A generalization of Hotelling's theorem for large p small n data

Piercesare Secchi, Aymeric Stamm, Simone Vantini

A GENERALIZATION OF HOTELLING'S THEOREM FOR LARGE p SMALL n DATA

Piercesare Secchi[1], Aymeric Stamm[2] and Simone Vantini[1]

[1] MOX - Department of Mathematics "Francesco Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy.
(e-mail: piercesare.secchi@polimi.it, simone.vantini@polimi.it)

[2] University of Rennes I, IRISA, UMR CNRS-6074 Campus de Beaulieu, F-35042 Rennes, France.
(e-mail: aymeric.stamm@irisa.fr)

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 (e.g., spectral data, micro-arrays, and functional data). The convergence rate of the new statistic to its p-asymptotic distribution is analyzed by means of MC simulations, as well as its power which is compared with that of two recent alternatives: a model-dependent test relying on stronger assumptions (Srivastava (2007)) and a model-free permutation test relying on weaker assumptions (Pesarin and Salmaso (2010)).

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 (2007) is pioneering in this direction. A generalization of the Hotelling's Theorem is there 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 providing a modeling perspective of little practical interest. 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 (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.

Permutation tests provide a further alternative approach to the inference for large p small n data. According to this approach, the extremity of a suitable test statistic is tested with respect to its permutational distribution (i.e., the discrete distribution obtained by randomly assigning the observed data to units). 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 (2010) recently proposed the use of permutation tests in the framework of multivariate analysis (even in the case of large p small n data).

Similarly to Srivastava (2007) and differently from Pesarin and Salmaso (2010), 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. In the last section, our proposal is empirically compared with the inferential tools proposed in Srivastava (2007) and in Pesarin and Salmaso (2010), considered as prototypes of two approaches to the same problem that are very close and far from ours, respectively.

2 GENERALIZING 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 1 (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. Theorem 1 and Corollary 1 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..

We thus 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. Let A^+ being the Moore-Penrose inverse of a real positive semi-definite matrix A.

Theorem 2 (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 $p \to \infty$:

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

Corollary 2 (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 $p \to \infty$:

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

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

The proof of Theorem 2 and of Corollary 2 can be found in Secchi et al.(2010).

Note that Corollary 2 is based on the univariate random quantity $n(\overline{\mathbf{X}}-\mu_p)'S^+(\overline{\mathbf{X}}-\mu_p)$. We named this quantity Generalized Hotelling's T^2 since it can be proven (Secchi et al. (2010)) that it generalizes Hotelling's $T^2 = n(\overline{\mathbf{X}}-\mu_p)'S^{-1}(\overline{\mathbf{X}}-\mu_p)$ that appears in Corollary 1. Indeed, Hotelling's T^2 is defined only for $n > p \geq 1$ while the Generalized Hotelling's T^2 is defined for any n and p such that $n \geq 2$ and $p \geq 1$, and it coincides with the former when $n > p$.

Corollary 2 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 variables is far larger than the number n of sample units (e.g., genetics) or even virtually infinite (e.g., functional data). See Secchi et al. (2010) for explicit representations of the corresponding confidence and rejection regions.

The corresponding confidence and rejection regions present some peculiar features that are worth a little discussion. Because S^+ is positive semi-definite, given a data set, the confidence region – 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 $Im(S)$. In particular, the confidence region 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 test statistic $\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{np}{n-1} (\overline{\mathbf{X}} - \mu_{0p})' S^+ (\overline{\mathbf{X}} - \mu_{0p})$ 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).

Theorem 2 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 3 (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") $\{\mathbf{X}_{ai}\}_{i=1,\ldots,n_a} \sim iid\ N_p(\mu_{pa}, \Sigma_p)$, $\{\mathbf{X}_{bi}\}_{i=1,\ldots,n_b} \sim iid\ N_p(\mu_{pb}, \Sigma_p)$ *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 $p \to \infty$:

$$\frac{\overline{\sigma}^2}{\overline{\sigma^2}} \frac{p}{n_a + n_b - 2} \left(\frac{1}{n_a} + \frac{1}{n_b} \right)^{-1} \cdot$$
$$\cdot \left((\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b}) - (\mu_{pa} - \mu_{pb}) \right)' S^+_{pooled} \left((\overline{\mathbf{X}}_\mathbf{a} - \overline{\mathbf{X}}_\mathbf{b}) - (\mu_{pa} - \mu_{pb}) \right) \xrightarrow{\mathcal{D}} \chi^2(n_a + n_b - 2) \ ,$$

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.

The construction of a confidence region for estimating the difference of the two means and rejection region for testing the difference of the two means comes naturally (Secchi et al. (2010)).

3 Comparison with Pesarin-Salmaso's and Srivastava's Tests

In this section, we estimate, by means of MC simulations (1000 replications), the power and the actual level of significance of the new test that can be derived from Corollary 3 (i.e., 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 $H_1 : \mu_a \neq \mu_b$ at significance level

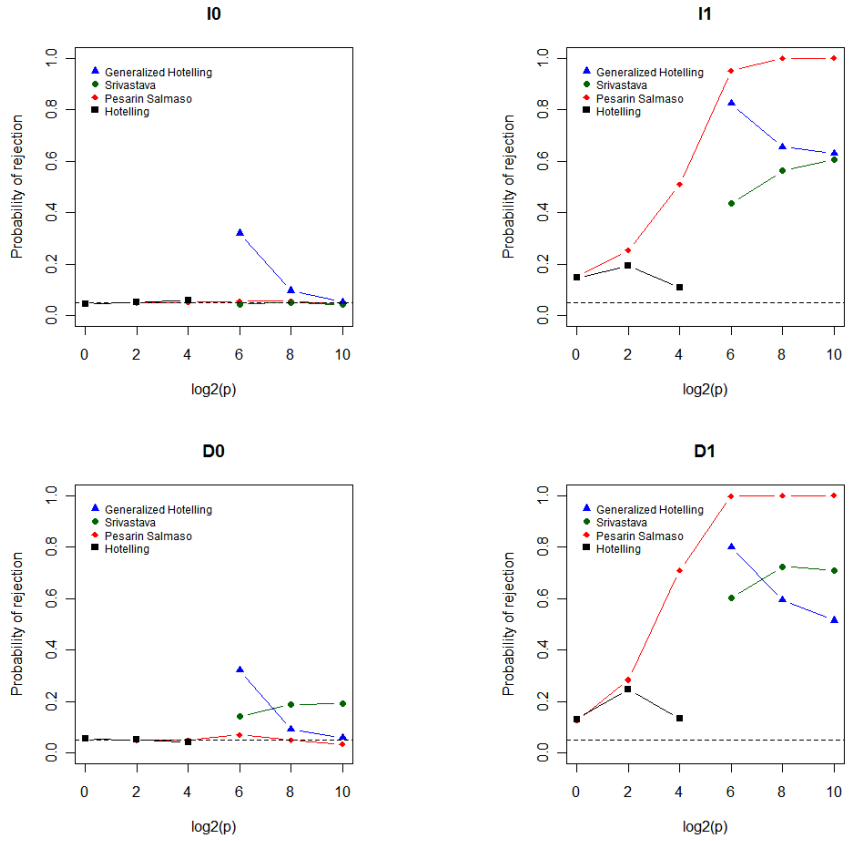

Figure 1. MC-estimates of the probability of rejecting $H_0 : \mu_a = \mu_b$ for different values of the number p of components. Each plot is associated to a different model (title) and each line to a different test (legend).

$\alpha = 5\%$ in four different cases of normal populations and for increasing values of the number p of components ranging between 2^0 and 2^{10} (i.e., 1 and 1024):

$$\begin{aligned}
\text{I0} &: \mu_a = \mathbf{0},\ \mu_b = \mathbf{0}, && \Sigma = I,\ n_a = 10,\ n_b = 10;\\
\text{D0} &: \mu_a = \mathbf{0},\ \mu_b = \mathbf{0}, && \Sigma = D,\ n_a = 10,\ n_b = 10;\\
\text{I1} &: \mu_a = \mathbf{0},\ \mu_b = 0.4 \cdot \mathbf{1},\ \Sigma = I,\ n_a = 10,\ n_b = 10;\\
\text{D1} &: \mu_a = \mathbf{0},\ \mu_b = 0.4 \cdot \mathbf{1},\ \Sigma = D,\ n_a = 10,\ n_b = 10
\end{aligned}$$

where I is the identity matrix; D is a diagonal matrix whose diagonal alternatively assumes the values 0.5 and 1.5. 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 (2010); the value for Σ used in cases I0 and I1 are the same used in Pesarin and Salmaso (2010), while the value used in cases D0 and D1 are meant to provide a less trivial situation (other cases are available in Secchi et al. (2010)).

In case I0, where H_0 is true and the assumptions supporting the Srivastava's test (i.e., independence and homoscedasticity of components) hold, the observed rate of rejection of the Srivastava's test clearly matches its nominal level of significance 5%; on the contrary, in case D0, where H_0 is still true but the assumptions 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. This comparison shows that the Srivastava's test can be become strongly biased even under small deviations from the homoscedasticity assumption.

The assumptions which the Generalized Hotelling's test is based on, hold for both I0 and D0, 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 same simulations also show that under the assumptions of independence and homoscedasticity of components, the Srivastava's test and the Generalized Hotelling's test are p-asymptotically equivalent; moreover they also show that the convergence rate of the latter 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 that is necessary to make the Generalized Hotelling's test reliable.

The Generalized Hotelling's test has been also compared with the Pesarin-Salmaso's test (Pesarin and Salmaso (2010)). Aim of this comparison is to see to what extent the model-based approach can compete with another promising and less traditional approach to the analysis of large p small n data: multivariate permutation test.

The actual level of significance of the Pesarin-Salmaso's test resembles the nominal level in both scenarios I0 and D0 for any value of p; moreover, 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 $p \to \infty$. The Pesarin-Salmaso's test also presents some 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. These issues are discussed in Secchi et al.(2010).

A comparison of the estimated power functions of the two tests (cases I1 and D1) shows 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.

REFERENCES

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

SECCHI, P., STAMM, A., VANTINI, S. (2010), "Large p Small n Data: Inference for the Mean," Tech. Rep. MOX 06/2011, Dept. of Mathematics, Politecnico di Milano.

SRIVASTAVA, M. S. (2007), "Multivariate theory for analyzing high dimensional data," *Journal of Japan Statistical Society*, 37, 53–86.