Revision submitted to JASA
February 2008

LIMITATIONS OF HIGH DIMENSION, LOW SAMPLE SIZE PRINCIPAL COMPONENTS FOR GAUSSIAN DATA

Keith E. Muller[1,*], Yueh-Yun Chi[1], Jeongyoun Ahn[2], and J. S. Marron[3]

[1]Department of Epidemiology and Health Policy Research,

University of Florida, Gainesville

[*]*e-mail:* Keith.Muller@biostat.ufl.edu

[2]Department of Statistics, University of Georgia, Athens

[3]Department of Statistics and Operations Research,

University of North Carolina, Chapel Hill

Muller's work supported in part by NCI P01 CA47982-04, R01 CA095749-01A1, R01 HL69808-01 and R01 CA67812-05. Chi's work supported by NCI P01 CA47 982-04.

KEYWORDS: Pseudo Wishart, Singular, Less than full rank, Principal component analysis, Estimating eigenvalues

ABSTRACT

Medical images and genetic assays typically generate High Dimension, Low Sample Size (HDLSS) data, namely more variables than independent sampling units. Scientists often use Principal Components Analysis (PCA) of sample covariance matrices to work around the limitations of HDLSS. We provide analytic results and Monte Carlo simulations for Gaussian data which strongly discourage the practice. All but a negligible fraction of population variation must occur in a set of dominant components far fewer in number than the number of independent sampling units. The results demonstrate why statisticians must assess the empirical performance of any analysis method in the HDLSS setting. Expressing HDLSS data in terms of underlying canonical forms helps develop analytic and sample properties.

1. INTRODUCTION

1.1 Motivation

Medical imaging scientists often use principal components analysis (PCA) to select a covariance model of High Dimension, Low Sample Size (HDLSS) data, i.e., more variables than independent observations. The resulting dimension reduction allows classical data analysis. However, results for other model selection methods lead to the suspicion that HDLSS will require extremely easy problems in order to succeed.

For Gaussian data, the sample covariance matrix provides an unbiased estimator of the population covariance matrix even with HDLSS. In contrast, for general population covariance, HDLSS disallows estimating all of the population eigenvalues. Statistically, PCA uses the sample-ordered eigenvalues of the sample covariance to estimate the largest population eigenvalues. PCA reduces the variable dimension by selecting the components associated with the largest few sample eigenvalues.

Good model selection by PCA requires the sample eigenvalues to be good estimators of the corresponding population eigenvalues (the component variances). Previous

theoretical and simulation results make it clear that allowing the number of variables to decrease from above toward the number of observations leads to poor performance. Here we extend the results by allowing the number of variables to decrease far below the number of observations, and therefore study the HDLSS setting.

1.2 PCA of Brain Anisotropy in a Child's Left Cerebellum

Cascio et al. (2008) used diffusion tensor imaging (DTI) to assess anisotropy (heterogeneity) in brain regions of interest for autistic, normal, and developmentally delayed children. The 22 normal and 10 developmentally delayed children served as a reference (control) group in the study. The left cerebellum provides one of many regions of interest in the brains of the children. Even considered by itself, the left cerebellum data display the HDLSS problem because for each of the 32 children have 387 values of fractional anisotropy (FA, a measure of heterogeneity), one per voxel (cube in 3-space). Although PCA can be computed with fewer degrees of freedom than variables, *how much confidence can we have that the results help us understand the population structure?* The same question arises in applying PCA in many other fields, including chemical spectroscopy, proteomics, metabolomics, etc.

A full rank multivariate linear model of the 387 response variables was fitted to the DTI data in order to adjust for age, gender, developmental group, and all their interactions. With 8 degrees of freedom for the model, the residuals were based on $\nu = 32-8 = 24$ error degrees of freedom, and a ratio of roughly 16 variables per degree of freedom. PCA of the residual covariance matrix gave Figure 1, which includes the traditional "scree" plot of sample-rank-order estimated eigenvalues $\left\{\widehat{\lambda}_j\right\}$, as well as the more informative plot of $\left\{\widehat{\lambda}_j^{1/2}\right\}$. Of the 24 components with non-zero variance, the first 18 accounted for 90% of the variation. The first component explained 13.9% of the variation, while the next 23 components accounted for 9.3% to 1.1%.

1.3 Literature Review

Many authors have described Wishart theory in describing the sample covariance distribution for Gaussian data (Johnson and Kotz, 1972; Anderson, 2004). Most results require nonsingular population covariance and more observations than variables. Khatri (1976) discussed relaxing the requirement of nonsingular population covariance. Uhlig (1994) discussed relaxing the second requirement (more observations than variables). Muller and Stewart (2006) described a Wishart taxonomy which includes all possible combinations of finite 1) variable dimensions, 2) population covariance ranks, and 3) sample sizes, including HDLSS cases.

The importance to statistical success of the ratio of observations to variables has been studied in many traditional low dimension settings. Simulation evidence emphasizes the importance of the ratio in determining the stability of factor analysis. The results suggest the same for PCA, a special case of factor analysis. MacCallum, Widaman, Zhang, and Hong (1999) supported that position for situations with more observations than variables. Preacher and MacCallum (2002) extended the conclusions to sample sizes as small as 10, including some HDLSS cases.

HDLSS data allow a variety of asymptotic scenarios. Johnstone (2001) allowed the number of variables and observations to increase at the same rate for HDLSS data. He described the limiting density of the largest sample eigenvalue when all population eigenvalues are equal. For population covariance matrices not far from the identity, if the ratio of variable dimension to sample size goes to a constant Baik et al. (2005) and Baik and Silverstein (2006) discovered that the sample eigenvalues behave as if the underlying covariance matrix were the identity matrix. Meinshausen and Bühlmann (2006) found that consistent selection of the mean model with HDLSS Gaussian data required a sparse covariance matrix among predictors. For the number of variables going to infinity with a fixed sample size, Ahn et al. (2007) described the geometric structure of HDLSS data and

discussed implications of using PCA in a high-dimensional space. All of the HDLSS results share a common feature: success requires a simple covariance structure, at least asymptotically.

1.4 Overview of PCA Results

We address the most basic questions about the quality of eigenvalue estimation for HDLSS and Gaussian data. We provide both analytic and numerical properties of sample covariance matrices, with either full or less-than-full rank population covariance matrix. We conclude that PCA of HDLSS data will fail unless the population covariance falls in a limited range of simple structures.

A key step involves expressing singular HDLSS data and covariance matrices in terms of nonsingular matrices. The expressions demonstrate that the population eigenvectors play no role whatsoever in the distribution of the sample eigenvalues for HDLSS data. Being able to ignore the eigenvectors greatly simplifies the design of simulations. Furthermore the full rank expressions make the calculations more accurate and faster.

The paper contains 6 more sections. Section 2 details all assumptions, while sections 3 and 4 cover analytic results for nonsingular and singular population covariance matrices. In section 5 analytic approximations lead to a formal conjecture that PCA works very poorly in HDLSS. Monte Carlo simulations in section 6 support the conjecture. We close with some suggestions for alternatives to PCA with HDLSS.

2. ASSUMPTIONS

2.1 Multivariate Gaussian Data

We follow Schott's (1997) notation for matrices and describe the $\nu \times p$ data matrix, $\boldsymbol{Y}$, as containing ν independent observations on p variables. The vector $\boldsymbol{y}_i = [\text{row}_i(\boldsymbol{Y})]'$ has a Gaussian distribution with mean zero, indicated $\boldsymbol{y}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. The assumption of zero means help simplify the exposition. Lemma 1 details how to relax the assumption and account for nonzero means and more general covariate adjustments, as usually

required in practice. Describing the distribution $\boldsymbol{Y}$ as a single matrix Gaussian, rather than a collection of vectors, greatly simplifies stating many of the results in the paper.

Definition 1. For symmetric $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ with no negative eigenvalues, $\boldsymbol{U}$ follows a *matrix Gaussian* distribution indicated $\boldsymbol{U} \sim \mathcal{N}_{\nu,p}(\boldsymbol{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$, if and only if $\text{vec}(\boldsymbol{U}) \sim \mathcal{N}_{\nu \cdot p}[\text{vec}(\boldsymbol{M}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Xi}]$ (Muller and Stewart, 2006, section 8.8).

When needed for clarity, writing $\mathcal{SN}_{n,p}(\boldsymbol{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$ explicitly indicates that one or both of $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ have one or more zero eigenvalues, and hence are singular matrices, which disallows the distribution from having a density. The notation $(\mathcal{S})\mathcal{N}_{n,p}(\boldsymbol{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$ indicates $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ may be singular or nonsingular.

Assumption 1. For integers $\nu > 0$ and $p > 0$, the data of interest follow a matrix Gaussian distribution, $\boldsymbol{Y} \sim (\mathcal{S})\mathcal{N}_{\nu,p}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{\Sigma})$. Although $\boldsymbol{I}_\nu$ is never singular $\boldsymbol{\Sigma}$ may be.

In practice, the sample covariance matrix must usually be adjusted for the mean and known covariates. Analysis of the DTI data illustrate the point. Applying PCA to the sample covariance would adjust only for the grand mean and not for mean differences due to gender, age, developmental group, nor for interactions. The presence of group differences would either distort the structure or define additional components in order to accommodate the variation among groups. Treating the mean model (first moment variation) separately from the covariance model (second moment variation) provides a more parsimonious, simpler, and easier to analyze covariance model. Achieving the simpler model requires accounting for mean structure in the data, as indicated in the following cautionary lemma.

Lemma 1. For Gaussian data the sample covariance matrix of the residuals from a multivariate linear model adjusting for the mean and any important covariates avoids considering the mean and covariates any further. Omitting the adjustment introduces additional sources of variability (often the largest) in the covariance matrix.

2.2 Wishart Sample Covariance

The sums of squares matrix for a Gaussian sample follows a Wishart distribution, and so lies at the heart of PCA theory. Muller and Stewart (2006) gave the following definition which explicitly allows population covariance of any rank (full rank or singular), as well as any ratio of sample size to number of variables, including HDLSS.

Definition 2. If $\boldsymbol{Y} \sim (\mathcal{S})\mathcal{N}_{\nu,p}(\mathbf{0}, \boldsymbol{I}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}'\boldsymbol{Y}$ follows a *central (integer) Wishart* distribution with (integer) $\nu > 0$ degrees of freedom, written $\boldsymbol{Y}'\boldsymbol{Y} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ for full rank $\boldsymbol{\Sigma}$, or $\mathcal{SW}_p(\nu, \boldsymbol{\Sigma})$ for singular $\boldsymbol{\Sigma}$, or $(\mathcal{S})\mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ for possibly singular $\boldsymbol{\Sigma}$. Considering the characteristic function allows extending the definition to all real ν.

The following assumption specifies the class of covariance matrices of interest by explicitly defining the $p \times p$ matrix of sums of squares. Even with HDLSS, the sample covariance matrix provides an unbiased estimator of the population covariance matrix.

Assumption 2. For integer $\nu > 0$, $\boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y} = \nu\widehat{\boldsymbol{\Sigma}} \sim (\mathcal{S})\mathcal{W}_p(\nu, \boldsymbol{\Sigma})$.

3. FULL RANK POPULATION COVARIANCE

3.1 Homogeneity of Component Variances, the Population Eigenvalues

With nonsingular population covariance, the complications in PCA for HDLSS data all stem from singularity of the sample covariance matrix, which occurs due to too few observations ($\nu < p$). The behavior of PCA for sample data varies as a function of the population pattern of eigenvalues. The average eigenvalue (first moment) can easily be seen to have no role in predicting performance of PCA for HDLSS except in limiting cases (average eigenvalue near zero or infinity). In contrast the second moment, the dispersion of the eigenvalue pattern, can tell us a great deal about the analytic and simulation performance of PCA for HDLSS. In order to facilitate the discussion, the next lemma defines a standard measure of eigenvalue dispersion. Special values of it lead to simple distributions of sample eigenvalues, even with HDLSS.

Lemma 2. For covariance matrix $\mathbf{\Sigma} = \mathbf{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\mathbf{\Upsilon}'$, the sphericity parameter

$$\begin{aligned}\epsilon &= \mathrm{tr}^2(\mathbf{\Sigma})/\left[p\mathrm{tr}\left(\mathbf{\Sigma}^2\right)\right] \\ &= \left(\sum_{j=1}^{p}\lambda_j\right)^2 \Big/ \left(p\sum_{j=1}^{p}\lambda_j^2\right) = \left(\overline{\lambda}\right)^2/\overline{\lambda^2}\end{aligned} \tag{1}$$

measures homogeneity of $\{\lambda_j\}$ and does not vary with $\mathbf{\Upsilon}$ or scale ($\boldsymbol{\lambda}$ and $c \cdot \boldsymbol{\lambda}$ give the same ϵ). Also $1/p \leq \epsilon \leq 1$ with $\epsilon = 1 \Leftrightarrow \lambda_j \equiv \overline{\lambda} > 0$ for complete homogeneity.

If $\mathrm{rank}(\mathbf{\Sigma}) = p$ and $\nu\widehat{\mathbf{\Sigma}} = \boldsymbol{S} \sim \mathcal{W}_p(\nu, \mathbf{\Sigma})$, then the p principal components underlying $\boldsymbol{S}$ are Gaussian and independent, with variances $\{\lambda_j\}$. Hence ϵ measures deviation from component sphericity (but not sphericity of the original variables, unless $\mathbf{\Upsilon} = \boldsymbol{I}$). The components have a spherical Gaussian distribution in p dimensions if and only if the eigenvalues are all equal ($\boldsymbol{\lambda} = \overline{\lambda}\mathbf{1}_p$), if and only if $\epsilon = 1$, whether $\nu < p$ (HDLSS) or $\nu \geq p$. If $\epsilon < 1$, then properties of sample eigenvalues differ for $\nu < p$ and $\nu \geq p$.

3.2 A Dual Matrix for the Sample Sums of Squares

Whenever the variable dimension exceeds the number of independent observations, analytic and computational problems abound for the sample covariance matrix. Fortunately the sample eigenvalues coincide with the eigenvalues of a smaller matrix of full rank. Even more felicitously the smaller dual matrix involves only independent random variables and does not depend on the population eigenvectors.

If $\nu < p$, then any realization of $\boldsymbol{S} = \nu\widehat{\mathbf{\Sigma}} = \boldsymbol{Y}'\boldsymbol{Y}$ is singular with ν nonzero eigenvalues. Necessarily, any realization of $\boldsymbol{S}$ has the same nonzero eigenvalues as the corresponding particular realization of the $\nu \times \nu$ and nonsingular matrix

$$\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}' \,. \tag{2}$$

Hence the distributions of nonzero eigenvalues of $\boldsymbol{S}$ and $\boldsymbol{S}_D$ coincide.

The dual matrix $\boldsymbol{S}_D$, switching the roles of columns and rows of a data matrix, is much simpler than $\boldsymbol{S}$. As stated in Theorem 1, although $\boldsymbol{S}_D$ and $\boldsymbol{S}$ share nonzero eigenvalues,

the matrix $\boldsymbol{S}_D$ depends on the eigenvalues of $\boldsymbol{\Sigma}$, but not on its eigenvectors. Furthermore $\boldsymbol{S}_D$ exactly equals a weighted sum of p independent Wishart matrices.

Theorem 1. If $\boldsymbol{Y} \sim \mathcal{N}_{\nu,p}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{\Sigma})$ for rank$(\boldsymbol{\Sigma}) = p$, then a) $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}\text{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$, b) $\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_p$, and c) $\boldsymbol{Y} = \boldsymbol{Z}\text{Dg}(\boldsymbol{\lambda})^{1/2}\boldsymbol{\Upsilon}'$ with $\boldsymbol{Z} \sim \mathcal{N}_{\nu,p}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{I}_p)$. d) If $\nu < p$ then the $p \times p$ matrix $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ has the same ν nonzero eigenvalues as the $\nu \times \nu$ matrix

$$\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}' = \boldsymbol{Z}\text{Dg}(\boldsymbol{\lambda})\boldsymbol{Z}' = \sum_{j=1}^{p} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j' = \sum_{j=1}^{p} \lambda_j \boldsymbol{W}_j, \tag{3}$$

with $\boldsymbol{z}_j$ being column j of $\boldsymbol{Z}$ and $\boldsymbol{W}_j \sim \mathcal{W}_\nu(1, \boldsymbol{I}_\nu)$ with $\boldsymbol{W}_j$ independent of $\boldsymbol{W}_{j'}$ ($j \neq j'$).

Corollary 1.1 The characteristic function is $\phi(\boldsymbol{T}; \boldsymbol{S}_D) = \prod_{j=1}^{p} |\boldsymbol{I}_\nu - 2\iota\lambda_j \boldsymbol{T}|^{-1/2}$, $\boldsymbol{T} = \boldsymbol{T}'$.

Corollary 1.2 The first moment of $\boldsymbol{S}_D$ is $\text{E}(\boldsymbol{S}_D) = \sum_{j=1}^{p} \lambda_j \boldsymbol{I}_\nu$. The variance of an element is $\mathcal{V}(\langle \boldsymbol{S}_D \rangle_{jj'}) = 2 \cdot \sum_{j=1}^{p} \lambda_j^2$ for $j = j'$, and $\mathcal{V}(\langle \boldsymbol{S}_D \rangle_{jj'}) = \sum_{j=1}^{p} \lambda_j^2$ for $j \neq j'$.

At first glance, equation 3 may look too simple and wrong. As often happens in statistical expressions, the equation may be interpreted as a statement about random variables or particular realizations. Therefore it helps to carefully attend to the distinctions among parameters (unknown constants), parameter estimates, a random variable (matrix), a particular realization of a random matrix, and a (known) constant. In particular, the random matrix $\boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ has rank$(\boldsymbol{\Sigma}) = p$ and hence population nonsingular covariance. If $\nu < p$ then any particular realization of $\boldsymbol{S}$, say $\boldsymbol{S}_*$, must be singular and hence gives singular sample covariance, $\widehat{\boldsymbol{\Sigma}}_*$.

3.3 Sample-Ordered Distributions of Estimated Eigenvalues

The sampling and computational process does not allow observing individual estimates of particular eigenvalues. Instead data analysts can observe only the sample order of eigenvalues from PCA of a sample covariance matrix. For PCA to succeed as a data analysis requires reasonable accuracy in using the observed sample-ordered eigenvalues to estimate the unobserved population-ordered eigenvalues. Previously known results about the distributions of *sample-ordered* eigenvalues from PCA of $\widehat{\boldsymbol{\Sigma}}$ based on

$\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ require *both* nonsingular $\boldsymbol{\Sigma}$ and $\nu \geq p$ (more observations than variables). Johnson and Kotz (1972) provided the most detailed discussion we know.

In contrast to the requirements for previous results on eigenvalue distributions, interest in HDLSS data involves fewer observations than variables, $\nu < p$. Theorem 2 extends the results known about the distribution of sample-ordered eigenvalues to some HDLSS cases with full rank population covariance, $\nu < \text{rank}(\boldsymbol{\Sigma}) = p$.

Theorem 2. For $\nu < p = \text{rank}(\boldsymbol{\Sigma})$ the smallest $p-\nu$ eigenvalues of $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ are zero. The distribution of the ν largest sample-ordered eigenvalues $\boldsymbol{t}_+ = [\, t_{(1)} \cdots t_{(\nu)} \,]'$ for $t_{(1)} \leq \ldots \leq t_{(\nu)}$ depends on population eigenvalues $\boldsymbol{\lambda}$ as follows.
a) If $\boldsymbol{\lambda} = \overline{\lambda}\mathbf{1}_p$ so $\epsilon = 1$, then 1) the eigenvalues of $\boldsymbol{S}$ exactly follow the distribution of the eigenvalues of the $\nu \times \nu$ matrix $\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}' = \overline{\lambda}\boldsymbol{Z}\boldsymbol{Z}' = \overline{\lambda}\sum_{j=1}^{p}\boldsymbol{W}_j \sim \mathcal{W}_\nu\big(p, \overline{\lambda}\boldsymbol{I}_\nu\big)$, 2) the density of $\boldsymbol{t}_+$ is

$$f_1(\boldsymbol{t}_+; \nu, p) = \frac{\pi^{\nu/2}}{\big(2\overline{\lambda}\big)^{\nu p/2}} \prod_{j=1}^{\nu} \frac{\Gamma(j/2)\, t_{(j)}^{(p-\nu-1)/2}}{\Gamma[(p-j+1)/2]} \exp\Big(-\frac{1}{2\overline{\lambda}}\sum_{j=1}^{\nu} t_{(j)}\Big) \prod_{k>j} \big(t_{(k)} - t_{(j)}\big)\,, \tag{4}$$

which corresponds to the previously known result with p and ν exchanged, and 3) the (marginal) density of the largest, $t_{(\nu)}$, is known, along with other properties (Johnson and Kotz, 1972, 188-192).

b) If $1/p < \epsilon < 1$, then the joint density of $\boldsymbol{t}_+$ is unknown.

c) If $1/p < \epsilon \leq 1$ and $\nu = 1$, then $\boldsymbol{S}_D$ and $\boldsymbol{S}$ have one nonzero eigenvalue which equals a weighted sum of independent chi squares. Also $W_j \sim \mathcal{W}_1(1, \boldsymbol{I}_1) \;\Leftrightarrow\; W_j \sim \chi^2(1)$ and $\boldsymbol{S}_D = \sum_{j=1}^{p}\lambda_j W_j$ is 1×1. Davies' algorithm (1980) gives exact probabilities.

4. LESS-THAN-FULL RANK POPULATION COVARIANCE

4.1 Describing Sample-Singular Covariance Matrices

Less-than-full rank population covariance complicates the discussion of PCA for HDLSS data. The sample covariance matrix may be singular due to a) $\nu < p$ (HDLSS), b) $\text{rank}(\boldsymbol{\Sigma}) = p_+ < p$, or c) both. In any case, loosening the condition $\nu \geq p$ to $\nu \geq p_+$

allows generalizing the preceding analytic results. Our simulation and analytic results demonstrate that $\nu \gg p_+$ suffices to provide good estimation of the population eigenvalues. The variable dimension, p, plays a role solely in determining the singularity of the sample covariance matrix.

A more general form of the sphericity parameter ϵ does a much better job of describing eigenvalue patterns and characterizing distributions of sample eigenvalues. The more general form reduces to the original version for a nonsingular covariance matrix.

Lemma 3. A $p \times p$ singular covariance matrix $\mathbf{\Sigma}$ with $1 \leq \text{rank}(\mathbf{\Sigma}) = p_+ \leq p$ has p_+ strictly positive eigenvalues $\boldsymbol{\lambda}_+ = [\, \lambda_1 \, \cdots \, \lambda_{p_+} \,]$ and $p - p_+$ zero eigenvalues.

a) While $\epsilon = \text{tr}^2(\mathbf{\Sigma})/[p\text{tr}(\mathbf{\Sigma}^2)]$ characterizes the entire set of p eigenvalues, the parameter $\epsilon_+ = \epsilon p/p_+$ characterizes the dispersion of the p_+ *nonzero* eigenvalues:

$$\begin{aligned} \epsilon_+ &= \text{tr}^2(\mathbf{\Sigma})/\left[p_+ \text{tr}\left(\mathbf{\Sigma}^2\right)\right] \\ &= \left(\sum_{j=1}^{p_+} \lambda_{+j}\right)^2 \Big/ \left(p_+ \sum_{j=1}^{p_+} \lambda_{+j}^2\right) = \left(\overline{\lambda}_+\right)^2 / \overline{\lambda_+^2} \, . \end{aligned} \tag{5}$$

b) If $p_+ = 1$ then $\epsilon = 1/p$ and $\epsilon_+ = 1$.

c) If $1 < p_+ < p$ then $1/p < \epsilon \leq p_+/p$ and $1/p_+ < \epsilon_+ \leq 1$.

The following observations illustrate the value of considering ϵ_+ rather than ϵ. If $\nu\widehat{\mathbf{\Sigma}} = \boldsymbol{S} \sim \mathcal{SW}_p(\nu, \mathbf{\Sigma})$ and the p_+ nonzero eigenvalues of $\mathbf{\Sigma}$ are all equal, then the corresponding components have a spherical Gaussian distribution in p_+ dimensions. In contrast, the entire set of p components have a singular distribution. In general, if $\epsilon_+ < 1$ then properties of sample eigenvalues differ for $\nu < p_+$ and $\nu \geq p_+$.

4.2 The Data and Sample Covariance in Terms of Full Rank Matrices

Expressing the data and sample covariance in terms of full rank matrices provides many important advantages. It greatly improves computational speed and also helps computational accuracy. Most importantly the full-rank expressions greatly simplify deriving and understanding analytic properties. The next lemma gives explicit forms.

Lemma 4. If rank($\boldsymbol{\Sigma}$) $= p_+ < p$, then $\boldsymbol{Y} \sim \mathcal{SN}_{\nu,p}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{\Sigma})$. Having the p_+ nonzero eigenvalues in $\boldsymbol{\lambda}_+ = [\,\lambda_1 \;\ldots\; \lambda_{p_+}\,]'$ gives $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}_+ \mathrm{Dg}(\boldsymbol{\lambda}_+)\boldsymbol{\Upsilon}'_+ = \boldsymbol{\Phi}_+\boldsymbol{\Phi}'_+$ with $\boldsymbol{\Phi}_+ = \boldsymbol{\Upsilon}_+ \mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}$. For $\boldsymbol{Z} \sim \mathcal{N}_{\nu,p_+}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{I}_{p_+})$ it follows that

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\Phi}'_+ = \boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}\boldsymbol{\Upsilon}'_+ \,, \tag{6}$$

(Muller and Stewart, 2006, section 8.8), $\boldsymbol{Y}'\boldsymbol{Y} = \nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}$ and

$$\boldsymbol{S} = \boldsymbol{\Upsilon}_+ \mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}\boldsymbol{Z}'\boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}\boldsymbol{\Upsilon}'_+ \,. \tag{7}$$

The last line expresses singular $\boldsymbol{S}$ in terms of three simple, nonsingular matrices. 1) The eigenvector matrix $\boldsymbol{\Upsilon}_+$ is $p \times p_+$ and orthonormal by columns. 2) The diagonal matrix $\mathrm{Dg}(\boldsymbol{\lambda}_+)$ is $p_+ \times p_+$, with diagonal elements the strictly positive eigenvalues of $\boldsymbol{\Sigma}$, the population variances of the principal components. 3) The random matrix $\boldsymbol{Z} = \{z_{ij}\}$ is $\nu \times p_+$, has fully independent elements $z_{ij} \sim \mathcal{N}(0,1)$ and a nonsingular distribution.

If $\boldsymbol{\Sigma}$ is singular ($p_+ < p$) and $\nu \geq p_+$, then $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y}$ is singular with rank p_+. In contrast if $\boldsymbol{\Sigma}$ is singular and $\nu < p_+$ then $\boldsymbol{S}$ is singular with rank ν. The $\nu \times \nu$ dual matrix $\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}'$ is defined in both cases, but is nonsingular only if $\nu \leq p_+$. As detailed in the following theorem, if $\nu < p_+$, then the dual matrix has a simple distribution because it equals a weighted sum of independent Wishart matrices.

Theorem 3. If $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{S} \sim \mathcal{SW}_p(\nu, \boldsymbol{\Sigma})$ for $\nu < p_+ = \mathrm{rank}(\boldsymbol{\Sigma}) < p$, $\boldsymbol{Z} \sim \mathcal{N}_{\nu,p_+}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{I}_{p_+})$, $\boldsymbol{z}_j$ being column j of $\boldsymbol{Z}$, $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}_+ \mathrm{Dg}(\boldsymbol{\lambda}_+)\boldsymbol{\Upsilon}'_+$, and $\boldsymbol{W}_j \sim \mathcal{SW}_\nu(1, \boldsymbol{I}_\nu)$ with $\boldsymbol{W}_j$ independent of $\boldsymbol{W}_{j'}$ ($j \neq j'$), then

$$\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}' = \boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda}_+)\boldsymbol{Z}' = \sum_{j=1}^{p_+} \lambda_j \boldsymbol{z}_j \boldsymbol{z}'_j = \sum_{j=1}^{p_+} \lambda_j \boldsymbol{W}_j \,. \tag{8}$$

Corollary 3.1 The characteristic function is $\phi(\boldsymbol{T}; \boldsymbol{S}_D) = \prod_{j=1}^{p_+} |\boldsymbol{I}_\nu - 2\iota\lambda_j \boldsymbol{T}|^{-1/2}$, $\boldsymbol{T} = \boldsymbol{T}'$.

Corollary 3.2 The first moment of $\boldsymbol{S}_D$ is $\mathrm{E}(\boldsymbol{S}_D) = \sum_{j=1}^{p_+} \lambda_j \boldsymbol{I}_\nu$. The variance of an element is $\mathcal{V}(\langle \boldsymbol{S}_D \rangle_{jj'}) = 2 \cdot \sum_{j=1}^{p_+} \lambda_j^2$ for $j = j'$, and $\mathcal{V}(\langle \boldsymbol{S}_D \rangle_{jj'}) = \sum_{j=1}^{p_+} \lambda_j^2$ for $j \neq j'$.

4.3 Distributions of Estimated Eigenvalues for Singular Population Covariance

The distribution results for estimated eigenvalues in Theorem 2 extend to allow singular population covariance $\mathbf{\Sigma}$ as long as $1 \leq \text{rank}(\mathbf{\Sigma}) = p_+ < p$. The most complex situation involves both HDLSS and singular population covariance. A variety of cases arise, with some cases having sample eigenvalues with relatively simple distributions. Theorem 4 covers $\nu \geq p_+$, which includes HDLSS cases with $p > \nu \geq p_+$, as well as more traditional cases with $\nu \geq p > p_+$. In contrast, Theorem 5 covers $\nu < p_+ < p$.

Theorem 4. If $\mathbf{\Sigma} = \mathbf{\Upsilon}\text{Dg}(\boldsymbol{\lambda}_+, \mathbf{0})\mathbf{\Upsilon}'$ with $\nu \geq \text{rank}(\mathbf{\Sigma}) = p_+ < p$ and $\nu\widehat{\mathbf{\Sigma}} = \boldsymbol{S} \sim \mathcal{SW}_p(\nu, \mathbf{\Sigma})$, then the smallest $p - p_+$ sample eigenvalues of $\boldsymbol{S}$ are zero with probability one. The p_+ largest sample-ordered eigenvalues $\boldsymbol{t}_+ = [\, t_{(1)} \ \ldots \ t_{(p_+)} \,]'$ have the following properties.

a) If $\boldsymbol{\lambda}' = [\, \overline{\lambda}\mathbf{1}'_{p_+} \ \mathbf{0}' \,]$, then $\epsilon_+ = 1$ and nonzero component sphericity holds. The density of $\boldsymbol{t}_+$ is $f_1(\boldsymbol{t}_+; p_+, \nu)$ with $f_1(\,\cdot\,)$ as in part a) of Theorem 2. Other properties are similarly available.

b) If $1/p_+ < \epsilon_+ < 1$, $\boldsymbol{t}_+$ properties generalize the usual results for $\nu \geq p = \text{rank}(\mathbf{\Sigma})$ as in Johnson and Kotz (1972, p170). If $\beta_j = \left(1/\overline{\lambda}_+ - 1/\lambda_{+j}\right)/2$, $\boldsymbol{\chi}(m)$ is a partition of m, and $\mathcal{C}_{\boldsymbol{\chi}(m)}(\boldsymbol{A})$ is the corresponding zonal polynomial, the joint density of $\boldsymbol{t}_+$ is

$$f_{\epsilon_+}(\boldsymbol{t}_+) = f_1(\boldsymbol{t}_+; p_+, \nu) \cdot \pi^{-p_+/2} \prod_{j=1}^{p_+} \left(\frac{\nu}{\pi}\frac{\overline{\lambda}_+}{\lambda_{+j}}\right)^{\nu/2} \sum_{m=0}^{\infty} \sum_{\boldsymbol{\chi}(m)} \frac{\mathcal{C}_{\boldsymbol{\chi}(m)}[\text{Dg}(\boldsymbol{\beta})]\mathcal{C}_{\boldsymbol{\chi}(m)}[\text{Dg}(\boldsymbol{t}_+)]}{m!\mathcal{C}_{\boldsymbol{\chi}(m)}[\text{Dg}(\boldsymbol{I}_{p_+})]}. \quad (9)$$

c) If $\epsilon = 1/p$, then $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$ have only one nonzero eigenvalue and rank $p_+ = 1$. Also $\nu > 0$ insures $\widehat{\lambda}_{(1)}/\lambda_1 \sim \chi^2(\nu)$.

As noted earlier, the most complex situation involves both HDLSS and singular population covariance. Although Theorem 5 covers the most difficult combinations of HDLSS and singular population covariance, $\nu < p_+ < p$, some cases have sample eigenvalues with relatively simple distributions.

Theorem 5. If $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} \sim \mathcal{SW}_p(\nu, \boldsymbol{\Sigma})$ with $\nu < \text{rank}(\boldsymbol{\Sigma}) = p_+ < p$, then the smallest $p - \nu$ sample-ordered eigenvalues of $\boldsymbol{S}$ are zero with probability one. The ν largest sample-ordered eigenvalues of $\boldsymbol{S}$, $t_{(1)} \leq \ldots \leq t_{(\nu)}$, depend on population eigenvalues $\boldsymbol{\lambda}_+$ as follows.

a) If $\boldsymbol{\lambda}' = \left[\overline{\lambda}_+ \mathbf{1}'_{p_+} \ \mathbf{0}' \right]$, then $\epsilon_+ = 1$ and nonzero component sphericity holds. 1) The eigenvalues of $\boldsymbol{S}$ exactly follow the distribution of the eigenvalues of the $\nu \times \nu$ matrix $\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}' = \overline{\lambda}_+ \sum_{j=1}^{p_+} \boldsymbol{z}_j \boldsymbol{z}'_j = \overline{\lambda}_+ \sum_{j=1}^{p_+} \boldsymbol{W}_j \sim \mathcal{W}_\nu\left(p_+, \overline{\lambda}_+ \boldsymbol{I}_\nu\right)$.

2) The density of $\boldsymbol{t}_+$ is the same as in part a) of Theorem 2, but with p replaced by p_+.

3) the (marginal) density of the largest, $t_{(\nu)}$, is known, along with other properties (Johnson and Kotz, 1972, 188-192).

b) If $1/p_+ < \epsilon_+ < 1$, the joint density of $\boldsymbol{t}_+$ is unknown.

c) If $\nu = 1$ then $\boldsymbol{S}_D$ and $\boldsymbol{S}$ have one nonzero eigenvalue which is a weighted sum of independent chi squares. In particular, $W_j \sim \mathcal{SW}_1(1, \boldsymbol{I}_1) \Leftrightarrow W_j \sim \chi^2(1)$ and $\boldsymbol{S}_D = \sum_{j=1}^{p_+} \lambda_j W_j$ is 1×1. Davies' algorithm (1980) computes exact probabilities.

5. APPROXIMATIONS AND CONJECTURES

Two tools provide information about the distribution of sample eigenvalues when $\nu < \text{rank}(\boldsymbol{\Sigma}) = p_+$ and $1/p_+ < \epsilon_+ < 1$. 1) Matrices with known eigenvalue properties provide simple approximations for the matrices of interest. 2) Monte Carlo simulations help determine when the sample eigenvalues distribution can not be distinguished from the corresponding distribution for the approximation. Simulations also illustrate the discrepancy between the population and sample eigenvalues.

Matching moments allows approximating $\boldsymbol{S}_D$ by a single spherical Wishart with different degrees of freedom. The approximation in Theorem 6 leads to the conjecture that HDLSS sample eigenvalues behave as though all population values have been averaged together (homogenized, a bad feature in the present setting).

Theorem 6. a) If rank$(\boldsymbol{\Sigma}) = p_+ \leq p$, then all elements of $\boldsymbol{S}_{*1} \sim \mathcal{W}_\nu\big(p_+, \overline{\lambda}_+ \boldsymbol{I}_\nu\big)$ have the same first moments as the corresponding elements of $\boldsymbol{S}_D$: $\mathrm{E}(\boldsymbol{S}_{*1}) = \mathrm{E}(\boldsymbol{S}_D)$.

b) If $p_* = p_+\epsilon_+$ (usually fractional) and $\lambda_* = \overline{\lambda}_+/\epsilon_+$, then corresponding elements of $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p_*, \lambda_* \boldsymbol{I}_\nu)$ and $\boldsymbol{S}_D$ have the same first and second moments. Also, 18 of 21 types of third order and 69 of 79 types of fourth order moments are zero for $\boldsymbol{S}_D$ and $\boldsymbol{S}_{*2}$.

Having individual elements of $\boldsymbol{S}_D$ and $\boldsymbol{S}_{*2}$ share moments suggests their sample eigenvalues may also share moments. Hence the following statements formally predict that PCA should be expected to fail with HDLSS.

Conjectures. If two Wishart matrices have the same degrees of freedom, and their population covariance matrices have the same rank and values of ϵ_+, then the lower moments of the ν nonzero sample-ordered eigenvalues will be essentially indistinguishable from each other. More precisely, for $j \in \{1, 2\}$, $\boldsymbol{S}_j \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma}_j)$, with $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2)$, $\mathrm{rank}(\boldsymbol{\Sigma}_j) = p_+$, $\nu < p_+ \leq p$, and $\epsilon_+(\boldsymbol{\Sigma}_1) = \epsilon_+(\boldsymbol{\Sigma}_2)$, while $\boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_2$, the sets of ν nonzero sample eigenvalues for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ will have essentially the same lower moments, as will the sample eigenvalues $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p_*, \lambda_* \boldsymbol{I}_\nu)$.

6. SIMULATIONS OF SAMPLE-ORDERED EIGENVALUES

6.1 Design Motivation and Constraints

We designed the simulations to assess the accuracy of the conjectures in close collaboration with our medical imaging colleagues. Each successive set focused on an increasingly narrower range of conditions. Our collaborators insisted that compelling evidence of poor performance by PCA required simulating populations consistent with their expectations of imaging data. Simulation 1 used a coarse grid of conditions across the range of eigenvalue patterns not covered by our analytic results. The results led the imaging scientists to requested more simulations for very small ϵ (analytic results cover the boundaries $\epsilon = 1/p$ and $\epsilon_+ = 1/p_+$). A further request for changes in population eigenvalue patterns led to simulation 3, which the imaging scientists deemed compelling

evidence for cases of interest to them. Simulation 4 completes the picture by approximating the sample eigenvalue pattern of the DTI data introduced in section 1.2.

We designed the simulations to examine a range of variable dimensions (p), sample size (ν), and their ratio p/ν typical of medical imaging research we have encountered. The ratio p/ν of 16, for example, occurs in the DTI data. We considered only diagonal population covariance matrices because Theorem 1 assures us that population eigenvectors play no role in the distribution of sample eigenvalues for HDLSS PCA.

The focus on HDLSS led to 6 constraints in defining sets of population eigenvalues. 1) Sorting population eigenvalues so $\lambda_k \geq \lambda_{k+1}$ meant only monotone decreasing functions held any interest (without loss of generality). 2) Each set was scaled to help accuracy and align features across conditions in a simulation ($\overline{\lambda} = 1$ for simulations 1 and 4, $\lambda_1 = 1$ for simulations 2 and 3). 3) We sought functions giving eigenvalue ratios which remained the same for any value of p (the number of variables). 4) Our medical imaging collaborators only care about eigenvalue patterns with a small number of dominant components. Hence we considered only concave functions. 5) Testing whether two eigenvalue population patterns with the same ϵ have indistinguishable distributions of sample eigenvalues required finding two eigenvalue functions that differed in shape. 6) We wanted functions able to define eigenvalue patterns for any value of $\epsilon \in (1/p, 1)$.

6.2 Data Generation and Analysis Methods

All simulations were conducted with SAS/IML® (SAS Institute, 1999). The NORMAL and RANGAM functions generated the pseudo-random numbers. The EIGVAL function computed the eigenvalues. In each condition 10,000 replications were stored.

HDLSS simulations raise serious concerns about speed and accuracy. Careful scaling along with varying the algorithm for generating the Wishart matrices greatly improved speed and helped accuracy. With roughly 14 digits of accuracy, a number smaller than

10^{-14} in absolute value, the size of some eigenvalues in many of our simulations, can often be indistinguishable from zero. Consequently we invested time in checking the computations by comparing analytic results with simulation results for alternate algorithms. We believe the larger eigenvalues were computed with sufficient accuracy for the purposes needed.

6.3 Simulation 1 Motivation and Design

Simulation 1 allowed comparing the distributions of sample eigenvalues for two Wishart matrices having the same spherical Wishart approximation ($\boldsymbol{S}_{*2}$) but with different population eigenvalues. It also allows comparisons to a spherical Wishart with different degrees of freedom and two moments matched. We expected that all three sets of sample eigenvalues would be indistinguishable.

For $p = 64$, Figure 2 displays the square roots of the eigenvalues implied by the two population eigenvalue-pattern functions used to define $\{\lambda_j\}$ in the first simulation. Function $g_1()$ decreases smoothly at a rate determined by π, which was selected iteratively to fix $\epsilon \in \{0.2, 0.5, 0.8\}$ for each p:

$$g_1(j; \pi) = [1 - (j-1)/p]^{\pi} \, . \tag{10}$$

Function $g_2()$ joins two decreasing linear pieces at (α, β), with γ defining the smallest eigenvalue. Also $\{\alpha, \beta, \gamma\}$ with $1 \leq \alpha \leq p$ and $0 \leq \gamma \leq \beta \leq 1$ were selected iteratively to fix $\epsilon \in \{0.2, 0.5, 0.8\}$ for each p:

$$g_2(j; \alpha, \beta, \gamma) = \begin{cases} [j(\beta - 1) + (\alpha - \beta)]/(\alpha - 1) & 1 \leq j \leq \alpha \\ [j(\gamma - \beta) + (p\beta - \gamma\alpha)]/(p - \alpha) & \alpha \leq j \leq p \, . \end{cases} \tag{11}$$

A four-way factorial design used factors $p \in \{64, 256, 1024\}$ so $\log_2(p) \in \{6, 8, 10\}$, $\nu \in \{4, 8, 16, 32\}$, and $\epsilon \in \{0.20, 0.50, 0.80\}$. Solving for π in $g_1()$ achieved ϵ values which agreed to roughly 3 significant digits. Solving for $\{\alpha, \beta, \gamma\}$ in $g_2()$ achieved ϵ values agreeing to nearly 2 significant digits and therefore within 1% of the target.

For fixed p and ϵ both $g_1()$ and $g_2()$ lead to the same parameters for $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p\epsilon, \lambda_* \boldsymbol{I}_\nu)$ because $\lambda_* = \overline{\lambda}/\epsilon$. Figure 2 shows square roots of eigenvalues from $g_1()$ and $g_2()$ when $p = 64$. For $\epsilon = 0.2$ and $\epsilon = 0.5$, the population eigenvalue pattern from $g_1()$ appears quite different from the $g_2()$ pattern, and differs greatly from sphericity of $\boldsymbol{S}_{*2}$.

6.4 Simulation 2 Motivation and Design

As noted earlier, simulation 2 was designed to have eigenvalue patterns meeting the suggestions of medical imaging scientists, which meant small ϵ. Eigenvalue pattern $g_3()$ defines a step function giving only two distinct population eigenvalues, with p_1 eigenvalues of 1 and $p - p_1$ eigenvalues of $1/32$ or $1/64$. For $p \in \{256, 1024\}$ and $\nu \in \{4, 8, 16, 32\}$, here $p_1 \in \{8, 16\}$ fixed the number of "signal" eigenvalues, with $0.06 \leq \epsilon \leq 0.17$ across the range of conditions simulated.

The changes requested for simulation 2 reflect a simplification of the covariance structure. In turn, the same statement holds for simulation 3 relative to simulation 2. Consequently the design changes steered the simulations toward easier problems.

6.5 Simulation 3 Motivation and Design

Simulation 3 met the imaging scientists' objections to simulation 2 by using two linearly declining groups of population eigenvalues (large and small) with a wider range of separation between the two groups:

$$g_4(j, \pi, p, p_1, \tau) = \begin{cases} (1-\tau) g_1(j, \pi, p) + \tau g_1(j, \pi, p) & j \leq p_1 \\ \tau g_1(j, \pi, p) & j > p_1 \, . \end{cases} \tag{12}$$

If $j \leq p_1$, then $g_4() = g_1()$, and if $j > p_1$, a discount factor τ (> 0) reduces the magnitude of the $g_1()$ eigenvalues. Changing τ changes the gap between the first and second groups of eigenvalues, the "signal" and "noise" eigenvalues, with p_1 the number of "signal" eigenvalues. Values of $p = 256$, $p_1 = 8$, $\pi = 8.5118$, $\nu \in \{4, 8, 16, 32\}$, and $\tau \in \{0.01, 0.05, 0.1, 0.2\}$ were considered in a two-way factorial. The parameter τ gave

$\epsilon \in \{0.033, 0.041, 0.053, 0.084\}$. The ratio of mean eigenvalues between the two groups was $(\sum_{j=1}^{p_1}\lambda_j/p_1)/[\sum_{j=p_1+1}^{p}\lambda_j/(p-p_1)] \in \{1079, 207, 98, 44\}$.

Simulation 4 Motivation and Design

Simulation 4 asked whether the sample eigenvalue pattern for the DTI data (section 1.2), as seen in Figure 1, could be approximated well by data with population sphericity. We computed sample-ordered eigenvalues from 10,000 replicates of $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p_*, \lambda_* \boldsymbol{I}_\nu)$. Given that the DTI data had $\widehat{\overline{\lambda}} = 0.009$ we assumed the population had $\overline{\lambda} = 0.009$ and chose $p_* \in \{0.10p, 0.15p, 0.20p\}$ with $\lambda_* \in \{\overline{\lambda}/0.10, \overline{\lambda}/0.15, \overline{\lambda}/0.20\}$. We chose the values $\{0.10, 0.15, 0.20\}$ by an empirical grid search.

6.6 Numerical Results

Simulation 1. Figure 3 displays population values and box plots ($\pm$ 1.5 IQR) of the square roots of the largest 16 sample eigenvalues, as a function of size and ϵ for $p = 256$ and $\nu = 16$. Throughout, $\widehat{\boldsymbol{\Sigma}}_{g1}$ and $\widehat{\boldsymbol{\Sigma}}_{g2}$ indicate sample covariance matrices with population eigenvalue patterns $g_1()$ and $g_2()$ respectively. Statistical properties of sample eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{*2}$ (the spherical approximation) remain the same whenever the same ϵ is targeted (for fixed ν and p). Hence the $\widehat{\boldsymbol{\Sigma}}_{*2}$ patterns in any row in Figure 3 coincide across columns except for randomness across samples.

We produced one plot like Figure 3 for each condition. As ν/p decreased the largest sample and population eigenvalues differed more. Most importantly, as long as $\nu/p < 1/2$ the largest sample eigenvalue from $\widehat{\boldsymbol{\Sigma}}_{g1}$, $\widehat{\boldsymbol{\Sigma}}_{g2}$, and $\widehat{\boldsymbol{\Sigma}}_{*2}$ (the spherical approximation) had essentially the same mean and variance. The same held true for $\nu/p \leq 1$ as long as $\epsilon \geq 0.50$.

Overall, the sample data did not allow distinguishing two distinct underlying eigenvalue populations from each other, or from sphericity. Hence a PCA data analysis typically fails in the conditions studied.

Simulation 2. All results agreed completely with results from the other simulations.

Simulation 3. Figure 4 displays box plots ($\pm$ 1.5 IQR) of the square roots of the largest 16 sample eigenvalues, as a function of size, for eigenvalue patterns $g_4()$ and the corresponding spherical approximation, with $p = 256$, $p_1 = 8$, $\nu = 16$ and $\tau \in \{0.01, 0.10\}$. Population values are also plotted. The figure illustrates the conclusion that if $p_1 < \nu < p$, then τ must be extremely small to allow separating the first p_1 sample eigenvalues from the remaining $p - p_1$.

Simulation 4. The PCA results for the HLDSS DTI data in section 1.2 seem to identify a set of dominant components. Figure 5 displays the DTI sample eigenvalues along with box plots of simulated sample-ordered eigenvalues from spherical populations. Clearly the simulated data closely track the DTI values. Hence the results make it hard to claim the DTI data reflect a population eigenvalue pattern very far from sphericity.

6.7 Implications and Additional Deductions

The results deviated from the conjectures only when ν exceeded the number of population eigenvalues controlling essentially all of the variance. As an illustration, for $p = 256$, $p_1 = 8$, $\nu = 16$ and $\tau = 0.10$ in simulation 3, $\sum_{k=1}^{8}\lambda_k / \sum_{k=1}^{p}\lambda_k \approx 0.9997$ implies $p_1 = 8$ eigenvalues control 99.97% of the generalized variance, the trace. We formalize our conclusions in the following lemma about a limiting form of the characteristic function corresponding to simulation 3, and also a corollary to Theorem 4.

Lemma 5. If $\nu\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} \sim (\mathcal{S})\mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$ of rank p_+, $\boldsymbol{\lambda}' = [\,\boldsymbol{\lambda}_1' \;\; \tau\boldsymbol{\lambda}_2' \;\; \mathbf{0}_{p-p_+}'\,]$, $\boldsymbol{\Upsilon} = [\,\boldsymbol{\Upsilon}_1 \;\; \boldsymbol{\Upsilon}_2 \;\; \boldsymbol{\Upsilon}_0\,]$, $p_1 \geq 1$ positive values in $\boldsymbol{\lambda}_1$ and $(p_+ - p_1) \geq 1$ positive values in $\boldsymbol{\lambda}_2$, then $\boldsymbol{\Sigma} = \boldsymbol{\Phi}_1\boldsymbol{\Phi}_1' + \tau\boldsymbol{\Phi}_2\boldsymbol{\Phi}_2'$ with $\boldsymbol{\Phi}_j = \boldsymbol{\Upsilon}_j\mathrm{Dg}(\boldsymbol{\lambda}_j)^{1/2}$. If $\tau \to 0$ $\boldsymbol{S}$ has characteristic function

$$\begin{aligned} \lim_{\tau\to 0}[\phi(\boldsymbol{T};\boldsymbol{S})] &= \lim_{\tau\to 0}(|\boldsymbol{I}_p - 2\iota\boldsymbol{T}\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1' - 2\tau\iota\boldsymbol{T}\boldsymbol{\Phi}_2\boldsymbol{\Phi}_2'|^{-\nu/2}) \\ &= |\boldsymbol{I}_p - 2\iota\boldsymbol{T}\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'|^{-\nu/2} \\ &= |\boldsymbol{I}_{p_1} - 2\iota\boldsymbol{\Phi}_1'\boldsymbol{T}\boldsymbol{\Phi}_1|^{-\nu/2}\,. \end{aligned} \tag{13}$$

The last line of the lemma reduces the dimensions to $p_1 \times p_1$, with $p_1 \leq p_+ \leq p$. Equivalently, only the small number of very large eigenvalues matter. We describe such situations with a very strong signal and almost no noise in the following corollary.

Corollary to Theorem 4. If $\lambda_k \geq \lambda_{k+1}$ and $\nu \geq p_1$ exists with $\kappa = \sum_{k=1}^{p_1} \lambda_k / \sum_{k=1}^{p_+} \lambda_k \approx 1$, then the largest p_1 eigenvalues are reliably identifiable and the conjecture is not true. The distribution of the sample-ordered eigenvalues will be essentially indistinguishable from the distribution in Theorem 4 with p_1 replacing p_+.

7. DISCUSSION

7.1 Conclusions

Four general conclusions apply. 1) PCA will succeed with HDLSS data only for very easy problems. 2) As a default, we believe data analysts should avoid using PCA with HDLSS data. 3) Statisticians must determine the validity of any traditional multivariate method for HDLSS data. 4) Describing the underlying canonical structure helps derive analytic characteristics and predict sample properties.

7.2 Why Use PCA Rather Than Factor Analysis?

PCA helps derive analytic properties and provide insights about the results. However, the covariance structures of most interest to imaging scientists (as in simulations 2 and 3) implicitly require the more general factor analysis model. The factor analysis model expresses the response variables as a sum of shared and unique latent variables, a formulation inherent to "mixed" models. Hence for *data analysis*, we agree with Widaman (1993) and avoid PCA in favor of factor analysis. We studied PCA because so many scientists rely on it.

7.3 Defensible Strategies for HDLSS

Four strategies seem credible for HDLSS data. First, developing new theory seems the most difficult but rewarding. However, rough approximations based on overly simplified

covariance models have little appeal for practice. Exact finite sample theory or results characterizing the accuracy of approximation with small N are needed.

Second, using credible structured covariance patterns has great appeal for estimation. We speculate that good estimation can be achieved as long as the number of independent sampling units (not observations) substantially exceeds the number of covariance parameters (not the dimension of the covariance matrix). An important caution comes from the observation that inference for small samples based on structured covariance models in Gaussian mixed models still needs improvement in many ways (Muller and Stewart, 2006, Chapter 18; Orelien and Edwards, 2008).

Third, we recommend scientifically informed reduction to summary statistics to avoid HDLSS. The fear of losing information creates a barrier. When valid, the approach can greatly increase precision, as well as greatly simplify analysis and interpretation.

Fourth, analyzing the response variables in meaningful groups can find a comfortable middle ground between the rock of multiple comparisons and the hard spot of HDLSS. Avoiding HDLSS allows applying classical multivariate theory with data dimensions for which validity of estimation and inference can be assured.

APPENDIX: PROOFS

Theorems 1 and 3. use the same structure. Inner and outer products of $\boldsymbol{Y}$ have the same ν nonzero eigenvalues. Constituent matrix decomposition gives the weighted sum. Independence of $\{\boldsymbol{W}_j\}$ follows from independence of distinct subsets of $\boldsymbol{Z} = \{z_{ij}\}$.

Corollary 1.1. Given $\boldsymbol{S}_D = \sum_{j=1}^{p_+} \lambda_j \boldsymbol{W}_j$ and $\phi(\boldsymbol{T}; \boldsymbol{W}_j) = |(\boldsymbol{I}_\nu - 2\iota\boldsymbol{T})|^{-1/2}$, statistical independence of $\{\lambda_j \boldsymbol{W}_j\}$ implies $\phi(\boldsymbol{T}; \boldsymbol{S}_D) = \prod_{j=1}^{p_+} \phi(\boldsymbol{T}; \lambda_j \boldsymbol{W}_j)$.

Corollary 1.2 and 3.2. Properties of independent sums, the special case of diagonal population covariance, and moments in Wishart (1928) lead directly to the moments.

Theorem 2 a) Inner ($\boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y}$) and outer ($\boldsymbol{S}_D = \boldsymbol{Y}\boldsymbol{Y}'$) products have the same eigenvalues. In turn $\boldsymbol{Y}\boldsymbol{Y}' = [\boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda})^{1/2}\boldsymbol{\Upsilon}'][\boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})^{1/2}\boldsymbol{Z}'] = \boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda})\boldsymbol{Z}'$ which has

rank ν, and $\boldsymbol{Z} \sim \mathcal{N}_{\nu,p}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{I}_p)$. If $\epsilon = 1$, then $\boldsymbol{S}_D = \overline{\lambda}\boldsymbol{Z}\boldsymbol{Z}' \sim \mathcal{W}_\nu\big(p, \overline{\lambda}\boldsymbol{I}_\nu\big)$.

Theorem 2 c) If $\nu = 1$ then $\boldsymbol{Y}\boldsymbol{Y}' = \boldsymbol{Z}\mathrm{Dg}(\boldsymbol{\lambda})\boldsymbol{Z}' = \sum_{j=1}^{p} z_j^2\lambda_j$ is a quadratic form with independent $z_j \sim \mathcal{N}(0,1)$ and $z_j^2 \sim \chi^2(1,0)$.

Corollary 3.1. The proof of Corollary 1.1 with p replaced by p_+ applies.

Theorem 4 statement. The eigenvalues of $\boldsymbol{S} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$, coincide with the eigenvalues of $\boldsymbol{\Upsilon}'\boldsymbol{S}\boldsymbol{\Upsilon} \sim \mathcal{W}_p[\nu, \mathrm{Dg}(\boldsymbol{\lambda})]$. With $\boldsymbol{Z} = [\,\boldsymbol{Z}_+ \;\boldsymbol{Z}_0\,]$,
$\boldsymbol{\Upsilon}'\boldsymbol{S}\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}'[\boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})^{1/2}\boldsymbol{Z}'[\,\boldsymbol{Z}_+ \;\boldsymbol{Z}_0\,]\mathrm{Dg}(\boldsymbol{\lambda})^{1/2}\boldsymbol{\Upsilon}']\boldsymbol{\Upsilon} = \begin{bmatrix} \mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}\boldsymbol{Z}_+'\boldsymbol{Z}_+\mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$.
Hence the p_+ nonzero eigenvalues of $p \times p$ $\boldsymbol{\Upsilon}'\boldsymbol{S}\boldsymbol{\Upsilon}$ coincide with those of $\mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2}\boldsymbol{Z}_+'\boldsymbol{Z}_+\mathrm{Dg}(\boldsymbol{\lambda}_+)^{1/2} \sim \mathcal{W}_{p_+}[\nu, \mathrm{Dg}(\boldsymbol{\lambda}_+)]$.

Theorem 4a). Like the proof of Theorem 2a). If $\epsilon_+ = 1$ and $\boldsymbol{Z}_* \sim \mathcal{N}_{\nu,p_+}(\boldsymbol{0}, \boldsymbol{I}_\nu, \boldsymbol{I}_{p_+})$, $\boldsymbol{Y}\boldsymbol{Y}' = \overline{\lambda}_+[\,\boldsymbol{Z}_* \;\boldsymbol{0}\,][\,\boldsymbol{Z}_* \;\boldsymbol{0}\,]' = \overline{\lambda}_+\begin{bmatrix} \boldsymbol{Z}_*\boldsymbol{Z}_*' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$ has the same nonzero eigenvalues as $\overline{\lambda}_+\boldsymbol{Z}_*\boldsymbol{Z}_*' \sim \mathcal{W}_\nu\big(p_+, \overline{\lambda}_+\boldsymbol{I}_\nu\big)$ with $p_+ > \nu$.

Theorem 4c). If $\lambda_1 > 0$, $\lambda_j \geq \lambda_{j+1}$ and $\epsilon = 1/p$, then $[\big(\sum_{j=1}^{p}\lambda_j\big)^2/\lambda_1^2]/[\big(\sum_{j=1}^{p}\lambda_j^2\big)/\lambda_1^2] = 1 = \big(1 + \sum_{j=2}^{p} r_j\big)^2/\big(1 + \sum_{j=2}^{p} r_j^2\big)$ for $r_j = \lambda_j/\lambda_1$. If $r_2 > 0$ and other $r_j = 0$ then $(1+2r_2+r_2^2)/(1 + r_2^2) > 1$ which is a contradiction. The same logic applies to $\{\widehat{\lambda}_{(j)}\}$. The special case $p_+ = 1$ in $\boldsymbol{Y}\boldsymbol{Y}' = \overline{\lambda}_+[\,\boldsymbol{Z}_* \;\boldsymbol{0}\,][\,\boldsymbol{Z}_* \;\boldsymbol{0}\,]'$ applies. Here $\overline{\lambda}_+\boldsymbol{z}_*\boldsymbol{z}_*' \sim \mathcal{W}_\nu\big(1, \overline{\lambda}_+\boldsymbol{I}_\nu\big)$, has one nonzero eigenvalue, $\overline{\lambda}_+\boldsymbol{z}_*'\boldsymbol{z}_*$, with $\boldsymbol{z}_*'\boldsymbol{z}_* \sim \chi^2(\nu)$.

Theorem 5. The results follow from generalizing combinations of previous results.

Theorem 6a) $\mathrm{E}(\boldsymbol{S}_{*1}) = p \cdot \overline{\lambda}_+\boldsymbol{I}_\nu = \mathrm{E}(\boldsymbol{S}_D)$.

Theorem 6b) Moments of elements are $\mathrm{E}(\langle\boldsymbol{S}_D\rangle_{jk}) = M_1(j,k)$ and $\mathrm{E}[(\langle\boldsymbol{S}_D\rangle_{jk})^2] = M_2(j,k)$, with $\langle\boldsymbol{S}_D\rangle_{kk}$, a weighted sum of independent χ^2. A Satterthwaite approximation (Mathai and Provost, 1992) matches $M_1(j,j)$ and $M_2(j,j)$ to X_* with $X_*/\lambda_* \sim \chi^2(p_*)$. Hence $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p_*, \lambda_*\boldsymbol{I}_\nu)$ gives $\mathrm{E}(\langle\boldsymbol{S}_D\rangle_{jj}) = \mathrm{E}(\langle\boldsymbol{S}_{*2}\rangle_{jj})$ and $\mathrm{E}[(\langle\boldsymbol{S}_D\rangle_{jj})^2] = \mathrm{E}[(\langle\boldsymbol{S}_{*2}\rangle_{jj})^2]$. For $j \neq k$ $\mathrm{E}(\boldsymbol{S}_{*2}) = p_*\lambda_*\boldsymbol{I}_\nu = p\overline{\lambda}\boldsymbol{I}_\nu = \mathrm{E}(\boldsymbol{S}_D)$ gives $\mathrm{E}(\langle\boldsymbol{S}_D\rangle_{jk}) = \mathrm{E}(\langle\boldsymbol{S}_{*2}\rangle_{jk})$, $\mathrm{E}[(\langle\boldsymbol{S}_D\rangle_{jk})^2] = \sum_{j=1}^{p}\lambda_j^2$ and $\mathrm{E}[(\langle\boldsymbol{S}_{*2}\rangle_{jk})^2] = p_*\lambda_*^2 = \sum_{j=1}^{p}\lambda_j^2$.

All other second order moments are zero for $\boldsymbol{S}_D$ and $\boldsymbol{S}_{*2}$ due to diagonal covariance for $\boldsymbol{S}_{*2}$ and $\{\boldsymbol{W}_j\}$, by equations 4 and 6-9 in Wishart (1928, p44). Also 18 of 21 types of order 3 and 69 of 79 types of order 4 moments are zero for $\boldsymbol{S}_D$ and $\boldsymbol{S}_{*2}$.

REFERENCES

Ahn, J., Marron, J. S., Muller, K. E., and Chi, Y. Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, **94**, 760-766.

Anderson, T. W. (2004) *An Introduction to Multivariate Statistical Analysis.* 3rd ed. New York: Wiley.

Baik, J., Ben, A. G., Peche, S. (2005). Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Annals of Probability* **33**, 1643-1697.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382-1408.

Cascio, C. J., Gribbin, M. J., Gouttard S., Smith R. G., Jomier M,, Poe, M. D., Graves, M., Hazlett, H. C., Muller, K. E., Gerig, G., and Piven, J. (2008) Decreased variability of fractional anisotropy in young children with autism, *in review*.

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29**, 323-333.

Johnson, N. L. and Kotz, S. (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics*, **29**, 295-327.

Khatri, C. G. (1976) A note on multiple and canonical correlation for a singular covariance matrix, *Psychometrika*, **41**, 465- 470.

MacCallum, R. C., Widaman, K. F., Zhang, S. and Hong, S. (1999) Sample size in factor analysis, *Psychological Methods*, **4**, 84-99.

Mathai, A. M. and Provost, S. B. (1992) *Quadratic Forms in Random Variables*. New York: Marcel Dekker.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the LASSO, *The Annals of Statistics*, **34**, 1436-1462.

Muller, K. E. and Stewart, P. W. (2006) *Linear Model Theory for Univariate, Multivariate and Mixed Models.* New York: Wiley.

Orelien, J. G. and Edwards, L. J. (2008). Fixed effect variable selection in linear mixed models using R^2 statistics. *Computational Statistics and Data Analysis*, **52**, 1896-1907.

Preacher, K. J. and MacCallum, R. C. (2002) Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes, *Behavior Genetics*, **32**, 153- 161.

SAS Institute (1999) *SAS/IML® Software*. Cary, North Carolina: SAS Institute.

Schott, J. R. (1997) *Matrix Analysis for Statistics*. New York: Wiley.

Uhlig, H. (1994) On singular Wishart and singular multivariate beta distributions, *Annals of Statistics*, **22**, 395-405.

Widaman, K. F. (1993) Common factor analysis versus principal component analysis: differential bias in representing model parameters? *Multivariate Behavioral Research*, **28**, 263-311.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**, 32-52.

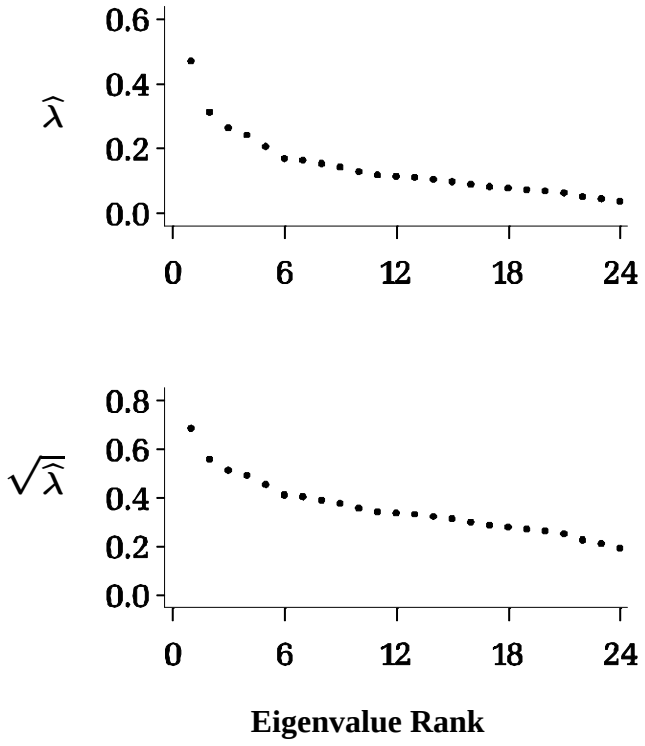

Figure 1. Sample ordered eigenvalues and their square roots for the residual sample covariance matrix of DTI data for $\nu = 24$ and $p = 387$.

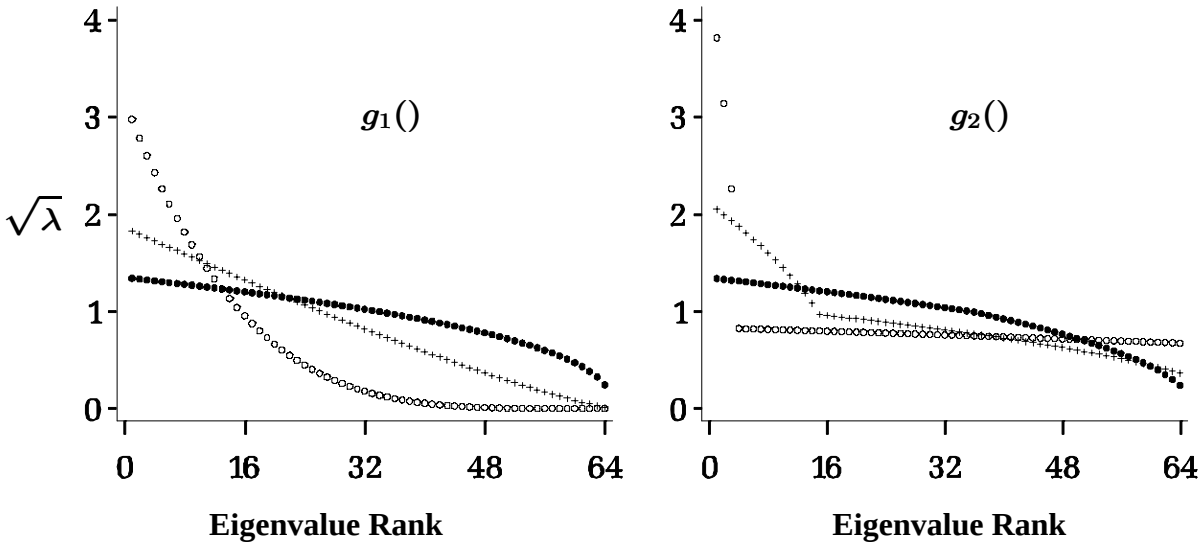

Figure 2. Square root of ordered eigenvalues of $\mathbf{\Sigma}$ as a function of order and ϵ: $\epsilon = 0.2$ open circle; $\epsilon = 0.5$ $+$; $\epsilon = 0.8$ solid circle.

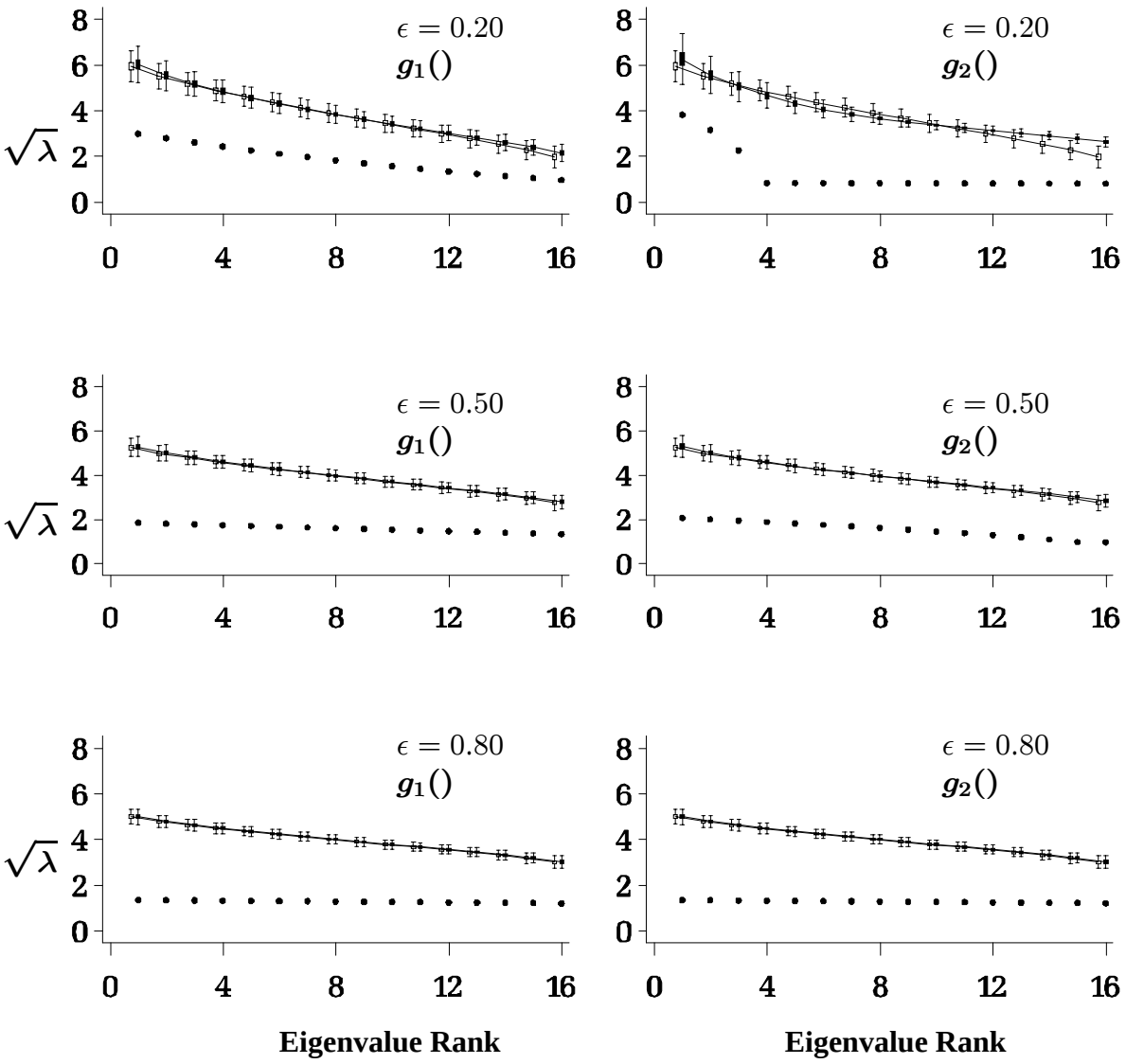

Figure 3. Box plots for sample-ordered square roots of eigenvalues for $\nu = 16$, $p = 256$. • for $\mathbf{\Sigma}$, solid boxes with — for $\widehat{\mathbf{\Sigma}}_{g1}$ in the left column and $\widehat{\mathbf{\Sigma}}_{g2}$ in the right column, open boxes for $\widehat{\mathbf{\Sigma}}_{*2}$.

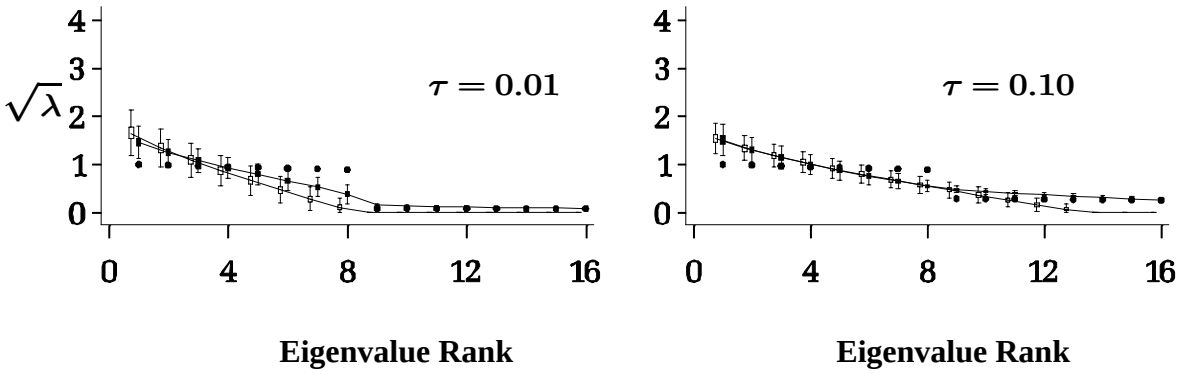

Figure 4. Square root of ordered eigenvalues for $\nu = 16$, $p = 256$, $p_1 = 8$ and $g_4(j, \pi, p, p_1, \tau)$.
• for $\mathbf{\Sigma}$, solid boxes with —— for **actual** $\widehat{\mathbf{\Sigma}}_{g4}$, **open** boxes for **approximation** $\widehat{\mathbf{\Sigma}}_{*2}$

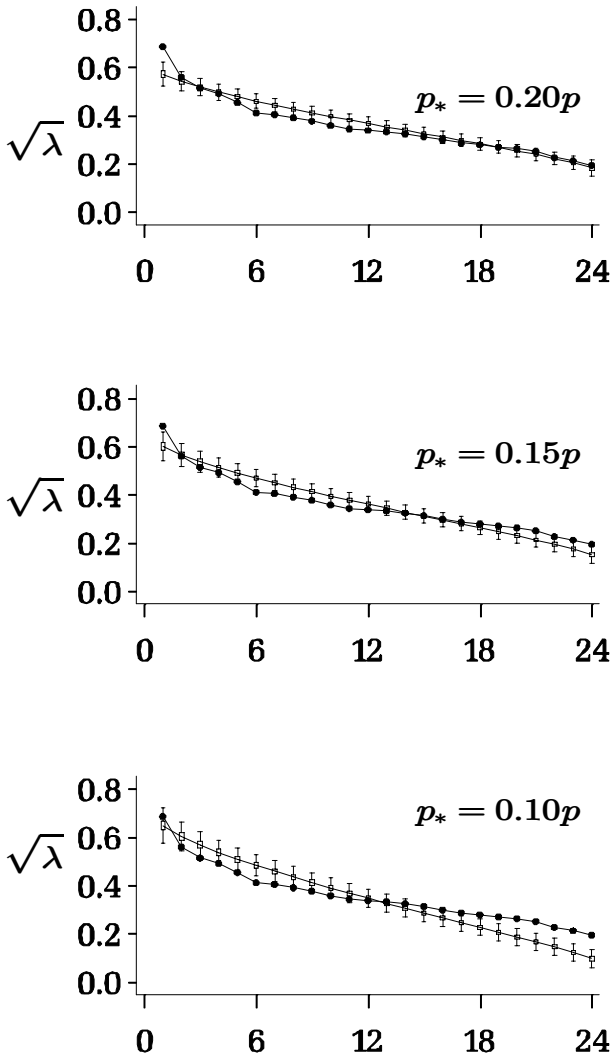

Figure 5. Square roots of sample ordered eigenvalues (black dots) for the residual covariance matrix of DTI data ($p = 387$, $\nu = 24$) and for 10,000 simulated samples of $\widehat{\boldsymbol{\Sigma}}_{*2}$ (box plots ± 1.5 IQR) with $\boldsymbol{S}_{*2} \sim \mathcal{W}_\nu(p_*, \lambda_* \boldsymbol{I}_\nu)$, $\overline{\lambda} = 0.009$, $p_* \in \{0.10p, 0.15p, 0.20p\}$, and $\lambda_* \in \{\overline{\lambda}/0.10, \overline{\lambda}/0.15, \overline{\lambda}/0.20\}$.