Published in final edited form as:
Commun Stat Simul Comput. 2013 ; 42(5): 1113–1125. doi:10.1080/03610918.2012.659819.

Two-Step Hypothesis Testing When the Number of Variables Exceeds the Sample Size

YUEH-YUN CHI[1] and **KEITH E. MULLER**[2]

[1]Department of Biostatistics, University of Florida, Gainesville, Florida, USA

[2]Department of Health Outcomes and Policy, University of Florida, Gainesville, Florida, USA

Abstract

Medical images and genetic assays typically generate data with more variables than subjects. Scientists may use a two-step approach for testing hypotheses about Gaussian mean vectors. In the first step, principal components analysis (PCA) selects a set of sample components fewer in number than the sample size. In the second step, applying classical multivariate analysis of variance (MANOVA) methods to the reduced set of variables provides the desired hypothesis tests. Simulation results presented here indicate that success of the PCA in the first step requires nearly all variation to occur in population components far fewer in number than the number of subjects. In the second step, multivariate tests fail to attain reasonable power except in restrictive, favorable cases. The results encourage using other approaches discussed in the article to provide dependable hypothesis testing with high dimension, low sample size data (HDLSS).

Keywords

Eigenvalues estimation; HDLSS; MANOVA; Principal component analysis

1. Introduction

1.1. *Motivation*

High Dimension, Low Sample Size data (HDLSS: more variables than independent sampling units) have become ubiquitous in medical imaging and biomedical research. Scientists may use a two-step process to test hypotheses about multivariate Gaussian means. First, Principal Components Analysis (PCA) selects a small number of components associated with the largest eigenvalues for dimension reduction. Second, classical multivariate analyses, that is, Multivariate Analysis of Variance (MANOVA), apply to the selected subset of components for global hypothesis testing. However, few results are available to support the validity of either step for HDLSS data, especially as a two-step process.

Address correspondence to Yueh-Yun Chi, Department of Biostatistics, University of Florida, Gainesville, FL, USA; yychi@ufl.edu.

One of our two goals was to characterize sample PCA performance for HDLSS data. PCA uses the sample-ordered eigenvalues and eigenvectors of the sample covariance or correlation matrix to estimate the population eigenvalues and eigenvectors. Eigenvalues quantify the relative importance of independent PCA directions, depicted by the corresponding eigenvectors. When the number of variable exceeds the sample size, both the sample covariance and correlation matrices become rank deficient which in turn affects estimation properties. Most importantly, no matter how large the number of variables becomes, fixing the sample size also fixes the number of nonzero eigenvalues that can be estimated as greater than zero. Clearly, the absolute and the relative dimensions of the data matrix have important effects on the performance of PCA. The underlying population covariance and correlation structure interacts with the dimensions to determine the reliability and validity of sample PCA analysis. Extensive simulations allow discerning conditions when PCA will succeed and when it will fail.

Our second goal was to evaluate the success of hypothesis testing with the two-step approach for HDLSS data. PCA reduces the dimensionality to allow applying MANOVA methods. Accurate and efficient testing requires the sample eigenvalues and eigenvectors from PCA in the first step to reliably identify the subspace corresponding to important population eigenvalues and important mean differences. Extensive simulations allow characterizing the performance of the multivariate hypothesis tests in the two-step approach.

1.2. An Example

Cascio et al. (2011) used diffusion tensor imaging (DTI) to compare heterogeneity in brain regions of interest in groups of 10 autistic, 22 normal, and 10 developmentally delayed children. Data for the left cerebellum (one of many regions of interest) display the HDLSS problem because each of the children have 387 values of fractional anisotropy (a measure of heterogeneity), one per voxel (cube in three-space). Although PCA can be computed on the data, can we have confidence that the results reflect the population structure?

A multivariate linear model adjusted the 387 response variables for age, gender, and their interactions for the 32 normal and developmentally delayed children. With four degrees of freedom for the model, the residuals were based on $\nu = 32 - 4 = 28$ error degrees of freedom, and roughly 14 variables per degree of freedom. PCA of the residual covariance matrix gave Fig. 1, which includes the traditional “scree” plot of sample-rankorder estimated eigenvalues $\{\hat{\lambda}_j\}$, as well as a plot of $\{\hat{\lambda}_j^{1/2}\}$. Of the 28 nonzero components, the first 21 accounted for 90% of the variation. Component 1 explained 11.9%, while the next 27 accounted for 8.6% to 0.9%. A MANOVA test on the 28 sample components gave *p*-value of 0.8286 indicating no heterogeneity in brain activity between normal and developmentally delayed children.

1.3. *Relevant Literature*

Johnson and Kotz (1972) and Anderson (2004) summarized Wishart theory as used in PCA with Gaussian data. Most results require nonsingular population covariance and more observations than variables. Khatri (1976) relaxed the first requirement, while Uhlig (1994) described relaxing the second (more observations than variables). Muller and Stewart (2006)

gave a Wishart taxonomy for all combinations of finite variable dimension, population covariance rank, and sample size, including HDLSS cases.

Simulation evidence emphasizes the importance of the ratio of observations to variables in determining the stability of factor analysis (MacCallum et al., 1999). Preacher and MacCallum (2002) extended the conclusions to sample sizes as small as 10, including some HDLSS cases. The results suggest the same conclusion holds for PCA, a special case of factor analysis. Widaman (1993) recommended factor analysis over PCA for dimension reduction.

Asymptotic methods for HDLSS data may be evaluated in two distinct performance domains: *Np*-asymptotic when $N \to \infty$, $p \to \infty$, $p/N \to c$ a constant, or *p*-asymptotic when $N < p \to \infty$. Johnstone (2001) described the limiting density of the largest sample eigenvalue with a *Np*-asymptotic approach when all population eigenvalues equal. For population covariance matrices near to an identity matrix, Baik et al. (2005) and Baik and Silverstein (2006) discovered that the sample eigenvalues behave as if the covariance matrix were the identity matrix as both $N, p \to \infty$ at a fixed rate. Johnstone and Lu (2009) described PCA consistency and sparseness conditions with an *Np*-asymptotic approach.

Hall et al. (2005) and Ahn et al. (2007) described the geometric structure of HDLSS data under *p*-asymptotic. Ahn et al. also discussed implications of using PCA with HDLSS data. Jung and Marron (2009) studied *p*-asymptotic behavior of the principal component directions and gave broad set of sufficient conditions for the performance of PCA. They proved that as $p \to \infty$ if the first few eigenvalues of population covariance matrix are large enough compared to the other, then the corresponding principal component directions are consistent or converge to the appropriate subspace and most other principal component directions are strongly inconsistent. All of the results share a common feature: success in HDLSS PCA requires a simple covariance structure, at least asymptotically.

In addition to the two-step approach of a PCA dimension reduction followed by multivariate tests, some hypothesis testing methods have been proposed to work around the limitations induced by HDLSS. We highlight the following as the best-developed. All require the *Np*-asymptotic assumption except the regularization method. Warton (2008) used regularized correlation or covariance matrices to compute an analog of the Hotelling-Lawley multivariate statistic. Srivastava and Fujikoshi (2006) proposed a statistic that remains well-defined with the less-than-full-rank sample covariance matrix. Srivastava (2007) developed multivariate tests based on the Moore-Penrose generalized inverse of the sample covariance matrix. Srivastava and Du (2008) inverted the diagonal matrix of sample variances and used it to compute an analog of the Hotelling-Lawley statistic.

1.4. *Overview of the Article*

The finite sample properties of the two-step hypothesis testing approach are investigated. Section 2 details numerical results about dimension reduction via PCA, the first step. Section 3 contains extensive simulations for the Type I error rate and statistical power of MANOVA tests on the reduced principal component set. The key result is that success of PCA as a visualization tool or an intermediate step for hypothesis testing require population

variation to be dominated by the first few principal components. The two-step hypothesis testing approach can have reasonable control of the Type I error rates. However, the approach has very low power unless the number of dominant components is sufficiently less than the sample size, group mean differences arise along the dominant principal component directions, and only a few sample principal components are retained. Section 4 summaries our conclusions and gives recommendation for making inferences with HDLSS data.

2. PCA with HDLSS

We use $\mathbf{1}_N$ to denote the $N \times 1$ vector of ones and $\boldsymbol{I}$ to denote the $p \times p$ identity matrix. Let the $N \times p$ outcome matrix $\boldsymbol{Y}$ contains N independent observations on p variables with mean $\mathcal{E}(\boldsymbol{Y}) = \mathbf{1}_N\boldsymbol{\mu}'$, covariance matrix $\mathcal{V}(\boldsymbol{Y}) = \Sigma$, and correlation matrix $\boldsymbol{R} = \mathrm{Dg}^{-1/2}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}\mathrm{Dg}^{1/2}(\boldsymbol{\Sigma})$. Without loss of generality, we focus primarily on full rank Σ for all the subsequent development. Parallel results can be derived for full rank $\boldsymbol{R}$. With $\boldsymbol{\lambda}$ the population eigenvalues and $\mathbf{Y}$ the corresponding orthonormal ($\mathbf{Y}'\mathbf{Y} = \mathbf{I}_p$) population eigenvectors, the eigenvalue decomposition gives $\Sigma = \mathbf{Y}\mathrm{Dg}(\boldsymbol{\lambda})\mathbf{Y}'$.

PCA uses the sample-ordered eigenvalues and corresponding eigenvectors of the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \left(\boldsymbol{Y} - 1\hat{\boldsymbol{\mu}}'\right)' \left(\boldsymbol{Y} - 1\hat{\boldsymbol{\mu}}'\right) / (N-1)$ or correlation matrix $\hat{\boldsymbol{R}} = \mathrm{D}g^{-1/2}\left(\hat{\boldsymbol{\Sigma}}\right)\hat{\boldsymbol{\Sigma}}Dg^{-1/2}\left(\hat{\boldsymbol{\Sigma}}\right)$ to estimate the population counterparts. For Gaussian distributed $\boldsymbol{Y}$ when the eigenvalues of Σ are distinct and $(N-1) \geq p$, the sample eigenvalues and eigenvectors are maximum likelihood estimators of the corresponding population parameters (Mardia et al., 1979). Consequently, the sample eigenvalues and eigenvectors are consistent and asymptotically unbiased.

The validity of PCA when $(N-1) < p$ lies on the validity of eigenvalue and eigenvector estimations. Jung and Marron (2009) studied the asymptotic behavior of sample eigenvectors, hence PCA directions, when $p \to \infty$, and documented their consistency when the first few population eigenvalues are large enough compared to the others. In contrast to Jung and Marron (2009), we set to examine the validity of eigenvalue estimation when $(N-1) < p$ and p is finite.

The relative importance of PCA directions is characterized by sample eigenvalues and varies with the population pattern of eigenvalues. The population average eigenvalue (first moment) has no role in predicting the accuracy of eigenvalue estimation except in limiting cases (population average eigenvalue near zero or infinity). In contrast, the second moment can tell us a great deal about the performance of the estimation. With eigenvalues $\boldsymbol{\lambda} = [\lambda_1,\ldots\lambda_p]'$, the standard sphericity parameter

$$\epsilon \equiv \frac{\left(\sum_{k=1}^{p} \lambda_k\right)^2}{p\sum_{k=1}^{p} \lambda_k^2} \quad (1)$$

quantifies the spread of population eigenvalues and the sphericity of the population components. Maximum sphericity requires $\epsilon = 1$ which occurs with all eigenvalues equal.

Minimal sphericity has $\epsilon = 1/p$ which occurs with one nonzero eigenvalue. Overall, $1/p \leq \epsilon \leq 1$.We will investigate eigenvalue estimation in relation to ϵ in the simulation.

Athree-way factorial design using factors $p \in \{64, 256, 1024\}$ so $\log_2(p) \in \{6, 8, 10\}$, $N \in \{4, 8, 16, 32\}$, and $\epsilon \in \{0.20, 0.50, 0.80\}$ were considered. For $i \in \{1,...N\}$, we had $y_i \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$. The following lemma allows simplification of our simulation designs by considering only diagonal $\boldsymbol{\Sigma}$ or $\boldsymbol{R}$.

Lemma 2.1

With $\boldsymbol{\Sigma} = \Upsilon Dg(\lambda)\Upsilon'$ *and* $\Upsilon'\Upsilon = \boldsymbol{I}_p$, $\hat{\boldsymbol{\Sigma}}$ and $\Upsilon' \hat{\boldsymbol{\Sigma}} \Upsilon$ *share the same eigenvalues. Similarly,* $\hat{\boldsymbol{R}}$ *and* $\Upsilon_R' \hat{\boldsymbol{R}} \Upsilon_R$ *share the same eigenvalues*

We assumed population eigenvalues $\{\lambda_j\}$ decrease smoothly at a rate determined by $\boldsymbol{\pi}$, which was selected iteratively to fix $\epsilon \in \{0.2, 0.5, 0.8\}$ for each p by having

$$\lambda_j = g_1(j, \pi, p) = [1 - (j-1)/p]^{\pi}. \quad (2)$$

All simulations were conducted with SAS/IML® (SAS Institute, 1999) and summarized from 10,000 replication in each condition. Fig. 2 displays population values and box plots (±1.5 IQR) of the square roots of the largest 16 sample eigenvalues, as a function of number of variables and sphericity ϵ for $N = 16$, and $p = 64$, 256,or 1024. The medians of the largest, nonzero 16 sample eigenvalues were away from their population counterparts except when $p = 64$ and $\epsilon = 0.2$, the condition that 93% of population variation was accounted for by the top 16 PCA directions. As the number of variables increased, so as the ratio of the number of variables to the sample size, the discrepancy between sample and population eigenvalues grew. Furthermore, as eigenvalue sphericity diminished, that is, ϵ decreased, the closer the sample eigenvalues were to the population parameters. The small the ϵ, the more likely a few largest PCA directions dominate and hence the less the bias from the sample eigenvalue estimation. Similar results were obtained for $N \in \{4, 8, 32\}$ and skipped for presentation.

The second simulation used eigenvalue patterns with major departures from sphericity (small ϵ). We used two nonlinearly decreasing groups of population eigenvalues with a wide separation between the groups:

$$\lambda_j = g_2(j, \pi, p, p_1, \tau) = \begin{cases} (1-\tau)\, g_1(j, \pi, p) + \tau g_1(j, \pi, p) & j \leq p_1 \\ \tau g_1(j, \pi, p) & j > p_1. \end{cases} \quad (3)$$

If $j \leq p_1$, then $g_2 = g_1$, and if $j > p_1$, a discount factor $\boldsymbol{\tau} > 0$ reduces the size of the g_1 eigenvalues. The parameter p_1 represents the number of dominant population components. Changing $\boldsymbol{\tau}$ changes the gap between the two groups and thus the dominance of the first group. A two-way factorial used $N \in \{4, 8, 16\}$, and $\boldsymbol{\tau} \in \{0.01, 0.05, 0.10, 0.20\}$ giving $\epsilon \in \{0.03, 0.04, 0.05, 0.08\}$ with fixed $p = 256$, $p_1 = 8$, and $\boldsymbol{\pi} = 8.5118$. The ratio of mean eigenvalues between the groups was $\left(\sum_{j=1}^{p_1} \lambda_j / p1\right) / \left[\sum_{j=p_1+1}^{p} \lambda_j / (p - p_1)\right] \in \{1079, 207, 98, 44\}$.

Figure 3 displays population values and box plots (±1.5 IQR) of the square roots of the nonzero sample eigenvalues, as a function of sample size and the parameter τ. Compared to Fig. 2, sample eigenvalues were closer to population eigenvalues as the small ϵ led to dominance of the first $p1 = 8$ PCA directions; however, sample estimates showed no clear indication of separation of the two population eigenvalue groups, except when $N = 16$ and $\tau = 0.01$. The result suggested that when $N < p$, the sample size must be greater than the number of highly dominant (τ very small) PCA directions ($p_1 < p$) to effectively separate the first p_1 sample eigenvalues from the remaining $p - p_1$. The following lemma and corollary are deduced from this result.

Lemma 2.2

If $\boldsymbol{\Sigma}=\boldsymbol{\Upsilon}Dg(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$ *with* $\boldsymbol{\Upsilon} = [\boldsymbol{\Upsilon}_1\boldsymbol{\Upsilon}_2]$, $\boldsymbol{\lambda}=\left[\boldsymbol{\lambda}_1'\tau\boldsymbol{\lambda}_2'\right]'$ *of* p_1 *positive values in* λ_1, *and* $(p - p_1)$ *positive values in* λ_2, *then* $\boldsymbol{\Sigma}=\boldsymbol{\Phi}\boldsymbol{\Phi}'$ *with* $\boldsymbol{\Phi} = \boldsymbol{\Upsilon}Dg(\lambda)^{1/2} = [\boldsymbol{\Phi}_1\boldsymbol{\tau}^{1/2}\boldsymbol{\Phi}_2]$, $\boldsymbol{\Phi}_j = \boldsymbol{\Upsilon}_jDg(\lambda_j)^{1/2}$. *As* $\tau\to 0$, $\boldsymbol{S}=\boldsymbol{YY}'$ *has characteristic function*

$$\begin{aligned}\lim_{\tau\to 0}[\phi(\boldsymbol{T};\boldsymbol{S})] &= \lim_{\tau\to 0}\left(|\boldsymbol{I}_p - 2\iota\boldsymbol{\Phi}'\boldsymbol{T}\boldsymbol{\Phi}|^{-N/2}\right)\\ &= \lim_{\tau\to 0}\left(\left|\boldsymbol{I}_p - 2\iota\begin{bmatrix}\boldsymbol{\Phi}_1'\boldsymbol{T}\boldsymbol{\Phi}_1 & \tau^{1/2}\boldsymbol{\Phi}_1'\boldsymbol{T}\boldsymbol{\Phi}_2\\ \tau^{1/2}\boldsymbol{\Phi}_1'\boldsymbol{T}\boldsymbol{\Phi}_2 & \tau\boldsymbol{\Phi}_2'\boldsymbol{T}\boldsymbol{\Phi}_2\end{bmatrix}\right|^{-N/2}\right)\\ &= |\boldsymbol{I}_{p1} - 2\iota\boldsymbol{\Phi}_1'\boldsymbol{T}\boldsymbol{\Phi}_1|^{-N/2},\end{aligned} \quad (4)$$

with $\langle\boldsymbol{T}\rangle_{jj} = u_{jj}$ *and* $\langle\boldsymbol{T}\rangle_{jk} = u_{jk}/2$ *for symmetric* $\boldsymbol{U}$. The last line of the lemma reduces the dimensions to $p_1 \times p_1$, *with* $p_1 \le p$. Equivalently, only the small number of very large eigenvalues matter as convergence in characteristic function leads to convergence in distribution. We describe such situations with a very strong signal and almost no noise in the following proposition.

Proposition 2.1

For $\lambda_k \ge \lambda_{k+1}$, *simulation results indicate that if* $N \ge p_1$ *and* $|\sum_{l=1}^{p_1}\lambda_k/\sum_{k=1}^{p}\lambda_k - 1| \le .03$, *then* p_1 *can be reliably identified in a data analysis*

3. MANOVA on Sample Principal Components

Many scientists with HDLSS data use selected sample principal components as surrogate outcomes in classical multivariate analysis. The importance of each component is determined by sample-ordered eigenvalues and can be transformed into percentage of total variation accounted for by the component. Common practice for deciding the number of components is to include the top three important components for the sake of simplicity and visualization, or to retain the smallest number of components that collectively account for at least 90% of total variation. To examine the accuracy of inference based on dimension reduction from PCA with either rule, we conducted a series of simulations to compute empirical type I error rates and statistical power.

HDLSS data were generated for a two-sample comparison with $N = 18$ (9 per group), $p \in \{4, 16, 64, 256\}$ and 10,000 replications. Two covariance structures were considered, one with autoregressive model of order one, AR(1). The AR(1) has a common variance of $\sigma^2 = 1$ and correlation $\rho_{AR} \in \{0.5, 0.8\}$. The second structure arises from a Kronecker product of a 2 × 2 unstructured covariance, $\mathbf{\Sigma}_u$, and an AR(1) of dimension $p/2$, giving $\mathbf{\Sigma}_u \otimes$AR(1; $p/2$), with $\boldsymbol{A} \otimes \boldsymbol{B} = \{a_{ij}\boldsymbol{B}\}$. Here, $\mathbf{\Sigma}_u$ has variances of $\sigma_1^2 = 1$ and $\sigma_2^2 \in \{2, 3\}$ to vary the ratio of σ_2^2 to σ_1^2 while $\rho_u = 0.5$. The AR(1; $p/2$) had $\rho_{AR} = 0.5$ with $\sigma^2 = 1$.

For $p \in \{4, 16\}$, PCA was not performed for the sample covariance matrices of full rank with error degrees of freedom of 16. For $p \in \{64, 256\}$, PCA provided the sample-ordered eigenvalues and corresponding eigenvectors, which were used for dimension reduction. Retaining fewer components than the error degrees of freedom allowed using a MANOVA test for overall mean differences. The number of components retained was either (1) fixed at 3, or (2) empirically chosen to be the number that led to at least 90% of total variation accounted, or (3) fixed at the maximal number 16. The third choice corresponded to using the Moore–Penrose generalized inverse in place of the sample covariance inverse in the calculation of MANOVA statistics.

Table 1 summarizes empirical type I error rates for 10,000 replications with $\alpha = 0.05$. The first four rows involve conditions with sample sizes large enough to allow applying MANOVA directly, without dimension reduction. The remaining rows involve $N < p$ and MANOVA tests performed on the reduced PCA dimensions. All conditions resulted in an acceptable control of type I error rate regardless of the covariance structure, number of variables, or number of PCA components retained.

We examined the power properties under three alternatives. For each combination of covariance structure and number of variables, we assumed that the mean vector is shifted:

1. along the first population PCA direction by $a_0\sqrt{\lambda_1}$ units, or
2. along the last population PCA direction by $a_0\sqrt{\lambda_p}$ units, or
3. along all population PCA directions by $a_0\sqrt{\lambda_j/p}$ units, respectively

with population-ordered eigenvalues $\boldsymbol{\lambda} = [\lambda_1\lambda_2\cdots\lambda_p]$. We chose $a_0 = 2$ for conditions 1 and 2, and $a_0 = 3$ for condition 3 to obtain moderate power in the comparisons.

Table 2 summarizes statistical power for 10,000 replications with $\alpha_0 = 0.05$ for condition 1. When the number of variables was sufficiently smaller than the error degrees of freedom, the MANOVA test had good power in detecting the overall group difference; whereas when the number of variables was equal to the error degrees of freedom, power was very poor. Despite its computability, MANOVA tests were very insensitive for the detection of overall group differences when the number of variables grew close to the sample size. As the number of variables increased and exceeded the error degrees of freedom, power decreased and the reduction varied inversely with the number of components retained. When the mean differences were highly concentrated in the first PCA direction, the more the dimension reduction by PCA, the better the statistical power. Performing MANOVA tests on the entire

sample principal components, that is, using a Moore–Penrose generalized inverse in the MANOVA statistics, gave very little power. The heterogeneity of the eigenvalues, as measured by ϵ, also played a role in determining power. Namely as ϵ (sphericity) decreased, dominant population components emerged and hence better power could be attained.

Table 3 lists power for condition 2, with the mean differences concentrated and shifted along the last PCA direction. Adequate power was attained only when the number of variables was sufficiently smaller than the sample size. As the number of variables grew close to or exceeded the sample size, statistical power for detecting differences on the least important component was very low. The MANOVA tests on the selected sample components became very inefficient when the group mean differences arose from least important components. Including a larger number of components gave some but rather limited improvement on power.

Lastly, Table 4 summarizes power when the mean differences were diffused and shifted along all PCA directions (condition 3). Power decreased rapidly as the number of variables exceeded the sample size. When PCA was performed (p = 64, 256), power for the MANOVA tests on the reduced dimensions varied inversely with the number of the dimensions retained. When the group mean differences spread across all population dimensions, better power could result from selecting fewer sample components for subsequent hypothesis testing. Like what had been inferred from Table 2 for condition 2 with mean differences concentrated along the most important component, the use of a Moore–Penrose generalized inverse in lieu of the nonexistent sample inverse resulted in very low statistical power.

4. Discussion

4.1. *Implications of the Results*

Three general conclusions apply. (1) Although PCA can succeed with HDLSS data in some favorable cases, otherwise PCA fails. (2) Data analysts should avoid using PCA for dimension reduction with HDLSS data unless a covariance structure dominated by a few components is defensibly expected. (3) The sensitivity of using traditional multivariate hypothesis testing methods on sample principal components varies with the population covariance structure, the population mean structure, the data dimensions, and the component selection process.

We studied PCA because so many scientists use it. However, the covariance structures of most interest to scientists may implicitly follow a factor analysis model. As noted in the introduction, we agree with Widaman (1993) and recommend factor analysis over PCA for dimension reduction. Unfortunately, simulations indicate that factor analysis handles HDLSS data no better than PCA (MacCallum et al., 1999; Preacher and MacCallum, 2002).

4.2. *Recommended Alternatives*

Four approaches are recommended for making inferences with HDLSS data. Each approach uses additional scientific and statistical thinking to ensure accurate inference.

1. Using a credible structured covariance pattern has great appeal, especially in the context of the general linear mixed model with more subjects than covariance parameters. In some settings, for example, time series covariance patterns (e.g., autoregressive, moving average, etc.) may apply. The Kenward–Roger approach provides the best inference in small samples with Gaussian data (Muller and Stewart, 2006, Ch. 18).
2. Taking advantage of logical structure in the data by using scientifically and statistically sufficient summary statistics helps avoid HDLSS problems. Rao et al. (2005) successfully used the strategy in analyzing kidney segmentation data.
3. Analyzing the response variables in scientifically meaningful groups can provide a valid approach. In the imaging example, scientists may be comfortable dividing the brain region of interest into subregions, based on a priori knowledge about structure and function. Avoiding HDLSS allows applying classical multivariate theory with data dimensions for which validity of estimation and inference can be assured.
4. The last recommended approach is to use recently developed methods specifically designed for HDLSS settings, and with known properties. We caution the reader that a vast number of suggestions have been made, with little data supporting most of the suggestions. In the end of Section 1.3, we have highlighted four articles which describe well-founded methods that seem most appealing.

Acknowledgments

Joint support for Chi and Muller came in part from a UF CTSI core grant via NCRR K30-RR022258, as well as NIDDK R01-DK072398, and NIDCR grant 1R01DE020832-01A1. Chi's support included NINDS R21-NS065098. Muller's support included NIDCR U54-DE019261, NCRR K30-RR022258, NHLBI R01-HL091005, and NIAAA R01-AA016549.

References

Ahn J, Marron JS, Muller KE, Chi YY. The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika. 2007; 94:760–766.

Anderson, TW. An Introduction to Multivariate Statistical Analysis. 3rd ed.. Wiley; New York: 2004.

Baik J, Ben AG, Peche S. Phase transition of the largest eigenvalue for non-null complex covariance matrices. Annals of Probability. 2005; 33:1643–1697.

Baik J, Silverstein JW. Eigenvalues of large sample covariance matrices of spiked population models. Journal of Multivariate Analysis. 2006; 97:1382–1408.

Cascio CJ, Gribbin MJ, Gouttard S, Smith RG, Jomier M, Poe MD, Graves M, Hazlett HC, Muller KE, Gerig G, Piven J. Decreased variability of fractional anisotropy in young children with autism. 2013 Manuscript submitted for publication.

Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society. 2005; 67:427–444.Series B

Johnson, NL.; Kotz, S. Distributions in Statistics: Continuous Multivariate Distributions. Wiley; New York: 1972.

Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Annals of Statistics. 2001; 29:295–327.

Johnstone IM, Lu AY. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association. 2009; 104:682–693. [PubMed: 20617121]

Jung S, Marron JS. PCA consistency in high dimension, low sample size context. The Annals of Statistics. 2009; 37:4104–4130.

Khatri CG. A note on multiple and canonical correlation for a singular covariance matrix. Psychometrika. 1976; 41:465–470.

MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychological Methods. 1999; 4:84–99.

Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. Academic Press, Inc; London: 1979.

Muller, KE.; Stewart, PW. Linear Model Theory for Univariate, Multivariate and Mixed Models. Wiley; New York: 2006.

Preacher KJ, MacCallum RC. Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes. Behavior Genetics. 2002; 32:153–161. [PubMed: 12036113]

Rao M, Stough J, Chi YY, Muller KE, Tracton GS, Pizer SM, Chaney EL. Comparison of human and automatic segmentations of kidneys from CT images. International Journal of Radiation Oncology, Biology and Physics. 2005; 61:954–960.

SAS Institute. SAS/IML® Software. SAS Institute; Cary, North Carolina: 1999.

Srivastava MS. Multivariate theory for analyzing high dimensional data. Journal of Japanese Statistical Society. 2007; 37:53–86.

Srivastava MS, Du M. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis. 2008; 99:386–402.

Srivastava MS, Fujikoshi Y. Multivariate analysis of variance with fewer observations than the dimension. Journal of Multivariate Analysis. 2006; 97:1927–1940.

Uhlig H. On singular Wishart and singular multivariate beta distributions. Annals of Statistics. 1994; 22:395–405.

Warton DI. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association. 2008; 103:340–349.

Widaman KF. Common factor analysis versus principal component analysis: differential bias in representing model parameters? Multivariate Behavioral Research. 1993; 28:263–311.

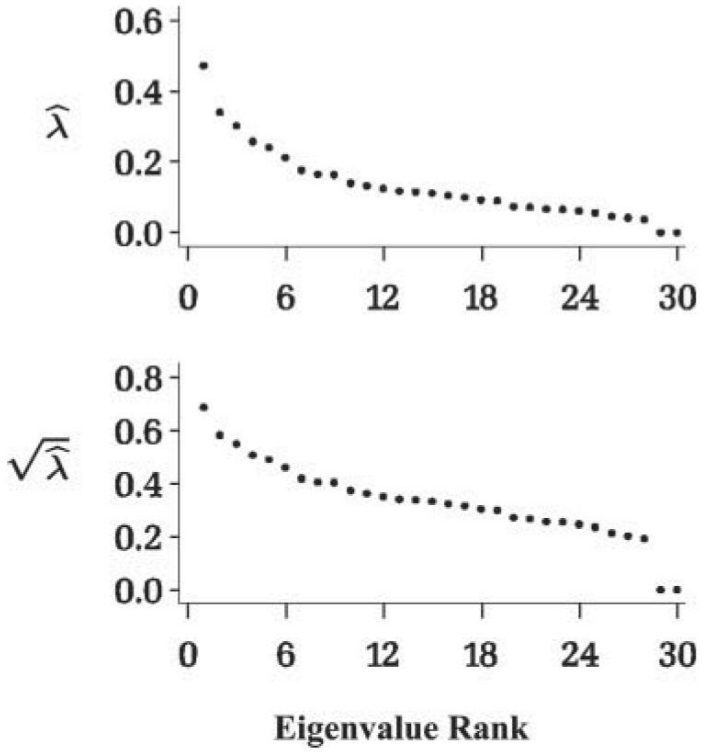

Figure 1.
Sample ordered eigenvalues (top) and their square roots (bottom) for the residual sample covariance matrix of DTI data for $\nu = 28$ and $p = 387$. The horizontal axes are eigenvalue orders.

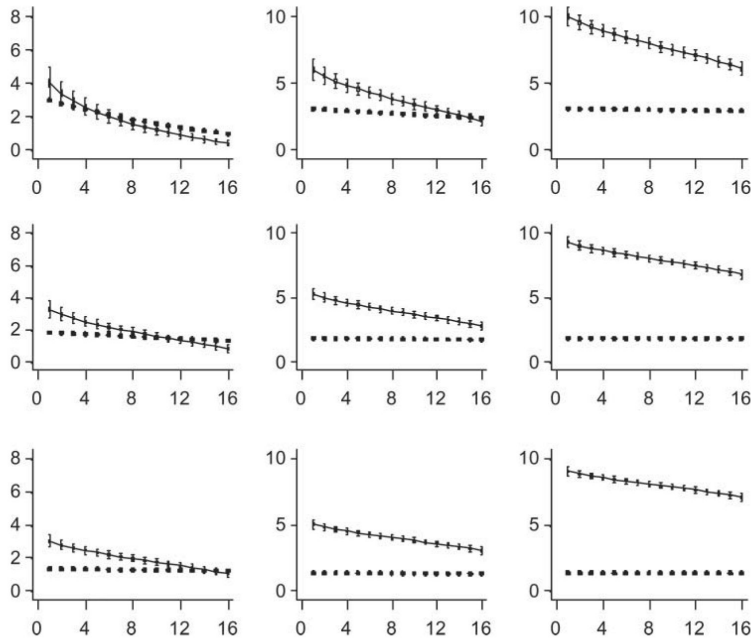

Figure 2.

Box plots for sample-ordered square roots of eigenvalues for N 16 and population with one smoothly decreasing eigenvalue pattern, g_1. Columns are (left to right) for p = 64, 256, and 1024. Rows are (top to bottom) ϵ = 0.2, 0.5, and 0.8. The vertical axes are square roots =of eigenvalues, and the horizontal axes are eigenvalue orders.

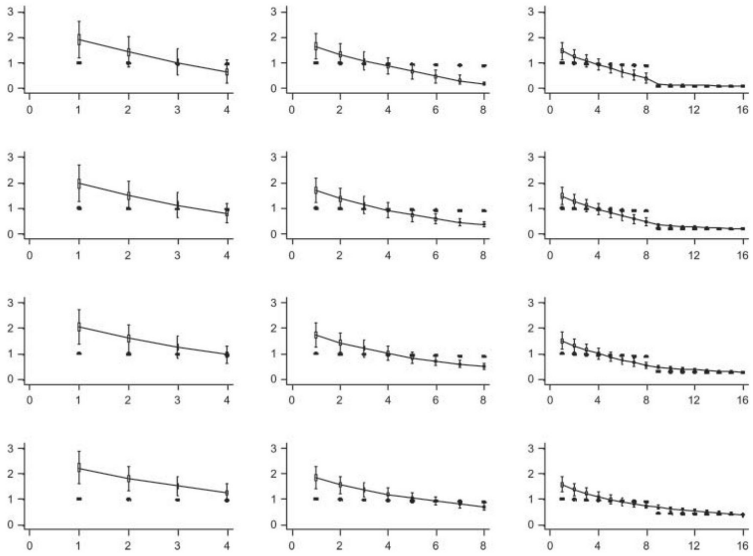

Figure 3.

Box plots for sample-ordered square roots of eigenvalues for p 256 and population with two smoothly decreasing eigenvalue patterns combined in g_2. Columns are (left to right) for N = 4, 8, and 16. Rows are (top to bottom) τ = 0.01, 0.05, 0.1, and 0.02. The vertical axes are square roots of eigenvalues, and the horizontal axes are eigenvalue orders.

Table 1

Empirical type I error rates for a two-sample comparison with 10,000 replications, $N = 18$ and $\alpha = 0.05$. The AR(1) covariance structure has variance of one and correlation of ρ_{AR}. The $\Sigma_u \otimes$AR(1; $p/2$) structure has Σ_u for a (2 × 2) unstructured covariance with variances of σ_1^2 and σ_2^2, AR(1)of dimension $p/2$ with variance of one, and $p_u = \rho_{AR} = 0.5$. The gray rows are conditions with number of variables fewer than error degrees of freedom. A test size is computed as the proportion of replications that have p-values obtained from performing MANOVA tests on the selected sample components less than or equal to 0.05 (a)

			AR1 Components retained					$\Sigma_u \otimes$AR(1; $p/2$) Components retained		
P	ρ_{AR}	ϵ	Mean #	Mean % Var	Test Size	σ_2^2 / σ_1^2	ϵ	Mean #	Mean % Var	Test size
4	0.5	0.69	4	100	0.0495	2	0.65	4	100	0.0495
4	0.8	0.40	4	100	0.0474	3	0.61	4	100	0.0474
16	0.5	0.62	16	100	0.0554	2	0.53	16	100	0.0554
16	0.8	0.25	16	100	0.0511	3	0.49	16	100	0.0511
64	0.5	0.61	3	39	0.0524	2	0.50	3	42	0.0505
64	0.5	0.61	12	92	0.0526	2	0.50	12	92	0.0502
64	0.5	0.61	16	100	0.0539	2	0.50	16	100	0.0511
64	0.8	0.23	3	55	0.0492	3	0.46	3	43	0.0511
64	0.8	0.23	10	92	0.0501	3	0.46	12	92	0.0509
64	0.8	0.23	16	100	0.0494	3	0.46	16	100	0.0517
256	0.5	0.60	3	30	0.0481	2	0.49	3	32	0.0467
256	0.5	0.60	15	95	0.0491	2	0.49	14	94	0.0499
256	0.5	0.60	16	100	0.0500	2	0.49	16	100	0.0452
256	0.8	0.22	3	37	0.0486	3	0.46	3	32	0.0512
256	0.8	0.22	13	93	0.0502	3	0.46	14	94	0.0528
256	0.8	0.22	16	100	0.0543	3	0.46	16	100	0.0491

Mean # and mean % variance are rounded to an integer.

Table 2

Empirical power for a two-sample comparison with mean differences concentrated along the first PCA direction, 10,000 replications, $N = 18$ and $\alpha = 0.05$. The AR(1) covariance structure has variance of one and correlation of ρ_{AR}. The $\Sigma_u \otimes$AR(1; $p/2$) structure has Σ_u for a (2 × 2) unstructured covariance with variances of σ_1^2 and σ_2^2, AR(1) of dimension $p/2$ with variance of one, and $\rho_u = \rho_{AR} = 0.5$. The gray rows are conditions with number of variables fewer than error degrees of freedom. A test size is computed as the proportion of replications that have p-values obtained from performing MANOVA tests on the selected sample components less than or equal to 0.05 (α)

			AR1 Components retained					$\Sigma_u \otimes$AR(1; $p/2$) Components retained		
P	ρ_{AR}	ϵ	Mean #	Mean % Var	Test Size	σ_2^2/σ_1^2	ϵ	Mean #	Mean % Var	Test Size
4	0.5	0.69	4	100	0.8215	2	0.65	4	100	0.8171
4	0.8	0.40	4	100	0.8276	3	0.61	4	100	0.8276
16	0.5	0.62	16	100	0.0756	2	0.53	16	100	0.0751
16	0.8	0.25	16	100	0.0737	3	0.49	16	100	0.0721
64	0.5	0.61	3	41	0.7236	2	0.50	3	45	0.8091
64	0.5	0.61	12	92	0.2257	2	0.50	12	92	0.2530
64	0.5	0.61	16	100	0.0712	2	0.50	16	100	0.0734
64	0.8	0.23	3	59	0.8663	3	0.46	3	46	0.8165
64	0.8	0.23	9	91	0.4270	3	0.46	12	92	0.2569
64	0.8	0.23	16	100	0.0749	3	0.46	16	100	0.0757
256	0.5	0.60	3	32	0.4283	2	0.49	3	32	0.5666
256	0.5	0.60	15	95	0.1047	2	0.49	14	94	0.1244
256	0.5	0.60	16	100	0.0630	2	0.49	16	100	0.0730
256	0.8	0.22	3	38	0.6963	3	0.46	3	33	0.5659
256	0.8	0.22	13	93	0.1809	3	0.46	14	94	0.1304
256	0.8	0.22	16	100	0.0783	3	0.46	16	100	0.0705

Mean # and mean % variance are rounded to an integer.

Table 3

Empirical power for a two-sample comparison with mean differences concentrated along the last PCA direction, 10,000 replications, $N = 18$ and $\alpha = 0.05$. The AR(1) covariance structure has variance of one and correlation of ρ_{AR}. The $\Sigma_u \in AR(1; p/2)$ structure has Σ_u for a (2 × 2) unstructured covariance with variances of σ_1^2 and σ_2^2, AR(1) of dimension $p/2$ with variance of one, and $\rho_u = \rho_{AR} = 0.5$. The gray rows are conditions with number of variables fewer than error degrees of freedom. A test size is computed as the proportion of replications that have p-values obtained from performing MANOVA tests on the selected sample components less than or equal to 0.05 (α)

			AR1 Components retained					$\Sigma_u \in AR(1; p/2)$ Components retained		
P	ρ_{AR}	ϵ	Mean #	Mean % Var	Power	σ_2^2/σ_1^2	ϵ	Mean #	Mean % Var	Power
4	0.5	0.69	4	100	0.8200	2	0.65	4	100	0.8218
4	0.8	0.40	4	100	0.8218	3	0.61	4	100	0.8144
16	0.5	0.62	16	100	0.0783	2	0.5316	16	100	0.0761
16	0.8	0.25	16	100	0.0748	3	0.49	16	100	0.0723
64	0.5	0.61	3	39	0.0863	2	0.50	3	42	0.0654
64	0.5	0.61	13	92	0.0998	2	0.50	12	92	0.0760
64	0.5	0.61	16	100	0.0589	2	0.50	16	100	0.0567
64	0.8	0.23	3	55	0.0542	3	0.46	3	43	0.0615
64	0.8	0.23	10	92	0.0663	3	0.46	12	92	0.0737
64	0.8	0.23	16	100	0.0565	3	0.46	16	100	0.0617
256	0.5	0.60	3	30	0.0652	2	0.49	3	32	0.0546
256	0.5	0.60	15	95	0.0624	2	0.49	14	94	0.0563
256	0.5	0.60	16	100	0.0556	2	0.49	16	100	0.0507
256	0.8	0.22	3	37	0.0502	3	0.46	3	32	0.0552
256	0.8	0.22	13	93	0.0542	3	0.46	14	94	0.0578
256	0.8	0.22	16	100	0.0545	3	0.46	16	100	0.0532

Mean # and mean % variance are rounded to an integer.

Commun Stat Simul Comput. Author manuscript; available in PMC 2014 May 20.

Table 4

Empirical power for a two-sample comparison with mean differences diffused along all PCA directions, 10,000 replications, N = 18and α = 0.05. The AR(1)covariance structure has variance of one and correlation of ρ. The $\Sigma_u \in$AR(1; p/2) structure has Σ_u for a (2 × 2) unstructured covariance with variances of σ_1^2 and σ_2^2, AR(1)of dimension p/2 with variance of one, and $\rho_u = \rho_{AR} = 0.5$. The gray rows are conditions with number of variables fewer than error degrees of freedom. A test size is computed as the proportion of replications that have p-values obtained from performing MANOVA tests on the selected sample components less than or equal to 0.05 (α)

			AR1 Components retained					$\Sigma_u \in$AR(1; p/2) Components retained		
P	ρ_{AR}	ϵ	Mean #	Mean % Var	Power	σ_2^2/σ_1^2	ϵ	Mean #	Mean % Var	Power
4	0.5	0.69	4	100	0.9952	2	0.65	4	100	0.9944
4	0.8	0.40	4	100	0.9965	3	0.61	4	100	0.9957
16	0.5	0.62	16	100	0.0952	2	0.53	16	100	0.0989
16	0.8	0.25	16	100	0.0963	3	0.49	16	100	0.0945
64	0.5	0.61	3	40	0.5903	2	0.50	3	42	0.5191
64	0.5	0.61	12	92	0.3079	2	0.50	12	92	0.3089
64	0.5	0.61	16	100	0.0823	2	0.50	16	100	0.0805
64	0.8	0.23	3	55	0.3281	3	0.46	3	43	0.5036
64	0.8	0.23	10	92	0.3132	3	0.46	12	92	0.3087
64	0.8	0.23	16	100	0.0773	3	0.46	16	100	0.0861
256	0.5	0.60	3	30	0.2856	2	0.49	3	32	0.2437
256	0.5	0.60	15	95	0.1074	2	0.49	14	94	0.1152
256	0.5	0.60	16	100	0.0697	2	0.49	16	100	0.0713
256	0.8	0.22	3	37	0.1494	3	0.46	3	32	0.2279
256	0.8	0.22	14	93	0.1050	3	0.46	14	94	0.1135
256	0.8	0.22	16	100	0.0662	3	0.46	16	100	0.0715

Mean # and mean % variance are rounded to an integer.