MODELLISTICA E CALCOLO SCIENTIFICO
MOX
MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 06/2011

Large p Small n Data: Inference for the Mean

Piercesare Secchi, A. Stamm, Simone Vantini

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Large p Small n Data: Inference for the Mean

P. Secchi [a], A. Stamm [b], S. Vantini [a]

[a] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica “F. Brioschi”
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
[b] University of Rennes I, IRISA, UMR CNRS-6074
Campus de Beaulieu, F-35042 Rennes, France
piercesare.secchi@polimi.it
aymeric.stamm@irisa.fr
simone.vantini@polimi.it

Keywords: Large p small n data, High-dimensional data, Inference for the mean, MANOVA.

AMS Subject Classification: 62H15, 62H10.

Abstract

We provide a generalization of Hotelling’s Theorem that enables inference (*i*) for the mean vector of a multivariate normal population and (*ii*) for the comparison of the mean vectors of two multivariate normal populations, when the number p of components is larger than the number n of sample units and the (common) covariance matrix is unknown. We find suitable test statistics and their p-asymptotic distributions that allow the inferential analysis of large p small n data.

1 Introduction

The advent and development of high precision data acquisition technologies in active fields of research (e.g., medicine, engineering, climatology, economics), that are able to capture real-time and/or spatially-referenced measures, have provided the scientific community with large amount of data that challenge the classical approach to data analysis.

Data sets are indeed increasingly becoming characterized by a number of random variables that is much larger than the number of sample units (large p small n data sets) in contrast to the “familiar” data sets where the number of sample units is often much larger than the number of random variables (small p large n data sets). This makes many classical inferential tools (e.g. Hotelling’s Theorem) almost useless in many fields at the forefront of scientific research and raises the demand for new inferential tools able to efficiently deal with this new kind of data.

The work of Srivastava Srivastava (2007) is pioneering in this direction. In it, a generalization of the Hotelling's Theorem is proposed: a generalized T^2 test statistic is found and its distribution law is computed for $p \geq n$ under the assumptions of normality and proportionality of the covariance matrix to the identity matrix (with the proportionality constant unknown); this assumption implies the independence among components (and among univariate test statistics as well), enabling classical inference procedures. We shall show that our results, which do not rely on the latter assumption, generalize this work in a much less stringent framework. In Srivastava Srivastava (2007), some inferential results non depending on strong assumptions on the covariance structure are presented as well, but, being asymptotic in both p and n, they are not suitable to perform inferential statistical analysis of large p small n data.

Other methods to deal with the analysis of large p small n data are objects of statistical investigation. Approaches based on the joint use of univariate test statistics for each component to build multivariate inference procedures have already appeared in the literature also under the assumption of dependence among components and thus among univariate test statistics as well. These approaches rely on the correction of each univariate significance (or confidence) levels such that the global significance (or confidence) level approximates a desired value. In particular, we can distinguish between a priori corrections, based on widely valid theoretical results (e.g., Bonferroni correction), and a posteriori corrections, based on the empirical distribution of the univariate p-values (e.g., Benjamini and Yekutieli Benjamini and Yekutieli (2001), Storey Storey (2003)).

Permutation tests provide a further alternative approach to the inference for large p small n data. According to this approach, a label based on an appropriate ranking among sample unit observations is associated to each observation; then, this label sequence is compared with all other possible non ranking-based label sequences and its extremity is actually tested. Permutation tests provide inferential procedures that are conditional (the focus is on the sample rather than on the population), and distribution-free (no strong assumption about the population distribution law is necessary). Pesarin and Salmaso Pesarin and Salmaso (2010, 2009) and Hall and Keilegorn Hall and Keilegorn (2007) recently proposed the use of permutation tests in the framework of multivariate and functional data analysis (an extreme case of large p small n data).

Functional Data Analysis (FDA) is indeed an active area of statistical research moving in this direction. In FDA, each sample unit is represented by means of a function (e.g., Ramsay and Silverman Ramsay and Silverman (2005), Ferraty and Vieu Ferraty and Vieu (2006)). The typical inferential approach of FDA is the projection of the n functions under investigation - virtually belonging to an ∞-dimensional functional space - onto a suitable finite p-dimensional functional subspace with p smaller than n where a classical inferential approach can still be used. Roughly speaking, the original FDA is replaced by a classical multivariate analysis that is expected to well approximate the former one. The choice for the finite p-dimensional functional subspace is often made a priori introducing arbitrariness in the outcome of the analysis; sometimes this choice is instead data driven (e.g., functional principal component analysis) but still used as if it was an a priori choice causing a systematic under estimate of the variability associated to the reduced representations of the data.

Finally, Bayesian statistics provides another natural framework where the analysis of large p small n data sets does not present - at least theoretically - any difficulty. Indeed, once chosen the conditional model for the data and the prior distribution for the unknown parameters, one observation (i.e., $n = 1$) is enough to obtain a likelihood function and thus a posterior distribution to make inference about the unknown parameters (e.g., the mean vector). On the other hand, from a computational point of view, a large number p of components can strongly affect the efficiency of the MCMC simulations often needed to come up with a usable posterior distribution. Moreover, from a practical point of view, a small sample size introduces a strong dependence of the posterior distribution on the prior making the Bayesian information updating quite ineffective.

Similarly to Srivastava Srivastava (2007) and differently from the other works presented above, our proposal is Hotelling-inspired. In particular, to overcome the impossibility of treating large p small n data by means of a classical model-based approach, our strategy focuses on the random "variability space explored by the data", i.e., the space generated by the first $n-1$ principal components. In this reduced space, the proposed analysis is almost classical with the important distinction that the randomness of this data-dependent reduced space is fully taken into account. A pairwise comparison between our new inferential procedure and the other approaches presented above is of sure interest. However, a comparison with the Bayesian perspective would be of uncertain interest since it would strongly depend on the subjective choice for the prior distribution. Thus, in the present work, just the theoretical and empirical comparison with the inferential approach proposed in Srivastava Srivastava (2007) and with the one proposed in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009) is carried out. These are two approaches to the same problem that are very close and far from ours, respectively.

The paper is outlined as follows: in Section 2, after the introduction of the probabilistic framework, a generalized version of the Hotelling's Theorem for $p \to \infty$ is proposed; part of its proof is reported in Appendix A. In Section 3, the previous generalized version of the Hotelling's Theorem is used for the inference for the mean vector of a p-variate normal population (and for the difference of the mean vectors of two p-variate normal populations) when the number p of components is far larger than the number n of sample units; a theoretical comparison with the classical Hotelling's Theorem is here undertaken. In Section 4, by means of MC simulations, our new inferential procedure is compared with the ones presented in Srivastava Srivastava (2007) and in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009). In the final Section, some hints for possible future investigations are reported; in particular, a conjecture supported either by the results of MC simulations and by its theoretical consistency with our generalization of the Hotelling's Theorem and with the results presented in Srivastava Srivastava (2007) is proposed.

2 Generalized Hotelling's Theorem

The classical approach to inference for the mean μ_p of a p-variate normal random vector with unknown full rank covariance matrix Σ_p relies on a famous corollary of the Hotelling's Theorem that holds when the number n of sample units is larger than the number p of random vector components.

Theorem 1 (Hotelling's Theorem)**.** *For $m \geq 1$ and $p \geq 1$, assume that:*

(i) $\mathbf{X} \sim N_p(\mu_p, \Sigma_p)$*;*

(ii) $W \sim Wishart_p(\Sigma_p, m)$*;*

(iii) $\mathbf{X}$ *and* W *are independent.*

Then, for $m \geq p$:

$$\frac{m-p+1}{p}(\mathbf{X}-\mu_p)'W^{-1}(\mathbf{X}-\mu_p) \sim F(p, m-p+1) \ .$$

Corollary 2 (Hotelling's T^2 Distribution Law)**.** *For $n \geq 2$ and $p \geq 1$, assume that:*

(i') $\{\mathbf{X}_i\}_{i=1,\ldots,n} \sim iid \ N_p(\mu_p, \Sigma_p)$.

Then, for $n > p$:

$$\frac{(n-p)n}{(n-1)p}(\overline{\mathbf{X}}-\mu_p)'S^{-1}(\overline{\mathbf{X}}-\mu_p) \sim F(p, n-p) \ ,$$

with $\overline{\mathbf{X}}$ and S being the sample mean and the sample covariance matrix, respectively.

The quantity $n(\overline{\mathbf{X}}-\mu_p)'S^{-1}(\overline{\mathbf{X}}-\mu_p)$ is known as Hotelling's T^2 due to its analogy with the squared of the univariate Student's t test statistic. Corollary 2 makes possible the development of inferential tools for the mean value of a p-variate normal random vector (e.g. confidence ellipsoidal regions or hypothesis testing) when the number n of sample units is larger than the number p of random vector components; there are no assumptions on the covariance matrix Σ_p that is only required to be positive definite. Proofs of Theorem 1 and Corollary 2 can be found, for instance, in Anderson Anderson (2003).

Theorem 1 and Corollary 2 become useless in applications where the covariance matrix is unknown and the number p of random vector components is larger than the number $n-1$, with n being the number of sample units. Indeed, in these cases, T^2 is not defined since S is not invertible because $rank(S) = \min(n-1, p)$ a.s. . Analogously to Srivastava Srivastava (2007), we decide to suitably generalize the inverse of S in order to obtain a suitable generalization of T^2. We considered the *Moore-Penrose Generalized Inverse* (Rao and Mitra Rao and Mitra (1971)) of a rectangular matrix, whose general definition can be found in Appendix A, since it always exists, it is unique, and it is equal to the inverse matrix when the latter is squared and invertible. Moreover, in the special case of a squared real positive semi-definite matrix A, the Moore-Penrose generalized inverse A^+ can be proved (see Appendix A) to be equal to:

$$A^+ = \sum_{i:\lambda_i \neq 0} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i' \ ,$$

with $\{\lambda_i\}_{i=1,\dots,p}$ and $\{\mathbf{e}_i\}_{i=1,\dots,p}$ being the eigenvalues and eigenvectors of A, respectively.

We now present a generalization of Hotelling's Theorem that can be used to make inference for the mean of a multivariate normal random vector when the sample size n is finite, the number of components p goes to infinity, and the covariance matrix is unknown.

Theorem 3 (Generalized Hotelling's Theorem)**.** *For $m \geq 1$ and $p \geq 1$, assume that:*

(i) $\mathbf{X} \sim N_p(\mu_p, \Sigma_p)$*;*

(ii) $W \sim Wishart_p(\Sigma_p, m)$ *;*

(iii) $\mathbf{X}$ *and* W *are independent;*

(iv) $0 < \overline{\sigma} = \lim_{p\to\infty} \frac{tr(\Sigma_p)}{p} < +\infty$ *and* $0 < \overline{\sigma^2} = \lim_{p\to\infty} \frac{tr(\Sigma_p^2)}{p} < +\infty$.

Then, for $p \to \infty$:

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} p(\mathbf{X} - \mu_p)' W^+ (\mathbf{X} - \mu_p) \xrightarrow{\mathscr{D}} \chi^2(m) \ .$$

The proof of Theorem 3 is based on the p-asymptotic distribution of three auxiliary random matrices Y, L, H that provide alternative useful representation of the random matrix W appearing in (*ii*). In particular, Y is a $p \times m$ random matrix whose m columns are independent $N_p(\mathbf{0}, \Sigma_p)$, i.e. we may represent $W = YY'$ since the two random matrices have the same law; L is a random diagonal matrix whose diagonal elements are the m non-zero ordered eigenvalues of W; and H is a $m \times p$ random matrix whose rows are the corresponding m eigenvectors (i.e. $W = H'LH$ with $HH' = I_m$). Random matrices Y, L, and H exist almost surely since W is a Wishart random matrix with m degrees of freedom. Also the diagonal matrix $\Lambda_p = diag(\lambda_1(\Sigma_p), \dots, \lambda_p(\Sigma_p))$ with $\lambda_1(\Sigma_p) \geq \dots \geq \lambda_p(\Sigma_p) > 0$ being the ordered eigenvalues of Σ_p always exists thanks to the positive definiteness of Σ_p.

To prove Theorem 3 we need the following

Lemma 4. *Under the assumptions (ii) and (iv) of Theorem 3:*

$$\begin{aligned} \tfrac{Y'Y}{p} &\xrightarrow[p\to\infty]{\mathscr{P}} \overline{\sigma} I_m \ , \\ \tfrac{L}{p} &\xrightarrow[p\to\infty]{\mathscr{P}} \overline{\sigma} I_m \ , \\ H\Lambda_p H' &\xrightarrow[p\to\infty]{\mathscr{D}} \tfrac{\overline{\sigma^2}}{\overline{\sigma}} I_m \ . \end{aligned}$$

The proof of Lemma 4 relies on Chebyshev's Inequality, Prokhorov's Theorem, Slutsky's Theorem, and on algebraic relations supported by the properties of Moore-Penrose general inverses. Because of its length it is reported in Appendix B.

Proof of Theorem 3. Let us define two auxiliary matrices $A = (H\Lambda_p H')^{-1/2}$ and $\mathbf{Z} = AH(\mathbf{X} - \mu_p)$. The conditional distribution of $\mathbf{Z}$ given H is

$$\mathbf{Z}|H \sim N_m(\mathbf{0}_m, AH\Lambda_p H'A) = N_m(\mathbf{0}_m, I_m) \ ,$$

since $\mathbf{X}$ is distributed as $N_p(\mu_p, \Lambda_p)$. The conditional distribution of $\mathbf{Z}$ given H does not depend on H: therefore, $\mathbf{Z}$ and H are independent and

$$\mathbf{Z} \sim N_m(\mathbf{0}_m, I_m) \ \text{ while } \ \mathbf{Z}'\mathbf{Z} \sim \chi^2(m) \ .$$

Thanks to Proposition 10 in Appendix A, the following equalities in distribution hold:

$$\begin{aligned} p(\mathbf{X} - \mu_p)'W^+(\mathbf{X} - \mu_p) &= p(\mathbf{X} - \mu_p)'H'L^{-1}H(\mathbf{X} - \mu_p) \\ &= p\mathbf{Z}'(ALA)^{-1}\mathbf{Z} \\ &= \left[\mathbf{Z}'\left(H\Lambda_p H'\right)^{1/2}\left(\frac{L}{p}\right)^{-1/2}\right]\left[\left(\frac{L}{p}\right)^{-1/2}\left(H\Lambda_p H'\right)^{1/2}\mathbf{Z}\right] \ . \end{aligned}$$

Because of Lemma 4 and the continuity of the maps $B \mapsto B^{-1/2}$ and $B \mapsto B^{1/2}$ over the set of positive definite matrices, we have that:

$$\begin{aligned} \left(\tfrac{L}{p}\right)^{-1/2} &\xrightarrow[p\to\infty]{\mathscr{P}} (\overline{\sigma})^{-1/2} I_m \ , \\ (H\Lambda_p H')^{1/2} &\xrightarrow[p\to\infty]{\mathscr{D}} \left(\tfrac{\overline{\sigma^2}}{\overline{\sigma}}\right)^{1/2} I_m \ . \end{aligned}$$

Thus, Slutsky's Theorem (e.g., Serfling Serfling (2002)) implies that:

$$\left(\frac{L}{p}\right)^{-1/2}\left(H\Lambda_p H'\right)^{1/2}\mathbf{Z} \xrightarrow[p\to\infty]{\mathscr{D}} \left(\frac{\overline{\sigma^2}}{\overline{\sigma}^2}\right)^{1/2}\mathbf{Z} \ .$$

Finally, since the Euclidean squared norm function on $\mathbb{R}^m$ is continuous,

$$p(\mathbf{X} - \mu_p)'W^+(\mathbf{X} - \mu_p) \xrightarrow[p\to\infty]{\mathscr{D}} \frac{\overline{\sigma^2}}{\overline{\sigma}^2}\mathbf{Z}'\mathbf{Z} \ ,$$

and thus

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} p(\mathbf{X} - \mu_p)'W^+(\mathbf{X} - \mu_p) \xrightarrow[p\to\infty]{\mathscr{D}} \chi^2(m) \ .$$

□

Remarks about Theorem 3

1. Note that in Theorem 3 the practical importance of "$p \to \infty$" (i.e. p-asymptoticity) is very general. For instance we might be considering the situation where we add extra components to a random normal vector $\mathbf{X}$ and infinite extra rows and columns to a random Wishart matrix W; this is for instance the case of discrete-time series when the time goes to infinity, or micro-array expressions when the number of genes goes to infinity. But "$p \to \infty$" can also be relevant in more complex situations where a sequence of random vectors $\mathbf{X}$ and of random matrices W of increasing dimensionality is investigated without any "nesting" property as p increases. This is for instance the case of subsequent finite-dimensional representations of functional data by means of subsequent non-necessarily nested basis (e.g. B-splines) whose dimension goes to infinity.

2. It is easy to prove that if the eigenvalues of Σ_p are uniformly bounded away from 0 and $+\infty$, i.e.:

   (iv') $\exists\, \underline{\lambda}, \overline{\lambda} : \forall p, 0 < \underline{\lambda} \leq \lambda_1(\Sigma_p) \leq \cdots \leq \lambda_p(\Sigma_p) \leq \overline{\lambda} < +\infty$,

   and at least one of the limits $\lim_{p\to\infty} \frac{tr(\Sigma_p)}{p}$ or $\lim_{p\to\infty} \frac{tr(\Sigma_p^2)}{p}$ exists, assumption (iv) of Theorem 3 is satisfied.

3 Inference for the Mean of Large p Small n Data

Provided that one can evaluate the ratio $\overline{\sigma}^2/\overline{\sigma^2}$ (this issue is tackled in subsection 3.3), Theorem 3 can be straightforwardly used to make inference for the mean of multivariate normal distribution when the number p of components is larger than the number n of sample units. Indeed, its natural consequence is the following.

Corollary 5 (Generalized Hotelling's T^2 p-asymptotic distribution law)**.** *For $n \geq 2$ and $p \geq 1$, assume that:*

(i') $\{\mathbf{X}_i\}_{i=1,\ldots,n} \sim$ *iid* $N_p(\mu_p, \Sigma_p)$*;*

(iv) $0 < \overline{\sigma} = \lim_{p\to\infty} \frac{tr(\Sigma_p)}{p} < +\infty$ *and* $0 < \overline{\sigma^2} = \lim_{p\to\infty} \frac{tr(\Sigma_p^2)}{p} < +\infty$.

Then, for $p \to \infty$:

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1} (\overline{\mathbf{X}} - \mu_p)' S^+ (\overline{\mathbf{X}} - \mu_p) \xrightarrow{\mathscr{D}} \chi^2(n-1) \,,$$

where $\overline{\mathbf{X}}$ and S are the sample mean and the sample covariance matrix, respectively.

Proof. It is a direct application of Theorem 3 since $\sqrt{n}(\bar{\mathbf{X}} - \mu_p) \sim N_p(\mathbf{0}_p, \Sigma_p)$, $(n-1)S \sim Wishart_p(\Sigma_p, n-1)$, and they are independent. □

The random quantity

$$T^2 = n(\overline{\mathbf{X}} - \mu_p)' S^+ (\overline{\mathbf{X}} - \mu_p) \tag{1}$$

can be naturally denoted as Generalized Hotelling's T^2 since it is defined for any n and p such that $n \geq 2$ and $p \geq 1$ and coincides with the classical Hotelling's $T^2 = n(\overline{\mathbf{X}} - \mu_p)' S^{-1} (\overline{\mathbf{X}} - \mu_p)$ when $p < n$. Despite the simplicity of this generalization, important differences occur between the new framework $p \geq n$ and the classical framework $p < n$. These differences involve:

- the connection between T^2 and the univariate Student's t test statistic (subsection 3.1);
- the invariance properties of T^2 (subsection 3.2);
- the distribution law of T^2 (subsection 3.3);
- the geometrical characteristics of the confidence regions and of the critical regions for the mean that can be derived from Corollaries 2 and 5 (subsection 3.4).

Remarks about Corollary 5

1. The p-asymptotic distribution law of T^2 strongly depends on assumptions (i') and (iv) involving the normal distribution of the observations and the p-asymptotic behavior of the covariance matrix, respectively. In particular, while the latter assumption could be probably relaxed (this issue is still under investigations by the authors), the former assumption cannot be weakened since there is no central limit theorem for $p \to \infty$ providing that $\sqrt{n}(\bar{\mathbf{X}} - \mu_p)$ is approximately normal and $(n-1)S$ is approximately a Wishart.
2. Even if our result covers a wide variety of real applications such as genomic data, micro arrays, functional data, and large p small n data in general, it does not cover the analysis of small p small n data with $p > n$.

3.1 Connections between the Generalized Hotelling's T^2 and the Student's t test Statistic

Student's t statistic comes natural in multivariate statistics if the $\mathbb{R}^p$-representations of the n sample units are projected along a certain direction $\mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$. Along this direction, the usual Student's t statistic can be computed:

$$t_{\mathbf{a}} = \sqrt{n} \frac{\mathbf{a}'(\bar{\mathbf{X}} - \mu_p)}{\sqrt{\mathbf{a}' S \mathbf{a}}} \sim t(n-1) \ ,$$

and univariate inference can be carried on along that direction.

Note that for all $\mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$, $t_{\mathbf{a}}$ is almost surely defined. Indeed, since $ker(S)$ has null Lebesgue measure on $\mathbb{R}^p$ and since X_i with $i = 1, \ldots, n$ are absolutely continuous random variables with respect to the same measure, the probability that $ker(S) \ni \mathbf{a}$ is equal to zero.

Maximization lemma of quadratic forms points out a strong relation between T^2 defined in (1) and the univariate $t_{\mathbf{a}}$. Indeed, one can show that

$$T^2 = \max_{\mathbf{a} \in Im(S) \setminus \{\mathbf{0}_p\}} t_{\mathbf{a}}^2 \; .$$

This means that making multivariate inference using T^2 at a certain confidence (significance) level is formally the same as making simultaneous univariate inference along any direction belonging to the "variability space explored by the data" (i.e. any direction $\mathbf{a} \in Im(S) \setminus \{\mathbf{0}_p\}$), while controlling the overall joint confidence (significance) level, ignoring all orthogonal directions (i.e. any direction $\mathbf{a} \in ker(S) \setminus \{\mathbf{0}_p\}$).

Note that $\mathbf{R}^p = Im(S) \oplus ker(S)$ for $n \geq 2$ and $p \geq 1$, with $Im(S) = \mathbf{R}^p$ and $ker(S) = \{\mathbf{0}_p\}$ almost surely if and only if $n > p$. Thus, when $n > p$, T^2 can be more simply defined as $\max_{\mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}} t_{\mathbf{a}}^2$; this is actually the most common way through which T^2 is introduced in the classical framework $n > p$. Unfortunately, in the general framework the latter definition does not hold since $t_{\mathbf{a}}^2$ is not uniformly bounded in $\mathbf{R}^p \setminus \{\mathbf{0}_p\}$ when $n \leq p$.

3.2 Invariance Properties of the Generalized Hotelling's T^2

Generalized Hotelling's T^2 is invariant under similarity transformations of the components (affine transformation preserving angles), i.e. affine transformation $A \bullet + \mathbf{b}$ such that $A \in \mathbb{R}^{p \times p}$ with $A = aO$, where $a \in \mathbb{R}_0^+$ and O is an orthogonal matrix, and $\mathbf{b} \in \mathbb{R}^p$. Under these assumptions we have that:

$$n((A\overline{\mathbf{X}} + \mathbf{b}) - (A\mu_p + \mathbf{b}))'(ASA')^+((A\overline{\mathbf{X}} + \mathbf{b}) - (A\mu_p + \mathbf{b})) = n(\overline{\mathbf{X}} - \mu_p)'S^+(\overline{\mathbf{X}} - \mu_p) \; .$$

The previous result relies on the fact, proven in Appendix A, that $(ASA')^+ = (A')^{-1}S^+A^{-1}$ only for $A = aO$, where $a \in \mathbb{R}_0^+$ and O is an orthogonal matrix.

Similarity transformations are also those transformations that do not affect assumption (iv) of Theorem 3 nor the value of the constant $\overline{\sigma}^2/\overline{\sigma^2}$.

It is easy to show that for $n > p$, T^2 is invariant under the wider class of affine transformations of the components, i.e. transformations $A \bullet + \mathbf{b}$ with $A \in \mathbb{R}^{p \times p}$ invertible and $\mathbf{b} \in \mathbb{R}^p$. This is due to the fact that $(ASA')^{-1} = (A')^{-1}S^{-1}A^{-1}$ for any invertible A.

3.3 On the p-asymptotic Law of the Generalized Hotelling's T^2

For $n > p$ the law of T^2 is independent from Σ_p (see Corollary 2); the p-asymptotic distribution law of T^2 depends instead on Σ_p through the constant ratio $\overline{\sigma}^2/\overline{\sigma^2}$. For Corollary 5 to have some impact for inferential purposes, the constant $\overline{\sigma}^2/\overline{\sigma^2}$ needs to be known or at least efficiently estimated. Two cases may occur in practical situations when Σ_p is not known:

(a) the constant $\overline{\sigma}^2/\overline{\sigma^2}$ is known even if Σ_p is not completely known. This case may occur when partial knowledge of Σ_p is available;

(b) the constant $\overline{\sigma}^2/\overline{\sigma^2}$ is not known and thus it needs to be estimated.

Case (a) covers a large variety of practical situations. For instance:

- For any $p > n$, the covariance matrix Σ_p is known up to an unknown multiplying factor: $\Sigma_p = \gamma V_p$ with V_p being a known $\mathbb{R}^{p\times p}$ positive definite matrix satisfying (iv) and γ an unknown positive constant. In this case, Σ_p implicitly satisfies (iv) and $\overline{\sigma}^2/\overline{\sigma^2} = \overline{\sigma_v}^2/\overline{\sigma_v^2}$ where $\overline{\sigma_v} = \lim_{p\to\infty} \frac{tr(V_p)}{p}$ and $\overline{\sigma_v^2} = \lim_{p\to\infty} \frac{tr(V_p^2)}{p}$. This situation includes the interesting case of independent homoscedastic components; in this case indeed, $V_p = \mathbf{I}_p$, $\overline{\sigma_v}^2 = 1$ and $\overline{\sigma_v^2} = 1$ and thus $\overline{\sigma}^2/\overline{\sigma^2} = 1$. This latter case has already been extensively considered in Srivastava Srivastava (2007); in particular in that work, the distribution of the random quantity $\frac{n(p-n+2)}{n-1}(\overline{\mathbf{X}} - \mu_p)'S^+(\overline{\mathbf{X}} - \mu_p)$ is computed for any $p > n$ under the assumption $\Sigma_p = \gamma \mathbf{I}_p$. The results presented in Srivastava Srivastava (2007) are consistent with Corollary 5. Indeed, when $\Sigma_p = \gamma \mathbf{I}_p$, the quantity $\frac{n(p-n+2)}{n-1}(\overline{\mathbf{X}} - \mu_p)'S^+(\overline{\mathbf{X}} - \mu_p)$ and the quantity $\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1}(\overline{\mathbf{X}} - \mu_p)'S^+(\overline{\mathbf{X}} - \mu_p)$ are p-asymptotically equivalent since their values converge to the same limit; moreover, as expected by Corollary 5, their limit distribution is a $\chi^2(n-1)$.

- For any $p > n$, the covariance matrix Σ_p is equal to $\Sigma_p = \widetilde{\Sigma}_p + \gamma V_p$, with $\widetilde{\Sigma}_p$ a positive definite (or even semi-definite) matrix such that $\lim_{p\to\infty} tr(\widetilde{\Sigma}_p) < +\infty$, γ an unknown positive constant, and V_p a known positive definite matrix satisfying (iv'). Indeed, also in this case it can be proven that $\overline{\sigma}^2/\overline{\sigma^2} = \overline{\sigma_v}^2/\overline{\sigma_v^2}$; the proof comes straightforward once it is noticed that, without loss of generality, V_p can be assumed to be diagonal. A covariance matrix of this form occurs for instance in all application where the observed p-variate random vectors are assumed to be generated by the sum of two independent terms: a structural term whose variability is concentrated on a finite number of components (or even infinite but with finite total variance) and a zero-mean nuisance term (due to background noise or measurement errors) satisfying (iv) acting on all components. If the covariance matrix of the nuisance term is assumed to be proportional to the identity matrix (as it often happens), also in this case we have $\overline{\sigma}^2/\overline{\sigma^2} = 1$. This assumption may hold for instance in genetics, where long array of genes are observed on a small number of patients, the variability of the array can indeed be assumed to be generated by two independent terms, an informative variability concentrated on a reduced number of positive/negative correlated genes and a nuisance homoscedastic error variability acting independently on each gene. Spectral data presents another situation where the latter assumption may hold; indeed, spectral data are characterized by the presence of nuisance background variability along the entire set of observed frequencies plus a series of independent sources of variability at some specific frequencies (bands) associated to the spectral firms of different molecules.

Case (b) is the case where the information about the covariance structure is sufficient to know that Σ_p satisfies (iv), but not sufficient to know the value of the constant

$\overline{\sigma}^2/\overline{\sigma^2}$. For instance, referring to the previous examples, we might know that V_p has a block structure with blocks $\ell \times \ell$ all equal to an unknown positive definite matrix B. In this case we know that $\overline{\sigma}$ and $\overline{\sigma^2}$ are for sure positive and finite without knowing their actual values $\overline{\sigma} = tr(B)/\ell$ and $\overline{\sigma^2} = tr(B^2)/\ell$.

In this second case having a good estimate of the constant $\overline{\sigma}^2/\overline{\sigma^2}$ becomes of primary importance. Once it is noticed that $\frac{tr(\Sigma_p)}{p} = \frac{\sum_{i=1}^p \lambda_i}{p}$ and $\frac{tr(\Sigma_p^2)}{p} = \frac{\sum_{i=1}^p \lambda_i^2}{p}$, Jensen's inequality provides an upper bound to the constant. Indeed, for any Σ_p satisfying (iv) we have that:

$$0 < \overline{\sigma}^2/\overline{\sigma^2} \leq 1\ ,$$

with equality holding when the p eigenvalues λ_i are all identical (this is the case $\Sigma_p = \gamma \mathbf{I_p}$). If the eigenvalues of Σ_p are uniformly bounded away from 0 and $+\infty$ by the constant $\underline{\lambda}$ and $\overline{\lambda}$, respectively (i.e. assumption (iv') in page 7), the constant $\overline{\sigma}^2/\overline{\sigma^2}$ can be further bounded:

$$\underline{\lambda}^2/\overline{\lambda}^2 \leq \overline{\sigma}^2/\overline{\sigma^2} \leq 1\ .$$

A more strict lower bound (useful for values of the ratio $\overline{\lambda}/\underline{\lambda}$ relatively close to 1) can be found by means of simple algebraic computations:

$$1 - \left(\overline{\lambda}/\underline{\lambda} - 1\right)^2 \leq \overline{\sigma}^2/\overline{\sigma^2} \leq 1. \tag{2}$$

This quadratic control may result very effective, for instance, if it is known that the maximum eigenvalue cannot exceed the minimum eigenvalue for more than the 10% (i.e. $\overline{\lambda}/\underline{\lambda} - 1 \leq 0.10$), the unknown constant is guaranteed not to be lower than 0.99.

Replacing $\overline{\sigma}^2/\overline{\sigma^2}$ with 1 takes to non-conservative inferential procedures, while replacing $\overline{\sigma}^2/\overline{\sigma^2}$ with a lower bound takes to conservative inferential procedures. Note that the lower/upper bound in (2) for $\overline{\sigma}^2/\overline{\sigma^2}$, provides also an outer/inner bound of the confidence regions that can be obtained by Corollary 5 (see Sections 3.4 and 3.5), and an upper/lower bound to the p-value associated to the hypothesis tests that can be obtained by Corollary 5 (see Sections 3.4 and 3.5).

A better estimate of $\overline{\sigma}^2/\overline{\sigma^2}$ can be obtained by using some estimates for $\frac{tr(\Sigma_p)}{p}$ and $\frac{tr(\Sigma_p^2)}{p}$. Indeed for $p \to \infty$ these quantities converge by definition to $\overline{\sigma}$ and $\overline{\sigma^2}$, and thus any unbiased estimator for $\frac{tr(\Sigma_p)}{p}$ (or $\frac{tr(\Sigma_p^2)}{p}$) for a given p is also a p-asymptotic unbiased estimator for $\overline{\sigma}$ (or $\overline{\sigma^2}$). The following estimators, defined for all $n \geq 3$ and $p \geq 1$, can be proven to satisfy this property:

$$\begin{array}{rcl} \hat{\overline{\sigma}}_p & := & \frac{trS}{p},\ and \\ \hat{\overline{\sigma^2_p}} & := & \frac{(n-1)^2}{(n-2)(n+1)}\left[\frac{trS^2}{p} - \frac{1}{n-1}\frac{(trS)^2}{p}\right]\ . \end{array} \tag{3}$$

Their properties are presented in Appendix C.

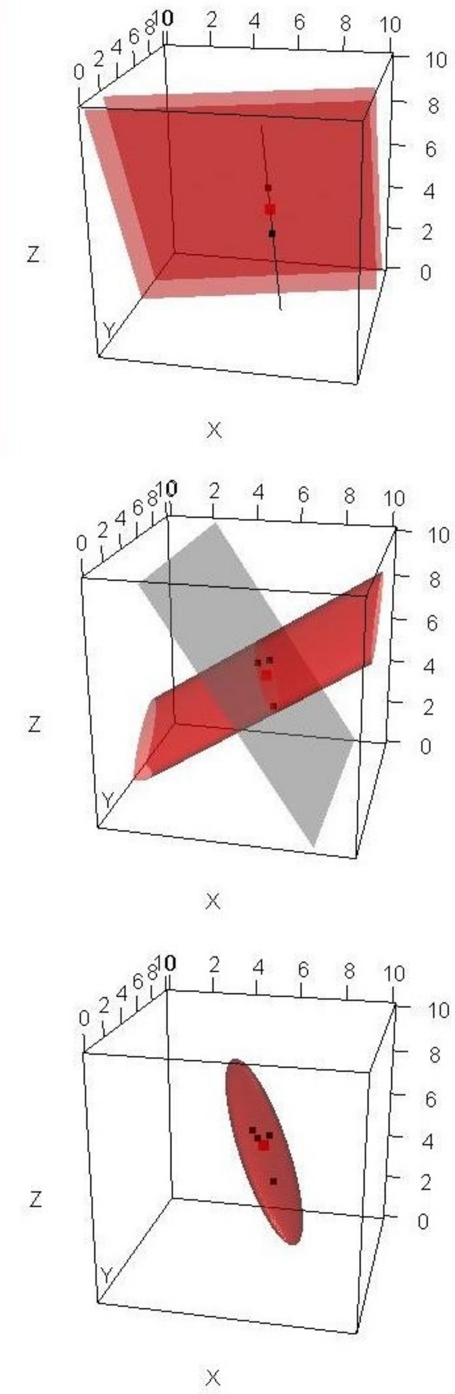

Figure 1: Examples of confidence regions for the mean vector when $p = 3$ and $n = 2, 3, 4$, respectively. Data points in black, sample mean in red, confidence regions in red, $Im(S)$ in gray.

3.4 p-asymptotic Confidence Region and Hypothesis Test for the Mean of a Normal Population when $p \gg n$

Corollary 5 turns out to be a useful tool for the construction of confidence regions and hypothesis tests for the mean in all practical situations where the number p of random vector components is far larger than the number n of sample units (e.g. genetics, spectral data) or even virtually infinite (e.g. functional data).

A p-asymptotic **Confidence Region** for the mean μ_p can be defined as follows:

$$CR_{1-\alpha}(\mu_p) := \left\{ \mathbf{m}_p \in \mathbf{R}^p : \frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1} (\mathbf{m}_p - \overline{\mathbf{X}})' S^+ (\mathbf{m}_p - \overline{\mathbf{X}}) \leq \chi^2_\alpha(n-1) \right\}, \qquad (4)$$

with $\chi^2_\alpha(n-1)$ being the upper α-quantile of a $\chi^2(n-1)$ random variable and $1-\alpha$ being the p-asymptotic confidence level.

Equivalently, a p-asymptotic **Hypothesis Test** for $H_0 : \mu_p = \mu_{0p}$ versus $H_1 : \mu_p \neq \mu_{0p}$ with p-asymptotic significance level α has the following rejection region:

Reject H_0 in favor of H_1 if:

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1} (\overline{\mathbf{X}} - \mu_{0p})' S^+ (\overline{\mathbf{X}} - \mu_{0p}) > \chi^2_\alpha(n-1) \; . \qquad (5)$$

Confidence region $CR_{1-\alpha}(\mu_p)$ is not of practical use for graphical purposes since a clear visual representation of it is not straightforward due to the large value of p. Similarly to the traditional multivariate framework, univariate projections of the confidence region along some directions (i.e. T^2-simultaneous confidence intervals) can give a rough idea about the location and shape of the confidence region, providing - in the case of rejection of H_0 - also some help in detecting the directions that have taken to the rejection of H_0. Since, by definition, we have that for $p \to \infty$

$$\mathbb{P}\left(\mathbf{a}'\mu_p \in \left[\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\frac{\mathbf{a}'S\mathbf{a}}{n}} \sqrt{\frac{\overline{\sigma^2}}{\overline{\sigma}^2} \frac{n-1}{p} \chi^2_\alpha(n-1)} \right], \forall \mathbf{a} \in Im(S) \right) \to 1 - \alpha \; ,$$

given a direction $\mathbf{a} \in Im(S) \setminus \{0\}$, the corresponding **$T^2$-simultaneous confidence interval** with p-asymptotic family-wise confidence $1-\alpha$ can be defined as follows:

$$\mathbf{a}'\mu_p \in \left[\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\frac{\mathbf{a}'S\mathbf{a}}{n}} \sqrt{\frac{\overline{\sigma^2}}{\overline{\sigma}^2} \frac{n-1}{p} \chi^2_\alpha(n-1)} \right] \; .$$

If $\mathbf{a} \notin ImS \setminus \{0\}$, then the corresponding T^2-simultaneous confidence interval is not bounded, i.e, equal to $[-\infty, +\infty]$.

Confidence region (4) and rejection region of test (5) present some peculiar features that are worth a little discussion.

Because S^+ is positive semi-definite, the confidence region $CR_{1-\alpha}(\mu_p)$ - which for $n > p$ is an ellipsoid subset of $\mathbb{R}^p$ - turns out to be a cylinder in $\mathbb{R}^p$ generated by the orthogonal extension in $ker(S)$ of an $n-1$-dimensional ellipsoid contained in $Im(S)$. As illustrative examples, three confidence regions for the mean vector when $p = 3$ and $n = 2,3,4$, respectively, are reported in Figure 1. In particular, as shown by the analytic expressions of the generalized T^2-simultaneous confidence intervals, $CR_{1-\alpha}(\mu_p)$ is bounded in all directions belonging to the random space $Im(S)$. These directions are easily identifiable since the first $n-1$ sample principal components provide an orthonormal basis for $Im(S)$.

Due to the non-null dimension of the random space $ker(S)$ and to the orthogonality between $ker(S)$ and $Im(S)$, we have that the statistic $\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1}(\overline{\mathbf{X}} - \mu_{0p})' S^+ (\overline{\mathbf{X}} - \mu_{0p})$ in the hypothesis test (5) does not change if μ_{0p} is replaced by $\mu_{0p} + \mathbf{m}_{ker(S)}$ with $\mathbf{m}_{ker(S)}$ being any vector belonging to $ker(S)$. This implies that H_0 might not be rejected even for values of the sample mean $\overline{\mathbf{X}}$ that are "really very far" from μ_{0p} in some direction within $ker(S)$. This is not surprising, because the use of S^+ implies an exclusive focus on the space $Im(S)$ (the variability space explored by the data), neglecting all $p-n+1$ directions associated to $ker(S)$ (the space orthogonal to the variability space explored by the data).

3.5 p-asymptotic Pooled Confidence Region and Hypothesis Test for Comparing the Means of Two Normal Populations when $p \gg n$

Theorem 3 can also be used to tackle the problem of comparing the means of two normal populations when the number p of components is larger than the number n of sample units. Indeed, under the same assumptions of the classical multivariate analysis of variance, we have that:

Corollary 6 (Generalized Pooled Hotelling's T^2_{pooled} p-asymptotic distribution law)**.** *For $n_a \geq 1$, $n_b \geq 1$, and $p \geq 1$, assume that:*

(i") $\{\mathbf{X}_{ai}\}_{i=1,\dots,n_a} \sim iid\, N_p(\mu_{pa}, \Sigma_p)$, $\{\mathbf{X}_{bi}\}_{i=1,\dots,n_b} \sim iid\, N_p(\mu_{pb}, \Sigma_p)$ and the two finite sequences are independent;

(iv) $0 < \overline{\sigma} = \lim_{p\to\infty} \frac{tr(\Sigma_p)}{p} < +\infty$ and $0 < \overline{\sigma^2} = \lim_{p\to\infty} \frac{tr(\Sigma_p^2)}{p} < +\infty$.

Then, for $n_a + n_b \geq 3$ and $p \to \infty$:

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{p}{n_a+n_b-2} \left(\frac{1}{n_a} + \frac{1}{n_b}\right)^{-1} \cdot$$

$$\cdot \left((\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b}) - (\mu_{pa} - \mu_{pb})\right)' S^+_{pooled} \left((\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b}) - (\mu_{pa} - \mu_{pb})\right) \quad \xrightarrow{\mathscr{D}} \quad \chi^2(n_a+n_b-2)\,,$$

where $\overline{\mathbf{X}}_\mathbf{a}$ and $\overline{\mathbf{X}}_\mathbf{b}$ are the two sample means, and S_{pooled} is the pooled sample covariance matrix.

Proof. It is another direct application of Theorem 3, since $\left(\frac{1}{n_a}+\frac{1}{n_b}\right)^{-1/2}\left((\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b})-(\mu_{pa}-\mu_{pb})\right)\sim N_p(\mathbf{0}_p,\Sigma_p)$, $(n_a+n_b-2)S_{pooled}=(n_a-1)S_a+(n_b-1)S_b\sim Wishart_p(\Sigma_p,n_a+n_b-2)$, and they are independent. □

It is natural to denote the following quantity as Generalized Pooled Hotelling's:

$$T^2_{pooled}=\left(\frac{1}{n_a}+\frac{1}{n_b}\right)^{-1}\left((\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b})-(\mu_{pa}-\mu_{pb})\right)'S^+_{pooled}\left((\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b})-(\mu_{pa}-\mu_{pb})\right). \tag{6}$$

Indeed, it is defined for any n_a, n_b, and p such that $n_a+n_b\geq 3$, $n_a\geq 1$, $n_b\geq 1$ and $p\geq 1$ and coincides with the classical definition of Pooled Hotelling's T^2_{pooled} when $p\leq n_a+n_b-2$. The similarities and the differences between the framework $p>n_a+n_b-2$ and the classical framework $p\leq n_a+n_b-2$ are analogous to the ones presented in Section 3 for the Generalized Hotelling's T^2.

In particular, also in the two-population framework we obtain a confidence region for estimating the difference of the two means and rejection region for testing the difference of the two means.

A p-asymptotic **Confidence Region** for difference of the means $\mu_{pa}-\mu_{pb}$ can be defined as follows:

$$\begin{aligned}CR_{1-\alpha}(\mu_{pa}-\mu_{pb})=\Bigg\{\Delta\mathbf{m_p}\in\mathbf{R}^p:\frac{\overline{\sigma}^2}{\overline{\sigma}^2}\frac{p}{n_a+n_b-2}\left(\frac{1}{n_a}+\frac{1}{n_b}\right)^{-1}\cdot\\ \cdot\left(\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b}-\Delta\mathbf{m_p}\right)'S^+_{pooled}\left(\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b}-\Delta\mathbf{m_p}\right)\leq\chi^2_\alpha(n_a+n_b-2)\Bigg\},\end{aligned} \tag{7}$$

with $1-\alpha$ being the p-asymptotic confidence level.

Equivalently, a p-asymptotic **Hypothesis Test** for $H_0:\mu_{pa}-\mu_{pb}=\Delta\mu_{0p}$ versus $H_1:\mu_{pa}-\mu_{pb}\neq\Delta\mu_{0p}$ with p-asymptotic significance level α has the following rejection region:

Reject H_0 in favor of H_1 if:

$$\begin{aligned}&\frac{\overline{\sigma}^2}{\overline{\sigma}^2}\frac{p}{n_a+n_b-2}\left(\frac{1}{n_a}+\frac{1}{n_b}\right)^{-1}\cdot\\ &\cdot\left(\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b}-\Delta\mu_{0p}\right)'S^+_{pooled}\left(\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b}-\Delta\mu_{0p})\right)>\chi^2_\alpha(n_a+n_b-2).\end{aligned} \tag{8}$$

Also the analytical expression of the T^2_{pooled}-simultaneous confidence intervals for the difference of the means comes naturally.

Given a direction $\mathbf{a}\in Im(S_{pooled})\setminus\{0\}$, the corresponding T^2_{pooled}**-simultaneous confidence interval** with p-asymptotic family-wise confidence $1-\alpha$ can be defined as follows:

$$\mathbf{a}'(\mu_{pa}-\mu_{pb})\in\left[\mathbf{a}'(\overline{\mathbf{X}}_\mathbf{a}-\overline{\mathbf{X}}_\mathbf{b})\pm\sqrt{\left(\frac{1}{n_a}+\frac{1}{n_b}\right)\mathbf{a}'S\mathbf{a}}\sqrt{\frac{\overline{\sigma}^2}{\overline{\sigma}^2}\frac{n_a+n_b-2}{p}\chi^2_\alpha(n_a+n_b-2)}\right].$$

If $\mathbf{a} \notin Im(S_{pooled}) \setminus \{0\}$, then the corresponding T^2_{pooled}-simultaneous confidence interval is not bounded, i.e, equal to $[-\infty, +\infty]$.

Recently, Srivastava and Yanagihara Srivastava and Yanagihara (2010) proposed a test for testing the equality of two covariance matrices of two normal distribution in the large p small n data framework, enabling to check for the homoscedasticity assumption which the previous results rely on.

4 Simulation Results

In this section, we estimate, by means of MC simulations, the power and the actual level of significance of the new test, presented in (8); in this section we will refer to it as the Generalized Hotelling's test. In particular, we estimate the probability of rejecting the null hypothesis $H_0 : \mu_a = \mu_b$ in favor of the alternative hypothesis $H_1 : \mu_a \neq \mu_b$ in eight different cases and for increasing values of the number p of components ranging between 2^0 and 2^{10} (i.e., 1 and 1024):

$$\begin{array}{rlllll}
\text{I0}: & \mu_a = \mathbf{0}, & \mu_b = \mathbf{0}, & \Sigma = I, & n_a = 10, & n_b = 10; \\
\text{D0}: & \mu_a = \mathbf{0}, & \mu_b = \mathbf{0}, & \Sigma = D, & n_a = 10, & n_b = 10; \\
\text{R0}: & \mu_a = \mathbf{0}, & \mu_b = \mathbf{0}, & \Sigma = R, & n_a = 10, & n_b = 10; \\
\text{S0}: & \mu_a = \mathbf{0}, & \mu_b = \mathbf{0}, & \Sigma = S, & n_a = 10, & n_b = 10; \\
\text{I1}: & \mu_a = \mathbf{0}, & \mu_b = 0.4 \cdot \mathbf{1}, & \Sigma = I, & n_a = 10, & n_b = 10; \\
\text{D1}: & \mu_a = \mathbf{0}, & \mu_b = 0.4 \cdot \mathbf{1}, & \Sigma = D, & n_a = 10, & n_b = 10; \\
\text{R1}: & \mu_a = \mathbf{0}, & \mu_b = 0.4 \cdot \mathbf{1}, & \Sigma = R, & n_a = 10, & n_b = 10; \\
\text{S1}: & \mu_a = \mathbf{0}, & \mu_b = 0.4 \cdot \mathbf{1}, & \Sigma = S, & n_a = 10, & n_b = 10;
\end{array}$$

where I is the identity matrix; D is a diagonal matrix whose diagonal alternatively assumes the values 0.5 and 1.5; R is a block matrix whose blocks are equal to the matrix $\left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$; S is a block matrix whose blocks are equal to the matrix $\left(\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right)$. Covariance matrices R and S can be simply obtained by D by means of an orthogonal transformation: 45° anticlockwise pairwise rotations and 45° clockwise pairwise rotations, respectively. The values for μ_a, μ_b, n_a, and n_b are the same used in the simulation study presented in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009); the value for Σ used in cases I0 and I1 are once again the same used in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009), while the value used in cases D0, D1, R0, R1, S0, and S1 are meant to provide less trivial situation where assumption (iv) still holds. Note that in the former two cases the constant $\overline{\sigma}^2 / \overline{\sigma^2} = 1$ while in the latter six $\overline{\sigma}^2 / \overline{\sigma^2} = 4/5$.

In details, for each case and for each value of the number p of components, 1000 synthetic data sets have been randomly generated according to the corresponding model and, for each one of these, the Generalized Hotelling's test has been performed at a nominal level of significance $\alpha = 0.05$. The relative number of times the null hypothesis has been rejected provides the estimate of either the actual level of significance of the Generalized Hotelling's test (cases I0, D0, R0, and S0) or its power (cases I1, D1, R1, and S1). The same data sets have been also used to perform two other tests recently appeared in the literature (Srivastava Srivastava (2007), Pesarin and Salmaso Pesarin and Salmaso (2010, 2009)) both suitable for dealing with large p small n data sets (in this section we will refer to them as the Srivastava's test and Pesarin-Salmaso's test, respectively).

Analogously to the Generalized Hotelling's test, also the Srivastava's test is based on the generalized T^2_{pooled}, but while the Generalized Hotelling's test uses a rejection region built from its p-asymptotic distribution under the assumption (iv), the Srivastava's test uses a rejection region built from its exact distribution under the assumption of independent and homoscedastic components.

The Pesarin-Salmaso's test is not a model-based test but a permutation test; the implementation used here is the same used in the simulation study presented in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009), i.e., the statistic used is actually a random weighted sum of the p univariate Student's t^2_{pooled} that can be written as $\left(\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b} - \Delta\mu_0\right)' S^{-1}_{diag} \left(\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b} - \Delta\mu_0\right)$, and its conditional distribution over the values observed within each data set is estimated by sampling 1000 random permutations of the $n_a + n_b = 20$ p-dimensional vectors making each data set. S_{diag} is the diagonal matrix whose diagonal elements are the p (non-pooled) sample variances of the p components.

The results of the simulation study are summarized in Figure 2. For completeness, in the cases in which $p \leq n_a + n_b - 2$, a traditional Hotelling's test has been implemented in place of both the Generalized Hotelling's test and Srivastava's test.

In subsections 4.1 and 4.2, the Generalized Hotelling's test is compared with the Srivastava's test and the Pesarin-Salmaso's test, respectively.

4.1 Comparison between the Generalized Hotelling's test and the Srivastava's test

In case I0, where H_0 is true and the hypotheses supporting the Srivastava's test hold, the observed rate of rejection of the Srivastava's test clearly matches its nominal level of significance 5%; on the contrary, in cases D0, R0, and S0, where H_0 is still true but the hypotheses supporting the Srivastava's test do not hold, the observed rate of rejection of the Srivastava's test significantly exceeds its nominal level of significance providing a strongly non conservative test.

The assumptions which the Generalized Hotelling's test is based on, hold instead for all cases, indeed for p "large enough" (in these cases 1024 seems to be a large enough value for p) the observed rate of rejection matches the nominal level of significance 5%. The almost identical patterns shown for cases D0, R0, and S0 confirm the invariance of the Generalized Hotelling's test under orthogonal transformations of the components. Further simulations, not reported here, for different values of n_a and n_b show a quicker (slower) convergence to the nominal level of significance for smaller (larger) values of n_a and n_b. For instance, for $n_a = n_b = 2$ (i.e. the smallest sample size we tested), the nominal value is already reached for $p = 64$. Mind the fact that, though small sample sizes increase the reliability of the Generalized Hotelling's test, they of course also reduce the power of the same test, as expected.

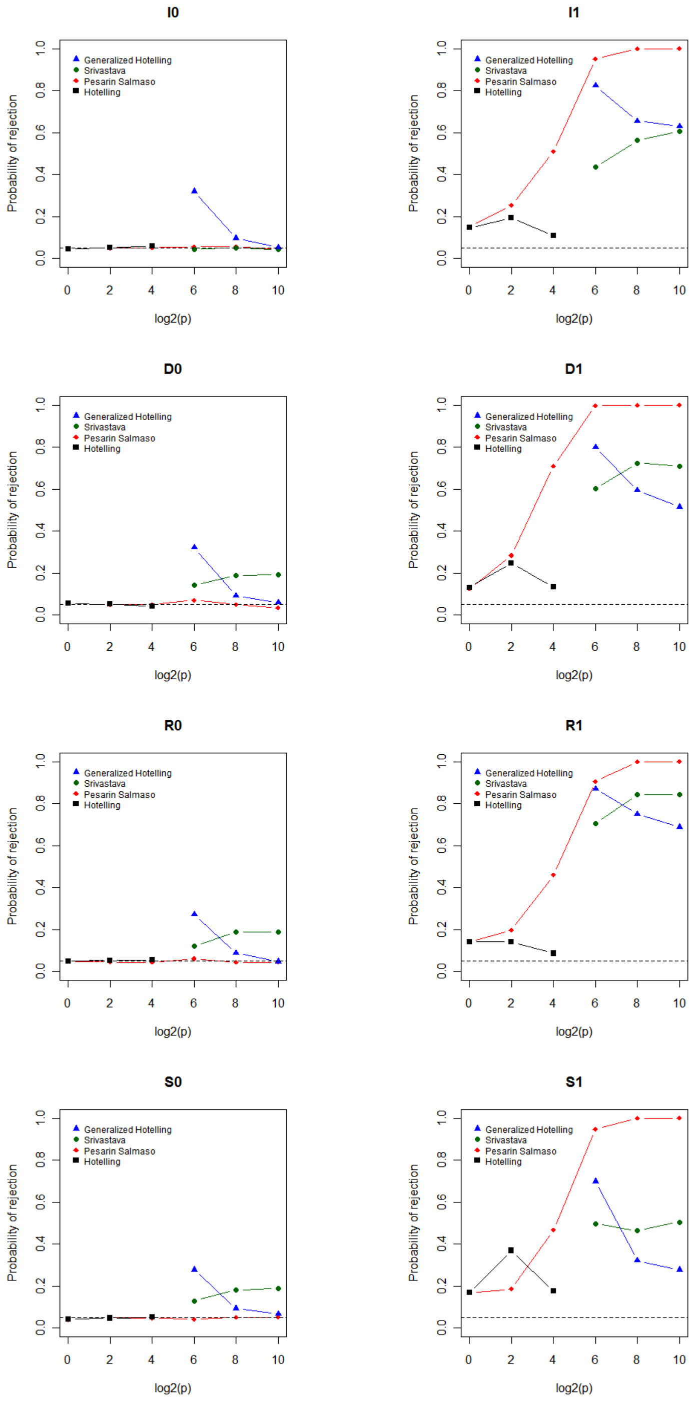

Figure 2: MC-estimates of the probability of rejecting $H_0 : \mu_a = \mu_b$ for different values of the number p of components. Each plot is associated to a different model (title) and each line to a different test (legend).

Fortunately, the same simulations also show that the convergence rate is independent from the value of the constant $\overline{\sigma}^2/\overline{\sigma^2}$. This fact enables an a-priori empirical measure, for a given sample size, of the minimal number p of random vector components (or given p, of the maximal sample size) that is necessary to make the Generalized Hotelling's test reliable.

Talking about the power under the alternative hypothesis $\Delta\mu = 0.4 \cdot \mathbf{1}$, in case I1 the superiority of the Generalized Hotelling's test is just apparent and due to the mismatch between its actual and its nominal level of significance for too small values of p. For large value of p (i.e. values for which the actual level of significance reaches its nominal value), the powers of the two tests appear almost identical confirming the p-asymptotical inferential equivalence of the two under the more stringent hypotheses of the Srivastava's test.

In cases D1, R1, and S1, the mismatch between the actual and the nominal level of significance completely affects the Srivastava's test providing meaningless power curves for this test. The only value of interest in these plots are the estimated powers of the Generalized Hotelling's test for $p = 1024$, that is the only case in which the nominal level of significance equals the actual one. Different values of that power are achieved in the three cases. In particular, a comparison of the Generalized Hotelling's test and the Hotelling's test across these tree cases points out an opposite behavior of the two: the power of the Generalized Hotelling's test is higher when the power of the Hotelling's test is lower and viceversa. More in detail, the power of the Generalized Hotelling's test is enhanced (reduced) and the power of the Hotelling's test reduced (enhanced) for alternative hypotheses providing a difference of the means with large (small) components in the directions of important (in terms of eigenvalues) principal components and small (large) components in the directions of minor (in terms of eigenvalues) principal components. Indeed, Hotelling's test is based on the Mahalanobis distance (induced by the inverse of the sample covariance matrix) between the sample difference of the means and the H_0 difference of the means; thus, in the Hotelling's test, the effect of differences occurring in the direction of the minor sample principal components is enlarged with respect to similar differences occurring in the direction of the important sample principal components. On the contrary, in the framework of the Generalized Hotelling's test the directions associated to the minor principal components are expected to be close to the directions detected by $ker(S)$ and thus any difference in these directions have a high probability to be annihilated by the Mahalanobis semi-distance induced by the generalized inverse of the sample covariance matrix used.

4.2 Comparison between the Generalized Hotelling's test and the Pesarin-Salmaso's test

The Generalized Hotelling's test has been also compared with the Pesarin-Salmaso's test (Pesarin and Salmaso Pesarin and Salmaso (2010, 2009)) by means of MC simulations. Aim of this comparison is to see to what extent the model-based approach, pioneered by Srivastava Srivastava (2007) and further developed in this work, can compete with another promising and less traditional approach to the analysis of large p

small n data: multivariate permutation test (Pesarin and Salmaso Pesarin and Salmaso (2010, 2009)).

The Pesarin-Salmaso's test presents some very interesting features (proven in Pesarin and Salmaso Pesarin and Salmaso (2010, 2009)): first of all it does not require the normality of data (test for multivariate normality is still an open problem); secondly, its actual level of significance resembles the nominal level in all simulated scenarios (cases I0, D0, R0, and S0) and for any value of p (i.e., it is not p-asymptotic); finally, under the alternative hypothesis $\mu_a = \mathbf{0}$ and $\mu_b = 0.4 \cdot \mathbf{1}$, its power is non-decreasing in p and has limit 1 for $p \to \infty$ (cases I1, D1, R1, and S1). The Pesarin-Salmaso's test also presents a couple of drawbacks due to the discrete nature of the permutational distribution and to the factorial growth of the number of permutation with respect to the sample size. Indeed, given sample sizes n_a and n_b, we have in general $(n_a + n_b)!$ possible permutations of data associated to $\frac{(n_a+n_b)!}{n_a!\, n_b!}$ possible different values of the test statistic.

The discrete nature of the permutational distribution is particulary evident for small sample sizes: in these cases, test randomization becomes mandatory in order to maintain a certain level of significance α; for instance, for $n_a = n_b = 2$, only 24 permutations are possible, the support of the permutational distribution is reduced to just six values, and thus the only non-randomized bilateral tests that can be performed are the ones carried out with level of significance $\alpha = 1/3$ or $\alpha = 2/3$. If the sample size grows, this issue becomes less relevant from a practical point of view, but on the meantime the number of possible permutations quickly increases making mandatory the use of an approximated permutational distribution based on a randomly selected subset of permutations; for instance, for $n_a = n_b = 10$, the number of possible permutations already exceeds 10^{18} and the support of the permutational distribution is made of more than 10^5 values.

On the whole, the fact that (in most cases) the statistical conclusions might change across different runs of the same analysis and of the same data set makes the permutation-based analysis non replicable. Though the randomness induced by the approximating permutational distribution is just due to computational limits and it can be overcome by increasing the size of the random subset of permutations, on the contrary, the randomness induced by the discrete nature of permutational distribution is not due to computational limits but is intrinsic to this approach and thus non reducible.

A comparison of the estimated power functions of the two tests (cases I1, D1, R1, and S1) shows: a substantial equivalence of the statistical power of the Pesarin-Salmaso's test and of the Hotelling's test in the univariate case (i.e., $p = 1$); a predominance of one of the two depending on the number p of random vector components and on the scenario for small values of p; a neat predominance of Pesarin-Salmaso's test over the Generalized Hotelling's test for larger values of p. In the setting described by the experiment (i.e., $\mu_a = \mathbf{0}$ and $\mu_b = 0.4$), our simulations suggest the permutation-based approach to be more suitable than a model-based approach for the analysis of large p small n data, at least when the sample sizes are large enough to avoid the use of a randomized permutation test. The fact that the latter conclusion would hold in a wider setting is still matter of investigation.

5 Discussion

In this paper we dealt with the problem of making inference for the mean vector of a p-variate normal random vector when the sample size is too small to enable the use of Hotelling's Theorem. The problem of making inference for the difference of the mean vectors of two p-variate normal random vectors when the sample sizes are too small is discussed as well. In particular, we provided a generalization of the Hotelling's Theorem for p going to ∞ and sample size remaining finite, based on the notion of Moore-Penrose generalized inverse of the sample covariance matrix, and that holds under weak assumptions guaranteeing the existence and non degeneracy of the corresponding limit statistic and distribution. Together with the theorem, we provided also explicit formulas to perform hypothesis test and to build confidence regions and T^2-simultaneous confidence intervals.

We tested our results by means of MC simulations performed for different values of the means, of the covariance matrix, of the sample size, and of the number p of random vector components. Simulations confirm our theoretical results and moreover enable the estimation of the statistical power of the test and of the rate of convergence to the p-asymptotic framework. Some interesting cases are presented in the paper and critically discussed in comparison with other two approaches presented in the literature: the Srivastava's test (Srivastava Srivastava (2007)) and the Pesarin-Salmaso's test (Pesarin and Salmaso Pesarin and Salmaso (2010, 2009)); further cases are available upon request of the reader.

The p-asymptotical inferential equivalence of the Srivastava's test and of the Generalized Hotelling's test under the more stringent hypothesis of the Srivastava's test, together with the independence from the ratio $\overline{\sigma}^2/\overline{\sigma^2}$ of the rate of convergence of the Generalized Hotelling's test to its nominal level of significance, suggest $\frac{(tr(\Sigma_p)/p)^2}{tr(\Sigma_p^2)/p}$ (i.e., the finite version of $\overline{\sigma}^2/\overline{\sigma^2}$) to be the right constant to correct the Srivastava's test statistic making it be distributed as a $F(n_a+n_b-2, p-n_a-n_b+3)$, that is the distribution that is known to follow when the covariance matrix is proportional to the identity matrix.

If this were true (further simulations not reported here seem to confirm it), the following conjectures - we could describe within a unique framework the classical Hotelling's test, the Srivastava's test, and the p-asymptotic Generalized Hotelling's test, as follows:

Conjecture 7 (Generalized Hotelling's T^2 distribution law). *For $n \geq 2$ and $p \geq 1$, assume that:*

(i') $\{\mathbf{X}_i\}_{i=1,\dots,n} \sim iid\, N_p(\mu_p, \Sigma_p)$.

Then:

$$\frac{tr(\mathbf{C}_p)^2}{tr(\mathbf{C}_p^2)} \frac{\nu_1}{(n-1)p\nu_2} T^2 \sim F(\nu_1, \nu_2) \ ,$$

with $\nu_1 = |(n-1)-p|+1$, $\nu_2 = \min(n-1,p)$, *and* $\mathbf{C}_p = \begin{cases} \mathbf{I}_p & \textit{for } p \leq n-1 \\ \Sigma_p & \textit{for } p > n-1 \end{cases}$.

Conjecture 8 (Generalized Pooled Hotelling's T^2_{pooled} distribution law)**.** *For* $n_a \geq 1$, $n_b \geq 1$, $n_a + n_b \geq 3$, *and* $p \geq 1$, *assume that:*

(i") $\{\mathbf{X}_{ai}\}_{i=1,\dots,n_a} \sim$ *iid* $N_p(\mu_{pa},\Sigma_p)$, $\{\mathbf{X}_{bi}\}_{i=1,\dots,n_b} \sim$ *iid* $N_p(\mu_{pb},\Sigma_p)$, *and the two finite sequences are independent.*

Then:

$$\frac{tr(\mathbf{C}_p)^2}{tr(\mathbf{C}_p^2)} \frac{\nu_1}{(n_a+n_b-2)p\nu_2} T^2_{pooled} \sim F(\nu_1,\nu_2) ,$$

with $\nu_1 = |(n_a+n_b-2)-p|+1$, $\nu_2 = \min(n_a+n_b-2,p)$, *and* $\mathbf{C}_p = \begin{cases} \mathbf{I}_p & \textit{for } p \leq n_a+n_b-2 \\ \Sigma_p & \textit{for } p > n_a+n_b-2 \end{cases}$.

The previous conjectures are proven to be correct when $p \leq n-1$ and $p \leq n_a+n_b-2$ respectively (Hotelling's Theorem), when $\Sigma_p = \mathbf{I}_p$ (Srivastava Srivastava (2007)) and when $p \to +\infty$ (Corollaries 5 and 6). The generic Σ_p finite p case is at the moment just supported by our simulations and by its consistence with the classical Hotelling's test, the Srivastava's test, and the Generalized Hotelling's test when restricted conditions are posed. Its general proof is still under investigation.

A Some useful properties of the Moore-Penrose Generalized Inverse

Many results presented in the paper rely on properties related to the Moore-Penrose inverse of positive semi-definite sample covariance matrices. In this appendix, all these properties are recalled.

Definition 1. Let A be an $q \times r$ matrix. The Moore-Penrose inverse of A, denoted by A^+, is the unique $r \times q$ matrix such that

1. $AA^+A = A$;
2. $A^+AA^+ = A^+$;
3. $(A^+A)^\star = A^+A$;
4. $(AA^+)^\star = AA^+$.

The first two properties let A^+ be a generalized inverse of A. The last two properties confer to A^+ its uniqueness. The symbol "$\star$" indicates the conjugate of a matrix. For our purposes, all matrices will have real entries and thus, it is equivalent to the symbol "$'$" indicating the transposed matrix.

The proof of the uniqueness of A^+ can be found, for instance, in Rao and Mitra Rao and Mitra (1971).

Moreover, it can be proven, by means of simple computations, that if A is a $p \times p$ symmetric matrix with real entries with rank $m \leq p$, then $A^+ = \sum_{i=1}^m \lambda_i^{-1} \mathbf{e}_i \mathbf{e}_i'$, with $\lambda_1, \dots, \lambda_m$ being the m non-zero eigenvalues of A and $\mathbf{e}_1, \dots, \mathbf{e}_m$ the corresponding eigenvectors. An immediate consequence of this result is that, if A is of full-rank, then $A^+ = A^{-1}$.

Hereby, we report some results necessary to the proof of Theorem 3.

Proposition 9. *Let A be a $\ell \times m$ matrix and B be an $m \times n$ matrix. If*

- *A has orthonormal columns, i.e, $A'A = I_m$; or,*
- *B has orthonormal rows, i.e, $BB' = I_m$; or,*
- *A is of full column rank m and B is of full row rank m,*

then, we have

$$(AB)^+ = B^+ A^+$$

The proof can be found in Rao and Mitra Rao and Mitra (1971).

Proposition 10. *Let A be a $\ell \times m$ matrix. Two particular cases are of interest:*

- *if A is of full column rank m, then $A'A$ is invertible and we get*

$$A^+ = (A'A)^{-1} A'$$

- *if A is of full row rank ℓ, then AA' is invertible and we get*

$$A^+ = A'(AA')^{-1}$$

The proof can be found in Rao and Mitra Rao and Mitra (1971).

Proposition 11. *With the same notations defined in the proof of Theorem 3, we have*

$$W^+ = H'L^{-1}H \; .$$

Proof. We first have

$$V^+ = \left(H'LH\right)^+ = \left[\left(H'L\right)H\right]^+$$

H has orthonormal rows since $HH' = I_m$. Thus, Proposition 9 holds and we have

$$V^+ = H^+ \left(H'L\right)^+$$

Now, we focus on the product $H'L$. H' has orthonormal columns since H has orthonormal rows. Therefore, once again, Proposition 9 holds and we obtain

$$V^+ = H^+ L^+ (H')^+$$

We now observe that

- L is invertible $\Rightarrow L^+ = L^{-1}$;
- H is of full row rank $\Rightarrow H^+ = H'(HH')^{-1}$ thanks to Proposition 10 (first part); then, since $HH' = I_m$, we obtain $H^+ = H'$;
- H' is of full column rank $\Rightarrow (H')^+ = (HH')^{-1}H$ thanks to Proposition 10 (second part); then, since $HH' = I_m$, we obtain $(H')^+ = H$.

and it ends the proof. □

B Proof of Lemma 4

Notation is the same used in the proof of Theorem 3. Lemma 4 states that under the assumptions of Theorem 3:

$$\frac{Y'Y}{p} \xrightarrow[p\to\infty]{\mathscr{P}} \overline{\sigma} I_m \;\; , \;\; \frac{L}{p} \xrightarrow[p\to\infty]{\mathscr{P}} \overline{\sigma} I_m \;\; , \; H\Lambda_p H' \xrightarrow[p\to\infty]{\mathscr{L}} \frac{\overline{\sigma^2}}{\overline{\sigma}} I_m \; .$$

Proof. Observe that, in distribution, the matrix $Y'Y$ is equal to the matrix $U'\Lambda_p U$, where the columns of U are iid $N_p(\mathbf{0}_p, I_p)$.
Note that

$$E[u_{ij}^2] = 1 \quad \text{and} \quad \text{var}[u_{ij}^2] = 2 \; .$$

A generic element of the matrix $Y'Y$ is then equal in distribution to

$$\mathbf{u}_i'\Lambda_p \mathbf{u}_j = \sum_{k=1}^{p} \lambda_k u_{ki} u_{kj} \; .$$

Thus,

$$E\left[\mathbf{u}_i'\Lambda_p \mathbf{u}_j\right] = \textstyle\sum_{k=1}^{p} \lambda_k \delta_{ij} \quad \text{and} \quad \text{var}\left[\mathbf{u}_i'\Lambda_p \mathbf{u}_j\right] = (1+\delta_{ij}) \textstyle\sum_{k=1}^{p} \lambda_k^2 \; .$$

Under assumptions (*iv*) of Theorem 3, it is immediate to see that

$$\begin{aligned} \lim_{p\to\infty} E\left[\tfrac{\mathbf{u}_i'\Lambda_p \mathbf{u}_j}{p}\right] &= \overline{\sigma}\delta_{ij} \; ; \\ \lim_{p\to\infty} \text{var}\left[\tfrac{\mathbf{u}_i'\Lambda_p \mathbf{u}_j}{p}\right] &= \lim_{p\to\,\infty} \tfrac{(1+\delta_{ij})}{p} \tfrac{tr\Lambda_p^2}{p} = 0 \; , \end{aligned}$$

which proves the first and second convergence results of Lemma 4.
Now, let $G = L^{1/2} H Y (Y'Y)^{-1}$. Note that

$$GG' = I_m \; , \quad Y = H' L^{1/2} G \quad \text{and} \quad GY' = L^{1/2} H \; .$$

Thus, in distribution,

$$H\Lambda_p H' = \left(\frac{L}{p}\right)^{-1/2} G \frac{U'\Lambda_p^2 U}{p} G' \left(\frac{L}{p}\right)^{-1/2} \; .$$

We now have

$$\begin{array}{rcl} \lim_{p\to\infty} E\left[\frac{\mathbf{u}_i'\Lambda_p^2\mathbf{u}_j}{p}\right] & = & \lim_{p\to\infty}\frac{tr\Lambda_p^2}{p}\delta_{ij}=\overline{\sigma^2}\delta_{ij}\ ;\ \mathit{and}, \\ \lim_{p\to\infty}\mathrm{var}\left[\frac{\mathbf{u}_i'\Lambda_p^2\mathbf{u}_j}{p}\right] & = & \lim_{p\to\infty}\frac{(1+\delta_{ij})}{p}\frac{tr\Lambda_p^4}{p}=0\ . \end{array}$$

Thus, in probability,

$$\lim_{p\to\infty}\frac{U'\Lambda_p^2U}{p}=\overline{\sigma^2}I_m\ .$$

Now, the root function and the inverse root function are both continuous on $\mathbb{R}^{\star+}$. Thus, in probability,

$$\begin{array}{rcl} \lim_{p\to\infty}\left(\frac{U'\Lambda_p^2U}{p}\right)^{1/2} & = & \left(\overline{\sigma^2}\right)^{1/2}I_m\ ; \\ \lim_{p\to\infty}\left(\frac{L}{p}\right)^{-1/2} & = & (\overline{\sigma})^{-1/2}I_m\ . \end{array}$$

Now, from the fact that $GG'=I_m$ and m is finite, we know that G, as a sequence indexed by p, is tight. From Prokhorov's Theorem (e.g., Billingsley Billingsley (1968)), we therefore know that there exists a subsequence of G which converges in distribution. Let then without loss of generality suppose that we work from now on with this subsequence.

Slutsky's Theorem (e.g., Serfling Serfling (2002)) assures that, in distribution,

$$\lim_{p\to\infty}\left(\frac{U'\Lambda_p^2U}{p}\right)^{1/2}G'=\left(\overline{\sigma^2}\right)^{1/2}G^{\star'}\ ,$$

where $G^{\star'}$ is the $m\times m$ limit matrix of the sequence $\{G_p'\}_{p\geq 1}$. It verifies in particular $G^{\star}G^{\star'}=I_m$.

Used a second time, Slutsky's Theorem implies that, in distribution,

$$\lim_{p\to\infty}\left(\frac{U'\Lambda_p^2U}{p}\right)^{1/2}G'\left(\frac{L}{p}\right)^{-1/2}=\left(\frac{\overline{\sigma^2}}{\overline{\sigma}}\right)^{1/2}G^{\star'}\ .$$

The function

$$\begin{array}{rll} N: & \mathbb{R}^{m\times m} & \to\mathbb{R}^{m\times m} \\ & A & \mapsto A'A \end{array}$$

is continuous and this proves the third convergence result of Lemma 4. □

C Properties of $\widehat{\overline{\sigma_p}}$ and $\widehat{\overline{\sigma_p^2}}$

The means and variances of estimators

$$\begin{array}{rcl} \widehat{\overline{\sigma_p}} & := & \frac{trS}{p},\ \text{and} \\ \widehat{\overline{\sigma_p^2}} & := & \frac{(n-1)^2}{(n-2)(n+1)}\left[\frac{trS^2}{p}-\frac{1}{n-1}\frac{(trS)^2}{p}\right]\ , \end{array} \tag{9}$$

can be trivially computed once introduced a set of p independent random variables $\mathbf{w}_i \sim N_{n-1}(\mathbf{0}_{n-1}, \mathbf{I}_{n-1})$, and noticed the following equalities in distribution (Srivastava Srivastava (2005)):

$$\begin{array}{rcl} (n-1)trS & = & \sum_{i=1}^{p} \lambda_i v_{ii} \,, \\ (n-1)^2(trS)^2 & = & \sum_{i=1}^{p} \lambda_i^2 v_{ii}^2 + 2\sum_{i<j}^{p} \lambda_i \lambda_j v_{ii} v_{jj} \,, \\ (n-1)^2 trS^2 & = & \sum_{i=1}^{p} \lambda_i^2 v_{ii}^2 + 2\sum_{i<j}^{p} \lambda_i \lambda_j v_{ij}^2 \,, \end{array} \tag{10}$$

where $v_{ij} = \mathbf{w}_i' \mathbf{w}_j$.

We can thus write from (9) and (10):

$$\begin{array}{rcl} \hat{\overline{\sigma_p}} & = & \frac{1}{p(n-1)} \sum_{i=1}^{p} \lambda_i v_{ii} \,, \\ \hat{\overline{\sigma_p^2}} & = & \frac{1}{p(n-2)(n+1)} \left[\frac{n-2}{n-1} \sum_{i=1}^{p} \lambda_i^2 v_{ii}^2 + 2\sum_{i<j}^{p} \lambda_i \lambda_j (v_{ij}^2 - \frac{1}{n-1} v_{ii} v_{jj}) \right] \,. \end{array} \tag{11}$$

It is easy, even if quite long, to show from (11) that, for all p, $\hat{\overline{\sigma_p}}$ and $\hat{\overline{\sigma_p^2}}$ are unbiased and consistent (for $n \to \infty$) estimators of $\frac{tr(\Sigma_p)}{p}$ and $\frac{tr(\Sigma_p^2)}{p}$, respectively.

Moreover, if $0 < \overline{\sigma} = \lim_{p\to\infty} \frac{tr(\Sigma_p)}{p} < +\infty$, $0 < \overline{\sigma^2} = \lim_{p\to\infty} \frac{tr(\Sigma_p^2)}{p} < +\infty$, and $0 < \overline{\sigma^4} = \lim_{p\to\infty} \frac{tr(\Sigma_p^4)}{p} < +\infty$, it can be shown that the unbiasedness and consistence hold p-asymptotically:

$$\begin{array}{rclcrcl} E\left[\lim_{p\to\infty} \hat{\overline{\sigma_p}}\right] & = & \overline{\sigma} \,, & & \lim_{n\to\infty} \mathbb{P}\left[\left|\lim_{p\to\infty} \hat{\overline{\sigma_p}} - \overline{\sigma}\right| \leq \varepsilon\right] & = & 1 \ \forall \varepsilon > 0 \,, \\ E\left[\lim_{p\to\infty} \hat{\overline{\sigma_p^2}}\right] & = & \overline{\sigma^2} \,, & & \lim_{n\to\infty} \mathbb{P}\left[\left|\lim_{p\to\infty} \hat{\overline{\sigma_p^2}} - \overline{\sigma^2}\right| \leq \varepsilon\right] & = & 1 \ \forall \varepsilon > 0 \,. \end{array}$$

Thus, for large values of p, $\hat{\overline{\sigma_p}}$ and $\hat{\overline{\sigma_p^2}}$ can be used to estimate $\overline{\sigma}$ and $\overline{\sigma^2}$, respectively.

Note that for $n \to \infty$, estimator $\hat{\overline{\sigma_p^2}}$ is inferentially equivalent to $\frac{tr(S^2)}{p}$ and thus also this simpler but biased estimator results consistent. So, for large values of n and p, $\frac{tr(S)}{p}$ and $\frac{tr(S^2)}{p}$ can be used to estimate $\overline{\sigma}$ and $\overline{\sigma^2}$, respectively.

Acknowledgements

All MC simulations and figures have been obtained by means of the statistical programming environment R (R Development Core Team R Development Core Team (2010)). Finally, we want to thank Prof. Marco Fuhrman (Dept. of Mathematics “F. Brioschi” - Politecnico di Milano) for his help in detailing some technical aspects of the work.

References

Anderson, T. W. (2003), *An introduction to multivariate statistical analysis*, Wiley Series in Probability and Statistics, John Wiley and Sons Inc, 3rd ed.

Benjamini, Y. and Yekutieli, D. (2001), “The control of the false discovery rate in multiple testing under dependency,” *Ann. Statist.*, 29, 1165–1188.

Billingsley, P. (1968), *Convergence of probability measures*, Wiley series in probability and mathematical statistics, Wiley, 2nd ed.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis*, Springer Series in Statistics, Springer, New York.

Hall, P. and Keilegorn, I. V. (2007), “Two-sample tests in Functional Data Analysis starting from discrete data,” *Statistica Sinica*, 17, 1511–1531.

Pesarin, F. and Salmaso, L. (2009), “Advances in Permutation Testing Approach,” Working Paper 6, Department of Management and Engineering, University of Padua.

— (2010), *Permutation Tests for Complex Data: Theory, Applications and Software*, Chichester: Wiley Series in Probability and Statistics.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer New York, 2nd ed.

Rao, C. R. and Mitra, S. K. (1971), *Generalized inverse of matrices and its applications*, Wiley Series in Probability and Statistics, John Wiley and Sons Inc.

Serfling, R. J. (2002), *Approximation Theorems of Mathematical Statistics*, vol. 413 of *Wiley series in probability and mathematical statistics*, Wiley.

Srivastava, M. S. (2005), “Some tests concerning the covariance matrix in high dimensional data,” *Journal of the Japan Statistical Society*, 35, 251–272.

— (2007), “Multivariate theory for analyzing high dimensional data,” *Journal of Japan Statistical Society*, 37, 53–86.

Srivastava, M. S. and Yanagihara, H. (2010), “Testing the equality of several covariance matrices with fewer observations than the dimension,” *J. Multivariate Anal.*, 101, 1319–1329.

Storey, J. D. (2003), “The positive false discovery rate: a Bayesian interpretation and the q-value,” *Ann. Statist.*, 31, 2013–2035.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

06/2011 Piercesare Secchi, A. Stamm, Simone Vantini:
Large p Small n Data: Inference for the Mean

05/2011 Gianni Arioli, Monica Gamba:
An algorithm for the study of parameter dependence for hyperbolic systems

04/2011 Francesca Ieva, Anna Maria Paganoni,
Davide Pigoli, Valeria Vitelli:
Multivariate functional clustering for the analysis of ECG curves morphology

03/2011 Giulia Garegnani, Giorgio Rosatti, Luca Bonaventura:
Mathematical and Numerical Modelling of Fully Coupled Mobile Bed Free Surface Flows

02/2011 Toni Lassila, Alfio Quarteroni, Gianluigi Rozza:
A reduced basis model with parametric coupling for fluid-structure interaction problems

01/2011 M. Dalla Rosa, Laura M. Sangalli, Simone Vantini:
Dimensional Reduction of Functional Data by means of Principal Differential Analysis

43/2010 Giancarlo Pennati, Gabriele Dubini,
Francesco Migliavacca, Chiara Corsini,
Luca Formaggia, Alfio Quarteroni,
Alessandro Veneziani:
Multiscale Modelling with Application to Paediatric Cardiac Surgery

42/2010 Stefano Baraldo, Francesca Ieva,
Anna Maria Paganoni, Valeria Vitelli:
Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients

41/2010 Davide Ambrosi, Gianni Arioli, Fabio Nobile,
Alfio Quarteroni:
Electromechanical coupling in cardiac dynamics: the active strain approach

40/2010 Carlo D'Angelo, Anna Scotti:
A Mixed Finite Element Method for Darcy Flow in Fractured Porous Media with non-matching Grids