{X_i\}_{i=1}^n$, the goal is to estimate M and the eigenvectors $\theta_1, \dots, \theta_M$. For simplicity, to derive the lower bounds, we first assume that M is known. In Section 5.2 we derive an estimator of M, which can be shown to be consistent under the assumed sparsity conditions. To assess the performance of any estimator, a minimax risk analysis approach is proposed. The first task is to specify a loss function $L(\widehat{\theta}_\nu, \theta_\nu)$ between the estimated and true eigenvector.

Eigenvectors are invariant to choice of sign, so we introduce a notation for the *acute* (angle) *difference* between unit vectors,

$$\mathbf{a} \ominus \mathbf{b} = \mathbf{a} - \operatorname{sign}(\langle \mathbf{a}, \mathbf{b} \rangle)\mathbf{b},$$

where $\mathbf{a}$ and $\mathbf{b}$ are $N \times 1$ vectors with unit l_2 norm. We consider the following loss function, also invariant to sign changes:

$$L(\mathbf{a}, \mathbf{b}) := 2\bigl(1 - |\langle \mathbf{a}, \mathbf{b} \rangle|\bigr) = \|\mathbf{a} \ominus \mathbf{b}\|^2. \tag{2.3}$$

An estimator $\widehat{\theta}_\nu$ is called consistent with respect to L, if $L(\widehat{\theta}_\nu, \theta_\nu) \to 0$ in probability as $n \to \infty$.

2.2. *Rate of convergence for ordinary PCA.* We first consider the asymptotic risk of the leading eigenvectors of the sample covariance matrix (henceforth referred to as the standard PCA estimators) when the ratio $N/n \to 0$ as $n \to \infty$. For future use, we define

$$h(\lambda) := \frac{\lambda^2}{1+\lambda}, \qquad \lambda > 0, \tag{2.4}$$

and

$$g(\lambda, \tau) = \frac{(\lambda - \tau)^2}{(1+\lambda)(1+\tau)}, \qquad \lambda, \tau > 0. \tag{2.5}$$

In Johnstone and Lu (2009) (Theorem 1) it was shown that under a single spike model, as $N/n \to 0$, the standard PCA estimator of the leading eigenvector is consistent. The following result, proven in the Appendix, is a refinement of that, as it also provides the leading error term.

THEOREM 2.1. *Let $\widehat{\theta}_{\nu,\mathrm{PCA}}$ be the eigenvector corresponding to the νth largest eigenvalue of* $\mathbf{S}$. *Assume that* (BA) *holds and $N, n \to \infty$ such that $N/n \to 0$. Then, for each $\nu = 1, \dots, M$,*

$$\sup_{\theta_\nu \in \mathbb{S}^{N-1}} \mathbb{E}L(\widehat{\theta}_{\nu,\mathrm{PCA}}, \theta_\nu) = \left[\frac{N-M}{nh(\lambda_\nu)} + \frac{1}{n}\sum_{\mu \neq \nu} \frac{1}{g(\lambda_\mu, \lambda_\nu)}\right]\bigl(1 + o(1)\bigr). \tag{2.6}$$

REMARK 2.1. Observe that Theorem 2.1 does not assume any special structure such as sparsity for the eigenvectors. The first term on the right-hand side of (2.6) is a nonparametric component which arises from the interaction of the noise terms with the different coordinates. The second term is "parametric" and results from the interaction with the remaining $M - 1$ eigenvectors corresponding to different eigenvalues. The second term shows that the closer the successive eigenvalues, the larger the estimation error. The upshot of (2.6) is that standard PCA yields a consistent estimator of the leading eigenvectors of the population covariance matrix when the dimension-to-sample-size ratio (N/n) is asymptotically negligible.

2.3. *l_q constraint on eigenvectors.* When $N/n \to c \in (0, \infty]$, standard PCA provides *inconsistent* estimators for the population eigenvectors, as shown by various authors [Johnstone and Lu (2009), Lu (2002), Nadler (2008), Onatski (2006), Paul (2007)]. In this subsection we consider the following model for approximate sparsity of the eigenvectors. For each $\nu = 1, \ldots, M$, assume that θ_ν belongs to an l_q ball with radius C, for some $q \in (0, 2)$, thus $\theta_\nu \in \Theta_q(C)$, where

$$\Theta_q(C) := \left\{ \mathbf{a} \in \mathbb{S}^{N-1} : \sum_{k=1}^{N} |a_k|^q \le C^q \right\}. \tag{2.7}$$

Note that our condition of sparsity is slightly different from that of Johnstone and Lu (2009). Since $0 < q < 2$, for $\Theta_q(C)$ to be nonempty, one needs $C \ge 1$. Further, if $C^q \ge N^{1-q/2}$, then the space $\Theta_q(C)$ is all of $\mathbb{S}^{N-1}$ because in this case, the least sparse vector $\frac{1}{\sqrt{N}}(1, 1, \ldots, 1)$ is in the parameter space.

The parameter space for $\boldsymbol{\theta} := [\theta_1 : \ldots : \theta_M]$ is denoted by

$$\Theta_q^M(C_1, \ldots, C_M) := \left\{ \boldsymbol{\theta} \in \prod_{\nu=1}^{M} \Theta_q(C_\nu) : \langle \theta_\nu, \theta_{\nu'} \rangle = 0, \text{ for } \nu \ne \nu' \right\}, \tag{2.8}$$

where $\Theta_q(C)$ is defined through (2.7), and $C_\nu \ge 1$ for all $\nu = 1, \ldots, M$. Thus Θ_q^M consists of sparse orthonormal M-frames, with sparsity measured in l_q. Note that in the analysis that follows we allow the C_ν's to increase with N.

REMARK 2.2. While our focus is on eigenvector sparsity, condition (2.8) also implies sparsity of the covariance matrix itself. In particular, for $q \in (0, 1)$, a spiked covariance matrix satisfying (2.8) also belongs to the class of sparse covariance matrices analyzed by Bickel and Levina (2008a), Cai and Liu (2011) and Cai and Zhou (2012). Indeed, Cai and Zhou (2012) obtained the minimax rate of convergence for covariance matrix estimators under the spectral norm when the rows of the population matrix satisfy a weak-l_q constraint. However, as we will show below, the minimax rate for estimation of the leading eigenvectors is faster than that for covariance estimation.

3. Lower bounds on the minimax risk. We now derive lower bounds on the minimax risk of estimating θ_ν under the loss function (2.3). To aid in describing and interpreting the lower bounds, we define the following two auxiliary parameters. The first is an *effective noise level per coordinate*

$$\tau_\nu^2 = 1/\big(nh(\lambda_\nu)\big) \tag{3.1}$$

and the second is an *effective dimension*

$$m_\nu := A_q(\bar{C}_\nu/\tau_\nu)^q, \tag{3.2}$$

where $a_q := (2/9)^{1-q/2}$, $c_1 := \log(9/8)$ and $A_q := 1/(a_q c_1^{q/2})$ and finally $\bar{C}_\nu^q := C_\nu^q - 1$.

The phrase *effective noise level per coordinate* is motivated by the risk bound in Theorem 2.1: dividing both sides of (2.6) by N, the expected "per coordinate" risk (or variance) of the PCA estimator is asymptotically τ_ν^2. Next, following Nadler (2009), let us provide a different interpretation of τ_ν. Consider a sparse θ_ν and an oracle that, regardless of the observed data, selects a set J_τ of all coordinates of θ_ν that are larger than τ in absolute value, and then performs PCA on the sample covariance restricted to these coordinates. Since $\theta_\nu \in \Theta_q(C_\nu)$, the maximal squared-bias is

$$\begin{aligned}\sup_{\theta_\nu \in \Theta_q(C_\nu)} \sum_{k \notin J_\tau} |\theta_{\nu k}|^2 &\asymp \sup\left\{\sum_{k=1}^N x_k^{2/q} : \sum_{k=1}^N x_k \le C_\nu^q, 0 \le x_k \le \tau^q\right\} \\ &\asymp C_\nu^q \tau^{2-q},\end{aligned}$$

which follows by the correspondence $x_k = |\theta_{\nu k}|^q$, and the convexity of the function $\sum_{k=1}^N x_k^{2/q}$. On the other hand, by Theorem 2.1, the maximal variance term of this oracle estimator is of the order $k_\tau/(nh(\lambda_\nu))$ where k_τ is the maximal number of coordinates of θ_ν exceeding τ. Again, $\theta_\nu \in \Theta_q(C_\nu)$ implies that $k_\tau \asymp C_\nu^q \tau^{-q}$. Thus, to balance the bias and variance terms, we need $\tau \asymp 1/\sqrt{nh(\lambda_\nu)} = \tau_\nu$. This heuristic analysis shows that τ_ν can be viewed as an *oracle threshold* for the coordinate selection scheme, that is, the best possible estimator of θ_ν based on individual coordinate selection can expect to recover only those coordinates that are above the threshold τ_ν.

To understand why m_ν is an *effective dimension*, consider the least sparse vector $\theta_\nu \in \Theta_q(C_\nu)$. This vector should have as many nonzero coordinates of equal size as possible. If $C_\nu^q > N^{1-q/2}$ then the vector with coordinates $\pm N^{-1/2}$ does the job. Otherwise, we set the first coordinate of the vector to be $\sqrt{1-r^2}$ for some $r \in (0,1)$ and choose all the nonzero coordinates to be of magnitude τ_ν. Clearly, we must have $r^2 = m\tau_\nu^2$, where $m+1$ is the maximal number of nonzero coordinates, while the l_q constraint implies that $(1-r^2)^{q/2} + m\tau_\nu^q \le C_\nu^q$. The last inequality shows that the maximal m is just a constant multiple of m_ν. This construction also constitutes the key idea in the proof of Theorems 3.1 and 3.2. Finally, we set

$$N' = N - M. \tag{3.3}$$

THEOREM 3.1. *Assume that* (BA) *holds*, $0 < q < 2$ *and* $n, N \to \infty$. *Then, there exists a constant* $B_1 > 0$ *such that for* n *sufficiently large*,

$$R_\nu^* := \inf_{\widehat{\theta}_\nu} \sup_{\Theta_q(\mathbf{C})} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu) \geq B_1 \delta_n, \tag{3.4}$$

where δ_n *is given by*

$$\delta_n = \begin{cases} \tau_\nu^2 N', & \text{if } \tau_\nu^2 N' < 1 \text{ and } N' < m_\nu \quad (\textit{dense setting}), \\ \tau_\nu^2 m_\nu, & \text{if } \tau_\nu^2 m_\nu < 1 \text{ and } m_\nu < N' \quad (\textit{sparse setting}), \\ 1, & \text{if } \tau_\nu^2 \cdot \min\{N', m_\nu\} > 1 \quad (\textit{weak signal}). \end{cases}$$

We may think of $m_{n,\nu} := \min\{N', m_\nu\}$ as the effective dimension of the least favorable configuration.

In the *thin* setting, $m_{n,\nu} = A_q \bar{C}_\nu^q [nh(\lambda_\nu)]^{q/2} < N$ (i.e., $\bar{C}_\nu^q n^{q/2} < c'N$ for some $c' > 0$), and the lower bound is of the order

$$\delta_n = A_q \bar{C}_\nu^q \tau_\nu^{2-q} = \frac{A_q \bar{C}_\nu^q}{[nh(\lambda_\nu)]^{1-q/2}} \asymp \frac{\bar{C}_{\nu,n}^q}{n^{1-q/2}}. \tag{3.5}$$

In the *dense* setting, on the other hand, $m_{n,\nu} = N - M$, and

$$\delta_n = \frac{N - M}{nh(\lambda_\nu)} \asymp \frac{N}{n}. \tag{3.6}$$

If $N/n \to c$ for some $c > 0$, then $\delta_n \asymp 1$, and so any estimator of the eigenvector θ_ν is inconsistent. If $N/n \to 0$, then equation (3.6) and Theorem 2.1 imply that the standard PCA estimator $\widehat{\theta}_{\nu,\mathrm{PCA}}$ attains the optimal rate of convergence.

A sharper lower bound is possible if $\bar{C}_\nu^q n^{q/2} = O(N^{1-\alpha})$ for some $\alpha \in (0, 1)$. We call this a *sparse* setting, noting that it is a special case of the thin setting. In this case the dimension N is much larger than the quantity $\bar{C}_\nu^q n^{q/2}$ measuring the effective dimension. Hence, we define a modified effective noise level per-coordinate

$$\bar{\tau}_\nu^2 = \frac{\alpha}{9} \frac{\log N}{nh(\lambda_\nu)},$$

and a modified effective dimension

$$\bar{m}_\nu = a_q^{-1} (\bar{C}_\nu / \bar{\tau}_\nu)^q.$$

THEOREM 3.2. *Assume that* (BA) *holds*, $0 < q < 2$ *and* $n, N \to \infty$ *in such a way that* $\bar{C}_\nu^q n^{q/2} = O(N^{1-\alpha})$ *for some* $\alpha \in (0, 1)$. *Then there exists a constant* B_1 *such that for* n *sufficiently large*, *the minimax bound* (3.4) *holds with*

$$\delta_n = \bar{m}_\nu \bar{\tau}_\nu^2 = a_q^{-1} C_\nu^q \left(\frac{\log N}{nh(\lambda_\nu)} \right)^{1-q/2} \qquad (\textit{sparse setting}) \tag{3.7}$$

so long as this quantity is ≤ 1.

Note that in the sparse setting δ_n is larger by a factor of $(\log N)^{1-q/2}$ compared to the thin setting [equation (3.5)].

It should be noted that for fixed signal strength λ_ν, for the corresponding eigenvector to be *thin*, *but not sparse* is somewhat of a "rarity," as the following argument shows: consider first the case $N = o(n)$. If $N' < A_q(h(\lambda_\nu))^{q/2}\bar{C}_\nu^q n^{q/2}$, then we are in the dense setting, since $\tau_\nu^2 N' \asymp N/n \to 0$. On the other hand, if $N = o(n)$ and $\bar{C}_\nu^q n^{q/2} = O(N^{1-\alpha})$ for some $\alpha \in (0,1)$, then θ_ν is sparse, according to the discussion preceding Theorem 3.2. So, if $N = o(n)$, for the eigenvector θ_ν to be *thin but not sparse*, we need $\bar{C}_\nu^q n^{q/2} \asymp N s_N$ where s_N is a term which may be constant or may converge to zero at a rate slower than any polynomial in N. Next, consider the case $n = o(N)$. For a meaningful lower bound, we require $\tau_\nu^2 m_\nu < 1$, which means that $\bar{C}_\nu^q n^{q/2} < c_{q,\nu} n$ for some constant $c_{q,\nu} > 0$. Thus, as long as $n = O(N^{1-\alpha})$ for some $\alpha \in (0,1)$, θ_ν cannot be *thin but not sparse*. Finally, suppose that $N \asymp n$, and let $\bar{C}_\nu^q = N^\beta$ for some $\beta \geq 0$. If $\beta < 1 - q/2$, then we are in the sparse case. On the other hand, if $\beta > 1 - q/2$, then there is no sparsity at all since when $C_\nu^q \geq N^{1-q/2}$ the entire $\mathbb{S}^{N-1}$ belongs to the relevant l_q ball for θ_ν. Hence, only if $\beta = 1 - q/2$ exactly, it is possible for θ_ν to be sparse. This analysis emphasizes the point that at least for a fixed signal strength, thin but not sparse is a somewhat special situation.

4. Risk of the diagonal thresholding estimator.

In this section, we analyze the convergence rate of the diagonal thresholding (DT) approach to sparse PCA proposed by Johnstone and Lu (2009) (JL). In this section and in Section 5, we assume for simplicity that $N \geq n$. Let the sample variance of the kth coordinate, the kth diagonal entry of $\mathbf{S}$, be denoted by $\mathbf{S}_{kk}$. Then DT consists of the following steps:

(1) Define $I = I(\gamma_n)$ to be the set of indices $k \in \{1, \ldots, N\}$ such that $\mathbf{S}_{kk} > 1 + \gamma_n$ for some threshold $\gamma_n > 0$.

(2) Let $\mathbf{S}_{II}$ be the submatrix of $\mathbf{S}$ corresponding to the coordinates I. Perform an eigen-analysis of $\mathbf{S}_{II}$ and denote its eigenvectors by $\mathbf{f}_i$, $i = 1, \ldots, \min\{n, |I|\}$.

(3) For $\nu = 1, \ldots, M$, estimate θ_ν by the $N \times 1$ vector $\widetilde{\mathbf{f}}_\nu$, obtained from $\mathbf{f}_\nu$ by augmenting zeros to all the coordinates in $I^c := \{1, \ldots, N\} \setminus I$.

Assuming that $\theta_\nu \in \Theta_q(C_\nu)$, and a threshold $\gamma_n = \gamma\sqrt{\log N/n}$ for some $\gamma > 0$, JL showed that DT yields a consistent estimator of θ_ν, but did not further analyze the risk. Indeed, as we prove below, the risk of the DT estimator is not rate optimal. This might be anticipated from the lower bound on the minimax risk (Theorems 3.1 and 3.2) which indicate that to attain the optimal risk, a coordinate selection scheme must select all coordinates of θ_ν of size at least $c\sqrt{\log N/n}$ for some $c > 0$. With a threshold of the form γ_n above, however, only coordinates of size $(\log N/n)^{1/4}$ are selected. Even for the case of a single signal, $M = 1$, this leads to a much larger lower bound.

THEOREM 4.1. *Suppose that* (BA) *holds with* $M = 1$. *Let* $C > 1$ (*may depend on* n), $0 < q < 2$ *and* $n, N \to \infty$ *be such that* $C^q n^{q/4} = o(\sqrt{n})$. *Then the diagonal thresholding estimator* $\widehat{\theta}_{1,\mathrm{DT}}$ *satisfies*

$$\sup_{\theta_1 \in \Theta_q(C)} \mathbb{E}L(\widehat{\theta}_{1,\mathrm{DT}}, \theta_1) \geq K_q \bar{C}^q n^{-(1-q/2)/2} \tag{4.1}$$

for a constant $K_q > 0$, *where* $\bar{C}^q = C^q - 1$.

A comparison of (4.1) with the lower bound (3.5), shows a large gap between the two rates, $n^{-1/2(1-q/2)}$ versus $n^{-(1-q/2)}$. This gap arises because DT uses only the diagonal of the sample covariance matrix **S**, ignoring the information in its off-diagonal entries. In the next section we propose a refinement of the DT scheme, denoted ASPCA, that constructs an improved eigenvector estimate using all entries of **S**.

In the *sparse* setting, the ITSPCA estimator of Ma (2011) attains the same asymptotic rate as the lower bound of Theorem 3.2, provided DT yields consistent estimates of the eigenvectors. The latter condition can be shown to hold if, for example, $C_\nu^q n^{q/4}(\log N)^{1-q/2} = o(\sqrt{n})$ for all $\nu = 1, \dots, M$. Thus, in the sparse setting, with this additional restriction, the lower bound on the minimax rate is sharp, and consequently, the DT estimator is not rate optimal.

5. A two-stage coordinate selection scheme. As discussed above, the DT scheme can reliably detect only those eigenvector coordinates k for which $|\theta_{\nu,k}| \geq c(\log N/n)^{1/4}$ (for some $c > 0$), whereas to reach the lower bound one needs to detect those coordinates for which $|\theta_{\nu,k}| \geq c(\log N/n)^{1/2}$.

To motivate an improved coordinate selection scheme, consider the single component (i.e., $M = 1$) case, and form a partition of the N coordinates into two sets A and B, where the former contains all those k such that $|\theta_{1k}|$ is "large" (selected by DT), and the latter contains the remaining smaller coordinates. Partition the matrix Σ as

$$\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

Observe that $\Sigma_{BA} = \lambda_1 \theta_{1,B} \theta_{1,A}^T$. Let $\widetilde{\theta}_1$ be a "preliminary" estimator of θ_1 such that $\lim_{n\to\infty} \mathbb{P}(\langle \widetilde{\theta}_{1,A}, \theta_{1,A}\rangle \geq \delta_0) = 1$ for some $\delta_0 > 0$ (e.g., $\widetilde{\theta}_1$ could be the DT estimator). Then we have the relationship

$$\Sigma_{BA}\widetilde{\theta}_{1,A} = \langle \widetilde{\theta}_{1,A}, \theta_{1,A}\rangle \lambda_1 \theta_{1,B} \approx c(\delta_0)\lambda_1\theta_{1,B}$$

for some $c(\delta_0)$ bounded below by $\delta_0/2$, say. Thus one possible strategy is to additionally select all those coordinates of $\Sigma_{BA}\widetilde{\theta}_{1,A}$ that are larger (in absolute value) than some constant multiple of $\sqrt{\log N}/\sqrt{nh(\lambda_1)}$. Neither Σ_{BA} nor λ_1 is

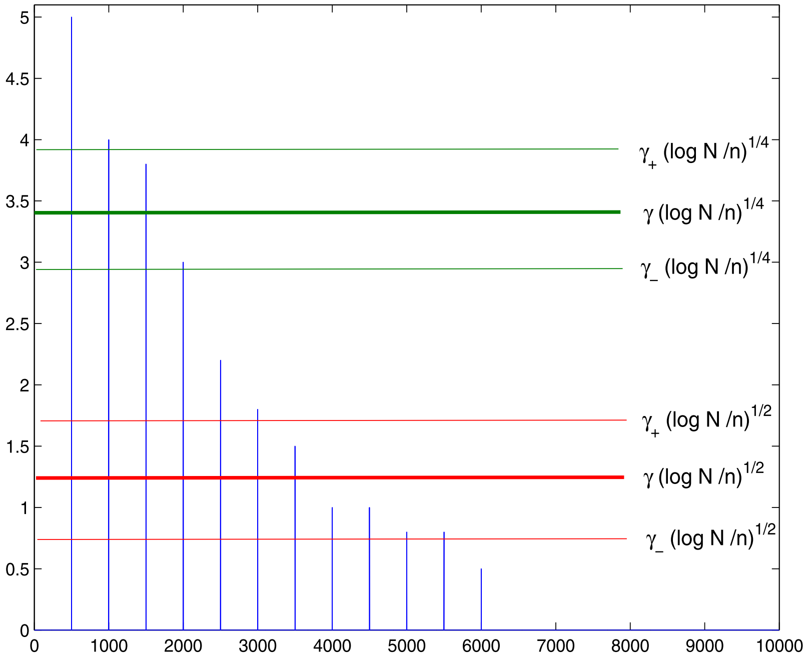

FIG. 1. *Schematic diagram of the DT and ASPCA thresholding schemes under the single component setting. The x-axis represents the indices of different coordinates of the first eigenvector and the vertical lines depict the absolute values of the coordinates. The threshold for the DT scheme is $\gamma(\log N/n)^{1/4}$ while the thresholds for the ASPCA scheme is $\gamma(\log N/n)^{1/2}$. For some generic constants $\gamma_+ > \gamma > \gamma_- > 0$, with high probability, the schemes select all coordinates above the upper limits (indicated by the multiplier γ_+) and discard all coordinates below the lower limits (indicated by the multiplier γ_-).*

known, but we can use $\mathbf{S}_{BA}$ as a surrogate for the former and the largest eigenvalue of $\mathbf{S}_{AA}$ to obtain an estimate for the latter. A technical challenge is to show that, with probability tending to 1, such a scheme indeed recovers all coordinates k with $|\theta_{1k}| > \gamma_+\sqrt{\log N}/\sqrt{nh(\lambda_1)}$, while discarding all coordinates k with $|\theta_{1k}| < \gamma_-\sqrt{\log N}/\sqrt{nh(\lambda_1)}$ for some constants $\gamma_+ > \gamma_- > 0$. Figure 1 provides a pictorial description of the DT and ASPCA coordinate selection schemes.

5.1. *ASPCA scheme.* Based on the ideas described above, we now present the ASPCA algorithm. It first makes two stages of coordinate selection, whereas the final stage consists of an eigen-analysis of the submatrix of $\mathbf{S}$ corresponding to the selected coordinates. The algorithm is described below.

For any $\gamma > 0$ define

$$I(\gamma) = \{k : \mathbf{S}_{kk} > 1 + \gamma\}. \tag{5.1}$$

Let $\gamma_i > 0$ for $i = 1, 2$ and $\kappa > 0$ be constants to be specified later.

Stage 1.

1^o Let $I = I(\gamma_{1,n})$ where $\gamma_{1,n} = \gamma_1\sqrt{\log N/n}$.

2^o Denote the eigenvalues and eigenvectors of $\mathbf{S}_{II}$ by $\widehat{\ell}_1 > \cdots > \widehat{\ell}_{m_1}$ and $\mathbf{f}_1, \ldots, \mathbf{f}_{m_1}$, respectively, where $m_1 = \min\{n, |I|\}$.

3^o Estimate M by $\widehat{M}$ defined in Section 5.2.

Stage 2.

4^o Let $\mathbf{E} = [\widehat{\ell}_1^{-1/2}\mathbf{f}_1 \cdots \widehat{\ell}_{\widehat{M}}^{-1/2}\mathbf{f}_{\widehat{M}}]$ and $\mathbf{Q} = \mathbf{S}_{I^cI}\mathbf{E}$.

5^o Let $J = \{k \notin I : (\mathbf{Q}\mathbf{Q}^T)_{kk} > \gamma_{2,n}^2\}$ for some $\gamma_{2,n} > 0$. Define $K = I \cup J$.

Stage 3.

6^o For $\nu = 1, \ldots, \widehat{M}$, denote by $\widehat{\theta}_\nu$ the νth eigenvector of $\mathbf{S}_{KK}$, augmented with zeros in the coordinates K^c.

REMARK 5.1. The ASPCA scheme is specified up to the choice of parameters γ_1 and $\gamma_{2,n}$ that determine its rate of convergence. It can be shown that choosing $\gamma_1 = 4$ and

$$\gamma_{2,n} = \kappa\sqrt{\frac{\log N}{n}} + \sqrt{\frac{\widehat{M}}{n}} \tag{5.2}$$

with $\kappa = \sqrt{3 + \varepsilon}$ for some $\varepsilon > 0$, results in an asymptotically optimal rate. Again, we note that for finite N, n, the actual performance in terms of the risk of the resulting eigenvector estimate may have a strong dependence on the threshold. In practice, a delicate choice of thresholds can be highly beneficial. This issue, as well as the analysis of the risk of the ASPCA estimator, are beyond the scope of this paper and will be studied in a separate publication.

5.2. *Estimation of M*. Estimation of the dimension of the signal subspace is a classical problem. If the signal eigenvalues are strong enough (i.e., $\lambda_\nu > c\sqrt{N/n}$ for all $\nu = 1, \ldots, M$, for some $c > 1$ independent of N, n), then nonparametric methods that do not assume eigenvector sparsity can asymptotically estimate the correct M; see, for example, Kritchman and Nadler (2008). When the eigenvectors are sparse, we can detect much weaker signals, as we describe below.

We estimate M by thresholding the eigenvalues of the submatrix $\mathbf{S}_{\bar{I}\bar{I}}$ where $\bar{I} := I(\bar{\gamma}\sqrt{\log N/n})$ for some $\bar{\gamma} \in (0, \gamma_1)$. Let $\bar{m} = \min\{n, |\bar{I}|\}$ and $\bar{\ell}_1 > \cdots > \bar{\ell}_{\bar{m}}$ be the nonzero eigenvalues of $\mathbf{S}_{\bar{I}\bar{I}}$. Let $\alpha_n > 0$ be a threshold of the form

$$\alpha_n = 2\sqrt{\frac{|\bar{I}|}{n}} + \left(1 + c_0\sqrt{\frac{\log n}{n}}\right)\frac{|\bar{I}|}{n}$$

for some user-defined constant $c_0 > 0$. Then, define $\widehat{M}$ by

$$\widehat{M} := \max\{1 \le k \le \bar{m} : \bar{\ell}_k > 1 + \alpha_n\}. \tag{5.3}$$

The idea is that, for large enough n, $I(\gamma_n) \subset \bar{I}$ with high probability and thus $|\bar{I}|$ acts as an upper bound on $|I(\gamma_{1n})|$. Using this and the behavior of the extreme eigenvalues of a Wishart matrix, it can be shown that, with a suitable choice of c_0 and $\bar{\gamma}$, $\widehat{M}$ is a consistent estimator of M.

6. Summary and discussion.

In this paper we have derived lower bounds on eigenvector estimates under three different sparsity regimes, denoted dense, thin and sparse. In the *dense* setting, Theorems 2.1 and 3.1 show that when $N/n \to 0$, the standard PCA estimator attains the optimal rate of convergence.

In the *sparse* setting, Theorem 3.1 of Ma (2011) shows that the maximal risk of the ITSPCA estimator proposed by him attains the same asymptotic rate as the corresponding lower bound of Theorem 3.2. This implies that in the sparse setting, the lower bound on the minimax rate is indeed sharp. In a separate paper, we prove that in the sparse regime, the ASPCA algorithm also attains the minimax rate. All these sparse setting results currently require the additional condition of consistency of DT—without this condition, the rate optimality question remains open.

Finally, our analysis leaves some open questions in the intermediate *thin* regime. According to Theorem 3.1, the lower bound in this regime is smaller by a factor of $(\log N)^{1-q/2}$, as compared to the sparse setting. Therefore, whether there exists an estimator (and in particular, one with low complexity), that attains the current lower bound, or whether this lower bound can be improved is an open question for future research. However, as we indicated at the end of Section 3, the eigenvector being thin but not sparse is a somewhat rare occurrence in terms of mathematical possibilities.

APPENDIX A: PROOFS

A.1. Asymptotic risk of the standard PCA estimator.

To prove Theorem 2.1, on the risk of the PCA estimator, we use the following lemmas. Throughout, $\|B\| = \sup\{x^T Bx : \|x\|_2 = 1\}$ denotes the spectral norm on square matrices.

Deviation of extreme Wishart eigenvalues and quadratic forms. In our analysis, we will need a probabilistic bound for deviations of $\|n^{-1}\mathbf{Z}\mathbf{Z}^T - I\|$. This is given in the following lemma, proven in Appendix B.

LEMMA A.1. *Let $\mathbf{Z}$ be an $N \times n$ matrix with i.i.d. $N(0,1)$ entries. Suppose $N < n$ and set $t_n = 8\sqrt{n^{-1}\log n}$ and $\gamma_n = N/n$. Then for any $c > 0$, there exists $n_c \geq 1$ such that for all $n \geq n_c$,*

$$\mathbb{P}(\|n^{-1}\mathbf{Z}\mathbf{Z}^T - I_N\| > \gamma_n + 2\sqrt{\gamma_n} + ct_n) \leq 2n^{-c^2}. \tag{A.1}$$

LEMMA A.2 [Johnstone (2001)]. *Let χ_n^2 denote a Chi-square random variable with n degrees of freedom. Then*

$$\mathbb{P}\bigl(\chi_n^2 > n(1+\varepsilon)\bigr) \le e^{-3n\varepsilon^2/16} \qquad \left(0 < \varepsilon < \frac{1}{2}\right), \tag{A.2}$$

$$\mathbb{P}\bigl(\chi_n^2 < n(1-\varepsilon)\bigr) \le e^{-n\varepsilon^2/4} \qquad (0 < \varepsilon < 1), \tag{A.3}$$

$$\mathbb{P}\bigl(\chi_n^2 > n(1+\varepsilon)\bigr) \le \frac{\sqrt{2}}{\varepsilon\sqrt{n}} e^{-n\varepsilon^2/4} \qquad (0 < \varepsilon < 1/2, n \ge 16). \tag{A.4}$$

LEMMA A.3 [Johnstone and Lu (2009)]. *Let $y_{1i}, y_{2i}, i = 1, \dots, n$, be two sequences of mutually independent, i.i.d. $N(0, 1)$ random variables. Then for large n and any b s.t. $0 < b \ll \sqrt{n}$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_{1i} y_{2i}\right| > \sqrt{b/n}\right) \le 2\exp\left\{-\frac{3b}{2} + O\bigl(n^{-1}b^2\bigr)\right\}. \tag{A.5}$$

Perturbation of eigen-structure. The following lemma, modified in Appendix B from Paul (2005), is convenient for risk analysis of estimators of eigenvectors. Several variants of this lemma appear in the literature, most based on the approach of Kato (1980). To state it, let the eigenvalues of a symmetric matrix A be denoted by $\lambda_1(A) \ge \cdots \ge \lambda_m(A)$, with the convention that $\lambda_0(A) = \infty$ and $\lambda_{m+1}(A) = -\infty$. Let P_s denote the projection matrix onto the possibly multidimensional eigenspace corresponding to $\lambda_s(A)$ and define

$$H_r(A) = \sum_{s \ne r} \frac{1}{\lambda_s(A) - \lambda_r(A)} P_s(A).$$

Note that $H_r(A)$ may be viewed as the resolvent of A "evaluated at $\lambda_r(A)$."

LEMMA A.4. *Let A and B be symmetric $m \times m$ matrices. Suppose that $\lambda_r(A)$ is a unique eigenvalue of A with*

$$\delta_r(A) = \min\bigl\{\bigl|\lambda_j(A) - \lambda_r(A)\bigr| : 1 \le j \ne r \le m\bigr\}.$$

Let $\mathbf{p}_r$ denote the unit eigenvector associated with the $\lambda_r(A)$. Then

$$\mathbf{p}_r(A + B) \ominus \mathbf{p}_r(A) = -H_r(A)B\mathbf{p}_r(A) + R_r, \tag{A.6}$$

where, if $4\|B\| \le \delta_r^{-1}(A)$,

$$\|R_r\| \le K\delta_r^{-1}(A)\bigl\|H_r(A)B\mathbf{p}_r(A)\bigr\|\,\|B\|, \tag{A.7}$$

and we may take $K = 30$.

Proof of Theorem 2.1. First we outline the approach. For notational simplicity, throughout this subsection, we write $\widehat{\theta}_\nu$ to mean $\widehat{\theta}_{\nu,\mathrm{PCA}}$. Recall that the loss function $L(\widehat{\theta}_\nu, \theta_\nu) = \|\widehat{\theta}_\nu \ominus \theta_\nu\|^2$. Invoking Lemma A.4 with $A = \Sigma$ and $B = \mathbf{S} - \Sigma$ we get

$$\widehat{\theta}_\nu \ominus \theta_\nu = -H_\nu \mathbf{S}\theta_\nu + R_\nu, \tag{A.8}$$

where

$$H_\nu \equiv H_\nu(\Sigma) := \sum_{1 \le \mu \ne \nu \le M} \frac{1}{\lambda_\mu - \lambda_\nu} \theta_\mu \theta_\mu^T - \frac{1}{\lambda_\nu} P_\perp \tag{A.9}$$

and $P_\perp = I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T$. Note that $H_\nu \theta_\nu = 0$ and that $H_\nu \Sigma \theta_\nu = 0$.

Let $\varepsilon_{n\nu} = K \delta_r^{-1}(\Sigma) \|\mathbf{S} - \Sigma\|$. We have from (A.7) that

$$\|R_\nu\| \le \|H_\nu \mathbf{S}\theta_\nu\| \delta'_{n\nu},$$

and we will show that as $n \to \infty$, $\varepsilon_{n\nu} \to 0$ with probability approaching 1 and that

$$\|H_\nu \mathbf{S}\theta_\nu\|^2 (1 - \varepsilon_{n\nu})^2 \le L(\widehat{\theta}_\nu, \theta_\nu) \le \|H_\nu \mathbf{S}\theta_\nu\|^2 (1 + \varepsilon_{n\nu})^2. \tag{A.10}$$

Theorem 2.1 then follows from an (exact, nonasymptotic) evaluation,

$$\mathbb{E}\big[\|H_\nu \mathbf{S}\theta_\nu\|^2\big] = \frac{N - M}{nh(\lambda_\nu)} + \frac{1}{n} \sum_{\mu \ne \nu} \frac{(1 + \lambda_\mu)(1 + \lambda_\nu)}{(\lambda_\mu - \lambda_\nu)^2}. \tag{A.11}$$

We begin with the evaluation of (A.11). First we derive a convenient representation of $H_\nu \mathbf{S}\theta_\nu$. In matrix form, model (2.2) becomes

$$\mathbf{X} = \sum_{\nu=1}^M \sqrt{\lambda_\nu} \theta_\nu v_\nu^T + \mathbf{Z}, \tag{A.12}$$

where $v_\nu = (v_{\nu i})_{i=1}^n$, for $\nu = 1, \ldots, M$. Also, define

$$z_\nu = \mathbf{Z}^T \theta_\nu, \qquad w_\nu = \mathbf{X}^T \theta_\nu = \sqrt{\lambda_\nu} v_\nu + z_\nu \tag{A.13}$$

and

$$\langle \mathbf{a}, \mathbf{b} \rangle_n = \frac{1}{n} \sum_{i=1}^n a_i b_i \qquad \text{for arbitrary } \mathbf{a}, \mathbf{b} \in \mathbb{R}^n. \tag{A.14}$$

Then we have

$$\mathbf{S}\theta_\nu = \frac{1}{n} \mathbf{X} w_\nu = \sum_{\mu=1}^M \sqrt{\lambda_\mu} \langle v_\mu, w_\nu \rangle_n \theta_\mu + \frac{1}{n} \mathbf{Z} w_\nu.$$

Using (A.13),

$$\frac{1}{n} H_\nu \mathbf{Z} w_\nu = \sum_{\mu \ne \nu} \frac{\langle z_\mu, w_\nu \rangle}{\lambda_\mu - \lambda_\nu} \theta_\mu - \frac{1}{n\lambda_\nu} P_\perp \mathbf{Z} w_\nu.$$

Using (A.9), $H_\nu \theta_\mu = (\lambda_\mu - \lambda_\nu)^{-1}\theta_\mu$ for $\mu \neq \nu$, and we arrive at the desired representation

$$H_\nu \mathbf{S}\theta_\nu = \sum_{\mu \neq \nu} \frac{\langle w_\mu, w_\nu\rangle_n}{\lambda_\mu - \lambda_\nu}\theta_\mu - \frac{1}{n\lambda_\nu} P_\perp \mathbf{Z} w_\nu. \tag{A.15}$$

By orthogonality,

$$\|H_\nu \mathbf{S}\theta_\nu\|^2 = \sum_{\mu \neq \nu} \frac{\langle w_\mu, w_\nu\rangle_n^2}{(\lambda_\mu - \lambda_\nu)^2} + \frac{1}{n^2\lambda_\nu^2} w_\nu^T \mathbf{Z}^T P_\perp \mathbf{Z} w_\nu. \tag{A.16}$$

Now we compute the expectation. One verifies that $z_\nu \sim N(0, I_n)$ independently of each other and of each $v_\nu \sim N(0, I_n)$, so that $w_\nu \sim N(0, (1+\lambda_\nu)I_n)$ independently. Hence, for $\mu \neq \nu$,

$$\begin{aligned} \mathbb{E}\bigl[\langle w_\mu, w_\nu\rangle_n^2\bigr] &= n^{-2}\mathbb{E}\,\mathrm{tr}\bigl(w_\nu w_\nu^T w_\mu w_\mu^T\bigr) \\ &= n^{-2}\,\mathrm{tr}\bigl((1+\lambda_\mu)(1+\lambda_\nu)I_n\bigr) \\ &= n^{-1}(1+\lambda_\mu)(1+\lambda_\nu). \end{aligned} \tag{A.17}$$

From (A.13),

$$\begin{aligned} \mathbb{E}\bigl[w_\nu^T \mathbf{Z}^T P_\perp \mathbf{Z} w_\nu | \mathbf{Z}\bigr] &= z_\nu^T \mathbf{Z}^T P_\perp \mathbf{Z} z_\nu + \lambda_\nu \mathbb{E}\bigl[v_\nu^T \mathbf{Z}^T P_\perp \mathbf{Z} v_\nu | \mathbf{Z}\bigr] \\ &= \mathrm{tr}\bigl(\mathbf{Z}\mathbf{Z}^T P_\perp \mathbf{Z}\mathbf{Z}^T \theta_\nu \theta_\nu^T\bigr) + \lambda_\nu\,\mathrm{tr}\bigl(P_\perp \mathbf{Z}\mathbf{Z}^T\bigr). \end{aligned}$$

Now, it can be easily verified that if $W := \mathbf{Z}\mathbf{Z}^T \sim W_N(n, I)$, then for arbitrary symmetric $N \times N$ matrices Q, R, we have

$$\mathbb{E}\bigl[\mathrm{tr}(WQWR)\bigr] = n\bigl[\mathrm{tr}(QR) + \mathrm{tr}(Q)\,\mathrm{tr}(R)\bigr] + n^2\,\mathrm{tr}(QR). \tag{A.18}$$

Taking $Q = P_\perp$ and $R = \theta_\mu \theta_\mu^T$ and noting that $QR = 0$, by (A.18) we have

$$\mathbb{E}\bigl[w_\nu^T \mathbf{Z}^T P_\perp \mathbf{Z} w_\nu\bigr] = n\,\mathrm{tr}(P_\perp) + n\lambda_\nu\,\mathrm{tr}(P_\perp) = n(N-M)(1+\lambda_\nu). \tag{A.19}$$

Combining (A.17) with (A.19) in computing the expectation of (A.16), we obtain the expression (A.11) for $\mathbb{E}\|H_\nu \mathbf{S}\theta_\nu\|^2$.

A.2. Bound for $\|\mathbf{S} - \boldsymbol{\Sigma}\|$. We begin with the decomposition of the sample covariance matrix $\mathbf{S}$. Introduce the abbreviation $\xi_\mu = n^{-1}\mathbf{Z} v_\mu$. Then

$$\begin{aligned} \mathbf{S} = {} & \sum_{\mu,\mu'=1}^{M} \sqrt{\lambda_\mu \lambda_{\mu'}} \langle v_\mu, v_{\mu'}\rangle_n \theta_\mu \theta_{\mu'}^T + \sum_{\mu=1}^{M} \sqrt{\lambda_\mu}\bigl(\theta_\mu \xi_\mu^T + \xi_\mu \theta_\mu^T\bigr) \\ & + n^{-1}\mathbf{Z}\mathbf{Z}^T \end{aligned} \tag{A.20}$$

and from (2.1), with $V_{\mu\mu'} = |\langle v_\mu, v_{\mu'}\rangle_n - \delta_{\mu\mu'}|$ and $\delta_{\mu\mu'}$ denoting the Kronecker symbol,

$$\|\mathbf{S} - \Sigma\| \le \sum_{\mu,\mu'=1}^{M} \sqrt{\lambda_\mu \lambda_{\mu'}} V_{\mu\mu'} + 2\sum_{\mu=1}^{M} \sqrt{\lambda_\mu}\|\xi_\mu\| + \|n^{-1}\mathbf{Z}\mathbf{Z}^T - I\|. \tag{A.21}$$

We establish a bound for $\|\mathbf{S}-\Sigma\|$ with probability converging to one. Introduce notation

$$\eta_n = \sqrt{N^{-1}\log n}, \qquad \bar{\eta}_n = \sqrt{n^{-1}\log n}, \qquad \gamma_n = N/n, \qquad \sqrt{\Lambda} = \sum_{\mu=1}^{M} \sqrt{\lambda_\mu}.$$

Fix $c > 0$ and assume that $\gamma_n \le 1$. Initially, we assume that $2c\eta_n \le 1/2$, which is equivalent to $N \ge 16c^2 \log n$.

We introduce some events of high probability under which (A.21) may be bounded. Thus, let D_1 be the intersection of the events

$$\begin{aligned} \big|\|v_\mu\|_n^2 - 1\big| &\le 2c\bar{\eta}_n, \qquad 1 \le \mu \le M, \\ \big|\langle v_\mu, v_{\mu'}\rangle_n\big| &\le c\bar{\eta}_n, \qquad 1 \le \mu \ne \mu' \le M, \\ N^{-1}\|\mathbf{Z}v_\mu\|^2/\|v_\mu\|^2 &\le 1 + 2c\eta_n, \qquad 1 \le \mu \le M, \end{aligned} \tag{A.22}$$

and let D_2 be the event

$$\|n^{-1}\mathbf{Z}\mathbf{Z}^T - I\| \le \gamma_n + 2\sqrt{\gamma_n} + 8c\bar{\eta}_n. \tag{A.23}$$

To bound the probability of D_1^c, in the case of the first line of (A.22), use (A.3) and (A.4) with $\varepsilon = 2c\bar{\eta}_n$. For the second, use (A.5) with $b = c^2 \log n$. For the third, observe that $\mathbf{Z}v_\mu/\|v_\mu\| \sim N_N(0, I)$, and again use (A.4), this time with $\varepsilon = 2c\eta_n < 1/2$. For D_2^c, we appeal to Lemma A.1. As a result,

$$\begin{aligned} \mathbb{P}(D_1^c) &\le 3Mn^{-c^2} + M(M-1)n^{-(3/2)c^2 + O(n^{-1}\log^2 n)}, \\ \mathbb{P}(D_2^c) &\le 2n^{-c^2}. \end{aligned} \tag{A.24}$$

To bound (A.21) on the event $D_1 \cap D_2$, we use bounds (A.22) and (A.23), and also write

$$\begin{aligned} \|\xi_\mu\| &= \sqrt{\gamma_n}\frac{\|\mathbf{Z}v_\mu\|}{\sqrt{N}\|v_\mu\|}\frac{\|v_\mu\|}{\sqrt{n}} \\ &\le \sqrt{\gamma_n}(1 + 2c\eta_n)^{1/2}(1 + 2c\bar{\eta}_n)^{1/2} \\ &= \sqrt{\gamma_n}H_n, \end{aligned} \tag{A.25}$$

say, and also noting that $\bar{\eta}_n \le \eta_n$, we obtain

$$\|\mathbf{S}-\Sigma\| \le \sqrt{\gamma_n}\big[2c\eta_n\Lambda + 2\sqrt{\Lambda}H_n + 4(1 + 2c\eta_n)\big]. \tag{A.26}$$

Now combine the bound $2c\eta_n < 1/2$ with $H_n \le 3/2$ and $2\sqrt{\Lambda} \le \Lambda + 1$ to conclude that on $D_1 \cap D_2$,

$$\|\mathbf{S}-\Sigma\| \le 2(\Lambda + 4)\sqrt{\gamma_n}$$

and so

$$\varepsilon_{n\nu} \le 2K\delta_\nu^{-1}(\Sigma)(\Lambda + 4)\sqrt{\gamma_n} \to 0$$

since $N/n \to 0$.

Let us now turn to the case $N \le 16c^2 \log n$. We can replace the last event in (A.22) by the event

$$N^{-1}\|\mathbf{Z}v_\mu\|^2/\|v_\mu\|^2 \le 2(c^2 \log n + \log N), \qquad 1 \le \mu \le M,$$

and the second bound holds for $\mathbb{P}(D_1^c)$ for sufficiently large n, using the bound $\mathbb{P}(N^{-1}\chi_N^2 > a) \le 2N(1-\Phi(\sqrt{a})) \le N\sqrt{2/a\pi}e^{-a/2}$ for any $a > 0$. In (A.25), we replace the term $(1+2c\eta_n)^{1/2}$ by $(2c^2 \log n + 2\log N)^{1/2}$ which may be bounded by $a_1\sqrt{\log n}$. As soon as $N \ge 4c^2$, we also have $2c\eta_n \le \sqrt{\log n}$ and so $1 + 2c\bar{\eta}_n \le 1 + \sqrt{\gamma_n \log n}$. This leads to a bound for the analog of H_n in (A.26) and so to

$$\|\mathbf{S} - \Sigma\| \le \sqrt{\gamma_n \log n}\{\Lambda + 2a_1\sqrt{\Lambda}(1 + \sqrt{\gamma_n \log n})^{1/2} + a_2\}.$$

When $N \le 16c^2 \log n$, we have $\sqrt{\gamma_n \log n} \le 4c \log n/\sqrt{n}$ and so

$$\varepsilon_{n\nu} \le a_3 K \delta_\nu^{-1}(\Sigma)(\Lambda + 1)\log n/\sqrt{n} \to 0.$$

To summarize, choose $c = \sqrt{2}$, say, so that $D_n = D_1 \cap D_2$ has probability at least $1 - O(n^{-2})$, and on D_n we have $\varepsilon_{n\nu} \to 0$. This completes the proof of (A.10).

Theorem 2.1 now follows from noticing that $L(\widehat{\theta}_\nu, \theta_\nu) \le 2$ and so

$$\mathbb{E}[L(\widehat{\theta}_\nu, \theta_\nu), D_n^c] \le 2\mathbb{P}(D_n^c) = O(n^{-2}) = o(\mathbb{E}\|H_\nu \mathbf{S}\theta_\nu\|^2)$$

and an additional computation using (A.16) which shows that

$$\mathbb{E}[\|H_\nu \mathbf{S}\theta_\nu\|^2, D_n^c] \le (\mathbb{E}[\|H_\nu \mathbf{S}\theta_\nu\|^4])^{1/2} P(D_n^c) = o(\mathbb{E}\|H_\nu \mathbf{S}\theta_\nu\|^2).$$

A.3. Lower bound on the minimax risk. In this subsection, we prove Theorems 3.1 and 3.2. The key idea in the proofs is to utilize the geometry of the parameter space in order to construct appropriate finite-dimensional subproblems for which bounds are easier to obtain. We first give an overview of the general machinery used in the proof.

Risk bounding strategy. A key tool for deriving lower bounds on the minimax risk is *Fano's lemma*. In this subsection, we use superscripts on vectors θ as indices, not exponents. First, we fix $\nu \in \{1, \dots, M\}$ and then construct a large finite subset $\mathcal{F}$ of $\Theta_q^M(C_1, \dots, C_M)$, such that for some $\delta > 0$, to be chosen

$$\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \mathcal{F} \quad \Longrightarrow \quad L(\theta_\nu^1, \theta_\nu^2) \ge 4\delta.$$

This property will be referred to as "4δ-distinguishability in θ_ν." Given any estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, based on data $\mathbf{X}_n = (X_1, \dots, X_n)$, define a new estimator $\phi(\mathbf{X}_n) = \boldsymbol{\theta}^*$, whose M components are given by $\theta_\mu^* = \arg\min_{\boldsymbol{\theta} \in \mathcal{F}} L(\widehat{\theta}_\mu, \theta_\mu)$, where $\widehat{\theta}_\mu$ is the μth column of $\widehat{\boldsymbol{\theta}}$. Then, by Chebyshev's inequality and the 4δ-distinguishability in θ_ν, it follows that

$$\sup_{\boldsymbol{\theta} \in \Theta_q^M(C_1,\dots,C_M)} \mathbb{E}_{\boldsymbol{\theta}} L(\widehat{\theta}_\nu, \theta_\nu) \ge \delta \sup_{\boldsymbol{\theta} \in \mathcal{F}} \mathbb{P}_{\boldsymbol{\theta}}(\phi(\mathbf{X}_n) \ne \boldsymbol{\theta}). \tag{A.27}$$

The task is then to find an appropriate lower bound for the quantity on the right-hand side of (A.27). For this, we use the following version of Fano's lemma, due to Birgé (2001), modifying a result of Yang and Barron (1999), pages 1570 and 1571.

LEMMA A.5. *Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability distributions on a common measurable space, where Θ is an arbitrary parameter set. Let $p_{\max}$ be the minimax risk over Θ, with the loss function $L'(\theta, \theta') = \mathbf{1}_{\theta \neq \theta'}$,*

$$p_{\max} = \inf_T \sup_{\theta \in \Theta} \mathbb{P}_\theta(T \neq \theta) = \inf_T \sup_{\theta \in \Theta} \mathbb{E} L'(\theta, T),$$

where T denotes an arbitrary estimator of θ with values in Θ. Then for any finite subset $\mathcal{F}$ of Θ, with elements $\theta_1, \ldots, \theta_J$ where $J = |\mathcal{F}|$,

$$p_{\max} \geq 1 - \inf_Q \frac{J^{-1} \sum_{i=1}^J K(P_i, Q) + \log 2}{\log J}, \tag{A.28}$$

where $P_i = \mathbb{P}_{\theta_i}$, and Q is an arbitrary probability distribution, and $K(P_i, Q)$ is the Kullback–Leibler divergence of Q from P_i.

The following lemma, proven in Appendix B, gives the Kullback–Leibler discrepancy corresponding to two different values of the parameter.

LEMMA A.6. *Let $\boldsymbol{\theta}^j := [\theta_1^j : \ldots : \theta_M^j]$, $j = 1, 2$ be two parameters (i.e., for each j, θ_k^j's are orthonormal). Let Σ_j denote the matrix given by (2.1) with $\boldsymbol{\theta} = \boldsymbol{\theta}^j$ (and $\sigma = 1$). Let P_j denote the joint probability distribution of n i.i.d. observations from $N(0, \Sigma_j)$ and let $\eta(\lambda) = \lambda/(1+\lambda)$. Then the Kullback–Leibler discrepancy of P_2 with respect to P_1 is given by*

$$\mathcal{K}_{1,2} := K(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) = \frac{n}{2}\left[\sum_{\mu=1}^M \eta(\lambda_\mu)\lambda_\mu - \sum_{\mu=1}^M \sum_{\mu'=1}^M \eta(\lambda_\mu)\lambda_{\mu'} \langle \theta_{\mu'}^1, \theta_\mu^2 \rangle^2\right]. \tag{A.29}$$

Geometry of the hypothesis set and sphere packing. Next, we describe the construction of a large set of hypotheses $\mathcal{F}$, satisfying the 4δ distinguishability condition. Our construction is based on the well-studied sphere-packing problem, namely how many unit vectors can be packed onto $\mathbb{S}^{m-1}$ with given minimal pair-wise distance between any two vectors.

Here we follow the construction due to Zong (1999) (page 77). Let m be a large positive integer, and $m_0 = \lfloor 2m/9 \rfloor$. Define Y_m^* as the maximal set of points of the form $\mathbf{z} = (z_1, \ldots, z_m)$ in $\mathbb{S}^{m-1}$ such that the following is true:

$$\sqrt{m_0} z_i \in \{-1, 0, 1\} \ \forall i \qquad \sum_{i=1}^m |z_i| \leq \sqrt{m_0}$$

and

$$\text{for } \mathbf{z}, \mathbf{z}' \in Y_m^* \qquad \|\mathbf{z} - \mathbf{z}'\| \geq 1.$$

For any $m \geq 1$, the maximal number of points lying on $\mathbb{S}^{m-1}$ such that any two points are at distance at least 1, is called the *kissing number* of an m-sphere. Zong (1999) used the construction described above to derive a lower bound on the *kissing number*, by showing that $|Y_m^*| \geq (9/8)^{m(1+o(1))}$ for m large.

Next, for $m < N - M$ we use the sets Y_m^* to construct our hypothesis set $\mathcal{F}$ of the same size, $|\mathcal{F}| = |Y_m^*|$. To this end, let $\{\mathbf{e}_\mu\}_{\mu=1}^N$ denote the standard basis of $\mathbb{R}^N$. Our initial set $\boldsymbol{\theta}^0$ is composed of the first M standard basis vectors, $\boldsymbol{\theta}^0 = [\mathbf{e}_1 : \ldots : \mathbf{e}_M]$. Then, for fixed ν, and values of m, r yet to be determined, each of the other hypotheses $\boldsymbol{\theta}^j \in \mathcal{F}$ has the same vectors as $\boldsymbol{\theta}^0$ for $k \neq \nu$. The difference is that the νth vector is instead given by

$$\theta_\nu^j = \sqrt{1 - r^2}\mathbf{e}_\nu + r \sum_{l=1}^m z_l^j \mathbf{e}_{M+l}, \qquad j = 1, \ldots, |\mathcal{F}|, \tag{A.30}$$

where $\mathbf{z}^j = (z_1^j, \ldots, z_m^j)$, $j \geq 1$, is an enumeration of the elements of Y_m^*. Thus θ_ν^j perturbs $\mathbf{e}_\nu$ in subsets of the fixed set of coordinates $\{M + 1, \ldots, M + m\}$, according to the sphere-packing construction for $\mathbb{S}^{m-1}$.

The construction ensures that $\theta_1^j, \ldots, \theta_M^j$ are orthonormal for each j. In particular, $\langle \theta_\mu^j, \mathbf{e}_{\mu'} \rangle$ vanishes unless $\mu = \mu'$, and so (A.29) simplifies to

$$K(\boldsymbol{\theta}^j, \boldsymbol{\theta}^0) = \tfrac{1}{2} n h(\lambda_\nu)(1 - \langle \theta_\nu^j, \theta_\nu^0 \rangle^2) = \tfrac{1}{2} n h(\lambda_\nu) r^2 \tag{A.31}$$

for $j = 1, \ldots, |\mathcal{F}|$. Finally, $\langle \theta_\nu^j, \theta_\nu^k \rangle = 1 - r^2 + r^2 \langle \mathbf{z}^j, \mathbf{z}^k \rangle$, and so by construction, for any $\boldsymbol{\theta}^j, \boldsymbol{\theta}^k \in \mathcal{F}$ with $j \neq k$, we have

$$L(\theta_\nu^j, \theta_\nu^k) \geq r^2. \tag{A.32}$$

In other words, the set $\mathcal{F}$ is r^2-distinguishable in θ_ν. Consequently, combining (A.27), (A.28) and (A.31) (taking $Q = P_{\boldsymbol{\theta}^0}$ in Lemma A.5), we have

$$R_\nu^* = \inf_{\widehat{\theta}_\nu} \sup_{\Theta_q(\mathbf{C})} \mathbb{E} L(\widehat{\theta}_\nu, \theta_\nu) \geq (r^2/4)[1 - a(r, \mathcal{F})] \tag{A.33}$$

with

$$a(r, \mathcal{F}) = \frac{n h(\lambda_\nu) r^2/2 + \log 2}{\log |\mathcal{F}|}. \tag{A.34}$$

Proof of Theorem 3.1. It remains to specify m and let $r \in (0, 1)$. Let Y_m^* be the sphere-packing set defined above, and let $\mathcal{F}$ be the corresponding set of hypotheses, defined via (A.30).

Let $c_1 = \log(9/8)$, then we have $\log|\mathcal{F}| \geq b_m c_1 m$, where $b_m \to 1$ as $m \to \infty$. Now choose $r = r(m)$ so that $a(r, \mathcal{F}) \leq 3/4$ asymptotically in the bound (A.33). To accomplish this, set

$$r^2 = \frac{c_1 m}{nh(\lambda_\nu)}. \tag{A.35}$$

Indeed, inserting this into (A.34) we find that

$$a(r, \mathcal{F}) \leq \frac{c_1 m/2 + \log 2}{b_m c_1 m}.$$

Therefore, so long as $m \geq m_*$, an absolute constant, we have $a(r, \mathcal{F}_0) \leq 3/4$ and hence $R_\nu^* \geq r^2/16 = (c_1/16) m \tau_\nu^2$.

We also need to ensure that $\theta_\nu^j \in \Theta_q(C_\nu)$. Since exactly m_0 coordinates are nonzero out of $\{M+1, \ldots, M+m\}$,

$$\|\theta_\nu^j\|_q^q = (1 - r^2)^{q/2} + r^q m_0^{1-q/2} \leq 1 + a_q r^q m^{1-q/2},$$

where $a_q = (2/9)^{1-q/2}$. A sufficient condition for $\theta_\nu^{(j)} \in \Theta_q(C_\nu)$ is that

$$a_q m (r^2/m)^{q/2} \leq \bar{C}_\nu^q. \tag{A.36}$$

Our choice (A.35) fixes r^2/m, and so, recalling that $A_q = 1/(a_q c_1^{q/2})$, the previous display becomes

$$m \leq A_q \bar{C}_\nu^q [nh(\lambda_\nu)]^{q/2}.$$

To simultaneously ensure that (i) $r^2 < 1$, (ii) m does not exceed the number of available co-ordinates, $N - M$ and (iii) $\theta_\nu^j \in \Theta_q(C_\nu)$, we set

$$m = \lfloor m' \rfloor, \qquad m' = \min\{nh(\lambda_\nu), N - M, A_q \bar{C}_\nu^q (nh(\lambda_\nu))^{q/2}\}.$$

Recalling (3.1), (3.2) and (3.3), we have

$$m' = \min\{\tau_\nu^{-2}, N', m_\nu\} = \tau_\nu^{-2} \min\{1, \tau_\nu^2 \cdot \min\{N', m_\nu\}\}.$$

To complete the proof of Theorem 3.1, set $B_1 = [(m_* + 1)/m_*] c_1/16$ and observe that

$$R_\nu^* \geq B_1 m' \tau_\nu^2.$$

Proof of Theorem 3.2. The construction of the set of hypotheses in the proof of Theorem 3.1 considered a fixed set of potential nonzero coordinates, namely $\{M+1, \ldots, M+m\}$. However, in the *sparse* setting, when the effective dimension is significantly smaller than the nominal dimension N, it is possible to construct a much larger collection of hypotheses by allowing the set of nonzero coordinates to span all remaining coordinates $\{M+1, \ldots, N\}$.

In the proof of Theorem 3.2 we shall use the following lemma, proven in Appendix B. Call $A \subset \{1, \ldots, N\}$ an *m-set* if $|A| = m$.

LEMMA A.7. *Let k be fixed, and let $\mathcal{A}_k$ be the maximal collection of m-sets such that the intersection of any two members has cardinality at most $k-1$. Then, necessarily,*

$$|\mathcal{A}_k| \geq \binom{N}{k} \Big/ \binom{m}{k}^2. \tag{A.37}$$

Let $k = [m_0/2] + 1$ and $m_0 = [\beta m]$ with $0 < \beta < 1$. Suppose that $m, N \to \infty$ with $m = o(N)$. Then

$$|\mathcal{A}_k| \geq \exp\bigl[N\mathcal{E}(\beta m/2N) - 2m\mathcal{E}(\beta/2)\bigr](1 + o(1)), \tag{A.38}$$

where $\mathcal{E}(x)$ is the Shannon entropy function,

$$\mathcal{E}(x) = -x\log(x) - (1-x)\log(1-x), \qquad 0 < x < 1.$$

Let π be an m-set contained in $\{M+1, \dots, N\}$, and construct a family $\mathcal{F}_\pi$ by modifying (A.30) to use the set π rather than the fixed set $\{M+1, \dots, M+m\}$ as in Theorem 3.1,

$$\theta_\nu^{(j,\pi)} = \sqrt{1-r^2}\mathbf{e}_\nu + r\sum_{l\in\pi} z_l^j \mathbf{e}_l, \qquad j = 1, \dots, |Y_m^*|.$$

We will choose m below to ensure that $\theta_\nu^{(j,\pi)} \in \Theta_q(C_\nu)$. Let $\mathcal{P}$ be a collection of sets π such that, for any two sets π and π' in $\mathcal{P}$, the set $\pi \cap \pi'$ has cardinality at most $m_0/2$. This ensures that the sets $\mathcal{F}_\pi$ are disjoint for $\pi \neq \pi'$, since each $\theta_\nu^{(j,\pi)}$ is nonzero in exactly $m_0 + 1$ coordinates. This construction also ensures that

$$\text{for all } \boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \bigcup_{\pi\in\mathcal{P}} \mathcal{F}_\pi \qquad L(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \geq \left(\frac{m_0}{2} + \frac{m_0}{2}\right)\left(\frac{r}{\sqrt{m_0}}\right)^2 = r^2.$$

Define $\mathcal{F} := \bigcup_{\pi\in\mathcal{P}} \mathcal{F}_\pi$. Then

$$|\mathcal{F}| = \left|\bigcup_{\pi\in\mathcal{P}} \mathcal{F}_\pi\right| = |\mathcal{P}||Y_m^*| \geq |\mathcal{P}|(9/8)^{m(1+o(1))}. \tag{A.39}$$

By Lemma A.7, there is a collection $\mathcal{P}$ such that $|\mathcal{P}|$ is at least $\exp([N\mathcal{E}(m/9N) - 2m\mathcal{E}(1/9)](1 + o(1)))$. Since $\mathcal{E}(x) \geq -x\log x$, it follows from (A.39) that

$$\frac{\log|\mathcal{F}|}{m} \geq \left(\frac{1}{9}\log\frac{9N}{m} - 2\mathcal{E}(1/9)\right) + \log(9/8)\bigl(1 + o(1)\bigr) \geq \frac{\alpha}{9}\log N + O(1),$$

since $m = O(N^{1-\alpha})$.

Proceeding as for Theorem 3.1, we have $\log|\mathcal{F}| \geq b_m(\alpha/9)m\log N$, where $b_m \to 1$. Let us set (with m still to be specified)

$$r^2 = m\frac{(\alpha/9)\log N}{nh(\lambda_\nu)} = m\bar{\tau}_\nu^2. \tag{A.40}$$

Again, so long as $m \geq m_*$, we have $a(r, \mathcal{F}) \leq 3/4$ and $R_\nu^* \geq r^2/16 = (1/16)m\bar{\tau}_\nu^2$. We also need to ensure that $\theta_\nu^{(j,\pi)} \in \Theta_q(C_\nu)$, which as before is implied by (A.36). Substituting (A.40) puts this into the form

$$m \leq \bar{m}_\nu = a_q^{-1}(\bar{C}_\nu/\bar{\tau}_\nu)^q.$$

To simultaneously ensure that (i) $r^2 < 1$, (ii) m does not exceed the number of available co-ordinates, $N - M$ and (iii) $\theta_\nu^j \in \Theta_q(C_\nu)$, we set

$$m = \lfloor m' \rfloor, \qquad m' = \min\{\bar{\tau}_\nu^{-2}, N - M, \bar{m}_\nu\}.$$

The assumption that $\bar{C}_\nu^q n^{q/2} = O(N^{1-\alpha})$ for some $\alpha \in (0, 1)$ is equivalent to the assertion $\bar{m}_\nu = O(N^{1-\alpha})$, and so for n sufficiently large, $\bar{m}_\nu \leq N - M$ and so $m' = \bar{m}_\nu$ so long as $\bar{m}_\nu \bar{\tau}_\nu^2 \leq 1$. Theorem 3.2 now follows from our bound on R_ν^*.

A.4. Lower bound on the risk of the DT estimator. To prove Theorem 4.1, assume w.l.o.g. that $\langle\widehat{\theta}_{1,\mathrm{DT}}, \theta_1\rangle > 0$, and decompose the loss as

$$L(\widehat{\theta}_{1,\mathrm{DT}}, \theta_1) = \|\theta_1 - \theta_{1,I}\|^2 + \|\widehat{\theta}_{1,\mathrm{DT}} - \theta_{1,I}\|^2, \tag{A.41}$$

where $I = I(\gamma_n)$ is the set of coordinates selected by the DT scheme and $\theta_{1,I}$ denotes the subvector of θ_1 corresponding to this set. Note that, in (A.41), the first term on the right can be viewed as a bias term while the second term can be seen as a variance term.

We choose a particular vector $\theta_1 = \theta_* \in \Theta_q(C)$ so that

$$\mathbb{E}\|\theta_* - \theta_{*,I}\|^2 \geq K\bar{C}^q n^{-(1-q/2)/2}. \tag{A.42}$$

This, together with (A.41), proves Theorem 4.1 since the worst case risk is clearly at least as large as (A.42). Accordingly, set $r_n = \bar{C}^{q/2} n^{-(1-q/2)/4}$, where $\bar{C}^q = C^q - 1$. Since $C^q n^{q/4} = o(n^{1/2})$, we have $r_n = o(1)$, and so for sufficiently large n, we can take $r_n < 1$ and define

$$\theta_{*,k} = \begin{cases} \sqrt{1 - r_n^2}, & \text{if } k = 1, \\ \dfrac{r_n}{\sqrt{m_n}}, & \text{if } 2 \leq k \leq m_n + 1, \\ 0, & \text{if } m_n + 2 \leq k \leq N, \end{cases}$$

where $m_n = \lfloor (1/2)\bar{C}^q n^{q/4} \rfloor$. Then by construction $\theta_* \in \Theta_q(C)$, since

$$\sum_{k=1}^N |\theta_{*,k}|^q = (1 - r_n^2)^{q/2} + r_n^q m_n^{1-q/2} < 1 + r_n^q m_n^{1-q/2} \leq 1 + \frac{\bar{C}^q}{2^{1-q/2}} < C^q,$$

where the last inequality is due to $q \in (0, 2)$ and $\bar{C}^q = C^q - 1$.

For notational convenience, let $\alpha_n = \gamma\sqrt{\log N/n}$. Recall that DT selects all coordinates k for which $\mathbf{S}_{kk} > 1 + \alpha_n$. Since $\mathbf{S}_{kk} \sim (1 + \lambda_1 \theta_{*,k}^2)\chi_n^2/n$, coordinate k is *not selected* with probability

$$p_k = \mathbb{P}(\mathbf{S}_{kk} < 1 + \alpha_n) = \mathbb{P}\big(\chi_n^2 < n(1 + \varepsilon_n)\big), \tag{A.43}$$

where $\varepsilon_n = (1+\alpha_n)/(1+\lambda_1\theta_{*,k}^2) - 1$. Notice that, for $k = 2, \dots, m_n + 1$, $p_k = p_2$ and $\theta_{*,k} = 0$ for $k > m_n + 1$. Hence,

$$\mathbb{E}\|\theta_* - \theta_{*,I}\|^2 = \sum_{k=1}^{N} p_k|\theta_{*,k}|^2 > p_2 \sum_{k=2}^{m_n+1} |\theta_{*,k}|^2 = p_2 r_n^2 = p_2 \bar{C}^q n^{-(1-q/2)/2}.$$

Now, use bound (A.3) to show that $n\varepsilon_n^2 \to \infty$ in (A.43) and hence that $p_2 \to 1$. Indeed $\theta_{*,2}^2 = r_n^2/m_n = 2n^{-1/2}(1 + o(1))$, and so

$$\varepsilon_n = \frac{\alpha_n - \lambda_1\theta_{*,k}^2}{1 + \lambda_1\theta_{*,k}^2} = \frac{1}{2\sqrt{n}}[\gamma\sqrt{\log N} - 2\lambda_1]$$

for sufficiently large n. Hence, $n\varepsilon_n^2 \to \infty$, and the proof is complete.

APPENDIX B: PROOFS OF RELEVANT LEMMAS

B.1. Proof of Lemma A.1. We use the following result on extreme eigenvalues of Wishart matrices from Davidson and Szarek (2001).

LEMMA B.1. *Let Z be a $p \times q$ matrix of i.i.d. $N(0,1)$ entries with $p \le q$. Let $s_{\max}(Z)$ and $s_{\min}(Z)$ denote the largest and the smallest singular value of Z, respectively. Then*

$$\mathbb{P}\big(s_{\max}(Z/\sqrt{q}) > 1 + \sqrt{p/q} + t\big) \le e^{-qt^2/2}, \tag{B.1}$$

$$\mathbb{P}\big(s_{\min}(Z/\sqrt{q}) < 1 - \sqrt{p/q} - t\big) \le e^{-qt^2/2}. \tag{B.2}$$

Observe first that

$$\Delta := \|n^{-1}\mathbf{Z}\mathbf{Z}^T - I_N\| = \max\{\lambda_1(n^{-1}\mathbf{Z}\mathbf{Z}^T) - 1, 1 - \lambda_N(\mathbf{Z}\mathbf{Z}^T)\}.$$

Let $s_\pm$ denote the maximum and minimum singular values of $N^{-1/2}\mathbf{Z}$. Define $\gamma(t) := \sqrt{N/n} + t$ for $t > 0$. Then since $\Delta = \max\{s_+^2 - 1, 1 - s_-^2\}$, and letting $\Delta_n(t) := 2\gamma(t) + \gamma(t)^2$, we have

$$\{\Delta > \Delta_n(t)\} \subset \{s_+ > 1 + \gamma(t)\} \cup \{s_- < 1 - \gamma(t)\}.$$

We apply Lemma B.1 with $p = N$ and $q = n$, and get

$$\mathbb{P}\big(\Delta > \Delta_n(t)\big) \le 2e^{-nt^2/2}.$$

We observe that, with $\gamma_n = N/n \le 1$,

$$\begin{aligned}\Delta_n(t) &= (N/n + 2\sqrt{N/n}) + t(2 + t + 2\sqrt{N/n}) \\ &\le \gamma_n + 2\sqrt{\gamma}_n + t(4+t).\end{aligned} \tag{B.3}$$

Now choose $t = c\sqrt{2\log n/n}$ so that tail probability is at most $2e^{-n^2t^2/2} = 2n^{-c^2}$. The result is now proved, since if $c\sqrt{\log n/n} \le 1$, then $t(4+t) \le ct_n$.

B.2. Proof of Lemma A.4. Paul (2005) introduced the quantities

$$\Delta_r := \frac{1}{2}[\|H_r(A)B\| + |\lambda_r(A+B) - \lambda_r(A)|\|H_r(A)\|], \tag{B.4}$$

$$\bar{\Delta}_r = \frac{\|B\|}{\min_{1\le j\ne r\le m}|\lambda_j(A) - \lambda_r(A)|} \tag{B.5}$$

and showed that the residual term R_r can be bounded by

$$\begin{aligned}\|R_r\| \le \min\biggl\{10\bar{\Delta}_r^2, \|H_r(A)B\mathbf{p}_r(A)\|\biggl[&\frac{2\Delta_r(1+2\Delta_r)}{1-2\Delta_r(1+2\Delta_r)} \\ &+ \frac{\|H_r(A)B\mathbf{p}_r(A)\|}{(1-2\Delta_r(1+2\Delta_r))^2}\biggr]\biggr\},\end{aligned} \tag{B.6}$$

where the second bound holds only if $\Delta_r < (\sqrt{5}-1)/4$.

We now show that if $\bar{\Delta}_r \le 1/4$, then we can simplify bound (B.6) to obtain (A.7). To see this, note that $|\lambda_r(A+B) - \lambda_r(A)| \le \|B\|$ and that $\|H_r(A)\| \le [\min_{j\ne r}|\lambda_j(A) - \lambda_r(A)|]^{-1}$, so that

$$\Delta_r \le \|H_r(A)\|\|B\| \le \bar{\Delta}_r.$$

Now, defining $\delta := 2\bar{\Delta}_r(1 + 2\bar{\Delta}_r)$ and $\beta := \|H_r(A)B\mathbf{p}_r(A)\|$, we have $10\bar{\Delta}_r^2 \le (5/2)\delta^2$, and the bound (B.6) may be expressed as

$$\|R_r\| \le \frac{\beta\delta}{1-\delta}\min\biggl\{\frac{5}{2}\frac{\delta(1-\delta)}{\beta}, 1 + \frac{\beta}{\delta(1-\delta)}\biggr\}.$$

For $x > 0$, the function $x \mapsto \min\{5x/2, 1 + 1/x\} \le 5/2$. Further, if $\bar{\Delta}_r < 1/4$, then $\delta < 3\bar{\Delta}_r < 3/4$, and so we conclude that

$$\|R_r\| \le 10\beta\delta \le 30\beta\bar{\Delta}_r.$$

B.3. Proof of Lemma A.6. Recall that, if distributions F_1 and F_2 have density functions f_1 and f_2, respectively, such that the support of f_1 is contained in the support of f_2, then the Kullback–Leibler discrepancy of F_2 with respect to F_1, to be denoted by $K(F_1, F_2)$, is given by

$$K(F_1, F_2) = \int \log \frac{f_1(y)}{f_2(y)} f_1(y)\,dy. \tag{B.7}$$

For n i.i.d. observations $X_i, i = 1, \dots, n$, the Kullback–Leibler discrepancy is just n times the Kullback–Leibler discrepancy for a single observation. Therefore, without loss of generality we take $n = 1$. Since

$$\Sigma^{-1} = I - \sum_{\mu=1}^{M} \eta(\lambda_\mu)\theta_\mu\theta_\mu^T, \tag{B.8}$$

the log-likelihood function for a single observation is given by

$$\begin{aligned} \log f(x|\boldsymbol{\theta}) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}x^T\Sigma^{-1}x \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{\mu=1}^{M}\log(1+\lambda_\mu) \\ &\quad - \frac{1}{2}\left(\langle x, x\rangle - \sum_{\mu=1}^{M}\eta(\lambda_\mu)\langle x, \theta_\mu\rangle^2\right). \end{aligned} \tag{B.9}$$

From (B.9), we have

$$\begin{aligned} \mathcal{K}_{1,2} &= \mathbb{E}_{\boldsymbol{\theta}^1}(\log f(X|\boldsymbol{\theta}^1) - \log f(X|\boldsymbol{\theta}^2)) \\ &= \frac{1}{2}\sum_{\mu=1}^{M}\eta(\lambda_\mu)[\mathbb{E}_{\boldsymbol{\theta}^1}(\langle X, \theta_\mu^1\rangle)^2 - \mathbb{E}_{\boldsymbol{\theta}^1}(\langle X, \theta_\mu^2\rangle)^2] \\ &= \frac{1}{2}\sum_{\mu=1}^{M}\eta(\lambda_\mu)[\langle\theta_\mu^1, \Sigma_1\theta_\mu^1\rangle - \langle\theta_\mu^2, \Sigma_1\theta_\mu^2\rangle] \\ &= \frac{1}{2}\sum_{\mu=1}^{M}\eta(\lambda_\mu)\left[\sum_{\mu'=1}^{M}\lambda_{\mu'}\{\langle\theta_{\mu'}^1, \theta_\mu^1\rangle^2 - \langle\theta_{\mu'}^1, \theta_\mu^2\rangle^2\}\right], \end{aligned}$$

which equals the RHS of (A.29), since the columns of $\boldsymbol{\theta}^j$ are orthonormal for each $j = 1, 2$.

B.4. Proof of Lemma A.7. Let $\mathcal{P}_m$ be the collection of all m-sets of $\{1, \dots, N\}$, clearly $|\mathcal{P}_m| = \binom{N}{m}$. For any m-set A, let $\mathcal{I}(A)$ denote the collection of "inadmissible" m-sets A' for which $|A \cap A'| \geq k$. Clearly

$$|\mathcal{I}(A)| \leq \binom{m}{k}\binom{N-k}{m-k}.$$

If $\mathcal{A}_k$ is maximal, then $\mathcal{P}_m = \bigcup_{A \in \mathcal{A}_k} \mathcal{I}(A)$, and so (A.37) follows from the inequality

$$|\mathcal{P}_m| \leq |\mathcal{A}_k| \max_A |\mathcal{I}(A)|$$

and rearrangement of factorials.

Turning to the second part, we recall that Stirling's formula shows that if k and $N \to \infty$,

$$\binom{N}{k} = \varsigma\left(\frac{N}{2\pi k(N-k)}\right)^{1/2} \exp\left\{N\mathcal{E}\left(\frac{k}{N}\right)\right\},$$

where $\varsigma \in (1-(6k)^{-1}, 1+(12N)^{-1})$. The coefficient multiplying the exponent in $\binom{N}{k}/\binom{m}{k}^2$ is

$$\sqrt{2\pi k}(1-k/N)^{-1/2}(1-k/m) \sim \sqrt{\pi\beta m}(1-\beta/2) \to \infty$$

under our assumptions, and this yields (A.38).

Acknowledgement. The authors thank the Editor and an anonymous referee for their thoughtful suggestions. The final draft was completed with the hospitality of the Institute of Mathematical Sciences at the National University of Singapore.

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

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148. MR0145620

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BIRGÉ, L. (2001). A new look at an old result: Fano's lemma. Technical report, Univ. Paris 6.

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

CAI, T. T., MA, Z. and WU, Y. (2012). Sparse PCA: Optimal rates and adaptive estimation. Technical report. Available at arXiv:1211.1309.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CAI, T. T. and ZHOU, H. H. (2012). Minimax esrimation of large covariance matrices under l_1 norm. *Statist. Sinica* **22** 1319–1378.

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

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, *Vol. I* (W. B. Johnson and J. Lindenstrauss, eds.) 317–366. North-Holland, Amsterdam. MR1863696

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

JOHNSTONE, I. M. (2001). Chi-square oracle inequalities. In *State of the Art in Probability and Statistics* (*Leiden*, 1999) (M. de Gunst, C. Klaassen and A. van der Waart, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 399–418. IMS, Beachwood, OH. MR1836572

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. (2002). *Principal Component Analysis*. Springer, Berlin.

KATO, T. (1980). *Perturbation Theory of Linear Operators*. Springer, New York.

KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems* **94** 19–32.

LU, A. Y. (2002). Sparse principal components analysis for functional data. Ph.D. thesis, Stanford Univ., Stanford, CA.

MA, Z. (2011). Sparse principal component analysis and iterative thresholding. Technical report, Dept. Statistics, The Wharton School, Univ. Pennsylvania, Philadelphia, PA.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

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

NADLER, B. (2009). Discussion of "On consistency and sparsity for principal component analysis in high dimensions." *J. Amer. Statist. Assoc.* **104** 694–697. MR2751449

ONATSKI, A. (2006). Determining the number of factors from empirical distribution of eigenvalues. Technical report, Columbia Univ.

PAUL, D. (2005). Nonparametric estimation of principal components. Ph.D. thesis, Stanford Univ. Stanford, CA.

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

PAUL, D. and JOHNSTONE, I. M. (2007). Augmented sparse principal component analysis for high dimensional data. Technical report, Univ. California, Davis. Available at arXiv:1202.1242.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

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. (2011). Consistency of sparse PCA in high dimension, low sample size contexts. Technical report. Available at http://arxiv.org/pdf/1104.4289v1.pdf.

TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 611–622. MR1707864

VAN TREES, H. L. (2002). *Optimum Array Processing*. Wiley, New York.

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. Technical report. Available at http://arxiv.org/pdf/1202.0786.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.

YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

ZONG, C. (1999). *Sphere Packings*. Springer, New York. MR1707318

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

A. BIRNBAUM
SCHOOL OF COMPUTER SCIENCE
AND ENGINEERING
HEBREW UNIVERSITY OF JERUSALEM
THE EDMOND J. SAFRA CAMPUS
JERUSALEM, 91904
ISRAEL
E-MAIL: aharob01@cs.huji.ac.il

I. M. JOHNSTONE
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: imj@stanford.edu

B. NADLER
DEPARTMENT OF COMPUTER SCIENCE
AND APPLIED MATHEMATICS
WEIZMANN INSTITUTE OF SCIENCE
P.O. BOX 26, REHOVOT, 76100
ISRAEL
E-MAIL: boaz.nadler@weizmann.ac.il

D. PAUL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
USA
E-MAIL: debashis@wald.ucdavis.edu