Augmented Sparse Principal Component Analysis for High Dimensional Data

Debashis Paul & Iain M. Johnstone

December 11, 2007

1 Introduction

Principal components analysis (PCA) has been a widely used technique in reducing dimensionality of multivariate data. A traditional setting where PCA is applicable is when one has repeated observations from a multivariate population that can be described reasonably well by its first two moments. When the dimension of sample observations, is fixed, distributional properties of the eigenvalues and eigenvectors of the sample covariance have been dealt with at length by various authors. Anderson (1963), Muirhead (1982) and Tyler (1983) are among standard references. Much of the "large sample" study of the eigen-structure of the sample covariance matrix is based on the fact that, sample covariance approximates population covariance matrix well when sample size is large. However, due to advances in data acquisition technologies, statistical problems, where the dimensionality of individuals are of nearly the same order of magnitude as (or even bigger than) the sample size, are increasingly common. The following is a representative list of areas and articles where PCA has been in use. In all these cases N denotes the dimension of an observation and n denotes the sample size.

- *Image recognition :* The face recognition problem is to identify a face from a collection of faces. Here each observation is a digitized image of the face of a person. So typically, with 128×128 pixel grids, one has to deal with a situation where $N \approx 1.6 \times 10^6$. Whereas, a standard image database, e.g. that of students of Brown University Wickerhauser (1994), may contain only a few hundred pictures.

- *Shape analysis :* Stegmann and Gomez (2002), Cootes, Edwards and Taylor (2001) outline a class of methods for analyzing the shape of an object based on repeated measurements that involves annotating the objects for landmarks. These landmarks act as features of the objects, and hence, can be thought of as the dimension of the observations. For a specific example relating to motion of hand Stegmann and Gomez (2002), the number of landmarks is 56 and sample size is 40.

- *Chemometrics :* In many chemometric studies, sometimes the data consists of several thousands of spectra measured at several hundred wavelength positions, e.g. data collected for calibration of spectrometers. Vogt, Dable, Cramer and Booksh (2004) give an overview of some of these applications.

- *Econometrics :* Large factor analysis models are often used in econometric studies, e.g. in dealing with hundreds of stock prices as a multivariate time series. Markowitz's theory of optimal portfolios ask this question. *Given a set of financial assets characterized by their average return and risk, what is the optimal weight of each asset, such that the overall*

portfolio provides the best return? Laloux, Cizeau, Bouchaud and Potters (2000) discuss several applications. Bai (2003) considers some inferential aspects.

- *Climate studies :* Measurements on atmospheric indicators, like ozone concentration etc. are taken at a number of monitoring stations over a number of time points. In this literature, principal components are commonly referred to as "empirical orthogonal functions". Preisendorfer (1988) gives a detailed treatment. EOFs are also used for model diagnostics and data summary Cassou, Deser, Terraty, Hurrell and Drévillon (2004).

- *Communication theory :* Tulino and Verdu (2004) give an extensive treatment to the connection between random matrix theory and vector channels used in wireless communications.

- *Functional data analysis :* Since observations are curves, which are typically measured at a large number of points, the data is high dimensional. Buja, Hastie and Tibshirani (1995) give an example of speech dataset consisting of 162 observations - each one is a periodogram of a "phoneme" spoken by a person. Ramsay and Silverman (2002) discuss other applications.

- *Microarray analysis :* Gene microarrays present data in the form expression profiles of several thousand genes for each subject under study. Bair, Hastie, Paul and Tibshirani (2006) analyze an example involving the study of survival times of 240 $(= n)$ patients with diffuse large B-cell lymphoma, with gene expression measurements for 7389 $(= N)$ genes.

Of late, researchers in various fields have been using different versions of non-identity covariance matrices of growing dimension. Among these, a particularly interesting model assumes that,

(*) the eigenvalues of the population covariance matrix Σ are (in descending order)

$$\ell_1, \ldots, \ell_M, \sigma^2, \ldots, \sigma^2,$$

where $\ell_M > \sigma^2 > 0$.

This has been deemed the "spiked population model" by Johnstone (2001). It has also been observed that for certain types of data, e.g. in speech recognition Buja, Hastie and Tibshirani (1995), wireless communication Telatar (1999), statistical learning (Hoyle and Rattray (2003, 2004)), a few of the sample eigenvalues have limiting behavior that is different from the behavior when the covariance is the identity. This paper deals with the issue of estimating the eigenvectors of Σ, when it has the structure described by (*), and the dimension N grows to infinity together with sample size n.

In many practical problems, at least the leading eigenvectors are thought to represent some underlying phenomena. This has been one of the reasons for their popularity in analysis of what can be characterized as functional data. For example, Zhao, Marron and Wells (2004) consider the "yeast cell cycle" data of Spellman *et al.* (1998), and argue that the first two components obtained by a functional PCA of the data represent systematic structure. In climate studies, empirical orthogonal functions are often used for identifying patterns in the data, as well as for

data summary. See for example Corti, Molteni and Palmer (1999). In many of these instances there is some idea about the structure of the eigenvectors of the covariance matrix, such as to the extent they are smooth, or oscillatory. At the same time, these data are often corrupted with a substantial amount of noise, which can lead to very noisy estimates of the eigen-elements. There is also a growing literature on functional response models in which the regressors are random functions and the responses are either vectors or functions (Chiou, Müller and Wang (2004), Hall and Horowitz (2004), Cardot, Ferraty and Sarda (2003)). Quite often a functional principal component regression is used to solve these problems. Thus, there are both practical and scientific interests in devising methods for estimating the eigenvectors and eigenvalues that can take advantage of the information about the structure of the population eigenvectors. At the same time, there is also a need to address this estimation problem from a broader statistical perspective.

In multivariate analysis, there is a huge body of work on estimation of population covariance, and in particular on developing optimal strategies for estimation from a decision theoretic point of view. Dey and Srinivasan (1985), Efron and Morris (1976), Haff (1980), Loh (1988) are some of the standard references in this field. However, a decision theoretic treatment of functional data analysis is still somewhat limited in its breadth. Hall and Horowitz (2004) and Tony Cai and Hall (2005) derive optimal rates of convergence of estimators of the regression function and fitted response in functional linear model context. Cardot (2000) gave upper bounds on the rate of convergence of a spline-based estimator of eigenvectors under some smoothness assumptions. Kneip (1994) also derived similar results in a slightly different context.

In this paper, the aim is to address the problem of estimating eigenvectors from a minimax risk analysis viewpoint. Henceforth, the observations will be assumed to have a Gaussian distribution. This assumption, though somewhat idealized, helps in bringing out some essential features of the estimation problem. Since algebraic manipulation of spectral elements of a matrix is rather difficult, it is not easy to make any precise finite sample statement about the risk properties of estimators. Therefore the analysis is mostly asymptotic in nature, even though efforts have been made to make the approximations to risk etc. as explicit as possible. The asymptotic regime considered here assumes a triangular array structure in which N, the dimensionality of individual observations, tends to ∞ with sample size n. This framework is partly motivated by similar analytical approaches to the problem of estimation of mean function in nonparametric regression context. In particular, a squared error type loss is proposed, and some l^q-type sparsity constraint is imposed on the parameters, which in our case are individual eigenvectors. Relevance of this sort of constraints in the context of functional data analysis is discussed in Section 3. The main results of this chapter are the following. Theorem 1 describes risk behavior of sample eigenvectors as estimators of their population counterparts. Theorem 2 gives a lower bound on the minimax risk. An estimation scheme, named *Augmented Sparse Principal Component Analysis (ASPCA)* is proposed and is shown to have the optimal rate of convergence over a class of l^q norm-constrained parameter spaces under suitable regularity conditions. Throughout it is assumed that the leading eigenvalues of the population covariance matrix are distinct, so the eigenvectors are identifiable. A more general framework, which looks at estimating the eigen-subspaces and allows for eigenvalues with arbitrary multiplicity, is beyond the scope of this paper.

3

2 Model

Suppose that, $\{X_i : i = 1, \ldots, n\}_{n \geq 1}$ is a triangular array, where the $N \times 1$ vectors $X_i := X_i^n, i = 1, \ldots, n$ are i.i.d. on a common probability space for each n. The dimension N is assumed to be a function of n and increases without bound as $n \to \infty$. The observation vectors are assumed to be i.i.d. as $N(\xi, \Sigma)$, where ξ is the mean vector; and Σ is the covariance matrix. The assumption on Σ is that, it is a finite rank perturbation of (a multiple of) the identity. In other words,

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

where $\lambda_1 > \lambda_2 > \ldots > \lambda_M > 0$, and the vectors $\theta_1, \ldots, \theta_M$ are orthonormal. Notice that strict inequality in the order relationship among the λ_ν's implies that the θ_ν are identifiable up to a sign convention. Notice that with this identifiability condition, θ_ν is the eigenvector corresponding to the ν-th largest eigenvalue, namely, $\lambda_\nu + \sigma^2$, of Σ. The term "finite rank" means that, M will remain fixed for all the asymptotic analysis that follows. This analysis involves letting both n and N increase to infinity simultaneously. Therefore, Σ, the λ_ν's and the θ_ν's should be thought of as being dependent on N.

The observations can be equivalently described in terms of the *factor analysis model* :

$$X_{ik} = \xi + \sum_{k=1}^{M} \sqrt{\lambda_\nu} v_{\nu i} \theta_{\nu k} + \sigma Z_{ik}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, N. \tag{2}$$

Here, for each n, $v_{\nu i}$, Z_{ik} are all independently and identically distributed as $N(0, 1)$. $M \geq 1$ is assumed fixed.

Since the eigenvectors of Σ are invariant to a scale change in the original observations, for simplifying notation, it is assumed that $\sigma = 1$. Notice that this also means that, $\lambda_1, \ldots, \lambda_M$ appearing in the results relating to the rates of convergence of various estimators of θ_ν should be changed to $\lambda_1/\sigma, \ldots, \lambda_M/\sigma$ when (1) holds with an arbitrary $\sigma > 0$.

Another simplifying assumption is that, $\xi = 0$. This is because, the main focus of the current exposition is on estimating the eigen-structure of Σ, and the unnormalized sample covariance matrix

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

where $\overline{X}$ is the sample mean, has the same distribution as that of the matrix

$$\sum_{i=1}^{n-1} Y_i Y_i^T,$$

where Y_i are i.i.d. $N(0, \Sigma)$. This means that, for estimation purposes, if the attention is restricted to the sample covariance matrix, then from an asymptotic analysis point of view, it is enough to assume $\xi = 0$, and to define the sample covariance matrix as $\mathbf{S} = \frac{1}{n} \mathbf{X}\mathbf{X}^T$, where $\mathbf{X} = [X_1 : \ldots : X_n]$.

The following condition, or *Basic Assumption* will be used frequently, and will be referred to as **BA**.

BA (2) and (1) hold, with $\xi = 0$ and $\sigma = 1$; $N = N(n) \to \infty$ as $n \to \infty$; $\lambda_1 > \ldots > \lambda_M > 0$.

For the estimation problem, it may be assumed that, as $n, N \to \infty$, $\theta_\nu := \theta_\nu^n \to \overline{\theta}_\nu$ in $l^2(\mathbb{R})$, though it is not strictly necessary. But this assumption is appropriate if the observation vectors are the vectors of first N coefficients of some noisy function in $L^2(D)$ (where D is an interval in $\mathbb{R}$), when represented in a suitable orthogonal basis for the $L^2(D)$ space. See Section **??** for more details. In such cases one can talk about estimating the eigenfunctions of the underlying covariance operator, and the term consistency has its usual interpretation. However, even if θ_ν^n does not converge in l^2, one can still use the term "consistency" of an estimator $\widehat{\theta}_\nu^n$ to mean that $L(\widehat{\theta}_\nu^n, \theta_\nu^n) \to 0$ in probability as $n \to \infty$, where L is an appropriate loss function.

2.1 Squared error type loss

The goal is, given data $X_1, X_2, \ldots, X_n$, to estimate θ_ν, for $\nu = 1, \ldots, M$. To assess the performance of any such estimator, a minimax risk analysis approach is proposed. The first task is to specify a loss function for this estimation problem. Observe that since the model is invariant under separate changes of sign of the θ_ν, it is necessary to specify a loss function that is also invariant under a sign change. We specify the following loss function :

$$L(\mathbf{a}, \mathbf{b}) = L([\mathbf{a}], [\mathbf{b}]) := 2(1 - |\langle \mathbf{a}, \mathbf{b} \rangle|) = \| \mathbf{a} - sign(\langle \mathbf{a}, \mathbf{b} \rangle)\mathbf{b} \|^2, \tag{3}$$

where $\mathbf{a}$ and $\mathbf{b}$ are $N \times 1$ vectors with l^2 norm 1; and $[\mathbf{a}]$ denotes the equivalence class of $\mathbf{a}$ under sign change. Note that, $L(\mathbf{a}, \mathbf{b})$ can also be written as $\min\{\| \mathbf{a} - \mathbf{b} \|^2, \| \mathbf{a} + \mathbf{b} \|^2\}$. There is another useful relationship with a different loss function, denoted by $L_s(\mathbf{a}, \mathbf{b}) := \sin^2 \angle(\mathbf{a}, \mathbf{b})$, for any two $N \times 1$ unit vectors $\mathbf{a}$ and $\mathbf{b}$. $\sin \angle(\cdot, \cdot)$ is a metric on the space $\mathbb{S}^{N-1}$, i.e. the unit sphere in $\mathbb{R}^N$. Also, $L_s(\mathbf{a}, \mathbf{b}) = \sin^2 \angle(\mathbf{a}, \mathbf{b}) = 1 - |\langle \mathbf{a}, \mathbf{b} \rangle|^2 = L(\mathbf{a}, \mathbf{b})(2 - L(\mathbf{a}, \mathbf{b}))$. Hence, if $L(\mathbf{a}, \mathbf{b}) \approx 0$, then these two quantities have approximately the same value. This implies that, the asymptotic risk bounds derived in terms of the loss function L remain valid, up to a constant factor, for the loss function L_s as well.

2.2 Rate of convergence for ordinary PCA

It is assumed that either λ_1 is fixed, or that it varies with n and N so that,

L1 as $n, N \to \infty$, $\frac{\lambda_\nu}{\lambda_1} \to \rho_\nu$ for $\nu = 1, \ldots, M$, where $1 = \rho_1 > \rho_2 > \ldots > \rho_M$;

L2 as $n, N \to \infty$, $\frac{N}{nh(\lambda_1)} \to 0$, where

$$h(\lambda) = \frac{\lambda^2}{1 + \lambda} . \tag{4}$$

Notice that, all four conditions (i)-(iv) below imply that $\frac{N}{nh(\lambda_1)} \to 0$ as $n \to \infty$.

(i) $\frac{N}{n} \to \gamma \in (0, \infty)$ and $\frac{N}{n\lambda_1} \to 0$

(ii) $\lambda_1 \to 0$, $\frac{N}{n} \to 0$ and $\frac{N}{n\lambda_1^2} \to 0$

(iii) $0 < \liminf_{n \to \infty} \lambda_1 \leq \limsup_{n \to \infty} \lambda_1 < \infty$ and $\frac{N}{n} \to 0$

(iv) $\frac{N}{n} \to \infty$, and $\frac{N}{n\lambda_1} \to 0$.

Remark : Condition **L1** is really an asymptotic identifiability condition which guarantees that at the scale of the largest "signal" eigenvalue, bigger eigenvalues are well-separated.

Theorem 1: *Suppose that the eigenvalues $\lambda_1, \ldots, \lambda_M$ satisfy **L1** and **L2**. If $\log(n \vee N) = o(n \wedge N)$, then for $\nu = 1, \ldots, M$,*

$$\sup_{\theta_\nu \in \mathbb{S}^{N-1}} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu) = \left[\frac{N-M}{nh(\lambda_\nu)} + \frac{1}{n} \sum_{\mu \neq \nu} \frac{(\lambda_\mu + 1)(\lambda_\nu + 1)}{(\lambda_\nu - \lambda_\mu)^2} \right] (1 + o(1)). \tag{5}$$

Remark : It is possible to relax some of the conditions stated in the theorem. On the other hand, with some reasonable assumptions on the decay of the eigenvalues, it is also possible to incorporate cases where M is no longer a constant, but increases with n. Then the issues would include, rates of growth of M and the rate of decay of eigenvalues that would result in the OPCA estimator retaining consistency and the expression for its asymptotic risk. These issues are not going to be addressed here. However, it is important to note that, such questions have been investigated - not necessarily for the Gaussian case - in the context of spectral decomposition of L^2 stochastic processes by, Hall and Horowitz (2004), Tony Cai and Hall (2005), Boente and Fraiman (2000), Hall and Hosseini-Nasab (2006) among others. However, these analyses do not deal with measurement errors. The condition $\frac{N}{nh(\lambda_\nu)} \to 0$ is a necessary condition for uniform convergence, as shown in Theorem 2. It should be noted that, there are results, proved under slightly different circumstances, that obtain the rates given by (5) as an upper bound on the rate of convergence of OPCA estimators (Bai (2003), Cardot (2000), Kneip (1994)). These analyses, while treating the problem under less restrictive assumptions than Gaussianity (essentially, finite eighth moment for the noise Z_{ik}), make the assumption that $\frac{N^2}{n} \to 0$, when the λ_ν's are considered fixed.

3 Sparse model for eigenvectors

In this section we discuss the concept of sparsity of the eigenvectors and impose some restrictions on the space of eigenvectors that lead to a sparse parametrization. This notion will be used later from a decision-theoretic view point in order to analyze the risk behavior of estimators of the eigenvectors. From now on, θ will be used to denote the matrix $[\theta_1, \ldots, \theta_M]$.

3.1 l^q constraint on the parameters

The parameter space is taken to be a class of M-dimensional positive semi-definite matrices satisfying the following criteria:

- $\lambda_1 > \ldots > \lambda_M$.

- For each $\nu = 1, \ldots, M$, $\theta_\nu \in \Theta_\nu$ for some $\Theta_\nu \subset \mathbb{S}^{N-1}$ that gives a sparse parametrization, in that most of the coefficients $\theta_{\nu k}$ are close to zero.

- $\theta_1, \ldots, \theta_M$ are orthonormal.

One way to formalize the requirement of sparsity is to demand, as in Johnstone and Lu (2004), that θ_ν belongs to a weak-l^q space $wl^q(C)$ where $C, q > 0$. This space is defined as follows. Suppose that the coordinates of a vector $\mathbf{x} \in \mathbb{R}^N$ are $|x|_{(1)}, \ldots, |x|_{(N)}$, where $|x|_{(k)}$ is the k-th largest element, in absolute value. Then

$$\mathbf{x} \in wl^q(C) \quad \Leftrightarrow \quad |x|_{(k)} \leq Ck^{-1/q}, \quad k = 1, 2, \ldots. \tag{6}$$

In the Functional Data Analysis context, one can think of the observations as the vectors of wavelet coefficients (when transformed in an orthogonal wavelet basis of sufficient regularity) of the observed functions. If the smoothness of a function g is measured by its membership in a Besov space $B_{q',r}^\alpha$, and if the vector of its wavelet coefficients, when expanded in a sufficiently regular wavelet basis, is denoted by $\mathbf{g}$, then from Donoho (1993),

$$g \in B_{q',r}^\alpha \quad \Longrightarrow \quad \mathbf{g} \in wl^q, \qquad q = \frac{2}{2\alpha + 1}, \ \text{if} \ \alpha > (1/q' - 1/2)_+.$$

One may refer to Johnstone (2002) for more details. Treating this as a motivation, instead of imposing a weak-l^q constraint on the parameter θ_ν, we rather impose an l^q constraint. Note that, for $C, q > 0$,

$$\mathbf{x} \in \mathbb{R}^N \cap l^q(C) \quad \Leftrightarrow \quad \sum_{k=1}^{N} |x_k|^q \leq C^q. \tag{7}$$

Since $l^q(C) \hookrightarrow wl^q(C)$, it is possible to derive lower bounds on the minimax risk of estimators when the parameter lies in a wl^q space by restricting attention to an l^q ball of appropriate radius.

For $C > 0$, define $\Theta_q(C)$ by

$$\Theta_q(C) = \{a \in \mathbb{S}^{N-1} : \sum_{k=1}^{N} |a_k|^q \leq C^q\}, \tag{8}$$

where $\mathbb{S}^{N-1}$ is the unit sphere in $\mathbb{R}^N$ centered at 0. One important fact is, if $0 < q < 2$, for $\Theta_q(C)$ to be nonempty, one needs $C \geq 1$, while for $q > 2$, the reverse inequality is necessary. Further, for $0 < q < 2$, if $C^q \geq N^{1-q/2}$, then the space $\Theta_q(C)$ reduces to $\mathbb{S}^{N-1}$ because in this case, the vector $(1/\sqrt{N}, 1/\sqrt{N}, \ldots, 1/\sqrt{N})$ is in the parameter space. Also, the only vectors that belong in the space when $C = 1$ are the poles, i.e. vectors of the form $(0, 0, \ldots, 0, \pm 1, 0, \ldots, 0)$, where the non-zero term appears in exactly one coordinate. Define, for $q \in (0, 2)$, m_C to be an integer ≥ 1 that satisfies

$$m_C^{1-q/2} \leq C^q < (m_C + 1)^{1-q/2}. \tag{9}$$

Then m_C is the largest dimension of a unit sphere, centered at 0, that fits inside the parameter space $\Theta_q(C)$.

3.2 Parameter space

The parameter space for $\theta := [\theta_1 : \dots : \theta_M]$ is denoted by

$$\Theta_q^M(C_1, \dots, C_M) = \{\theta \in \prod_{\nu=1}^M \Theta_q(C_\nu) \; : \; \langle \theta_\nu, \theta_{\nu'} \rangle = 0, \;\; \text{for} \;\; \nu \neq \nu'\}, \tag{10}$$

where $\Theta_q(C)$ is defined through (8), and $C_\nu \geq 1$ for all $\nu = 1, \dots, M$.

Remark : If $M > 1$, one can describe the sparsity of the eigenvectors in a different way. Consider the sequence $\zeta := \zeta^N = (\sqrt{\sum_{\nu=1}^M \lambda_\nu \theta_{\nu k}^2} : k = 1, 2, \dots, N)$. One may demand that the vector ζ be sparse in an l^q or weak-l^q sense. This particular approach to sparsity has some natural interpretability, since the quantity $\zeta_k^2 = \sum_{\nu=1}^M \lambda_\nu \theta_{\nu k}^2$, where ζ_k is the k-th coordinate of ζ, is the variance of the k-th coordinate of the "signal" part of the vector X. There is a connection between this model and the model we intend to study. If (10) holds, then $\zeta \in l_N^q(\overline{C}_\lambda)$, where $\overline{C}_\lambda^q = \sum_{\nu=1}^M \lambda_\nu^{q/2} C_\nu^q$. On the other hand, l^q (weak-l^q) sparsity of ζ implies l^q (weak-l^q) sparsity of θ_ν for all $\nu = 1, \dots, M$.

3.3 Lower bound on the minimax risk

In this section a lower bound on the minimax risk of estimating θ_ν over the parameter space (10) is derived when $0 < q < 2$, under the loss function defined through (3). The result is stated under some simplifying assumptions that make the asymptotic analysis more transparent. Define

$$g(\lambda, \tau) = \frac{(\lambda - \tau)^2}{(1 + \lambda)(1 + \tau)}, \qquad \lambda, \tau > 0. \tag{11}$$

A1 There exists a constant $C_0 > 0$ such that $C_0^q < C_\mu^q - 1$ for all $\mu = 1, \dots, M$, for all N.

A2 As $n, N \to \infty$, $nh(\lambda_\nu) \to \infty$.

A3 As $n, N \to \infty$, $nh(\lambda_\nu) = O(1)$.

A4 As $n, N \to \infty$, $ng(\lambda_\mu, \lambda_\nu) \to \infty$ for all $\mu = 1, \dots, \nu - 1, \nu + 1, \dots, M$.

A5 As $n, N \to \infty$, $n \max_{1 \leq \mu \neq \nu \leq M} g(\lambda_\mu, \lambda_\nu) = O(1)$.

Conditions **A4** and **A5** are applicable only when $M > 1$. In the statement of the following theorem, the infimum is taken over all estimators $\widehat{\theta}_\nu$, estimating θ_ν, satisfying $\| \widehat{\theta}_\nu \| = 1$.

Theorem 2: *Let $0 < q < 2$ and $1 \leq \nu \leq M$. Suppose that **A1** holds.*

(a) *If **A3** holds, then there exists $B_1 > 0$ such that*

$$\liminf_{n \to \infty} \inf_{\widehat{\theta}_\nu} \sup_{\theta \in \Theta_q^M(C_1, \dots, C_M)} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu) \geq B_1. \tag{12}$$

(b) If **A2** holds, then there exists $B_2 > 0$, $A_q > 0$, and $c_1 \in (0,1)$, such that

$$\liminf_{n\to\infty} \delta_n^{-1} \inf_{\widehat{\theta}_\nu} \sup_{\theta \in \Theta_q^M(C_1,\dots,C_M)} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu) \geq B_2, \tag{13}$$

where δ_n is defined by

$$\delta_n = \begin{cases} c_1 & if \quad nh(\lambda_\nu) \leq \min\{c_1(N-M), A_q \overline{C}_\nu^q (nh(\lambda_\nu))^{q/2}\} \\ c_1 \frac{N-M}{nh(\lambda_\nu)} & if \quad c_1(N-M) \leq \min\{nh(\lambda_\nu), A_q \overline{C}_\nu^q (nh(\lambda_\nu))^{q/2}\} \\ A_q \frac{\overline{C}_\nu^q}{(nh(\lambda_\nu))^{1-q/2}} & if \quad A_q \overline{C}_\nu^q (nh(\lambda_\nu))^{q/2} \leq \min\{nh(\lambda_\nu), c_1(N-M)\} \end{cases} \tag{14}$$

and

$$\delta_n = (c_2(\alpha))^{1-q/2} \frac{\overline{C}_\nu^q (\log N)^{1-q/2}}{(nh(\lambda_\nu))^{1-q/2}}, \quad if \quad A_{q,\alpha} \overline{C}_\nu^q \left(\frac{nh(\lambda_\nu)}{\log N}\right)^{q/2} \leq \min\left\{\frac{nh(\lambda_\nu)}{\log N}, KN^{1-\alpha}\right\}, \tag{15}$$

for some $K > 0$, $\alpha \in (0,1)$, $c_q(\alpha) \in (0,1)$ and $A_{q,\alpha} > 0$. Here $\overline{C}_\nu^q := C_\nu^q - 1$. Also, one can take $c_1 = \log(9/8)$, $A_q = \left(\frac{9c_1}{2}\right)^{1-q/2}$, $A_{q,\alpha} = (\alpha/2)^{1-q/2}$, $c_q(\alpha) = (\alpha/9)^{1-q/2}$, $B_2 = \frac{1}{8}$ and $B_3 = (8e)^{-1}$.

(c) Suppose that $M > 1$. If **A4** holds, then there exists $B_3 > 0$ such that

$$\liminf_{n\to\infty} \overline{\delta}_n^{-1} \inf_{\widehat{\theta}_\nu} \sup_{\theta \in \Theta_q^M(C_1,\dots,C_M)} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu) \geq B_3, \tag{16}$$

where

$$\overline{\delta}_n = \frac{1}{n} \max_{\mu \in \{1,\dots,M\}\setminus\{\nu\}} \frac{1}{g(\lambda_\mu, \lambda_\nu)} . \tag{17}$$

One can take $B_3 = \frac{1}{8e}$. However, if **A5** holds, then (12) is true.

Remark : In the statement of Theorem 2, there is much flexibility in terms of what values the "hyperparameters" $C_1, \dots, C_M$ and the eigenvalues $\lambda_1, \dots, \lambda_M$ can take. In particular, they can vary with N, subject to the modest requirement that **A1** is satisfied. However, the constants appearing in equations (13) and (16) are not optimal.

Remark : Another notable aspect is that, as the proof later shows, the rate lower bounds in Part (b) are all of the form $\frac{m}{nh(\lambda_\nu)}$, where m is the "effective" number of "significant" coordinates. This phrase becomes clear if one notices further that, in the construction that leads to the lower bound (see Section 6.7), the vector θ_ν in a near-worst case scenario has overwhelming number of coordinates of size $const. \frac{1}{\sqrt{nh(\lambda_\nu)}}$, or, in the case (15), of size $const. \frac{\sqrt{\log N}}{\sqrt{nh(\lambda_\nu)}}$. Here m is of the same order as the number of these "significant" coordinates. This suggests that, an estimation strategy that is able to extract coordinates of θ_ν of the stated size, would have the right rate of convergence, subject to possibly some regularity conditions. The estimator described later (ASPCA) is constructed by following this principle.

Part (a) and the second statement of Part (c) of Theorem 2 depict situations under which there is no estimator that is asymptotically uniformly consistent over $\Theta_q^M(C_1, \ldots, C_M)$. Moreover, the first part of Part (b), and Theorem 1 readily yield the following corollary.

Corollary 1: *If the conditions of Theorem 1 hold, and if* **A1** *holds, together with the condition that*

$$\liminf_{n \to \infty} \frac{\overline{C}_\nu^q (nh(\lambda_\nu))^{q/2}}{N} > c_1^{q/2} A_q^{-1},$$

then the usual PCA-based estimator of $\widehat{\theta}_\nu$, i.e. the eigenvector corresponding to the ν-th largest eigenvalue of $\mathbf{S}$, has asymptotically the best rate of convergence.

Remark : A closer look at the proof of Theorem 1 reveals that the method of proof explicitly made use of condition **L1** to ensure that the contribution of $\lambda_1, \ldots, \lambda_M$ to the residual term of the second order expansion of $\widehat{\theta}_\nu$ is bounded. However, the condition $n \max_{\mu \neq \nu} g(\lambda_\mu, \lambda_\nu) \to \infty$ is certainly much weaker than that. The method of proof pursued here fails to settle the question as to whether this is sufficient to get the asymptotic rate (5). It is conjectured that this is the case.

4   Estimation scheme

This section outlines an estimation strategy for the eigenvectors θ_ν, $\nu = 1, \ldots, M$. Model (2) is assumed throughtout for observations X_i, $i = 1, \ldots, n$. We propose estimators is for the case when the noise variance σ^2 is known. Therefore, without loss of generality, it can be taken to be 1. Henceforth, for simplicity of notations, it is also assumed that $\xi = 0$. In practice, one may have to estimate σ^2 from data. The median of the diagonal entries of the sample covariance matrix $\mathbf{S} := \frac{1}{n} \mathbf{X} \mathbf{X}^T$ serves as a reasonable (although slightly biased) estimator of σ^2, if the true model is sparse. In the latter case, the data are rescaled by multiplying each observation by $\widehat{\sigma}^{-1}$, and the resultant covariance matrix is called, with a slight abuse of notation, $\mathbf{S}$. Note that, in this case, the estimates of eigenvalues of Σ are $\widehat{\sigma}^2$ times the corresponding eigenvalues of $\mathbf{S}$.

4.1   Sparse Principal Components Analysis (SPCA)

In order to motivate the approach that is described in what follows, consider first the SPCA estimation scheme studied by Johnstone and Lu (2004). To that end, let $\mathbf{S} = \frac{1}{n} \mathbf{X} \mathbf{X}^T$ denote the sample covariance matrix. Suppose that the sample variances of coordinates (i.e., diagonal terms of $\mathbf{S}$) are denoted by $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2$.

- Define $\widehat{I}_n$ to be the set of indices $k \in \{1, \ldots, N\}$ such that $\hat{\sigma}_k^2 > \gamma_n$ for some threshold $\gamma_n > 0$.

- Let $\mathbf{S}_{\widehat{I}_n, \widehat{I}_n}$ be the submatrix of $\mathbf{S}$ corresponding to the coordinates $\widehat{I}_n$. Perform an eigenanalysis of $\mathbf{S}_{\widehat{I}_n, \widehat{I}_n}$. Denote the eigenvectors by $\mathbf{e}_1, \ldots, \mathbf{e}_{\min\{n, |\widehat{I}_n|\}}$.

- For $\nu = 1, \ldots, M$, estimate θ_ν by $\widetilde{\mathbf{e}}_\nu$ where $\widetilde{\mathbf{e}}_\nu$, an $N \times 1$ vector, is obtained from $\mathbf{e}_\nu$ by augmenting zeros to all the coordinates that are in $\{1, \ldots, N\} \setminus \widehat{I}_n$.

Johnstone and Lu (2004) showed that, if one chooses an appropriate threshold γ_n, then the estimate of θ_ν is consistent under the weak-l^q sparsity constraint on θ_ν. However, Paul and Johnstone (2004) showed that even with the best choice of γ_n, the rate of convergence of the risk of this estimate is not optimal. Indeed, Paul and Johnstone (2004) demonstrate an estimator which has a better rate of convergence in the single component ($M = 1$) situation.

4.2 Augmented Sparse PCA (ASPCA)

We now propose the ASPCA estimation scheme. This scheme is a refinement of the SPCA scheme of Johnstone and Lu (2004), and can be viewed as a generalization of the estimation scheme proposed by Paul and Johnstone (2004) in the single component ($M = 1$) case.

The key idea behind this estimation scheme is that, in addition to using the coordinates having large variance, if one also uses the covariance structure appropriately, then under the assumption of a sparse structure of the eigenvectors, one will be able to extract a lot more information and thereby get more accurate estimate of the eigenvalues and eigenvectors. Notice that SPCA only focuses on the diagonal of the covariance matrix and therefore ignores the covariance structure. This renders this scheme suboptimal from an asymptotic minimax risk analysis point of view. To make this point clearer, it is instructive to analyze the covariance matrix in the $M = 1$ case. In view of the second Remark after the statement of Theorem 2 one expects to be able to recover coordinates k for which $|\theta_{1k}| \gg \frac{1}{\sqrt{nh(\lambda_1)}}$. However, the best choice for γ_n for SPCA is $\gamma\sqrt{\frac{\log n}{n}}$, for some constant $\gamma > 0$, which is way too large. On the other hand, suppose that one divides the coordinates into two sets A and B, where the former contains all those k such that $|\theta_k|$ is "large", and the latter contains smaller coordinates. Partition the matrix Σ as

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

Here $\Sigma_{BA} = \lambda_1 \theta_{1,B} \theta_{1,A}^T$. Assume that, there is a "preliminary" estimator of θ_1, say $\widetilde{\theta}_1$ such that, $\langle \widetilde{\theta}_{1,A}, \theta_{1,A} \rangle \to 1$ in probability as $n \to \infty$. Then one can use this estimator as a "filter", in a way described below, to recover the "informative ones" among the smaller coordinates. This can be seen from the following relationship

$$\Sigma_{BA} \widetilde{\theta}_{1,A} = \langle \widetilde{\theta}_{1,A}, \theta_{1,A} \rangle \lambda_1 \theta_{1,B} \approx \lambda_1 \theta_{1,B}.$$

In this manner one can extract some information about those coordinates of θ_1 that are in set B. The algorithm described below is a generalization of this idea. It has three stages. First two stages will be referred to as "coordinate selection" stages. The final stage consists of an eigen-analysis of the submatrix of $\mathbf{S}$ corresponding to the selected coordinates, followed by a hard thresholding of the estimated eigenvectors.

Let $\gamma_i > 0$ for $i = 1, 2, 3$ and $\kappa > 0$ be four constants to be specified later. Define $\gamma_{1,n} = \gamma_1 \sqrt{\frac{\log(n \vee N)}{n}}$.

1° Select coordinates k such that $\widehat{\sigma}_{kk} := \mathbf{S}_{kk} > 1 + \gamma_{1,n}$. Denote the set of selected coordinates by $\widehat{I}_{1,n}$.

2° Perform spectral decomposition of $\mathbf{S}_{\widehat{I}_{1,n},\widehat{I}_{1,n}}$. Denote the eigenvalues by $\widehat{\ell}_1 > \ldots > \widehat{\ell}_{m_1}$ where $m_1 = \min\{n, |\widehat{I}_{1,n}|\}$, and corresponding eigenvectors by $\mathbf{e}_1, \ldots, \mathbf{e}_{m_1}$.

3° Estimate M by $\widehat{M}$ defined in Section 4.3. Estimate λ_j by $\widetilde{\lambda}_j = \widehat{\ell}_j - 1$, $j = 1, \ldots, \widehat{M}$.

4° Define $E = [\frac{1}{\sqrt{\widehat{\ell}_1}}\mathbf{e}_1 : \ldots : \frac{1}{\sqrt{\widehat{\ell}_{\widehat{M}}}}\mathbf{e}_{\widehat{M}}]$. Compute $\mathbf{Q} = \mathbf{S}_{\widehat{I}_{1,n}^c, \widehat{I}_{1,n}} E$.

5° Denote the diagonal of the matrix $\mathbf{Q}\mathbf{Q}^T$ by T. Define $\widehat{I}_{2,n}$ to be the set of coordinates $k \in \{1, \ldots, N\} \setminus \widehat{I}_{1,n}$ such that $|T_k| > \gamma_{2,n}^2$ where

$$\gamma_{2,n} = \gamma_2 \left(\sqrt{\frac{\log(n \vee N)}{n}} + \frac{1}{\kappa}\sqrt{\frac{\widehat{M}}{n}} \right).$$

6° Take the union $\widehat{I}_n := \widehat{I}_{1,n} \bigcup \widehat{I}_{2,n}$. Perform spectral decomposition of $\mathbf{S}_{\widehat{I}_n, \widehat{I}_n}$. Estimate θ_ν by augmenting the ν-th eigenvector, with zeros in the coordinates $\{1, \ldots, N\} \setminus \widehat{I}_n$, for $\nu = 1, \ldots, \widehat{M}$. Call this vector $\widehat{\theta}_\nu$.

7° Perform a coordinatewise "hard" thresholding of $\widehat{\theta}_\nu$ at threshold

$$\gamma_{3,n} := \gamma_3 \sqrt{\frac{\log(n \vee N)}{nh(\widetilde{\lambda}_\nu)}} ,$$

and then normalize the thresholded vectors to get the final estimate $\overline{\theta}_\nu$.

Remark : The scheme is specified except for the "tuning parameters" $\gamma_1, \gamma_2, \gamma_3$ and κ. The choice of γ_i's is discussed in the context of deriving upper bounds on the risk of the estimator. It will be shown that, it suffices to take $\gamma_1 = 4$, $\kappa = 2 + \epsilon$ for a small $\epsilon > 0$, and $\gamma_2 = \sqrt{\frac{3}{2}}\kappa$. An analysis of the thresholding scheme is not done here, but in practice $\gamma_3 = 3$ works well enough, and some calculations suggest that $\gamma_3 = 2$ suffices asymptotically.

4.3 Estimation of M

Let $\overline{\gamma}_1, \gamma_1' > 0$ be such that $\overline{\gamma}_1 > \gamma_1'$. Define

$$\widehat{\overline{I}}_{1,n} = \{k : \mathbf{S}_{kk} > 1 + \overline{\gamma}_{1,n}\} \text{ where } \overline{\gamma}_{1,n} = \overline{\gamma}_1 \sqrt{\frac{\log(n \vee N)}{n}} , \tag{18}$$

$$\widehat{I}'_{1,n} = \{k : \mathbf{S}_{kk} > 1 + \gamma'_{1,n}\} \text{ where } \gamma'_{1,n} = \gamma'_1 \sqrt{\frac{\log(n \vee N)}{n}} . \tag{19}$$

Define

$$\alpha_n = 2\sqrt{\frac{|\widehat{\overline{I}}_{1,n}|}{n}} + \frac{|\widehat{I}'_{1,n}|}{n} + 6\left(\frac{|\widehat{I}'_{1,n}|}{n} \vee 1\right) \sqrt{\frac{\log(n \vee |\widehat{I}'_{1,n}|)}{n \vee |\widehat{I}'_{1,n}|}} . \tag{20}$$

12

Let $\widehat{\widehat{\ell}}_1 > \ldots > \widehat{\widehat{\ell}}_{\overline{m}_1}$, where $\overline{m}_1 = \min\{n, |\widehat{\overline{I}}_{1,n}|\}$, be the nonzero eigenvalues of $\mathbf{S}_{\widehat{\overline{I}}_{1,n}, \widehat{\overline{I}}_{1,n}}$. Define $\widehat{M}$ by

$$\widehat{M} = \max\{1 \le k \le \overline{m}_1 : \widehat{\widehat{\ell}}_k > 1 + \alpha_n\}. \tag{21}$$

The choice of γ'_1 and $\overline{\gamma}_1$ is discussed in Section 8.5.

Remark : Sparsity of the eigenvectors is an implicit assumption for ASPCA scheme. However, in practice, and specifically with only moderately large samples, it is not always the case that ASPCA is able to select the significant coordinates. More importantly, the scheme produces a *bona fide* estimator only when $\widehat{I}_{1,n}$ is non-empty. If this is not the case, then one may use the ν-th eigenvector of $\mathbf{S}$ as the estimator of θ_ν. However, determination of M in this situation is a difficult issue, and without recourse to additional information, one may set $\widehat{M} = 0$.

5 Rates of convergence

In this section we describe the asymptotic risk of ASPCA estimators under some regularity conditions. The risk is analyzed under the loss function (3), and it is assumed that condition **BA** of Section 2 holds. Further, the parameter space for $\theta = [\theta_1 : \ldots : \theta_M]$, over which the risk is maximized, is taken to be $\Theta_q^M(C_1, \ldots, C_M)$ defined through (10) in Section 3.2, where $0 < q < 2$ and $C_1, \ldots, C_M > 1$.

5.1 Sufficient conditions for convergence

The following conditions are imposed on the "hyperparameters" of the parameter space $\Theta_q(C_1, \ldots, C_M)$. Suppose that $\rho_1, \ldots, \rho_M$ are as in **C1** given below. Define

$$\rho_q(C) := \sum_{\nu=1}^{M} \rho_\nu^{q/2} C_\nu^q. \tag{22}$$

Observe that, since $C_\nu \ge 1$ for all $\nu = 1, \ldots, M$, $\rho_q(C) \ge \sum_{\nu=1}^{M} \rho_\nu^{q/2} \ge 1$.

C1 $\lambda_1, \ldots, \lambda_M$ are such that, as $n \to \infty$, $\frac{\lambda_\nu}{\lambda_1} \to \rho_\nu$ where $1 \equiv \rho_1 > \rho_2 > \ldots > \rho_M$.

C2 $\log N \asymp \log n$ and $\frac{(\log n)^2}{n \lambda_1^2} \to 0$ as $n \to \infty$.

C3 $\frac{\rho_q(C)(\log N)^{1/2-q/4}}{\lambda_1^{1-q/2} n^{1/2-q/4}} \to 0$ as $n \to \infty$.

We discuss briefly the importance of these conditions. **C1** is a repetition of **L1**. **C2** is a convenient and very mild technical assumption that should hold in most practical situations. Second part of **C2** is non-trivial only when $\lambda_1 \to 0$ as $n \to \infty$. **C3** requires some explanation. It will become increasingly clear that, in order to get a uniformly consistent estimate of the eigenvectors from the preliminary SPCA step, one needs **C3** to hold. Indeed, the sequence described in **C3** has the same asymptotic order as a common upper bound for the rate of

13

convergence of the supremum risk of the SPCA estimators of all the θ_ν's. So, the implication is that if **C3** holds then the SPCA scheme of Johnstone and Lu (2004) gives consistent estimates.

Remark : Note that, $\frac{1}{nh(\lambda)} \leq \frac{1+c}{n\lambda^2}$ if $\lambda \in (0, c)$ and $\frac{1}{nh(\lambda)} \leq \frac{1}{\eta(c)n\lambda}$ if $\lambda \geq c$, for any $c > 0$. Since $\rho_q(C) \geq 1$, **C3** guarantees that

$$\frac{\rho_q(C)(\log N)^{1-q/2}}{(nh(\lambda_1))^{1-q/2}} = o(1), \qquad \text{as } n \to \infty. \tag{23}$$

In fact, if $\liminf_{n\to\infty} \lambda_1 \geq c > 0$, then the upper bound in (23) can be replaced by $o((\frac{\log N}{n})^{1/2-q/4})$. It will be shown that this is a common (and near-optimal) upper bound on the rate of convergence of the ASPCA estimate of θ_ν's. If one compares this with the lower bound given by Theorem 2, it is conjectured that (23) should also be a sufficient condition for establishing that the lower bound defined through (15) is also the upper bound on the minimax risk, at the level of rates. However, since our method depends on finding a preliminary consistent estimator of the eigenvectors (in our case SPCA), the somewhat stronger condition **C3** becomes necessary to establish rates of convergence of the ASPCA estimator.

5.2 Statement of the result

Now we state the main result of this section. The asymptotic analysis of risk is conducted only for the estimator $\widehat{\theta}_\nu$ for eigenvector θ_ν, and not for the thresholding estimator $\widetilde{\theta}_\nu$. Derivation of the results for $\widetilde{\theta}_\nu$ requires additional technical work, but can be carried out. It can be shown that in certain circumstances the latter has a slightly better asymptotic risk property. In practice, the thresholding estimator seems to work better when the eigenvalues are well-separated. The following theorem describes the asymptotic behavior of the risk of the ASPCA estimator $\widehat{\theta}_\nu$ under the loss function L defined through (3). $g(\cdot, \cdot)$ is defined by (11).

Theorem 3: *Assume that* **BA** *and conditions* **C1**-**C3** *hold. Then, there are constants* $K := K(q, \gamma_1, \gamma_2, \kappa)$ *and* $K' := K'(q, M, \gamma_1, \gamma_2, \kappa)$ *such that, as* $n \to \infty$, *for all* $\nu = 1, \ldots, M$,

$$\sup_{\theta \in \Theta_q^M(C_1, \ldots, C_M)} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu)$$

$$\leq \left[K\left(C_\nu^q + K'\rho_\nu^{-q} \frac{\rho_q(C)}{\log(n \vee N)}\right) \left(\frac{\log(n \vee N)}{nh(\lambda_\nu)}\right)^{1-q/2} + \sum_{\mu \neq \nu}^M \frac{1}{ng(\lambda_\mu, \lambda_\nu)} \right] (1 + o(1)) \quad (24)$$

Remark : The expression in the upper bound is somewhat cumbersome, but the significance of each of the terms in (24) will become clear in the course of the proof. However, notice that, if the parameters $C_1, \ldots, C_M$ of the space $\Theta_q(M)(C_1, \ldots, C_M)$ are such that,

$$\exists \; 0 < \underline{C} < \overline{C} < \infty, \quad \text{such that} \quad \underline{C} \leq \frac{\max_{1 \leq \mu \leq M} C_\mu}{\min_{1 \leq \mu \leq M} C_\mu} \leq \overline{C}, \quad \text{for all} \; n, \tag{25}$$

then, Theorem 3 and Theorem 2 together imply that, under conditions **BA**, **C1**-**C3**, **A1** and the condition on the hyperparameters given by (15), the ASPCA estimator $\widehat{\theta}_\nu$ has the optimal

rate of convergence. The condition (25) is satisfied in particular if $C_1, \ldots, C_M$ are all bounded above.

It is important to emphasize that (24) is an asymptotic result in the following sense. It is possible to give finite a sample bound on $\sup_{\theta \in \Theta_q^M(C_1, \ldots, C_M)} \mathbb{E}L(\widehat{\theta}_\nu, \theta_\nu)$. However, this upper bound involves many additional terms whose total contribution is smaller than a prescribed $\epsilon > 0$ only when $n \geq n_\epsilon$, say, where n_ϵ depends on the hyperparameters, apart from ϵ.

Remark : It is instructive to compare the asymptotic supremum risk of ASPCA with that of OPCA (or usual PCA based) estimator of θ_ν. A closer inspection of the proof reveals that, if for all sufficiently large n,

$$N \leq K''(\frac{\rho_q(C)}{\log(n \vee N)}) \left(\frac{\log(n \vee N)}{nh(\lambda_\nu)} \right)^{-q/2},$$

then for some constant $K'' > 0$, under **BA** and **C1-C3**, one can replace the upper bound in (24) by

$$\overline{K} \left[\frac{N \log(n \vee N)}{nh(\lambda_\nu)} + \sum_{\mu \neq \nu} \frac{1}{ng(\lambda_\mu, \lambda_\nu)} \right] (1 + o(1)),$$

for some constant $\overline{K}$. This rate is greater than that of OPCA estimator by a factor of at most $\log(n \vee N)$. However, observe that, the bound on the risk of OPCA estimator holds under weaker conditions. In particular, Theorem 1 does not assume any particular structure for the eigenvectors.

6 Proof of Theorem 2

The proof requires a closer look at the geometry of the parameter space, in order to obtain good finite dimensional subproblems that can then be used as inputs to the general machinery, to come up with the final expressions.

6.1 Risk bounding strategy

A key tool for our proof the lower bound on the minimax risk is Fano's lemma. Thus, it is necessary to derive a general expression for the Kullback-Leibler discrepancy between the probability distributions described by two separate parameter values.

Proposition 1: *Let* $\theta^{(j)} = [\theta_1^{(j)} : \ldots : \theta_M^{(j)}]$, $j = 1, 2$ *be two parameters. Let* $\Sigma_{(j)}$ *denote the matrix given by (1) with* $\theta = \theta^{(j)}$ *(and* $\sigma = 1$*). Let* P_j *denote the joint probability distribution of* n *i.i.d. observations from* $N(0, \Sigma_{(j)})$*. Then the Kullback-Leibler discrepancy of* P_2 *from* P_1, *to be denoted by* $K_{1,2} := K(\theta^{(1)}, \theta^{(2)})$, *is given by*

$$K_{1,2} \stackrel{def}{=} K(\theta^{(1)}, \theta^{(2)}) = n \left[\frac{1}{2} \sum_{\nu=1}^{M} \eta(\lambda_\nu)\lambda_\nu - \frac{1}{2} \sum_{\nu=1}^{M} \sum_{\nu'=1}^{M} \eta(\lambda_\nu)\lambda_{\nu'} |\langle \theta_{\nu'}^{(1)}, \theta_\nu^{(2)} \rangle|^2 \right], \quad (26)$$

where

$$\eta(\lambda) = \frac{\lambda}{1+\lambda}, \qquad \lambda > 0. \tag{27}$$

6.2 Use of Fano's lemma

We outline the general approach pursued in the rest of this section. The idea is to bound the supremum of the risk on the entire parameter space by the maximum risk over a finite subset of it, and then to use some variant of Fano's lemma to provide a lower bound for the latter quantity.

Thus, the goal is to find an appropriate *finite* subset $\mathcal{F}_0$ of $\Theta_q^M(C_1, \ldots, C_M)$, such that the following properties hold.

(1) If $\theta^{(1)}, \theta^{(2)} \in \mathcal{F}_0$, then $L(\theta_\nu^{(1)}, \theta_\nu^{(2)}) \geq 4\delta$, for some $\delta > 0$ (to be chosen). This property will be referred to as "4δ-distinguishability in θ_ν".

(2) The element $\theta \in \mathcal{F}_0$ is a unique representative of the equivalence class $[\theta]$, where $[\theta]$ is defined to be the class of $N \times M$ matrices whose ν-th column is either θ_ν or $-\theta_\nu$.

(3) Subject to (1), the quantity $\sup_{i \neq j: \; \theta^{(i)}, \theta^{(j)} \in \mathcal{F}_0} K(\theta^{(i)}, \theta^{(j)}) + K(\theta^{(j)}, \theta^{(i)})$ is as small as possible.

Given any estimator $\widehat{\theta}$ of θ, based on data $\mathbf{X}_n = (X_1, \ldots, X_n)$, define a new estimator $\phi(\mathbf{X}_n)$ (an $N \times M$ matrix) as $\phi(\mathbf{X}_n) = \theta^*$ if $\theta^* = \arg\min_{\theta \in \mathcal{F}_0} L(\theta_\nu, \widehat{\theta}_\nu)$, where $\widehat{\theta}_\nu$ is the ν-th column of $\widehat{\theta}$ (i.e., estimate of θ_ν). Then, by Chebyshev's inequality,

$$\begin{aligned}
\sup_{\theta \in \Theta_q^M(C_1, \ldots, C_M)} \mathbb{E}_\theta L(\theta_\nu, \widehat{\theta}_\nu) &\geq \delta \sup_{\theta \in \Theta_q^M(C_1, \ldots, C_M)} \mathbb{P}_\theta(L(\theta_\nu, \widehat{\theta}_\nu) \geq \delta) \\
&\geq \delta \sup_{\theta \in \mathcal{F}_0} \mathbb{P}_\theta(L(\theta_\nu, \widehat{\theta}_\nu) \geq \delta) \\
&\geq \delta \sup_{\theta \in \mathcal{F}_0} \mathbb{P}_\theta([\phi(\mathbf{X}_n)] \neq [\theta]). \tag{28}
\end{aligned}$$

The last inequality is because, if $L(\theta_\nu^{(j)}, \widehat{\theta}_\nu) < \delta$ for any $\theta^{(j)} \in \mathcal{F}_0$, then by the "$4\delta$-distinguishability in θ_ν" (property (1) above), it follows that $[\phi_\nu(\mathbf{X}_n)] = [\theta_\nu^{(j)}]$, and hence $[\phi(\mathbf{X}_n)] = [\theta^{(j)}]$.

Two versions of Fano's lemma are found to be useful in this context. The following version, due to Birgé (2001), of a result of Yang and Barron (1999) (p.1570-71), is most suitable when $\mathcal{F}_0$ can be chosen to be large.

Lemma 1: *Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability distributions on a common measurable space, where Θ is an arbitrary parameter space. Suppose that a loss function for the estimation problem is given by $L'(\theta, \theta') = \mathbf{1}_{\theta \neq \theta'}$. Define the minimax risk over Θ by*

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

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 $J = |\mathcal{F}|$,

$$p_{max} \geq 1 - \inf_Q \ \frac{J^{-1} \sum_{i=1}^{J} K(P_i, Q) + \log 2}{\log J} \tag{29}$$

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.

To use Lemma 1 choose P_i to be $P_{\Sigma_{(i)}} \equiv P_{\theta^{(i)}} := N^{\otimes n}(0, \Sigma_{(i)})$, where $\Sigma_{(i)}$ is the matrix $\sum_{\nu=1}^{M} \lambda_\nu \theta_\nu^{(i)} \theta_\nu^{(i)T} + I$, and $\theta^{(i)} \in \mathcal{F}_0$ $i = 1, \ldots, |\mathcal{F}_0|$, are the distinct values of parameter θ that constitute the set $\mathcal{F}_0$. Then set $Q_0 = P_{\theta^{(0)}}$, for some appropriately chosen $\theta^{(0)} \in \Theta_q^M(C_1, \ldots, C_M)$ such that the following condition is satisfied.

$$ave_{1 \leq i \leq |\mathcal{F}_0|} K(\theta^{(i)}, \theta^{(0)}) \asymp \sup_{1 \leq i \leq |\mathcal{F}_0|} K(\theta^{(i)}, \theta^{(0)}), \tag{30}$$

where the notation "$\asymp$" means that the both sides are are within constant multiples of each other. Then it follows from (28) and Lemma 1 that,

$$\delta^{-1} \sup_{\theta \in \Theta_q^M(C_1, \ldots, C_M)} \mathbb{E}_\theta L(\theta_\nu, \widehat{\theta}_\nu) \geq 1 - \frac{ave_{1 \leq i \leq |\mathcal{F}_0|} K(\theta^{(i)}, \theta^{(0)}) + \log 2}{\log |\mathcal{F}_0|} \ . \tag{31}$$

To complete the picture it is desirable that

$$\frac{ave_{1 \leq i \leq |\mathcal{F}_0|} K(\theta^{(i)}, \theta^{(0)}) + \log 2}{\log |\mathcal{F}_0|} \ \approx \ c, \tag{32}$$

where c is a number between 0 and 1.

A different version of Fano's lemma, due to Birgé (2001), is needed when $\mathcal{F}_0$ consists of only two elements $\theta^{(1)}$ and $\theta^{(2)}$, so that the classification problem reduces to a test of hypothesis of P_1 against P_2.

Lemma 2: *Let α_T and β_T denote respectively the Type I and Type II errors associated with an arbitrary test T between the two simple hypotheses P_1 and P_2. Define, $\pi_{mis} = \inf_T(\alpha_T + \beta_T)$, where the infimum is taken over all test procedures.*

$$K(P_1, P_2) \geq -\log[\pi_{mis}(2 - \pi_{mis})]. \tag{33}$$

6.3 Geometry of the parameter space

We view the space $\Theta_q(C)$, for $0 < q < 2$, as the N-dimensional unit sphere centered at the origin, from which some parts have been chopped off, symmetrically in each coordinate, such that there is some portion left at each pole (i.e., a point of the form $(0, \ldots, 0, \pm 1, 0, \ldots, 0)$,

where the non-zero term appears only once). In this connection, we define an object that is central to the proof of Theorem 3.3.

Definition : *Let $0 < r < 1$ and $N > m \geq 1$. An (N, m, r) **polar sphere** at pole k_0, on set $J = \{j_1, \ldots, j_m\}$, where $1 \leq k_0 \leq N$ and $j_l \in \{1, \ldots, N\} \setminus \{k_0\}$ for $l = 1, \ldots, m$, is a subset of $\mathbb{S}^{N-1}$ given by*

$$\mathcal{S}(N, m, r, k_0, J) := \{\mathbf{x} \in \mathbb{S}^{N-1} : x_{k_0} = \sqrt{1 - r^2}, \sum_{l=1}^{m} x_{j_l}^2 = r^2\}. \tag{34}$$

So, an (N, m, r) polar sphere is centered at the point $(0, \ldots, 0, \sqrt{1 - r^2}, 0, \ldots, 0)$, (which is not in $\mathbb{S}^{N-1}$), has radius r, and has dimension m. Note that, the largest sphere of any given dimension m, such that $C^q < m^{1-q/2}$ (equivalently, $m > m_C$, where m_C is defined through (9)), that can be inscribed inside $\Theta_q(C)$ is an (N, m, r) polar sphere. The radius r of such a polar sphere, given $C^q < m^{1-q/2}$ (or $m > m_C$), to be denoted by $r_m(C)$, satisfies

$$\{1 - (r_m(C))^2\}^{q/2} + m^{1-q/2}\{r_m(C)\}^q = C^q. \tag{35}$$

Of course, if $C^q \geq m^{1-q/2}$ (or $m_C \geq m$) then as a convention, $r_m(C) = 1$. Condition (35) ensures that all the points lying on an (N, m, r) polar sphere such that $r \in (0, r_m(C))$, are inside $\Theta_q(C)$.

6.4   A common recipe for Part (a) and Part (b)

In the proof of Part (a) and Part (b) of the theorem, there is a common theme in the construction of $\mathcal{F}_0$. Let $\mathbf{e}_\mu$ denote the N-vector whose μ-th coordinate is 1 and rest are all zero. In either case, if $\{\theta^{(j)}, j = 1, \ldots, |\mathcal{F}_0|\}$ is an enumeration of the elements of $\mathcal{F}_0$, then the following are true.

(F1) There is an $N \times M$ matrix $\theta^{(0)}$, such that $\theta_\nu^{(0)} = \mathbf{e}_\nu$.

(F2) $\theta_\mu^{(j)} = \mathbf{e}_\mu$ for $\mu = 1, \ldots, \nu - 1, \nu + 1, \ldots, M$, for all $j = 0, 1, \ldots, |\mathcal{F}_0|$.

(F3) $\theta_\nu^{(j)} \in \mathcal{S}(N, m, r, \nu, J)$ for some m, r and J. m and r are fixed for all $1 \leq j \leq |\mathcal{F}_0|$, but J may be different for different j, depending on the situation.

The $\theta^{(0)}$ in (F1) is the same $\theta^{(0)}$ appearing in (31). Also, (26) simplifies to

$$K(\theta^{(j)}, \theta^{(0)}) = \frac{1}{2}nh(\lambda_\nu)(1 - (\langle \theta_\nu^{(j)}, \theta_\nu^{(0)} \rangle)^2) = \frac{1}{2}nh(\lambda_\nu)r^2, \quad j = 1, \ldots, |\mathcal{F}_0|. \tag{36}$$

Moreover, in either case, the points $\theta^{(j)}$ are so chosen that

$$L(\theta_\nu^{(j)}, \theta_\nu^{(k)}) \geq r^2, \quad \text{for all} \quad 1 \leq j \neq k \leq |\mathcal{F}_0|. \tag{37}$$

In other words, the set $\mathcal{F}_0$ is r^2 distinguishable in θ_ν.

18

6.5 Proof of Part (a)

Construct $\mathcal{F}_0$ satisfying (F1)-(F3), with

$$\theta_\nu^{(j)} = \sqrt{1 - r^2}\mathbf{e}_\nu + r\mathbf{e}_j, \qquad j = M+1, \ldots, N,$$

where $r \in (0,1)$ is such that $(1 - r^2)^{q/2} + r^q \le C_\nu^q$. Thus, $|\mathcal{F}_0| = N - M$. Verify that (37) holds, in fact the lower bound is $2r^2$, with an equality. Therefore, (31) applies, with $\delta = \frac{r^2}{2}$. Since $nh(\lambda_\nu)$ is bounded above, and $\log(N - M) \to \infty$ as $n \to \infty$, (12) follows from (36).

6.6 Connection to "Sphere packing"

Our proof of Part (b) of Theorem 2 depends crucially on the following construction due to Zong (1999).

Let m be a large positive integer, and $m_0 = \left[\frac{2m}{9}\right]$ (the largest integer $\le \frac{2m}{9}$). 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.

$$\sqrt{m_0}z_i \in \{-1, 0, 1\} \; \forall \, i, \quad \sum_{i=1}^m |z_i| = \sqrt{m_0} \;\; \text{and, for} \;\; \mathbf{z}, \mathbf{z}' \in Y_m^*, \quad \| \mathbf{z} - \mathbf{z}' \| \ge 1. \tag{38}$$

For any $m \ge 1$, the maximal number of points lying on $\mathbb{S}^{m-1}$ such that any two points are at distance at least 1, is exactly same as the *kissing number* of an m-sphere. It is known that this number is $\le 3^m$ and $\ge (9/8)^{m(1+o(1))}$. Zong (1999) uses the construction described above to derive the lower bound, by showing that $|Y_m^*| \ge (9/8)^{m(1+o(1))}$ for m large.

6.7 Proof of Part (b)

Structures of $\mathcal{F}_0$ for the three cases in (14) are similar. Set $m \le (N - M)$, large. Set $c_1 = \log(9/8)$, $A_q = (9c_1/2)^{1-q/2}$. Choose $r \approx \sqrt{\delta_n}$, and define the set $\mathcal{F}_0$ satisfying (F1)-(F3) and the following construction.

Set $|\mathcal{F}_0| = |Y_m^*|$, where Y_m^* is the set defined in Section 6.6. Set,

$$\theta_\nu^{(j)} = \sqrt{1 - r^2}\mathbf{e}_\nu + r \sum_{l=1}^m z_l^{(j)}\mathbf{e}_{l+M}, \quad j = 1, \ldots, |\mathcal{F}_0|, \tag{39}$$

where $\mathbf{z}^{(j)} = (z_1^{(j)}, \ldots, z_m^{(j)})$, $j \ge 1$, is an enumeration of the elements of Y_m^*. Observe that, for all $j \ge 1$,

$$\theta_\nu^{(j)} \in \mathcal{S}(N, m, r, \nu, \{M+1, \ldots, M+m\}) \bigcap \mathcal{S}(N, m_0, r, \nu, supp(\mathbf{z}^{(j)})), \tag{40}$$

where $supp(\mathbf{z}^{(j)})$ is the set of nonzero coordinates of $\mathbf{z}^{(j)}$. Therefore, (37) and (36) hold for all $j \ge 1$.

6.7.1 Case : $nh(\lambda_\nu) \leq \min\{c_1(N-M), A_q\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2}\}$

Take $m = [nh(\lambda_\nu)]$ and $r^2 = c_1$. Observe that, for all $j \geq 1$,

$$\| \theta_\nu^{(j)} \|_q^q = (1 - r^2)^{q/2} + m_0^{1-q/2}r^q \leq 1 + (2/9)^{1-q/2}(nh(\lambda_\nu))^{1-q/2}c_1^{q/2} \leq 1 + c_1\overline{C}_\nu^q < C_\nu^q.$$

Thus, $\mathcal{F}_0 \subset \Theta_q^M(C_1, \ldots, C_M)$. Further, since $nh(\lambda_\nu) \to \infty$, $\log|\mathcal{F}_0| \geq c_1 nh(\lambda_\nu)(1+o(1))$. Since (37) and (36) hold, with $\delta = \frac{r^2}{4}$, from (31) the result follows, because

$$\limsup_{n\to\infty} \frac{ave_{1\leq j\leq|\mathcal{F}_0|}K(\theta^{(j)}, \theta^{(0)}) + \log 2}{\log|\mathcal{F}_0|} \leq \limsup_{n\to\infty} \frac{\frac{1}{2}c_1 nh(\lambda_\nu) + \log 2}{c_1 nh(\lambda_\nu)} = \frac{1}{2} \ .$$

6.7.2 Case : $c_1(N-M) \leq \min\{nh(\lambda_\nu), A_q\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2}\}$

Take $m = N - M$ and $r^2 = \frac{c_1(N-M)}{nh(\lambda_\nu)}$. Then, for all $j \geq 1$,

$$\| \theta_\nu^{(j)} \|_q^q \leq 1 + (2/9)^{1-q/2}(N-M)c_1^{q/2}(nh(\lambda_\nu))^{-q/2} \leq 1 + \overline{C}_\nu^q = C_\nu^q.$$

The result follows by arguments similar to those used for the case $nh(\lambda_\nu) \leq \min\{c_1(N - M), A_q\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2}\}$.

6.7.3 Case : $A_q\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2} \leq \min\{nh(\lambda_\nu), c_1(N-M)\}$

Take $m = [c_1^{-q/2}(9/2)^{1-q/2}\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2}]$ and $r^2 = c_1\frac{m}{nh(\lambda_\nu)}$. Again, verify that $m \to \infty$ as $n \to \infty$ (by **A1**), and for $j \geq 1$,

$$\| \theta_\nu^{(j)} \|_q^q \leq 1 + (2/9)^{1-q/2}m^{1-q/2}c_1^{q/2}(\frac{m}{nh(\lambda_\nu)})^{q/2} \leq 1 + \overline{C}_\nu^q = C_\nu^q,$$

and the result follows by familiar arguments.

6.7.4 Proof of (15)

The construction in all three previous cases assumes that the set of non-zero coordinates is held fixed (in our case $\{M+1, \ldots, M+m\}$) for every fixed m. However, it is possible to get a bigger set $\mathcal{F}_0$ satisfying the requirements, if this condition is relaxed.

Suppose that $A_{q,\alpha} = (\alpha/2)^{1-q/2}$, and the condition in (15) holds for some $\alpha \in (0,1)$. Set $m = [(\alpha/9)^{-q/2}(9/2)^{1-q/2}\overline{C}_\nu^q(nh(\lambda_\nu))^{q/2}(\log N)^{-q/2}]$ and $r^2 = (\alpha/9)\frac{m}{nh(\lambda_\nu)}$. Take $c_q(\alpha) = (\alpha/9)^{1-q/2}$. Observe that $m \to \infty$ as $n \to \infty$, $m = O(N^{1-\alpha})$ and $r \in (0,1)$. Set $\theta^{(0)} = [\mathbf{e}_1 : \ldots : \mathbf{e}_M]$. For every set $\pi \subset \{M+1, \ldots, N\}$ of size m, construct $\mathcal{F}_\pi$ satisfying (F1)-(F3) such that,

$$\theta_\nu^{(j)} = \sqrt{1 - r^2}\mathbf{e}_\nu + r\sum_{l\in\pi} z_l^{(j)}\mathbf{e}_l, \quad j = 1, \ldots, |Y_m^*|. \tag{41}$$

As before, $\mathcal{F}_\pi \subset \Theta_q^M(C_1,\ldots,C_M)$, for all π, so that (36) and (37) are satisfied. Let $\mathcal{P}$ to be a collection of such sets π such that, for any two sets π and π' in $\mathcal{P}$, the set $\pi \cap \pi'$ has cardinality at most $\frac{m_0}{2}$. This ensures that

$$\text{for} \quad \mathbf{y}, \mathbf{y}' \in \bigcup_{\pi \in \mathcal{P}} \mathcal{F}_\pi, \quad L(\mathbf{y}, \mathbf{y}') \geq r^2.$$

This also ensures that the sets $\mathcal{F}_\pi$ are disjoint for $\pi \neq \pi'$, since each $\theta_\nu^{(j)}$ for $\theta^{(j)} \in \mathcal{F}_0$ is nonzero in exactly $m_0 + 1$ coordinates. Define $\mathcal{F}_0 = \bigcup_{\pi \in \mathcal{P}} \mathcal{F}_\pi$. Then

$$|\mathcal{F}_0| = |\bigcup_{\pi \in \mathcal{P}} \mathcal{F}_\pi| = |\mathcal{P}| \; |Y_m^*| \geq |\mathcal{P}|(9/8)^{m(1+o(1))}. \tag{42}$$

By Lemma 7, stated in Section 9.4, there is a collection $\mathcal{P}$ such that $|\mathcal{P}|$ is at least $\exp([N\mathcal{E}(m/9N) - 2m\mathcal{E}(1/9)](1 + o(1)))$, where $\mathcal{E}(x)$ is the Shannon entropy function :

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

Since $\mathcal{E}(x) \sim -x\log x$ when $x \to 0+$, it follows from (42) that,

$$\frac{\log|\mathcal{F}_0|}{m} \geq [\frac{1}{9}(\log N - \log m) - 2\mathcal{E}(1/9) + \log 9 + \log(9/8)](1 + o(1)) \geq \frac{\alpha}{9}\log N(1 + o(1)),$$

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

$$\limsup_{n \to \infty} \frac{ave_{\theta^{(j)} \in |\mathcal{F}_0|}K(\theta^{(j)}, \theta^{(0)}) + \log 2}{\log|\mathcal{F}_0|} \leq \limsup_{n \to \infty} \frac{\frac{1}{2}(\alpha/9)m\log N}{(\alpha/9)m\log N} = \frac{1}{2}$$

and use (31) to finish argument.

6.8   Proof of Part (c)

Consider first the proof of (16). Fix a $\mu \in \{1,\ldots,M\} \setminus \{\nu\}$. Define $\theta^{(1)}$ and $\theta^{(2)}$ as follows. Set $r^2 = \frac{2}{ng(\lambda_1, \lambda_2)}$ (assume w.l.o.g. that $r < 1 \wedge C_0$). Take $\theta_{\mu'}^{(j)} = \mathbf{e}_{\mu'}$, $j = 1,2$ for all $\mu' \neq \mu, \nu$. Define

$$\theta_\nu^{(1)} = \mathbf{e}_\nu, \quad \theta_\nu^{(2)} = \sqrt{1-r^2}\mathbf{e}_\nu + r\mathbf{e}_\mu, \quad \theta_\mu^{(1)} = \mathbf{e}_\mu, \quad \theta_\mu^{(2)} = -r\mathbf{e}_\nu + \sqrt{1-r^2}\mathbf{e}_\mu. \tag{43}$$

Observe that $\theta_\mu^{(j)} \perp \theta_\nu^{(j)}$, $j = 1,2$, $\langle\theta_\nu^{(1)}, \theta_\nu^{(2)}\rangle = \sqrt{1-r^2} = \langle\theta_\mu^{(1)}, \theta_\mu^{(2)}\rangle$ and $\langle\theta_\mu^{(1)}, \theta_\mu^{(2)}\rangle = r = -\langle\theta_\nu^{(1)}, \theta_\mu^{(2)}\rangle$. Also, by **A1**, $\theta^{(j)} \in \Theta_q^M(C_1,\ldots,C_M)$, for $j = 1,2$.

Let $P_j = N^{\otimes n}(0, \Sigma_{(j)})$. Then

$$
\begin{aligned}
K(P_1, P_2) + K(P_2, P_1) &= n[h(\lambda_\mu)(1 - |\langle\theta_\mu^{(1)}, \theta_\mu^{(2)}\rangle|^2) + h(\lambda_\nu)(1 - |\langle\theta_\nu^{(1)}, \theta_\nu^{(2)}\rangle|^2) \\
&\quad - \frac{1}{2}(\lambda_\mu\eta(\lambda_\nu) + \lambda_\nu\eta(\lambda_\mu))\{|\langle\theta_\mu^{(1)}, \theta_\nu^{(2)}\rangle|^2 + |\langle\theta_\nu^{(1)}, \theta_\mu^{(2)}\rangle|^2\}] \\
&= n[(h(\lambda_\mu) + h(\lambda_\nu))r^2 - \frac{1}{2}(\lambda_\mu\eta(\lambda_\nu) + \lambda_\nu\eta(\lambda_\mu))r^2] \\
&= ng(\lambda_\mu, \lambda_\nu)r^2. \tag{44}
\end{aligned}
$$

21

Apply Lemma 2 for testing P_1 against P_2. Define $p_{mis} = \inf_T(\alpha_T \vee \beta_T)$ and observe that $p_{mis} \leq \pi_{mis} \leq 2p_{mis}$. Since the lower bound in (33) is symmetric w.r.t. π_{mis}, and π_{mis} is symmetric w.r.t. P_1 and P_2, it follows that

$$ng(\lambda_\mu, \lambda_\nu)r^2 = K(P_1, P_2) + K(P_2, P_1) \geq -2\log(\pi_{mis}(2 - \pi_{mis})).$$

This implies that
$$e^{-\frac{n}{2}g(\lambda_\mu, \lambda_\nu)r^2} \leq \pi_{mis}(2 - \pi_{mis}) \leq 2\pi_{mis} \leq 4p_{mis}$$

Since, $L(\theta^{(1)}, \theta^{(2)}) = 2(1 - \sqrt{1 - r^2}) \geq r^2$, and $r^2 = \frac{2}{ng(\lambda_\mu, \lambda_\nu)}$, use (28) with $\mathcal{F}_0 = \{\theta^{(1)}, \theta^{(2)}\}$ and $\delta = r^2$ to get,

$$\sup_{\theta \in \Theta_q^M(\theta_1, \dots, \theta_M)} \mathbb{E}_\theta L(\theta_\nu, \widehat{\theta}_\nu) \geq \frac{1}{8e} \frac{1}{ng(\lambda_\mu, \lambda_\nu)}.$$

Now, let μ vary over all the indices $1, \dots, \nu - 1, \nu + 1, \dots, M$ and the result follows.

In the situation where $\bar{\delta}_n \not\to 0$, as $n \to \infty$, simply take μ ($\neq \nu$) to be the index for which $g(\lambda_\mu, \lambda_\nu)$ is minimum. Then apply the same procedure as in above with $r \in (0, C_0)$ fixed.

7   Proof of Theorem 1

We require two main tools in the proof of Theorem 1 - one (Lemma 5) is concerned with the deviations of the extreme eigenvalues of a Wishart(N, n) matrix and the other (Lemma 6) relates to the change in the eigen-structure of a symmetric matrix caused by a small, additive perturbation. Sections 9.1 and 9.2 are devoted to them. The importance of Lemma 6 is that, in order to bound the risk of an estimator of θ_ν one only needs to compute the expectation of squared norm of a quantity that is linear in $\mathbf{S}$ (or a submatrix of this, in case of ASPCA estimator). The second bound in (132) then ensures that the remainder is necessarily of smaller order of magnitude. This fact is used explicitly in deriving (66).

Remark : In view of Lemma 6, $H_\nu(\Sigma)$ becomes a key quantity in the analysis of the risk of any estimator of θ_ν. Observe that,

$$H_\nu := H_\nu(\Sigma) = \sum_{1 \leq \nu' \neq \nu \leq M} \frac{1}{\lambda_{\nu'} - \lambda_\nu} \theta_{\nu'}\theta_{\nu'}^T - \frac{1}{\lambda_\nu}\left(I - \sum_{\nu'=1}^M \theta_{\nu'}\theta_{\nu'}^T\right), \qquad \nu = 1, \dots, M. \tag{45}$$

Expand matrix $\mathbf{S}$ as follows.

$$\mathbf{S} = \sum_{\mu=1}^M \frac{\| v_\mu \|^2}{n}\lambda_\mu\theta_\mu\theta_\mu^T + \sum_{\mu=1}^M \sqrt{\lambda_\mu}\left(\theta_\mu(\frac{1}{n}\mathbf{Z}v_\mu)^T + \frac{1}{n}\mathbf{Z}v_\mu\theta_\mu^T\right)$$
$$+ \sum_{\mu \neq \mu'} \frac{\langle v_\mu, v_{\mu'}\rangle}{n}\sqrt{\lambda_\mu\lambda_{\mu'}}\theta_\mu\theta_{\mu'}^T + \frac{1}{n}\mathbf{Z}\mathbf{Z}^T. \tag{46}$$

In order to use Lemma 6, an expression for $H_\nu \mathbf{S}\theta_\nu$ is needed. Use the fact that $H_\nu \theta_\nu = 0$ and $\theta_\nu^T \theta_\mu = \delta_{\mu\nu}$ (Kronecker's symbol), to conclude that

$$
\begin{aligned}
H_\nu \mathbf{S}\theta_\nu &= \sum_{\mu \neq \nu} \left(\sqrt{\lambda_\mu} \frac{1}{n} \langle \mathbf{Z}v_\mu, \theta_\nu \rangle + \sqrt{\lambda_\mu \lambda_\nu} \frac{1}{n} \langle v_\mu, v_\nu \rangle \right) H_\nu \theta_\mu \\
&\quad + \sqrt{\lambda_\nu} H_\nu \frac{1}{n} \mathbf{Z} v_\nu + H_\nu \frac{1}{n} \mathbf{Z}\mathbf{Z}^T \theta_\nu.
\end{aligned}
\tag{47}
$$

Further, from (45) it follows that, $H_\nu \theta_\mu = \frac{1}{\lambda_\mu - \lambda_\nu} \theta_\mu$, if $\mu \neq \nu$. Also,

$$
H_\nu \mathbf{Z} v_\nu = -\frac{1}{\lambda_\nu}(I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T) \mathbf{Z} v_\nu + \sum_{\mu \neq \nu} \frac{1}{\lambda_\mu - \lambda_\nu} \langle \mathbf{Z} v_\nu, \theta_\mu \rangle \theta_\mu,
\tag{48}
$$

and

$$
H_\nu \mathbf{Z}\mathbf{Z}^T \theta_\nu = -\frac{1}{\lambda_\nu}(I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T) \mathbf{Z}\mathbf{Z}^T \theta_\nu + \sum_{\mu \neq \nu} \frac{1}{\lambda_\mu - \lambda_\nu} \langle \mathbf{Z}^T \theta_\mu, \mathbf{Z}^T \theta_\nu \rangle \theta_\mu.
\tag{49}
$$

From (47), (48) and (49), it follows that

$$
\begin{aligned}
H_\nu \mathbf{S}\theta_\nu &= \sum_{\mu \neq \nu} \frac{1}{\lambda_\mu - \lambda_\nu} \left(\sqrt{\lambda_\mu} \frac{1}{n} \langle \mathbf{Z}v_\mu, \theta_\nu \rangle + \sqrt{\lambda_\nu} \frac{1}{n} \langle \mathbf{Z}v_\nu, \theta_\mu \rangle \right) \theta_\mu \\
&\quad + \sum_{\mu \neq \nu} \frac{1}{\lambda_\mu - \lambda_\nu} \left(\sqrt{\lambda_\mu \lambda_\nu} \frac{1}{n} \langle v_\mu, v_\nu \rangle + \frac{1}{n} \langle \mathbf{Z}^T \theta_\mu, \mathbf{Z}^T \theta_\nu \rangle \right) \theta_\mu \\
&\quad - \frac{1}{n\lambda_\nu}(I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T) \mathbf{Z}\mathbf{Z}^T \theta_\nu - \frac{1}{n\sqrt{\lambda_\nu}}(I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T) \mathbf{Z} v_\nu.
\end{aligned}
\tag{50}
$$

Let Γ be an $N \times (N - M)$ matrix such that $\Gamma^T \Gamma = I$, and $\Gamma\Gamma^T = (I - \sum_{\mu=1}^M \theta_\mu \theta_\mu^T)$. Then, $\Gamma\theta_\mu = 0$ for all $\mu = 1, \ldots, M$.

A crucial fact here is that, since v_μ has i.i.d. $N(0, 1)$ entries, and is independent of $\mathbf{Z}$, for any $D \in \mathbb{R}^{m \times n}$, $D\mathbf{Z}\frac{v_\mu}{\|v_\mu\|}$ has a $N_m(0, DD^T)$ distribution, and is independent of v_μ. Furthermore, since θ_μ are orthonormal, and $\Gamma\theta_\mu = 0$ for all μ, it follows that $\mathbf{Z}^T \theta_\mu$ has a $N_n(0, I)$ distribution; $\{\mathbf{Z}^T \theta_\mu\}_{\mu=1}^M$ are mutually independent and are independent of $\Gamma\mathbf{Z}$.

Next, we compute some expectations that will lead to the final expression for $\mathbb{E} \| H_\nu \mathbf{S}\theta_\nu \|^2$.

$$
\begin{aligned}
&\mathbb{E}\left(\sqrt{\lambda_\mu} \frac{1}{n} \langle \mathbf{Z}v_\mu, \theta_\nu \rangle + \sqrt{\lambda_\nu} \frac{1}{n} \langle \mathbf{Z}v_\nu, \theta_\mu \rangle \right)^2 \\
&= \frac{1}{n^2} \left[\lambda_\mu \mathbb{E}(\langle \mathbf{Z}v_\mu, \theta_\nu \rangle)^2 + \lambda_\nu \mathbb{E}(\langle \mathbf{Z}v_\nu, \theta_\mu \rangle)^2 + 2\sqrt{\lambda_\mu \lambda_\nu} \mathbb{E}(\langle \mathbf{Z}v_\mu, \theta_\nu \rangle \langle \mathbf{Z}v_\nu, \theta_\mu \rangle) \right]^2 \\
&= \frac{\lambda_\mu + \lambda_\nu}{n},
\end{aligned}
\tag{51}
$$

23

since the cross product term vanishes, which can be verified by a simple conditioning argument. By similar calculations,

$$\mathbb{E}\left(\sqrt{\lambda_\mu\lambda_\nu}\frac{1}{n}\langle v_\mu, v_\nu\rangle + \frac{1}{n}\langle \mathbf{Z}^T\theta_\mu, \mathbf{Z}^T\theta_\nu\rangle\right)^2 = \frac{\lambda_\nu\lambda_\mu + 1}{n},\tag{52}$$

and

$$\mathbb{E}\left(\sqrt{\lambda_\mu}\frac{1}{n}\langle \mathbf{Z}v_\mu, \theta_\nu\rangle + \sqrt{\lambda_\nu}\frac{1}{n}\langle \mathbf{Z}v_\nu, \theta_\mu\rangle\right)\left(\sqrt{\lambda_\mu\lambda_\nu}\frac{1}{n}\langle v_\mu, v_\nu\rangle + \frac{1}{n}\langle \mathbf{Z}^T\theta_\mu, \mathbf{Z}^T\theta_\nu\rangle\right) = 0.\tag{53}$$

Since $trace(\Gamma\Gamma^T) = N - M$, from the remark made above, it follows that,

$$\mathbb{E}\parallel (I - \sum_{\mu=1}^{M}\theta_\mu\theta_\mu^T)\mathbf{Z}\mathbf{Z}^T\theta_\nu \parallel^2 = \mathbb{E}[(\theta_\nu^T\mathbf{Z})\mathbf{Z}^T\Gamma\Gamma^T\mathbf{Z}(\mathbf{Z}^T\theta_\nu)] = n(N - M),\tag{54}$$

$$\mathbb{E}\parallel (I - \sum_{\mu=1}^{M}\theta_\mu\theta_\mu^T)\mathbf{Z}v_\nu \parallel^2 = \mathbb{E}\parallel v_\nu \parallel^2 \mathbb{E}\parallel \Gamma^T\mathbf{Z}\frac{v_\nu}{\parallel v_\nu \parallel} \parallel^2 = n(N - M),\tag{55}$$

and

$$\mathbb{E}\langle(I - \sum_{\mu=1}^{M}\theta_\mu\theta_\mu^T)\mathbf{Z}\mathbf{Z}^T\theta_\nu, (I - \sum_{\mu=1}^{M}\theta_\mu\theta_\mu^T)\mathbf{Z}v_\nu\rangle = \mathbb{E}[v_\nu^T\mathbf{Z}^T\Gamma\Gamma^T\mathbf{Z}(\mathbf{Z}^T\theta_\nu)] = 0.\tag{56}$$

Use (50), and equations (51) - (56), together with the orthonormality of θ_μ's and the fact that $\Gamma\theta_\mu = 0$ for all μ to conclude that,

$$\mathbb{E}\parallel H_\nu\mathbf{S}\theta_\nu \parallel^2 = \frac{N - M}{nh(\lambda_\nu)} + \frac{1}{n}\sum_{\mu\neq\nu}\frac{(1 + \lambda_\mu)(1 + \lambda_\nu)}{(\lambda_\mu - \lambda_\nu)^2}.\tag{57}$$

The next step in the argument is to show that, $\max_{0\leq\mu\leq M}(\lambda_\mu - \lambda_{\mu+1})^{-1}\parallel \mathbf{S} - \Sigma \parallel$ is small with a very high probability. Here, by convention, $\lambda_0 = \infty$ and $\lambda_{M+1} = 0$. From (46),

$$\parallel \mathbf{S} - \Sigma \parallel \leq \sum_{\mu=1}^{M}\lambda_\mu|\frac{\parallel v_\mu \parallel^2}{n} - 1| + 2\sum_{\mu=1}^{M}\sqrt{\lambda_\mu}\frac{1}{n}\parallel \mathbf{Z}v_\mu \parallel$$

$$+ \sum_{\mu\neq\mu'}\sqrt{\lambda_\mu\lambda_{\mu'}}|\frac{\langle v_\mu, v_{\mu'}\rangle}{n}| + \parallel \frac{1}{n}\mathbf{Z}\mathbf{Z}^T - I \parallel.\tag{58}$$

Define, for any $c > 0$, $D_{1,n}(c)$ to be the set

$$D_{1,n}(c) = \bigcap_{\mu=1}^{M}\{|\frac{\parallel v_\mu \parallel^2}{n} - 1| \leq 2c\sqrt{\frac{\log(n\vee N)}{n}}\}$$

$$\bigcap_{\mu=1}^{M}\{\frac{\parallel \mathbf{Z}v_\mu \parallel}{n} \leq \left(1 + 2c\sqrt{\frac{\log(n\vee N)}{n\wedge N}}\right)\sqrt{\frac{N}{n}}\}$$

$$\bigcap_{1\leq\mu<\mu'\leq M}\{|\frac{\langle v_\mu, v_{\mu'}\rangle}{n}| \leq c\sqrt{\frac{\log(n\vee N)}{n}}\}.\tag{59}$$

24

Use Lemmas 14 and 15 to prove that,

$$1 - \mathbb{P}(D_{1,n}(c)) \le 3M(n \vee N)^{-c^2} + M(M-1)(n \vee N)^{-\frac{3}{2}c^2 + O(\log(n \vee N)/n)}. \qquad (60)$$

Define $D_{2,n}(c)$ as

$$D_{2,n}(c) = \{\| \frac{1}{n}\mathbf{Z}\mathbf{Z}^T - I \| \le 2\sqrt{\frac{N}{n}} + \frac{N}{n} + ct_n\}, \qquad (61)$$

with t_n as in Lemma 5. From (58), (60) and (123), it follows that for $n \ge n_c$,

$$\begin{aligned}
\mathbb{P}(\| \mathbf{S} - \Sigma \| > \epsilon_{n,N}(c,\lambda)) &\le 1 - \mathbb{P}(D_{1,n}(c) \cap D_{2,n}(c)) \\
&\le (3M+2)(n \vee N)^{-c^2} + M(M-1)(n \vee N)^{-\frac{3}{2}c^2 + O(\log(n \vee N)/n)},
\end{aligned} \qquad (62)$$

where

$$\begin{aligned}
\epsilon_{n,N}(c,\lambda) &= 2c(\sum_{\mu=1}^{M} \lambda_\mu)\sqrt{\frac{\log(n \vee N)}{n}} + 2(\sum_{\mu=1}^{M} \sqrt{\lambda_\mu})\left(1 + 2c\sqrt{\frac{\log(n \vee N)}{n \wedge N}}\right)\sqrt{\frac{N}{n}} \\
&\quad + c(\sum_{1 \le \mu \ne \mu' \le M} \sqrt{\lambda_\mu \lambda_{\mu'}})\sqrt{\frac{\log(n \vee N)}{n}} + 2\sqrt{\frac{N}{n}} + \frac{N}{n} + ct_n.
\end{aligned} \qquad (63)$$

Define

$$\delta_{n,N,\nu} = \max\{(\lambda_\nu - \lambda_{\nu+1})^{-1}, (\lambda_{\nu-1} - \lambda_\nu)^{-1}\}\epsilon_{n,N}(\sqrt{2}, \lambda), \qquad (64)$$

and observe that $\delta_{n,N,\nu} \to 0$ as $n \to \infty$ under **L1** and **L2**.

To complete the proof of (5), write

$$\widehat{\theta}_\nu - sign(\theta_\nu^T \widehat{\theta}_\nu)\theta_\nu = -H_\nu \mathbf{S}\theta_\nu + R_\nu. \qquad (65)$$

Since $\delta_{n,N,\nu} \to 0$, by (132), (130), (131) and (62), and the fact that $\Delta_r \le \overline{\Delta}_r$, for sufficiently large n, on $D_{1,n}(\sqrt{2}) \cap D_{2,n}(\sqrt{2})$,

$$\| H_\nu \mathbf{S}\theta_\nu \|^2 (1 - \delta'_{n,N,\nu})^2 \le L(\theta_\nu, \widehat{\theta}_\nu) \le \| H_\nu \mathbf{S}\theta_\nu \|^2 (1 + \delta'_{n,N,\nu})^2, \qquad (66)$$

where

$$\delta'_{n,N,\nu} = \frac{\delta_{n,N,\nu}}{(1 - 2\delta_{n,N,\nu}(1 + 2\delta_{n,N,\nu}))^2}[1 + 2(1 + \delta_{n,N,\nu})(1 - 2\delta_{n,N,\nu}(1 + 2\delta_{n,N,\nu}))], \qquad (67)$$

and $\delta'_{n,N,\nu} \to 0$ as $n \to \infty$. Since $L(\theta_\nu, \widehat{\theta}_\nu) \le 2$, (62), (66) and (57) together imply (5).

8   Proof of Theorem 3

In some respect the proof of Theorem 3 bears resemblance to the proof of Theorem 4 in Johnstone and Lu (2004). The basic idea in both these cases is to first provide a "bracketing relation". This means that, if $\widehat{I}_n$ denotes the set of selected coordinates, and $\underline{I}_n$ and $\overline{I}_n$ are two *non-random* sets with suitable properties, then an inequality of the form $\mathbb{P}(\underline{I}_n \subset \widehat{I}_n \subset \overline{I}_n) \ge 1 - b_n$ holds,

where b_n converges to zero at least polynomially in n. Once this relationship is established, one can utilize it to study the eigen-structure of the submatrix $\mathbf{S}_{\widehat{I}_n,\widehat{I}_n}$ of $\mathbf{S}$. The advantage of this is that the bracketing relation ensures that the quantities involved in the perturbation terms for the eigenvectors and eigenvalues can be controlled, except possibly on a set of probability at most b_n.

The proof of Theorem 3 follows this principle. However, there are several technical aspects in both the steps that require much computation. The first step, namely, establishing a bracketing relation for $\widehat{I}_n$, is done in Sections 8.2 - 8.6. The second step follows more or less the approach taken in the proof of Theorem 1, in that, on a set of high probability, an upper bound on $L(\widehat{\theta}_\nu, \theta_\nu)$ is established that is of the form $\| H_\nu \mathbf{S}_2 \theta_\nu \|^2 (1 + \delta_n)$, where H_ν is as in (45), $\mathbf{S}_2$ is the matrix defined through equation (93), and $\delta_n \to 0$. Then, by a careful examination of the different terms in an expansion of $H_\nu \mathbf{S}_2 \theta_\nu$, it is shown that an upper bound on $\mathbb{E} \| H_\nu \mathbf{S}_2 \theta_\nu \|^2$ is asymptotically same as the RHS of (24). This is done in Section 8.9. Some results related to the determination of correct asymptotic order of the terms in the aforementioned expansion are given in Section 9.5. Before going into the detailed analysis, it is necessary to fix some notation.

8.1 Notation

For any symmetric matrix D, $\lambda_k(D)$ will denote the k-th largest eigenvalue of D. Frequently, the set $\{1, \ldots, N\}$ will be divided into complementary sets A and B. Here A may refer to the set of coordinates selected either in the first stage, or in the second stage, or in a combination of both. $\mathbf{S}$ will be partitioned as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{AA} & \mathbf{S}_{AB} \\ \mathbf{S}_{BA} & \mathbf{S}_{BB} \end{bmatrix} \tag{68}$$

where $\mathbf{S}_{AB}$ is the submatrix of $\mathbf{S}$ whose row indices are from set A, and column indices are from set B. Any $N \times 1$ vector $\mathbf{x}$ may similarly be partitioned as $\mathbf{x} = (\mathbf{x}^T : \mathbf{y}^T)^T$. And for an $N \times k$ matrix $\mathbf{Y}$, $\mathbf{Y}_A$ and $\mathbf{Y}_B$ will denote the parts corresponding to rows with indices from set A and B, respectively. It should be clear, however, that no specific order relation among these indices is assumed, and in fact the order of the rows is unchanged in all of these situations. Expressions like (68) are just for convenience of writing.

8.2 Bracketing relations

In this section the bracketing relationship is established. The proof involves several parts. It essentially boils down to probabilistic analysis of 1° - 5° of the ASPCA algorithm. This is done in several stages. The coordinate selection step in 1° and 2° are jointly referred to as the *first stage*, and steps 3°, 4° and 5° are jointly referred to as the *second stage*.

8.3 First stage coordinate selection

In this section 1°, i.e., the first stage of the coordinate selection scheme, is analyzed. Define

$$\zeta_k = \sum_{\nu=1}^{M} \lambda_\nu \theta_{\nu k}^2, \quad k = 1, \ldots, M. \tag{69}$$

26

For $0 < a_- < 1 < a_+$, define

$$I_{1,n}^{\pm} = \{k : \zeta_k > a_{\mp}\gamma_1\sqrt{\frac{\log(N \vee n)}{n}}\}. \tag{70}$$

It is shown that $\widehat{I}_{1,n}$ satisfies the bracketing relation (74).

Let $\sigma_k^2 := \zeta_k + 1$. The selected coordinates are

$$\widehat{I}_{1,n} = \{k : \mathbf{S}_{kk} > 1 + \gamma_1\sqrt{\frac{\log(N \vee n)}{n}}\}. \tag{71}$$

Note that, $\mathbf{S}_{kk} \sim \sigma_k^2\chi_{(n)}^2/n$. Then,

$$
\begin{aligned}
\mathbb{P}(I_{1,n}^- \not\subset \widehat{I}_{1,n}) &= \mathbb{P}(\cup_{k \in I_{1,n}^-}\{\mathbf{S}_{kk} \leq 1 + \gamma_{1,n}\}) \leq \sum_{k \in I_{1,n}^-}\mathbb{P}(\mathbf{S}_{kk} \leq 1 + \gamma_{1,n}) \\
&\leq \sum_{k \in I_{1,n}^-}\mathbb{P}(\frac{\mathbf{S}_{kk}}{\sigma_k^2} \leq \frac{1 + \gamma_{1,n}}{1 + a_+\gamma_{1,n}}) \\
&\leq |I_{1,n}^-|\mathbb{P}(\frac{\chi_{(n)}^2}{n} - 1 \leq -\frac{\gamma_{1,n}(a_+ - 1)}{1 + a_+\gamma_{1,n}}), \qquad (\text{ since, } \mathbf{S}_{kk} \sim \sigma_k^2\chi_{(n)}^2/n) \\
&\leq |I_{1,n}^-|\exp\left(-\frac{n\gamma_{1,n}^2(a_+ - 1)^2}{4(1 + a_+\gamma_{1,n})^2}\right), \qquad (\text{by } (145)) \\
&\leq |I_{1,n}^-|(N \vee n)^{-(\gamma_1^2(a_+ - 1)^2/4)(1 + o(1))}. \tag{72}
\end{aligned}
$$

Similarly, if $n \geq 16$ then,

$$
\begin{aligned}
\mathbb{P}(\widehat{I}_{1,n} \not\subset I_{1,n}^+) &= \mathbb{P}(\cup_{k \notin I_{1,n}^+}\{\mathbf{S}_{kk} > 1 + \gamma_{1,n}\}) \leq \sum_{k \notin I_{1,n}^+}\mathbb{P}(\mathbf{S}_{kk} > 1 + \gamma_{1,n}) \\
&\leq \sum_{k \notin I_{1,n}^+}\mathbb{P}(\frac{\mathbf{S}_{kk}}{\sigma_k^2} > \frac{1 + \gamma_{1,n}}{1 + a_-\gamma_{1,n}}) \leq N\mathbb{P}(\frac{\chi_{(n)}^2}{n} - 1 > \frac{\gamma_{1,n}(1 - a_-)}{1 + a_-\gamma_{1,n}}) \\
&\leq N\frac{\sqrt{2}}{\gamma_1\sqrt{\log(N \vee n)}}\exp\left(-\frac{n\gamma_{1,n}^2(1 - a_-)^2}{4(1 + a_-\gamma_{1,n})^2}\right), \qquad (\text{by } (146)) \\
&\leq N(N \vee n)^{-(\gamma_1^2(1 - a_-)^2/4)(1 + o(1))}. \tag{73}
\end{aligned}
$$

Combine (72) and (73) to get, as $n \to \infty$,

$$
\begin{aligned}
&1 - \mathbb{P}(I_{1,n}^- \subset \widehat{I}_{1,n} \subset I_{1,n}^+) \\
&\leq |I_{1,n}^-|(N \vee n)^{-(\gamma_1^2(a_+ - 1)^2/4)(1 + o(1))} + N(N \vee n)^{-(\gamma_1^2(1 - a_-)^2/4)(1 + o(1))}. \tag{74}
\end{aligned}
$$

For future use, it is important to have an upper bound on the size of the sets $I_{1,n}^{\pm}$. To this end, let $\mathbf{c} = (c_1, \dots, c_M)$ be such that $c_\nu > 0$ for all ν and $\sum_{\nu=1}^M c_\nu^2 = 1$.

$$I_{1,n}^{\pm} = \{k \in \{1, \dots, N\} : \sum_{\nu=1}^M \lambda_\nu\theta_{\nu k}^2 > a_{\mp}\gamma_{1,n}\} \subset \bigcup_{\nu=1}^M\{k \in \{1, \dots, N\} : |\theta_{\nu k}| > c_\nu\sqrt{\frac{a_{\mp}\gamma_{1,n}}{\lambda_\nu}}\}.$$

Since $\theta \in \Theta_q^M(C_1, \ldots, C_q)$, and $l^q(C) \hookrightarrow wl^q(C)$, it follows from above that,

$$|I_{1,n}^{\pm}| \le J_{1,n}(\mathbf{c}, \gamma_1, a_{\mp}) := a_{\mp}^{-q/2} \gamma_1^{-q/2} \Big(\sum_{\nu=1}^M c_\nu^{-q} \lambda_\nu^{q/2} C_\nu^q \Big) \frac{n^{q/4}}{(\log(N \vee n))^{q/4}}. \tag{75}$$

In fact, the upper bound is of the form $J_{1,n}(\mathbf{c}, \gamma_1, a_{\mp}) \wedge N$, since there are altogether N coordinates. Set $\mathbf{c} = (M^{-1/2}, \ldots, M^{-1/2})$, and denote the corresponding $J_{1,n}(\mathbf{c}, \gamma_1, a_{\mp})$ by $J_{1,n}(\gamma_1, a_{\mp})$. Whenever there is no ambiguity about the choice of γ_1 and $a_{\mp}$, $J_{1,n}(\gamma_1, a_{\mp})$ will be denoted by $J_{1,n}^{\pm}$. Notice that **C1** and **C2** imply that $J_{1,n}^+ \to \infty$ as $n \to \infty$. And **C3** implies that $\frac{J_{1,n}^+}{nh(\lambda_1)} \to 0$.

Remark : From now onwards, the set $\{I_{1,n}^- \subset \widehat{I}_{1,n} \subset I_{1,n}^+\}$ will be denoted by $G_{1,n}$. Observe that $G_{1,n}$ depends on θ. However, from (74), it follows that, if $\gamma_1 = 4$, $a_+ > 1 + \frac{1}{\sqrt{2}}$ and $0 < a_- < 1 - \frac{1}{\sqrt{2}}$, then there is an $\epsilon_0 > 0$ and an $n_0 \ge 1$, that depend on a_+ and a_-, such that for $n \ge n_0$,

$$\mathbb{P}(G_{1,n}^c) \le (N \vee n)^{-1-\epsilon_0}, \tag{76}$$

uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$.

8.4   Eigen-analysis of $\mathbf{S}_{\widehat{I}_{1,n}, \widehat{I}_{1,n}}$

Throughout we follow the convention that $\langle \mathbf{e}_\nu, \theta_{\nu, \widehat{I}_{1,n}} \rangle \ge 0$. Define

$$\widetilde{\mathbf{S}}_1 := \begin{bmatrix} \mathbf{S}_{\widehat{I}_{1,n}\widehat{I}_{1,n}} & O \\ O & O \end{bmatrix} \qquad \mathbf{S}_1 := \begin{bmatrix} \mathbf{S}_{\widehat{I}_{1,n}\widehat{I}_{1,n}} & O \\ O & I \end{bmatrix}. \tag{77}$$

Let $\widetilde{\mathbf{e}}_k$ be the eigenvector associated with eigenvalue $\widehat{\ell}_k$ of $\widetilde{\mathbf{S}}_1$, for $k = 1, \ldots, m_1$, where $m_1 = (n \wedge |\widehat{I}_{1,n}|)$. Eigenvalues of $\mathbf{S}_1$ belong to the set $\{\widehat{\ell}_1, \ldots, \widehat{\ell}_{m_1}\} \cup \{1\}$; and the eigenvector corresponding to the eigenvalue $\widehat{\ell}_k$ is $\widetilde{\mathbf{e}}_k$, $1 \le k \le m_1$. Note that, $\widehat{\ell}_k$ is not necessarily the k-th largest eigenvalue of $\mathbf{S}_1$. However, the analysis here will show that this happens with very high probability for sufficiently large n.

Let $t_{1,n}^+ = 6(J_{1,n}^+/n \vee 1)\sqrt{\log(n \vee J_{1,n}^+)/(n \vee J_{1,n}^+)}$. Define,

$$
\begin{aligned}
\varepsilon_{1,n} &= \frac{2\sqrt{2}}{\lambda_1} \sum_{\nu=1}^{M} \lambda_\nu \sqrt{\frac{\log(n \vee J_{1,n}^+)}{n}} \\
\varepsilon_{2,n} &= \frac{2}{\lambda_1} \sum_{\nu=1}^{M} \sqrt{\lambda_\nu} \left(1 + 2\sqrt{2}\sqrt{\frac{\log(n \vee J_{1,n}^+)}{n \wedge J_{1,n}^+}} \right) \sqrt{\frac{J_{1,n}^+}{n}} \\
\varepsilon_{3,n} &= \frac{\sqrt{2}}{\lambda_1} \sum_{\nu \neq \nu'} \sqrt{\lambda_\nu \lambda_{\nu'}} \sqrt{\frac{\log n}{n}} \\
\varepsilon_{4,n} &= \frac{1}{\lambda_1} \left(2\sqrt{\frac{J_{1,n}^+}{n}} + \frac{J_{1,n}^+}{n} + \sqrt{2} t_{1,n}^+ \right) \\
\varepsilon_{5,n} &= c_q a_+^{1-q/2} \gamma_1^{1-q/2} \left(\sum_{\nu=1}^{M} \lambda_\nu^{q/2} C_\nu^q \right) \frac{(\log(N \vee n))^{1/2-q/4}}{\lambda_1 n^{1/2-q/4}}
\end{aligned}
\tag{78}
$$

where $c_q = \frac{2}{2-q}$. Observe that, under conditions **C1**-**C3**, $\max_{1 \leq j \leq 5} \varepsilon_{j,n} \to 0$ as $n \to \infty$. Set $A = \widehat{I}_{1,n}$, $A_+ = I_{1,n}^+$, $B = \widehat{I}_{1,n}^c = \{1, \ldots, N\} \setminus \widehat{I}_{1,n}$, and define

$$
\begin{aligned}
G_{2,n} &= \bigcap_{\nu=1}^{M} \{ |\frac{\| v_\nu \|^2}{n} - 1| \leq 2\sqrt{2}\sqrt{\frac{\log n}{n}} \} \\
&\cap \bigcap_{\nu=1}^{M} \{ \frac{\| \mathbf{Z}_{A_+} v_\nu \|}{n} \leq \left(1 + 2\sqrt{2}\sqrt{\frac{\log(n \vee J_{1,n}^+)}{n \wedge J_{1,n}^+}} \right) \sqrt{\frac{J_{1,n}^+}{n}} \} \\
&\cap \bigcap_{1 \leq \nu < \nu' \leq M} \{ |\frac{\langle v_\nu, v_{\nu'} \rangle}{n}| \leq \sqrt{2}\sqrt{\frac{\log n}{n}} \},
\end{aligned}
\tag{79}
$$

and

$$
G_{3,n} = \{ \frac{1}{n} \| \mathbf{Z}_{A_+} \mathbf{Z}_{A_+}^T - I \| \leq 2\sqrt{\frac{J_{1,n}^+}{n}} + \frac{J_{1,n}^+}{n} + \sqrt{2} t_{1,n}^+ \}.
\tag{80}
$$

Then the following results hold.

Lemma 3: *Under conditions* **C1**-**C3**,

$$
\bigcap_{j=1}^{3} G_{j,n} \subset \{ |\lambda_\nu(\mathbf{S}_1) - (1 + \lambda_\nu)| \leq \lambda_1 \sum_{j=1}^{5} \varepsilon_{j,n} \},
\tag{81}
$$

$$
\mathbb{P}((G_{2,n} \cap G_{3,n})^c) \leq 3M(n \vee J_{1,n}^+)^{-2} + M(M-1)n^{-3+O(\frac{\log n}{n})} + 2(n \vee J_{1,n}^+)^{-2}.
\tag{82}
$$

Lemma 4: *Let* $\widetilde{t}_{1,n} = 6(|I_{1,n}^+|/n \vee 1)\sqrt{\log(n \vee |I_{1,n}^+|)/(n \vee |I_{1,n}^+|)}$. *Under conditions* **C1**-**C3**,

$$\mathbb{P}(\widehat{\ell}_{M+1} > (1 + \sqrt{\frac{|I_{1,n}^+|}{n}})^2 + \sqrt{2}\widetilde{t}_{1,n}, \ \widehat{I}_{1,n} \subset I_{1,n}^+) \leq 2(n \vee |I_{1,n}^+|)^{-2}. \tag{83}$$

Remark : Let $G_{4,n} = \{\widehat{\ell}_{M+1} \leq (1 + \sqrt{\frac{|I_{1,n}^+|}{n}})^2 + \sqrt{2}\widetilde{t}_{1,n}\}$, where $\widetilde{t}_{1,n}$ is as in Lemma 4. Observe that $G_{4,n}$ depends on θ; however, $\mathbb{P}(G_{1,n} \cap G_{4,n}^c) \leq 2n^{-2}$ for all $\theta \in \Theta_q^M(C_1, \ldots, C_M)$. It is easy to check that, under **C1**-**C3**,

$$2\sqrt{\frac{J_{1,n}^+}{n}} + \frac{J_{1,n}^+}{n} = o(\lambda_1) \quad \text{as} \ \ n \to \infty. \tag{84}$$

Therefore, from Lemma 3 and Lemma 4 it follows that, for sufficiently large n, uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$,

$$\mathbb{P}(\max_{1 \leq \nu \leq M} |\widehat{\ell}_\nu - (1 + \lambda_\nu)| > \lambda_1 \sum_{j=1}^5 \varepsilon_{j,n}, \ G_{1,n}) \leq K_1(M)n^{-2}, \tag{85}$$

for some constant $K_1(M)$ that does not depend on θ.

8.5 Consistency of $\widehat{M}$

Proposition 2: *Under conditions* **C1**-**C3**, *and with* α_n *defined through (20)*, $\widehat{M}$ *is a consistent estimator of* M. *In particular, if* $\overline{\gamma}_1 = 9$, $\gamma_1' = 3$, *then there are constants* $\overline{a}_+ > 1 > \overline{a}_- > 0$, $1 > a' > 0$, *and an* n_{*0} *such that for* $n \geq n_{*0}$, *uniformly in* $\theta \in \Theta_q^M(C_1, \ldots, C_M)$,

$$\mathbb{P}(\widehat{M} \neq M) \leq K_2(M)n^{-1-\epsilon_1}, \tag{86}$$

for some constants $K_2(M) > 0$ *and* $\epsilon_1 := \epsilon_1(\overline{\gamma}_1, \gamma_1', \overline{a}_\pm, a') > 0$ *independent of* θ.

8.6 Second stage coordinate selection

Steps 4^0 and 5^0 of the ASPCA scheme are analyzed in this subsection. For future reference, it is convenient to denote the event $\bigcap_{j=1}^4 G_{j,n} \cap \{\widehat{M} = M\}$ by $\overline{G}_{1,n}$. The ultimate goal of this section is to establish (92). Throughout, it is assumed that **BA** and **C1**-**C3** are valid. Observe that, by definition (see 4° and 5° of ASPCA scheme), $T_k = \sum_{\mu=1}^M Q_{k\mu}^2$ if $k \notin \widehat{I}_{1,n}$, and define it to be zero otherwise.

8.6.1 A preliminary bracketing relation

First, define

$$\widetilde{\zeta}_k = \sum_{\nu=1}^M h(\lambda_\nu)\theta_{\nu k}^2, \quad k = 1, \ldots, N. \tag{87}$$

Define, for $0 < \gamma_{2,-} < \gamma_2 < \gamma_{2,+}$,

$$I_n^\pm = \{k : \widetilde{\zeta}_k > \gamma_{2,\mp}^2 \, \frac{\log(N \vee n)}{n}\}. \tag{88}$$

Observe that $\widetilde{\zeta}_k \geq \eta(\lambda_M)\zeta_k$. This implies that, for some $n_{*1} \geq n_{*0} \vee n'_{*0}$, for all $n \geq n_{*1}$, $I_{n,1}^+ \subset I_n^-$, uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$. Note that

$$\mathbb{P}(\{I_n^- \subset \widehat{I}_{1,n} \cup \widehat{I}_{2,n} \subset I_n^+\}^c, \overline{G}_{1,n}) \leq \mathbb{P}(I_n^- \not\subset \widehat{I}_{1,n} \cup \widehat{I}_{2,n}, \overline{G}_{1,n}) + \mathbb{P}(\widehat{I}_{1,n} \cup \widehat{I}_{2,n} \not\subset I_n^+, \overline{G}_{1,n}).$$

In the following, D is a generic measurable set w.r.t. the σ-algebra generated by $\mathbf{Z}$ and $v_1, \ldots, v_M$. Then, for $n \geq n_{*1}$,

$$\mathbb{P}(\widehat{I}_{1,n} \cup \widehat{I}_{2,n} \not\subset I_n^+, \, \overline{G}_{1,n} \cap D) \;\; = \;\; \mathbb{P}(\cup_{k \notin I_n^+}\{k \in \widehat{I}_{1,n} \cup \widehat{I}_{2,n}\}, \, \overline{G}_{1,n})$$

$$= \mathbb{P}(\cup_{k \notin I_n^+}\{k \in \widehat{I}_{2,n} \cap \widehat{I}_{1,n}^c\}, \, \overline{G}_{1,n} \cap D) \;\; \leq \;\; \sum_{k \notin I_n^+} \mathbb{P}(k \in \widehat{I}_{2,n} \cap \widehat{I}_{1,n}^c, \, \overline{G}_{1,n} \cap D)$$

$$= \sum_{k \notin I_n^+} \mathbb{P}(T_k > \gamma_{2,n}^2, \, \overline{G}_{1,n} \cap D), \tag{89}$$

where the last equality is from the inclusion $\widehat{I}_{1,n} \subset I_{1,n}^+ \subset I_n^- \subset I_n^+$. Similarly,

$$\mathbb{P}(I_n^- \not\subset \widehat{I}_{1,n} \cup \widehat{I}_{2,n}, \, \overline{G}_{1,n} \cap D) \;\; = \;\; \mathbb{P}(\cup_{k \in I_n^-}\{k \notin \widehat{I}_{1,n} \cup \widehat{I}_{2,n}\}, \, \overline{G}_{1,n} \cap D)$$

$$= \mathbb{P}(\cup_{k \in I_n^- \setminus I_{1,n}^-}\{k \in \widehat{I}_{1,n}^c \cap \widehat{I}_{2,n}^c\}, \, \overline{G}_{1,n} \cap D) \;\; \leq \;\; \sum_{k \in I_n^- \setminus I_{1,n}^-} \mathbb{P}(k \notin \widehat{I}_{1,n}, k \notin \widehat{I}_{2,n}, \, \overline{G}_{1,n} \cap D)$$

$$= \sum_{k \in I_n^- \setminus I_{1,n}^-} \mathbb{P}(T_k \leq \gamma_{2,n}^2, k \notin \widehat{I}_{1,n}, \, \overline{G}_{1,n} \cap D). \tag{90}$$

8.6.2 Final bracketing relation

It can be shown using some rather lengthy technical arguments (provided in the technical note) that, given appropriate γ_2, $\gamma_{2,+}$ and $\gamma_{2,-}$, for all sufficiently large n, except on a set of negligible probability, uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$,

$$\begin{cases} T_k < \gamma_{2,n}^2 & if \;\; k \notin I_n^+, \\ T_k > \gamma_{2,n}^2 & if \;\; k \in I_n^- \setminus I_{1,n}^-. \end{cases} \tag{91}$$

Once (91) is established, it follows from (89), (90), and some probabilistic bounds (also given in the technical note) that there exists n_{*6} such that for all $n \geq n_{*6}$,

$$\mathbb{P}(I_n^- \subset \widehat{I}_{1,n} \cup \widehat{I}_{2,n} \subset I_n^+, \, \overline{G}_{1,n}) \geq 1 - K_6(M)n^{-1-\epsilon_2(\kappa)}, \tag{92}$$

for some $K_6(M) > 0$ and $\epsilon_2(\kappa) > 0$. Moreover, the bound (92) is uniform in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$.

8.7 Second stage : perturbation analysis

The rest of this section deals with the part of the proof of Theorem 3 that involves analyzing the behavior of the submatrix of $\mathbf{S}$ that corresponds to the set of selected coordinates. To begin with, define $\widehat{I}_n := \widehat{I}_{1,n} \cup \widehat{I}_{2,n}$, and $\overline{G}_{3,n} := \{I_n^- \subset \widehat{I}_n \subset I_n^+\} \cap \overline{G}_{2,n}$. Then define

$$\widetilde{\mathbf{S}}_2 = \begin{bmatrix} \mathbf{S}_{\widehat{I}_n,\widehat{I}_n} & O \\ O & O \end{bmatrix} \qquad \mathbf{S}_2 = \begin{bmatrix} \mathbf{S}_{\widehat{I}_n,\widehat{I}_n} & O \\ O & I \end{bmatrix}. \tag{93}$$

In this section A will denote the set $\widehat{I}_n$, $B = \{1,\ldots,N\} \setminus A =: A^c$, $A_\pm = I_n^\pm$, $\overline{A}_- = A_- \setminus A$, $B_- = \{1,\ldots,N\} \setminus A_- =: A_-^c$. The first task before us is to derive an equivalent of Lemma 3. This is done in Section 8.8. The vector $H_\nu \mathbf{S}_2 \theta_\nu$ is expanded, and then the important terms are isolated in Section 8.9. Finally, the proof is completed in Section 8.10.

8.8 Eigen-analysis of $\mathbf{S}_2$

$\widehat{\theta}_\nu$, $\nu = 1,\ldots,M$ are the eigenvectors corresponding to the M largest (in decreasing order) eigenvalues of $\mathbf{S}_2$. As a convention $\langle \widehat{\theta}_\nu, \theta_\nu \rangle \geq 0$ for all $\nu = 1,\ldots,M$. Let the first M eigenvalues of $\widetilde{\mathbf{S}}_2$ be $\widetilde{\ell}_1 > \ldots > \widetilde{\ell}_M$. Then arguments similar to what are used in Section 8.4 establishes the following results.

On $\overline{G}_{3,n}$, for all $\mu = 1,\ldots,M$,

$$\| \theta_{\mu,A_-^c} \|^2 \leq \overline{\tau}_{n,\mu}^2 := c_q \gamma_{2,+}^{2-q} \frac{C_\mu^q (\log(n \vee N))^{1-q/2}}{(nh(\lambda_\mu))^{1-q/2}}, \tag{94}$$

and

$$|I_n^\pm| \leq J_{2,n}^\pm := \gamma_{2,\mp}^{-q} M^{q/2} (\sum_{\mu=1}^M h(\lambda_\mu)^{q/2} C_\mu^q) \left(\frac{n}{\log(n \vee N)} \right)^{q/2}. \tag{95}$$

Under **C1-C3**, as $n \to \infty$, for all $\nu = 1,\ldots,M$,

$$\frac{J_{2,n}^\pm}{nh(\lambda_\nu)} \leq \gamma_{2,\mp}^{-q/2} M^{q/2} \frac{\lambda_1^q}{\lambda_\nu^q} \frac{(\sum_{\mu=1}^M (\frac{\lambda_\mu}{\lambda_1})^{q/2} C_\mu^q)(\log(n \vee N))^{-q/2}}{(nh(\lambda_\nu))^{1-q/2}} \to 0; \tag{96}$$

and $\overline{\tau}_n := \max_{1 \leq \mu \leq M} \overline{\tau}_{n,\mu} \to 0$. Again, check that $|I_n^\pm|$ is bounded by N, and $\| \theta_{\mu,(I_n^-)^c} \|^2$ is bounded by $\gamma_{2,+}^2 N \log(n \vee N)(nh(\lambda_\mu))^{-1}$. This observation leads to the fact alluded to in Remark 5.2.

For $j = 1,\ldots,4$, define $\overline{\varepsilon}_{j,n}$ as $\varepsilon_{j,n}$ is defined in (78), with $J_{1,n}^+$ replaced by $J_{2,n}^+$. Then define

$$\overline{\varepsilon}_{5,n} = c_q \gamma_{2,+}^{1-q/2} \left[\sum_{\mu=1}^M \left(\frac{\eta(\lambda_1)}{\eta(\lambda_\mu)} \right)^{1-q/2} \left(\frac{\lambda_\mu}{\lambda_1} \right)^{q/2} C_\mu^q \right] \left(\frac{\log(n \vee N)}{nh(\lambda_1)} \right)^{1-q/2}. \tag{97}$$

It follows that $\max_{1 \leq j \leq 5} \overline{\varepsilon}_{j,n} \to 0$ as $n \to \infty$. Define

$$\overline{\Delta}_{n,\nu} = \frac{\lambda_1}{\max\{\lambda_{\nu-1} - \lambda_\nu, \lambda_\nu - \lambda_{\nu+1}\}} \left[\sum_{j=1}^5 \overline{\varepsilon}_{j,n} + \sqrt{\sum_{\mu=1}^M \frac{\lambda_\mu}{\lambda_1}} \sqrt{\overline{\varepsilon}_{5,n}} \right], \tag{98}$$

32

and $\overline{\Delta}_n = \max_{1 \leq \nu \leq M} \overline{\Delta}_{n,\nu}$. A result that summarizes the behavior of the first M eigenvalues of $\mathbf{S}_2$ can now be stated.

Proposition 3: *There is a measurable set $\overline{G}_{4,n} \subset \overline{G}_{3,n}$, and an integer $n_{*7} \geq n_{*6}$, such that, for all $n \geq n_{*7}$ the following relations hold, uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$.*

$$\overline{G}_{4,n} \quad \subset \quad \bigcap_{\nu=1}^M \{\widetilde{\ell}_\nu = \lambda_\nu(\mathbf{S}_2) \text{ and } |\widetilde{\ell}_\nu - (1 + \lambda_\nu)| \leq \lambda_1 \sum_{j=1}^5 \overline{\varepsilon}_{j,n}\}, \tag{99}$$

$$\overline{G}_{4,n} \quad \subset \quad \{\| \mathbf{S}_2 - \Sigma \| \leq \lambda_1 (\sum_{j=1}^5 \overline{\varepsilon}_{j,n} + \sqrt{\sum_{\mu=1}^M \frac{\lambda_\mu}{\lambda_1} \sqrt{\overline{\varepsilon}_{5,n}}})\} \tag{100}$$

$$1 - \mathbb{P}(\overline{G}_{4,n}) \quad \leq \quad K_7(M) n^{-1-\epsilon_3}, \tag{101}$$

for some constants $K_7(M) > 0$ and $\epsilon_3 > 0$. ϵ_3 depends of γ_1, $\overline{\gamma}_1$, γ_1', $a_\pm$, γ_2, $\gamma_{2,\pm}$, and κ.

At this point it is useful to define a quantity that will play an important role in the analysis in Section 8.9. Define,

$$\vartheta_{n,\mu}^2 = \overline{\tau}_{n,\mu}^2 + \frac{J_{2,n}^+}{nh(\lambda_\mu)} + \sum_{\mu' \neq \mu} \frac{1}{ng(\lambda_{\mu'}, \lambda_\mu)}, \quad \mu = 1, \ldots, M. \tag{102}$$

Then define $\vartheta_n = \max_{1 \leq \mu \leq M} \vartheta_{n,\mu}$ and observe that, under **C1-C2**, $\vartheta_n \to 0$ as $n \to \infty$.

We argue that, for $n \geq n_{*8}$, say, on a set $\overline{G}_{5,n}$ with probability approaching 1 sufficiently fast,

$$L(\widehat{\theta}_\nu, \theta_\nu) \leq \| H_\nu \mathbf{S}_2 \theta_\nu \|^2 (1 + \overline{\delta}_{n,N,\nu}), \tag{103}$$

where $\overline{\delta}_{n,N,\nu} = o(1)$. Therefore, it remains to show that, $\mathbb{E} \| H_\nu \mathbf{S}_2 \theta_\nu \|^2 \mathbf{1}_{\overline{G}_{5,n}}$ is bounded by the quantity appearing on the RHS of (24).

8.9 Analysis of $H_\nu \mathbf{S}_2 \theta_\nu$

In this section, as in Section 8.10, ν is going to be a fixed index in $\{1, \ldots, M\}$. Before an analysis of $H_\nu \mathbf{S}_2 \theta_\nu$ is carried out, a few important facts are stated below. Here C is any subset of $\{1, \ldots, N\}$ satisfying $A_- \subset C$.

$$|\delta_{\mu\nu} - \langle \theta_{\mu,C}, \theta_{\nu,C} \rangle| \quad = \quad |\langle \theta_{\mu,C^c} \theta_{\nu,C^c} \rangle| \leq \overline{\tau}_{n,\mu} \overline{\tau}_{n,\nu} \leq (\vartheta_{n,\mu} \vee \vartheta_{n,\mu}) \vartheta_n, \tag{104}$$

$$\max_{1 \leq \mu \leq M} \frac{|I_n^\pm|}{nh(\lambda_\mu)} \quad \leq \quad \frac{h(\lambda_\nu)}{h(\lambda_M)} \vartheta_{n,\nu}^2. \tag{105}$$

Further,

$$\max_{1 \leq \mu, \mu' \leq M} \| \theta_{\mu,C} \| \sqrt{\frac{\log n}{nh(\lambda_{\mu'})}} \leq \overline{\tau}_n \sqrt{\frac{\log n}{nh(\lambda_M)}} = O(\frac{\log n \sqrt{J_{2,n}^+}}{nh(\lambda_\nu)}) = o(\vartheta_{n,\nu}), \tag{106}$$

which follows from **C1**, **C2**, (94), (96), and (102).

Next, observe that $H_\nu \theta_\nu = 0$ implies that

$$H_\nu \mathbf{S}_2 \theta_\nu = H_\nu (\mathbf{S}_2 - \Sigma) \theta_\nu = \begin{bmatrix} H_{\nu,AA}(\mathbf{S}_{AA} - I)\theta_{\nu,A} \\ H_{\nu,BA}(\mathbf{S}_{AA} - I)\theta_{\nu,A} \end{bmatrix} = \Psi, \quad \text{say.} \tag{107}$$

Then Ψ_A and Ψ_B have the general form, for $C = A, B$,

$$
\begin{aligned}
\Psi_C &= \sum_{\mu=1}^M \frac{\| v_\mu \|^2}{n} \lambda_\mu \langle \theta_{\mu,A}, \theta_{\nu,A} \rangle H_{\nu,CA} \theta_{\mu,A} + \sum_{\mu=1}^M \sqrt{\lambda_\mu} \frac{1}{n} \langle \mathbf{Z}_A v_\mu, \theta_{\nu,A} \rangle H_{\nu,CA} \theta_{\mu,A} \\
&+ \sum_{\mu=1}^M \sqrt{\lambda_\mu} \langle \theta_{\mu,A}, \theta_{\nu,A} \rangle H_{\nu,CA} \frac{1}{n} \mathbf{Z}_A v_\mu + \sum_{\mu \neq \mu'} \frac{\langle v_\mu, v_{\mu'} \rangle}{n} \sqrt{\lambda_\mu \lambda_{\mu'}} \langle \theta_{\mu',A}, \theta_{\nu,A} \rangle H_{\nu,CA} \theta_{\mu,A} \\
&+ H_{\nu,CA}(\frac{1}{n} \mathbf{Z}_A \mathbf{Z}_A^T - I)\theta_{\nu,A}.
\end{aligned}
\tag{108}
$$

When C is either A or B, and δ_{CA} is 1 or 0 according as whether $C = A$ or not,

$$
\begin{aligned}
H_{\nu,CA}\theta_{\mu,A} &= \sum_{\nu' \neq \nu} \frac{1}{\lambda_{\nu'} - \lambda_\nu} \langle \theta_{\nu',A}, \theta_{\mu,A} \rangle \theta_{\nu',C} + \frac{1}{\lambda_\nu} \sum_{\nu' \neq \mu}^M \langle \theta_{\nu',A}, \theta_{\mu,A} \rangle \theta_{\nu',C} \\
&\qquad\qquad - \frac{1}{\lambda_\nu}(\delta_{CA} - \| \theta_{\mu,A} \|^2)\theta_{\mu,C};
\end{aligned}
\tag{109}
$$

$$
\begin{aligned}
H_{\nu,CA}\mathbf{Z}_A v_\mu &= \sum_{\nu' \neq \nu} \frac{1}{\lambda_{\nu'} - \lambda_\nu} \langle \mathbf{Z}_A v_\mu, \theta_{\nu',A} \rangle \theta_{\nu',C} \\
&\qquad\qquad - \frac{1}{\lambda_\nu}(\delta_{CA} I - \sum_{\nu'=1}^M \theta_{\nu',C}\theta_{\nu',A})\mathbf{Z}_A v_\mu;
\end{aligned}
\tag{110}
$$

$$
\begin{aligned}
H_{\nu,CA}(\frac{1}{n}\mathbf{Z}_A \mathbf{Z}_A^T - I)\theta_{\nu,A} &= \sum_{\nu' \neq \nu} \frac{1}{\lambda_{\nu'} - \lambda_\nu} \left(\frac{1}{n} \langle \mathbf{Z}_A^T \theta_{A,\nu'}, \mathbf{Z}_A^T \theta_{A,\nu} \rangle - \langle \theta_{\nu',A}, \theta_{\nu,A} \rangle \right) \theta_{\nu',C} \\
&\qquad - \frac{1}{\lambda_\nu}(\delta_{CA} I - \sum_{\nu'=1}^M \theta_{\nu',C}\theta_{\nu',A})(\frac{1}{n}\mathbf{Z}_A \mathbf{Z}_A^T - I)\theta_{\nu,A}.
\end{aligned}
\tag{111}
$$

A further expansion of terms Ψ_A and Ψ_B can be computed, but at this point it is beneficial to isolate the important terms in the expansion. Accordingly, use Lemmas 9 - 13, together with (104), (105) and (106) to deduce that, there is a measurable set $\overline{G}_{5,n} \subset \overline{G}_{4,n}$, constants $K_8(M) > 0$, $\epsilon_4 > 0$ and an $n_{*8} \geq n_{*7}$ such that, $1 - \mathbb{P}(\overline{G}_{5,n}) \leq K_8(M)n^{-1-\epsilon_4}$, for $n \geq n_{*8}$, and

$$\Psi = \Psi_0 + \Psi_I + \Psi_{II} + \Psi_{III} + \Psi_{IV} + \Psi_{rem}, \tag{112}$$

where $\| \Psi_{rem} \| \leq b_n \vartheta_{n,\nu}$, with $b_n = o(1)$, and the other elements are described below.

$$\Psi_{0,A} = 0 \quad \text{and} \quad \Psi_{0,B} = \theta_{\nu,B}. \tag{113}$$

34

$\Psi_I = \sum_{\mu \neq \nu}^{M} w_{\mu\nu} \theta_\mu$ where $w_{\mu\nu}$ equals

$$\frac{\sqrt{\lambda_\mu}}{\lambda_\mu - \lambda_\nu} \frac{1}{n} \langle \mathbf{Z}_{A_-} v_\mu, \theta_{\nu,A_-} \rangle + \frac{\sqrt{\lambda_\nu}}{\lambda_\mu - \lambda_\nu} \frac{1}{n} \langle \mathbf{Z}_{A_-} v_\nu, \theta_{\mu,A_-} \rangle$$

$$+ \frac{\sqrt{\lambda_\mu \lambda_\nu}}{\lambda_\mu - \lambda_\nu} \frac{\langle v_\mu, v_\nu \rangle}{n} + \frac{1}{\lambda_\mu - \lambda_\nu} \left(\frac{1}{n} \langle \mathbf{Z}_{A_-}^T \theta_{A_-,\mu}, \mathbf{Z}_{A_-}^T \theta_{A_-,\nu} \rangle - \langle \theta_{\mu,A_-}, \theta_{\nu,A_-} \rangle \right) \qquad (114)$$

$$\Psi_{II} = -\frac{1}{\lambda_\nu}(I - \sum_{\mu=1}^{M} \theta_\mu \theta_\mu^T)(\frac{1}{n}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T - \Xi)\theta_\nu, \qquad (115)$$

where $\widetilde{\mathbf{Z}}_{A_-} = \mathbf{Z}_{A_-}$ and $\widetilde{\mathbf{Z}}_{A_-^c} = O$, i.e. a matrix whose entries are all 0; and Ξ is a $N \times N$ matrix whose (A_-, A_-) block is identity and the rest are all zero.

$$\Psi_{III} = -\frac{1}{\sqrt{\lambda_\nu}}(I - \sum_{\mu=1}^{M} \theta_\mu \theta_\mu^T)\frac{1}{n}\widetilde{\mathbf{Z}}v_\nu. \qquad (116)$$

Ψ_{IV} is such that $\Psi_{IV,A_-} = 0$, $\Psi_{IV,B} = 0$, and

$$\Psi_{IV,\overline{A}_-} = -\frac{1}{n}\mathbf{Z}_{\overline{A}_-}\left(\frac{1}{\sqrt{\lambda_\nu}}v_\nu + \frac{1}{\lambda_\nu}\mathbf{Z}_{A_-}^T\theta_{\nu,A_-}\right). \qquad (117)$$

8.10 Completion of the proof of Theorem 3

Suppose without loss of generality that n_{*8} in Section 8.9 is large enough so that $\overline{\Delta}_n < \frac{\sqrt{5}-1}{4}$. Since on $\overline{G}_{5,n}$, $\| \mathbf{S}_2 - \Sigma \| \leq \min\{\lambda_\nu - \lambda_{\nu+1}, \lambda_{\nu-1} - \lambda_\nu\}\overline{\Delta}_n$, where $\lambda_0 = \infty$ and $\lambda_{M+1} = 0$, argue that, by Lemma 6, for $n \geq n_{*8}$, on $\overline{G}_{5,n}$,

$$L(\widehat{\theta}_\nu, \theta_\nu) \leq \| H_\nu \mathbf{S}_2 \theta_\nu \|^2 (1 + \overline{\delta}_{n,N,\nu}), \qquad (118)$$

where $\overline{\delta}_{n,N,\nu} = o(1)$. Therefore, it remains to show that, $\mathbb{E} \| H_\nu \mathbf{S}_2 \theta_\nu \|^2 \mathbf{1}_{\overline{G}_{5,n}}$ is bounded by the quantity appearing on the RHS of (24). In view of the fact that, this upper bound is within a constant multiple of $\vartheta_{n,\nu}^2$, and $\| \Psi_{rem} \| = o(\vartheta_{n,\nu})$ on $\overline{G}_{5,n}$, it is enough that the same bound holds for $\mathbb{E} \| \Psi - \Psi_{rem} \|^2 \mathbf{1}_{\overline{G}_{5,n}}$.

Observe that Ψ_I, Ψ_{II}, and Ψ_{III} are mutually uncorrelated vectors. Also, by (94), $\mathbb{E} \| \Psi_0 \|^2 \mathbf{1}_{\overline{G}_{5,n}} \leq \overline{\tau}_{n,\nu}^2$. Therefore,

$$\mathbb{E} \| \Psi - \Psi_{rem} \|^2 \mathbf{1}_{\overline{G}_{5,n}}$$
$$\leq \overline{\tau}_{n,\nu}^2 + \mathbb{E} \| \Psi_I \|^2 + \mathbb{E} \| \Psi_{II} \|^2 + \mathbb{E} \| \Psi_{III} \|^2 + \mathbb{E} \| \Psi_{IV} \|^2 \mathbf{1}_{\overline{G}_{5,n}}$$
$$+ 2\mathbb{E}|\langle \Psi_0, \Psi_I + \Psi_{II} + \Psi_{III}\rangle|\mathbf{1}_{\overline{G}_{5,n}} + 2\mathbb{E}|\langle \Psi_{IV}, \Psi_0 + \Psi_I + \Psi_{II} + \Psi_{III}\rangle|\mathbf{1}_{\overline{G}_{5,n}} \qquad (119)$$

Observe that, $\Psi_{II,A_-} = -\frac{1}{\lambda_\nu}(I - \sum_{\mu=1}^{M} \theta_{\mu,A_-}\theta_{\mu,A_-}^T)(\frac{1}{n}\mathbf{Z}_{A_-}\mathbf{Z}_{A_-} - I)\theta_{\nu,A_-}$,

$$\Psi_{II,A_-^c} = \frac{1}{\lambda_\nu} \sum_{\mu=1}^{M} \left(\frac{1}{n} \langle \mathbf{Z}_{A_-}^T \theta_{\mu,A_-}, \mathbf{Z}_{A_-}^T \theta_{\nu,A_-} \rangle - \langle \theta_{\mu,A_-}, \theta_{\nu,A_-} \rangle \right) \theta_{\mu,A_-^c},$$

35

and

$$\Psi_{III,A_-} = -\frac{1}{\sqrt{\lambda_\nu}}(I - \sum_{\mu=1}^{M} \theta_{\mu,A_-}\theta_{\mu,A_-}^T)\frac{1}{n}\mathbf{Z}_{A_-}v_\nu, \ \Psi_{III,A_-^c} = \frac{1}{\sqrt{\lambda_\nu}}\sum_{\mu=1}^{M}\frac{1}{n}\langle\mathbf{Z}_{A_-}v_\nu, \theta_{\mu,A_-}\rangle\theta_{\mu,A_-^c}.$$

Thus, by a further application of Lemmas 9-12, it can be checked that, there is an integer $n_{*9} \geq n_{*8}$, and an event $\overline{G}_{6,n} \subset \overline{G}_{5,n}$ such that, for $n \geq n_{*9}$, on $\overline{G}_{6,n}$,

$$|\langle\Psi_0, \Psi_I + \Psi_{II} + \Psi_{III}\rangle| + |\langle\Psi_{IV}, \Psi_0 + \Psi_I + \Psi_{II} + \Psi_{III}\rangle| \leq b_n'\vartheta_{n,\nu}^2, \tag{120}$$

with $b_n' = o(1)$; and $\mathbb{P}(\overline{G}_{6,n}^c \cap \overline{G}_{5,n}) \leq n^{-2(1+\epsilon_5)}$, for some constants $K_9(M) > 0$ and $\epsilon_5 > 0$. On $\overline{G}_{5,n}$,

$$\| \Psi_{IV} \|^2 \leq \frac{1}{n^2} \| \mathbf{Z}_{A_{+/-}}\left(\frac{1}{\sqrt{\lambda_\nu}}v_\nu + \frac{1}{\lambda_\nu}\mathbf{Z}_{A_-}^T\theta_{\nu,A_-}\right) \|^2,$$

where $A_{+/-} := A_+ \setminus A_-$, and the (unrestricted) expectation of the random variable appearing in the upper bound is bounded by $\frac{|I_n^+| - |I_n^-|}{nh(\lambda_\nu)}$. From this, and some expectation computations similar to those in Section 7, deduce that,

$$\overline{\tau}_{n,\nu}^2 + \mathbb{E} \| \Psi_I \|^2 + \mathbb{E} \| \Psi_{II} \|^2 + \mathbb{E} \| \Psi_{III} \|^2 + \mathbb{E} \| \Psi_{IV} \|^2 \mathbf{1}_{\overline{G}_{5,n}} \leq \vartheta_{n,\nu}^2(1 + o(1)). \tag{121}$$

Finally, express the event $\overline{G}_{5,n}$ as (disjoint) union of $\overline{G}_{5,n} \cap \overline{G}_{6,n}$ and $\overline{G}_{5,n} \cap \overline{G}_{6,n}^c$; apply the bound (120) for the first set, and use Cauchy-Schwartz inequality for the second set, to conclude that,

$$\mathbb{E}|\langle\Psi_0, \Psi_I + \Psi_{II} + \Psi_{III}\rangle|\mathbf{1}_{\overline{G}_{5,n}} + \mathbb{E}|\langle\Psi_{IV}, \Psi_0 + \Psi_I + \Psi_{II} + \Psi_{III}\rangle|\mathbf{1}_{\overline{G}_{5,n}}$$
$$\leq \ b_n'\vartheta_{n,\nu}^2 + 2\sqrt{K_9(M)}n^{-1-\epsilon_5}\vartheta_n = o(\vartheta_{n,\nu}^2). \tag{122}$$

Combine (119), (121) and (122) to complete the proof.

9 Appendix

Some results that are needed to prove the three theorems are presented here.

9.1 Deviation of extreme eigenvalues

The goal is to provide a probabilistic bound for deviations of $\| \frac{1}{n}\mathbf{Z}\mathbf{Z}^T - I \|$. This is achieved through the following lemma.

Lemma 5: *Let $t_n = 6(\frac{N}{n} \vee 1)\sqrt{\frac{\log(n \vee N)}{n \vee N}}$. Then, for any $c > 0$, there exists $n_c \geq 1$ such that, for all $n \geq n_c$,*

$$\mathbb{P}\left(\| \frac{1}{n}\mathbf{Z}\mathbf{Z}^T - I \| > 2\sqrt{\frac{N}{n}} + \frac{N}{n} + ct_n\right) \leq 2(n \vee N)^{-c^2}. \tag{123}$$

Proof : By definition,

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

From Proposition 4 (due to Davidson and Szarek (2001)), and its consequence, Corollary 2, given below, it follows that,

$$\mathbb{P}\left(\| \frac{1}{n}\mathbf{Z}\mathbf{Z}^T - I \| > 2\sqrt{\frac{N}{n}} + \frac{N}{n} + ct_n \right)$$

$$\leq \exp\left(-\frac{nc^2t_n^2}{8(ct_n + (1 + \sqrt{N/n})^2)} \right) + \exp\left(-\frac{nc^2t_n^2}{8(ct_n + (1 - \sqrt{N/n})^2)} \right). \qquad (124)$$

First suppose that $n \geq N$. Then for n large enough, $ct_n < \frac{1}{2}$, so that

$$\frac{nc^2t_n^2}{8(ct_n + (1 + \sqrt{N/n})^2)} \geq \frac{nc^2t_n^2}{36}, \quad \text{and} \quad \frac{nc^2t_n^2}{8(ct_n + (1 - \sqrt{N/n})^2)} \geq \frac{nc^2t_n^2}{12}.$$

Since in this case $nt_n^2 = 36\log n$, (123) follows from (124). If $N > n$, then $\lambda_N(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T) = 0$, and

$$\frac{nc^2t_n^2}{8(ct_n + (1 \pm \sqrt{N/n})^2)} = \frac{Nc^2(\frac{n}{N}t_n)^2}{8(c\frac{n}{N}t_n + (1 \pm \sqrt{n/N})^2)} \ ,$$

and therefore, (123) follows if the roles of n and N are reversed.

Proposition 4: *Let Z be a $p \times q$ matrix of i.i.d. $N(0,1)$ entries with $p \leq q$. Let $s_{max}(Z)$ and $s_{min}(Z)$ denote the largest and the smallest singular value of Z, respectively. Then,*

$$\mathbb{P}(s_{max}(\frac{1}{\sqrt{q}}Z) > 1 + \sqrt{p/q} + t) \leq e^{-qt^2/2}, \qquad (125)$$

$$\mathbb{P}(s_{min}(\frac{1}{\sqrt{q}}Z) < 1 - \sqrt{p/q} - t) \leq e^{-qt^2/2}. \qquad (126)$$

Corollary 2: *Let $\mathbf{S} = \frac{1}{q}ZZ^T$ where Z is as in Proposition 4, with $p \leq q$. Let $m_1(p,q) := (1 + \sqrt{\frac{p}{q}})^2$ and $m_p(p,q) := (1 - \sqrt{\frac{p}{q}})^2$. Let $\lambda_1(\mathbf{S})$ and $\lambda_p(\mathbf{S})$ denote the largest and the smallest eigenvalues of $\mathbf{S}$. Then, for $t > 0$,*

$$\mathbb{P}(\lambda_1(\mathbf{S}) - m_1(p,q) > t) \leq \exp\left(-\frac{q}{2}(\sqrt{t + m_1(p,q)} - \sqrt{m_1(p,q)})^2 \right)$$

$$\leq \exp\left(-\frac{qt^2}{8(t + m_1(p,q))} \right), \qquad (127)$$

and

$$\mathbb{P}(\lambda_p(\mathbf{S}) - m_p(p,q) < -t) \leq \exp\left(-\frac{q}{2}(\sqrt{t + m_p(p,q)} - \sqrt{m_p(p,q)})^2 \right)$$

$$\leq \exp\left(-\frac{qt^2}{8(t + m_p(p,q))} \right). \qquad (128)$$

9.2 Perturbation of eigen-structure

The following lemma is most convenient for the risk analysis of estimators of θ_ν. Several variants of this lemma appear in the literature (Kneip and Utikal (2001), Tyler (1983), Tony Cai and Hall (2005)) and most of them implicitly use the approach proposed by Kato (1980).

Lemma 6: *For some $T \in \mathbb{N}$, let A and B be two symmetric $T \times T$ matrices. Let the eigenvalues of matrix A be denoted by $\lambda_1(A) \geq \ldots \geq \lambda_T(A)$. Set $\lambda_0(A) = \infty$ and $\lambda_{T+1}(A) = -\infty$. For any $r \in \{1, \ldots, T\}$, if $\lambda_r(A)$ is a unique eigenvalue of A, i.e., if $\lambda_{r-1}(A) > \lambda_r(A) > \lambda_{r+1}(A)$, then denoting by $\mathbf{p}_r$ the eigenvector associated with the r-th eigenvalue,*

$$\mathbf{p}_r(A+B) - sign(\mathbf{p}_r(A+B)^T\mathbf{p}_r(A))\mathbf{p}_r(A) = -H_r(A)B\mathbf{p}_r(A) + R_r \tag{129}$$

where $H_r(A) := \sum_{s \neq r} \frac{1}{\lambda_s(A) - \lambda_r(A)} P_{\mathcal{E}_s}(A)$ and $P_{\mathcal{E}_s}(A)$ denotes the projection matrix onto the eigenspace $\mathcal{E}_s$ corresponding to eigenvalue $\lambda_s(A)$ (possibly multi-dimensional). Define Δ_r and $\overline{\Delta}_r$ as

$$\Delta_r := \frac{1}{2}[\| H_r(A)B \| + |\lambda_r(A+B) - \lambda_r(A)| \, \| H_r(A) \|] \tag{130}$$

$$\overline{\Delta}_r = \frac{\| B \|}{\min_{1 \leq j \neq r \leq T} |\lambda_j(A) - \lambda_r(A)|} \, . \tag{131}$$

Then, the residual term R can be bounded by

$$\| R_r \| \leq \min\{10\overline{\Delta}_r^2, \quad \| H_r(A)B\mathbf{p}_r(A) \| \left[\frac{2\Delta_r(1+2\Delta_r)}{1 - 2\Delta_r(1+2\Delta_r)} + \frac{\| H_r(A)B\mathbf{p}_r(A) \|}{(1 - 2\Delta_r(1+2\Delta_r))^2} \right]\} \tag{132}$$

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

9.3 Proof of Proposition 1

Proof : For n i.i.d. observations $X_i, i = 1, \ldots, n$, the KL discrepancy of the data is just n times the KL discrepancy for a single observation. Therefore, w.l.o.g. take $n = 1$. Direct computation yields

$$\Sigma^{-1} = (I - \sum_{\nu=1}^{M} \eta(\lambda_\nu)\theta_\nu\theta_\nu^T). \tag{133}$$

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

$$\begin{aligned}
\log f(x|\theta) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}x^T\Sigma^{-1}x \\
&= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{\nu=1}^{M}\log(1+\lambda_\nu) - \frac{1}{2}(\langle x, x \rangle - \sum_{\nu=1}^{M}\eta(\lambda_\nu)\langle x, \theta_\nu \rangle^2).
\end{aligned} \tag{134}$$

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 from F_1, to be denoted by $K(F_1, F_2)$, is given by

$$K(F_1, F_2) = \int \log \frac{f_1(y)}{f_2(y)} f_1(y) dy. \tag{135}$$

38

Hence, from (134),

$$
\begin{aligned}
K_{1,2} &= \mathbb{E}_{\theta^{(1)}}(\log f(X|\theta^{(1)} - \log f(X|\theta^{(2)})) \\
&= \frac{1}{2}\sum_{\nu=1}^{M}\eta(\lambda_\nu)[\mathbb{E}_{\theta^{(1)}}(\langle X, \theta_\nu^{(1)}\rangle)^2 - \mathbb{E}_{\theta^{(1)}}(\langle X, \theta_\nu^{(2)}\rangle)^2] \\
&= \frac{1}{2}\sum_{\nu=1}^{M}\eta(\lambda_\nu)[\langle \theta_\nu^{(1)}, \Sigma_{(1)}\theta_\nu^{(1)}\rangle - \langle \theta_\nu^{(2)}, \Sigma_{(1)}\theta_\nu^{(2)}\rangle] \\
&= \frac{1}{2}\sum_{\nu=1}^{M}\eta(\lambda_\nu)[(\|\theta_\nu^{(1)}\|^2 - \|\theta_\nu^{(2)}\|^2)^2 + \sum_{\nu'=1}^{M}\lambda_{\nu'}\{\|\theta_{\nu'}^{(1)}\|^2 - (\langle\theta_{\nu'}^{(1)}, \theta_\nu^{(2)}\rangle)^2\}],
\end{aligned}
$$

which equals the RHS of (26), since the columns of $\theta^{(j)}$ are orthonormal for each $j = 1, 2$.

9.4 A counting lemma

Lemma 7: *Suppose that m, N are positive integers, such that $m \to \infty$ as $N \to \infty$ and $m = o(N)$. Let $\widetilde{Z}$ be the maximal set of points in $\mathbb{R}^N$ satisfying the following conditions:*

(i) for each $\mathbf{z} = (z_1, \ldots, z_N) \in \widetilde{Z}$, $z_i \in \{0,1\}$ for all $i = 1, \ldots, N$,

(ii) for each $\mathbf{z} \in \widetilde{Z}$, exactly m of coordinates of $\mathbf{z}$ are 1,

(iii) for every pair $\mathbf{z}$ and $\mathbf{z}'$ in $\widetilde{Z}$, $z_i = z_i'$ for at most $\left[\frac{m_0}{2}\right] =: k(m_0) - 1$ (i.e. $k(m_0)$ is the largest integer $\leq m_0/2 + 1$) nonzero coordinates, where $m_0 = [\beta m]$, for some $\beta \in (0,1)$.

Then cardinality of $\widetilde{Z}$ is at least $\exp([N\mathcal{E}(\frac{\beta m}{2N}) - 2m\mathcal{E}(\frac{\beta}{2})](1 + o(1)))$ where $\mathcal{E}(x)$ is the Shannon entropy function.

Proof : Trivially, $\widetilde{Z} \subset Z^*$, where Z^* is the set of *all* points $\mathbf{z}$ satisfying (i) and (ii). Thus, $|\widetilde{Z}| < |Z^*| = \binom{N}{m}$. On the other hand, for every point $\mathbf{z} \in Z^*$ there are *at most*

$$
g(N, m, m_0) = \binom{m}{k(m_0)}\binom{N - k(m_0)}{m - k(m_0)}
$$

points $\mathbf{w} \in Z^*$ such that *at least* $k(m_0)$ nonzero coordinates of $\mathbf{z}$ and $\mathbf{w}$ match. This is because, one can fix the m nonzero coordinates of $\mathbf{z}$ and demand that in $k(m_0)$ of those coordinates w_i must equal 1. Other $m - k(m_0)$ nonzero coordinates of $\mathbf{w}$ can therefore be chosen from the rest

$N - k(m_0)$ coordinates. Then, by the maximality of $\widetilde{Z}$, as $N \to \infty$,

$$
\begin{aligned}
|\widetilde{Z}| &\geq \binom{N}{m} g(N, m, m_0)^{-1} \\
&= \frac{N!}{(N-m)!m!} \frac{k(m_0)!(m-k(m_0))!}{m!} \frac{(m-k(m_0))!(N-m)!}{(N-k(m_0))!} \\
&= \frac{N!}{k(m_0)!(N-k(m_0))!} \left(\frac{k(m_0)!(m-k(m_0))!}{m!} \right)^2 \\
&\sim \sqrt{2\pi}\sqrt{k(m_0)} \frac{(m-k(m_0))N^{1/2}}{m(N-k(m_0))^{1/2}} \left(\frac{N}{k(m_0)}\right)^{k(m_0)} \left(\frac{N}{N-k(m_0)}\right)^{N-k(m_0)} \\
&\quad \times \left[\left(\frac{k(m_0)}{m}\right)^{k(m_0)}\left(\frac{m-k(m_0)}{m}\right)^{m-k(m_0)}\right]^2 \quad \text{(by Stirling's formula)} \\
&= \sqrt{2\pi}\sqrt{\frac{\beta m}{2}} \exp\left[N\mathcal{E}\left(\frac{\beta m}{2N}\right)(1+o(1))\right] \exp\left[-2m\mathcal{E}\left(\frac{\beta}{2}\right)(1+o(1))\right]. \quad (136)
\end{aligned}
$$

Where the last equality is because, for large m, $\frac{m_0}{m} \sim \frac{\beta}{2}$.

9.5 Some auxiliary lemmas

In the following lemmas we provide probabilistic bounds for the deviations of certain quadratic forms that arise in the analysis of the residual terms in the expansion of $\widehat{\theta}_\nu$. Many of these involve the random sets, either $\widehat{I}_{1,n}$ or $\widehat{I}_{2,n}$, of coordinates that are selected under the ASPCA scheme. It will be assumed that the quantities involved are all measurable w.r.t. the joint distribution of $\mathbf{Z}$ and $v_1, \ldots, v_M$, though it will not be made explicit in the description or the proof of the lemmas. The bounds hold uniformly in $\theta \in \Theta_q^M(C_1, \ldots, C_M)$.

Lemma 8: *Let $\epsilon_n > 0$. Let A denote the random set $\widehat{I}_{n,1}$, and $A_- = I_{1,n}^-$ and $A_+ = I_{1,n}^+$. Assume that $A_- \subset A_+ \setminus \{k\}$, for some $1 \leq k \leq N$. For any subset C of $\{1, \ldots, N\}$, let $Y_C := Y_C(\mathbf{Z}_C, V)$ be a random vector jointly measurable w.r.t. $\mathbf{Z}_C$ and $V = [v_1 : \ldots : v_M]$. Assume that for each C, either $\mathbb{P}_V(Y_C = 0) = 0$ a.e. V, or $\mathbb{P}_V(Y_C = 0) = 1$ a.e. V, where $\mathbb{P}_V$ denotes the conditional probability w.r.t. V. Let $W_{k,C} = \langle Z_k, \frac{Y_C}{\|Y_C\|} \rangle$ if $Y_C \neq 0$, and $W_{k,C} = 0$ otherwise. Then,*

$$
\mathbb{P}\left(|W_{k,A}| > \epsilon_n, A_- \subset A \subset A_+ \setminus \{k\}, \|V\| \leq \beta_n\right) \leq \frac{2}{a_n} \Phi(-\epsilon_n), \quad (137)
$$

where β_n is such that, on $\{\|V\| \leq \beta_n\}$, a.e. V,

$$
\mathbb{P}_V(\widehat{\sigma}_{kk} \leq 1 + \gamma_{1,n}) \geq a_n > 0. \quad (138)
$$

Lemma 9: *Let A be a random subset of $\{1, \ldots, N\}$ and $A_- \subset A_+$ be two non-random subsets of $\{1, \ldots, N\}$. Let, $k_\pm$ denote the size of the set $A_\pm$, and*

$$
\epsilon_n = \sqrt{c_1 \log n} + \|\theta_{\nu, A_-^c}\| \sqrt{c_1 \log n + 2k_+ \log 2},
$$

for some $c_1 > 0$. Then, for all $1 \le \mu \le M$,

$$\mathbb{P}\left(|\langle \frac{\mathbf{Z}_A v_\mu}{\| v_\mu \|}, \theta_{\nu,A}\rangle| > \epsilon_n, A_- \subset A \subset A_+\right) \le \frac{4n^{-c_1/2}}{\sqrt{2\pi}\sqrt{c_1 \log n}}. \tag{139}$$

Lemma 10: *Let A, $A_\pm$, $k_\pm$, and $\overline{A}_-$ be as in Lemma 9. Let*

$$\epsilon_n = \| \theta_{\mu,A_-^c} \| (1 + \sqrt{\frac{k_+}{n}} + \sqrt{\frac{c_2 \log n}{n}})\sqrt{\frac{c_1 \log n + 2k_+ \log 2}{n}},$$

where $c_1, c_2 > 0$. Then

$$\mathbb{P}\left(|\frac{1}{n}\langle \mathbf{Z}_{A_-}^T \theta_{\nu,A_-}, \mathbf{Z}_{\overline{A}_-}^T \theta_{\mu,\overline{A}_-}\rangle| > \epsilon_n, \ A_- \subset A \subset A\right) \le \frac{2n^{-c_1/2}}{\sqrt{2\pi}\sqrt{c_1 \log n}} + n^{-c_2/2}. \tag{140}$$

Lemma 11: *Let A, $A_\pm$, $k_\pm$ be as in Lemma 9. Let, $t_n = 6(\frac{k_+}{n} \vee 1)\sqrt{\frac{\log(n \vee k_+)}{n \vee k_+}}$. Let*

$$\begin{aligned}
\epsilon_n &= \sqrt{\frac{c_1 \log n}{n}} + \| \theta_{\nu,A_-^c} \|^2 (2\sqrt{\frac{k_+}{n}} + \frac{k_+}{n} + \sqrt{c_2/2}t_n) \\
&\quad + 2 \| \theta_{\nu,A_-^c} \| (1 + \sqrt{\frac{k_+}{n}} + \sqrt{\frac{c_2 \log n}{n}})\sqrt{\frac{c_2 \log n + 2k_+ \log 2}{n}},
\end{aligned}$$

for some $c_1, c_2 > 0$. Then there is an $n(c_2) \ge 16$ such that, for $n \ge n(c_2)$, $\sqrt{c_2/2}t_n < 1/2$, and

$$\mathbb{P}\left(|\frac{1}{n} \| \mathbf{Z}_A^T \theta_{\nu,A} \|^2 - \| \theta_{\nu,A} \|^2 | > \epsilon_n, \ A_- \subset A \subset A_+\right)$$
$$\le 2n^{-c_1/4} + \frac{2n^{-c_2/2}}{\sqrt{2\pi}\sqrt{c_1 \log n}} + n^{-c_2/2} + 2(n \vee k_+)^{-c_2/2}. \tag{141}$$

Lemma 12: *Let A, $A_\pm$, $k_\pm$ be as in Lemma 9. Let, $\mu \ne \nu$, and for some $t > 0$,*

$$\begin{aligned}
\epsilon_n &= \sqrt{\frac{c_1 \log n}{n}} + (\| \theta_{\mu,A_-^c} \| + \| \theta_{\nu,A_-^c} \|)(1 + \sqrt{\frac{k_+}{n}} + \sqrt{\frac{c_3 \log n}{n}})\sqrt{\frac{c_2 \log n + 2k_+ \log 2}{n}} \\
&\quad + \| \theta_{\nu,A_-^c} \|\| \theta_{\mu,A_-^c} \| \sqrt{\frac{c_2 \log n}{n}} + \| \theta_{\nu,A_-^c} \|\| \theta_{\mu,A_-^c} \| (2\sqrt{\frac{k_+}{n}} + \frac{k_+}{n} + \sqrt{c_3/2}t_n),
\end{aligned}$$

where $c_1, c_2, c_3 > 0$ and t_n is as in Lemma 11. Then, there is $n(c_3) \ge 16$ such that, for $n \ge n(c_3)$, $\sqrt{c_3}t_n < \frac{1}{2}$, and

$$\mathbb{P}\left(|\frac{1}{n}\langle \mathbf{Z}_A^T \theta_{\nu,A}, \mathbf{Z}_A^T \theta_{\mu,A}\rangle - \langle \theta_{\nu,A}, \theta_{\mu,A}\rangle| > \epsilon_n, A_- \subset A \subset A_+\right)$$
$$\le 2n^{-3c_1/2+O(\frac{\log n}{n})} + 2n^{-c_2/4} + \frac{2n^{-c_2/2}}{\sqrt{2\pi}\sqrt{c_1 \log n}} + n^{-c_3/2} + 2(n \vee k_+)^{-c_3/2}. \tag{142}$$

Lemma 13: *Let A, $A_\pm$, $k_\pm$ be as in Lemma 9. Let,*

$$\epsilon_n \;=\; 2 \parallel \theta_{\mu,A_-^c} \parallel (1 + \sqrt{\frac{k_+}{n}} + \sqrt{\frac{c_2 \log n}{n}}) \sqrt{\frac{k_+}{n}} \left(1 + \sqrt{\log 2 + \frac{c_1 \log n}{4k_+}}\right)^{1/2},$$

where $c_1, c_2 > 0$. Also, suppose that $k_+ \geq 16$. Then,

$$\mathbb{P}(\frac{1}{n} \parallel \mathbf{Z}_{A_-} \mathbf{Z}_{\overline{A}_-}^T \theta_{\mu,\overline{A}_-} \parallel > \epsilon_n, A_- \subset A \subset A_+) \leq n^{-c_1/4} + n^{-c_2/2}. \tag{143}$$

9.6 Deviation of quadratic forms

The following lemma is due to Johnstone (2001b).

Lemma 14: *Let $\chi^2_{(n)}$ denote a Chi-square random variable with n degrees of freedom. Then,*

$$\mathbb{P}(\chi^2_{(n)} > n(1+\epsilon)) \;\leq\; e^{-3n\epsilon^2/16} \qquad (0 < \epsilon < \frac{1}{2}), \tag{144}$$

$$\mathbb{P}(\chi^2_{(n)} < n(1-\epsilon)) \;\leq\; e^{-n\epsilon^2/4} \qquad (0 < \epsilon < 1), \tag{145}$$

$$\mathbb{P}(\chi^2_{(n)} > n(1+\epsilon)) \;\leq\; \frac{\sqrt{2}}{\epsilon\sqrt{n}} e^{-n\epsilon^2/4} \qquad (0 < \epsilon < n^{1/16}, n \geq 16). \tag{146}$$

The following lemma is from Johnstone and Lu (2004).

Lemma 15: *Let $y_{1i}, y_{2i}, i = 1, \ldots, n$ be two sequences of mutually independent, i.i.d. $N(0,1)$ random variables. Then for large n and any b s.t. $0 < b \ll \sqrt{n}$,*

$$\mathbb{P}(|\frac{1}{n} \sum_{i=1}^{n} y_{1i} y_{2i}| > \sqrt{b/n}) \leq 2\exp\{-\frac{3b}{2} + O(n^{-1}b^2)\}. \tag{147}$$

Reference

1. Anderson, T. W. (1963) : Asymptotic theory of principal component analysis, *Annals of Mathematical Statistics*, **34**, 122-148.

2. Bai, J. (2003) : Factor models for large dimensions, *Econometrica*, **71**, 135-171.

3. Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) : Prediction by supervised principal components, *Journal of the American Statistical Association*, **101**, 119-137.

4. Birgé, L. (2001) : A new look at an old result : Fano's lemma, *Technical Report*, Université Paris 6.

5. Boente, G. and Fraiman, R. (2000) : Kernel-based functional principal components, *Statistics and Probability Letters*, **48**, 335-345.

6. Buja, A. and Hastie, T. and Tibshirani, R. (1995) : Penalized discriminant analysis, *Annals of Statistics*, **23**, 73-102.

7. Cassou, C., Deser, C., Terraty, L., Hurrell, J. W. and Drévillon, M. (2004) : Summer sea surface temperature conditions in the North Atlantic and their impact upon the atmospheric circulation in early winter, *Journal of Climate*, **17**, 3349-3363.

8. Cardot, H. (2000) : Nonparametric estimation of smoothed principal components analysis of sampled noisy functions, *Journal of Nonparametric Statistics*, **12**, 503-538.

9. Cardot, H., Ferraty, F. and Sarda, P. (2003) : Spline estimators for the functional linear model, *Statistica Sinica*, **13**, 571-591.

10. Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2004) : Functional response model, *Statistica Sinica*, **14**, 675-693.

11. Cootes, T. F., Edwards, G. J. and Taylor, C. J. (2001) : Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 681-685.

12. Corti, S., Molteni, F. and Palmer, T. N. (1999) : Signature of recent climate change in frequencies of natural atmospheric circulation regimes, *Nature*, **398**, 799-802.

13. Davidson, K. R. and Szarek, S. (2001) : Local operator theory, random matrices and Banach spaces, in *Handbook on the Geometry of Banach Spaces*, **1**, *Eds. Johnson, W. B. and Lendenstrauss, J.*, 317-366, Elsevier Science.

14. Dey, D. K. and Srinivasan, C. (1985) : Estimation of a covariance matrix under Stein's loss, *Annals of Statistics*, **13**, 1581-1591.

15. Donoho, D. L. (1993) : Unconditional bases are optimal bases for data compression and statistical estimation, *Applied and Computational Harmonic Analysis*, **1**, 100-115.

16. Eaton, M. L. and Tyler, D. E. (1991) : On Wielandt's inequality and its application to the asymptotic distribution of a random symmetric matrix, *Annals of Statistics*, **19**, 260-271.

17. Efron, B. and Morris, C. (1976) : Multivariate empirical Bayes estimation of covariance matrices, *Annals of Statistics*, **4**, 22-32.

18. Haff, L. R. (1980) : Empirical Bayes estimation of the multivariate normal covariance matrix, *Annals of Statistics*, **8**, 586-597.

19. Hall, P. (1992) : *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

20. Hall, P. and Horowitz, J. L. (2004) : Methodology and convergence rates for functional linear regression, *Manuscript*.

21. Hall, P. and Hosseini-Nasab, M. (2006) : On properties of functional principal components analysis, *Journal of Royal Statistical Society, Series B*, **68**, 109-125.

22. Tony Cai, T. and Hall, P. (2005) : Prediction in functional linear regression, *Manuscript.*

23. Hoyle, D. and Rattray, M. (2003) : Limiting form of the sample covariance eigenspectrum in PCA and kernel PCA, *Advances in Neural Information Processing Systems*, **16**.

24. Hoyle, D. and Rattray, M. (2004) : Principal component analysis eigenvalue spectra from data with symmetry breaking structure, *Physical Review E*, **69**, 026124.

25. Johnstone, I. M. (2001) : On the distribution of the largest principal component, *Annals of Statistics*, **29**, 295-327.

26. Johnstone, I. M. (2001b) : Chi square oracle inequalities, in *Festschrift for William R. van Zwet*, **36**, *Eds. de Gunst, M., Klaassen, C. and Waart, A. van der*, 399-418, Institute of Mathematical Statistics.

27. Johnstone, I. M. (2002) : *Function estimation and gaussian sequence models*, Book Manuscript.

28. Johnstone, I. M. and Lu, A. Y. (2004) : Sparse principal component analysis, *Technical Report*, Stanford University.

29. Kato, T. (1980) : *Perturbation Theory of Linear Operators*, Springer-Verlag.

30. Kneip, A. (1994) : Nonparametric estimation of common regressors for similar curve data, *Annals of Statistics*, **22**, 1386-1427.

31. Kneip, A. and Utikal, K. J. (2001) : Inference for density families using functional principal component analysis, *Journal of the American Statistical Association*, **96**, 519-542.

32. Laloux, L., Cizeau, P., Bouchaud, J. P. and Potters, M. (2000) : Random matrix theory and financial correlations, *International Journal of Theoretical and Applied Finance*, **3**.

33. Loh, W.-L. (1988) : Estimating covariance matrices, *Ph. D. Thesis*, Stanford University.

34. Lu, A. Y. (2002) : Sparse principal component analysis for functional data, *Ph. D. Thesis*, Stanford University.

35. Muirhead, R. J. (1982) : *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, Inc.

36. Paul, D. (2005) : Nonparametric estimation of principal components, *Ph. D. Thesis*, Stanford University.

37. Paul, D. and Johnstone, I. M. (2004) : Estimation of principal components through coordinate selection, *Technical Report*, Stanford University.

38. Preisendorfer, R. W. (1988) : *Principal component analysis in meteorology and oceanography*, Elsevier, New York.

39. Ramsay, J. O. and Silverman, B. W. (1997) : *Functional Data Analysis*, Springer-Verlag.

40. Ramsay, J. O. and Silverman, B. W. (2002) : *Applied Functional Data Analysis : Methods and Case Studies*, Springer-Verlag.

41. Spellman, P.T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) : Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273-3297.

42. Stegmann, M. B. and Gomez, D. D. (2002) : A brief introduction to statistical shape analysis, *Lecture notes*, Technical University of Denmark.

43. Telatar, E. (1999) : Capacity of multi-antenna Gaussian channels, *European Transactions on Telecommunications*, **10**, 585-595.

44. Tulino, A. M. and Verdu, S. (2004) : Random matrices and wireless communications, *Foundations and Trends in Communications and Information Theory*, **1**.

45. Tyler, D. E. (1983) : The asymptotic distribution of principal component roots under local alternatives to multiple roots, *Annals of Statistics*, **11**, 1232-1242.

46. Vogt, F., Dable, B., Cramer, J. and Booksh, K. (2004) : Recent advancements in chemometrics for smart sensors, *The Analyst*, **129**, 492-502.

47. Wickerhauser, M. V. (1994) : *Adapted Wavelet Analysis from Theory to Software*, A K Peters, Ltd.

48. Yang, Y. and Barron, A. (1999) : Information-theoretic determination of minimax rates of convergence, *Annals of Statistics*, **27**, 1564-1599.

49. Zhao, X., Marron, J. S. and Wells, M. T. (2004) : The functional data analysis view of longitudinal data, *Statistica Sinica*, **14**, 789-808.

50. Zong, C. (1999) : *Sphere Packings*, Springer-Verlag.