Classification and Data Analysis 2009

Book of Short Papers

7° Meeting of the Classification and Data Analysis Group

of the Italian Statistical Society

Catania – September 9-11, 2009

Editors

Salvatore Ingrassia and Roberto Rocci

cleup

Table of contents

KEYNOTE PAPERS

INVITED PAPERS

Specialized Session 1
Multivariate Relations for General Data
Organizer: Carles M. Cuadras

Specialized Session 2

Estimating Causal Effects from Observational Data using Large Databases and Multivariate Matching

Organizer: Fabrizia Mealli

Specialized Session 3

New Developments in Complex Data Modeling for Evaluation

Organizer: Giorgio Vittadini

Specialized Session 4

Classification of Measurement Errors: Theory, ISO Standards and Applications

Organizer: Laura Deldossi, Diego Zappa

Specialized Session 5
Visualization of Relationships by Multidimensional Scaling
Organizer: Akinori Okada

Specialized Session 6
Computational and Theoretical Aspects in Mixture Models
Organizer: Donatella Vicari

Specialized Session 7
Recent Trends in Cluster Analysis
Organizers: Yves Lechevallier

Specialized Session 8
Models for the Analysis of Volatility in Financial Markets
Organizer: Edoardo Otranto

Specialized Session 9
Model selection in Advanced Classification and Regression Methods
Organizers: Roberta Siciliano, Claudio Conversano

Specialized Session 10
Operative Problems of Data Mining and Statistical Solutions
Organizer: Furio Camillo

Specialized Session 11
Statistical Methods for the Assessment of Scientific Research Communities
Organizer: Giuseppe Giordano

Specialized Session 12
Stable Likelihood Methods in Mixture Models
Organizer: Wilfried Seidel

Specialized Session 13

Statistical Analysis of Microarray Data

Organizer: Angelo M. Mineo

Specialized Session 14

Robust Methods for the Analysis of Complex Data

Organizer: Marco Riani

Solicited Session 1

Environmental Assessment and Air Quality Indicators

Organizer: Antonella Plaia

Solicited Session 2

Variable Selection in Multivariate Analysis

Organizer: Gabriele Soffritti

Solicited Session 3

Statistical Models for Electoral Flow Data

Organizer: Venera Tomaselli

Solicited Session 4

Partitioning Methods for Statistical Learning

Organizer: Massimo Aria

Solicited Session 5

Identification and Estimation of Causal Effects in the Presence of Complications

Organizer: Leonardo Grilli

Solicited Session 6

Statistical Analysis of Social Networks of Enterprises

Organizer: Francesco Palumbo

Solicited Session 7

Methodology for Evaluating the Effectiveness of the Italian University System

Organizer: Luigi Fabbris

Solicited Session 8
Advances in Data Collection for Surveys
Organizer: Pier Francesco Perri

Solicited Session 9
Functional Data Analysis
Organizer: Rosanna Verde

Solicited Session 10
Econometric Analysis of Data on Market Behaviors
Organizer: Roberto Cellini

Solicited Session 11
From Market Segmentation to Different Customer Loyalties
Organizer: Mario Montinaro

Solicited Session 12
Markov and Hidden Markov Processes for Data Analysis and Simulation
Organizer: Isabella Morlini

CONTRIBUTED PAPERS

Multivariate tests for patterned covariance matrices

Francesca Greselin, Salvatore Ingrassia and Antonio Punzo

Abstract In many real datasets, the same variables are measured on objects from different groups, and the covariance structure may vary from group to group. Oftentimes, the underlying population covariance matrices are not identical, yet they still have a common basic structure, e.g. there exists some rotation that diagonalizes simultaneously all covariance matrices in the groups, or all covariances can be made congruent by some translation and/or dilation. The purpose of this paper is to show how a test of homoscedasticity can be made more informative by performing a separate check for equality between shapes and equality between orientations of the concentration ellipsoids. This approach, combined with parsimonious parametrization in mixture data modeling, provides a formal hypothesis testing procedure for model assessment.

Key words: Homoscedasticity, Spectral decomposition, Principal component analysis, F-G algorithm, Multiple tests, UI tests, IU tests.

1 Introduction and basic definitions

In multivariate analysis it is customary to employ an unbiased version of the likelihood ratio (LR) test to ascertain equality of the covariance matrices referred to several groups. However, in many applications, data groups escape from homoscedasticity; nevertheless, one can observe some sort of common basic structure among covariance matrices, f.i. they share orientation or shape or size. To this aim, their spectral decomposition is employed, i.e. the *eigenvalues* and *eigenvectors* representation, as suggested by many authors in the literature, see [4]). In clustering males and females blue crabs [3] by a mixture model-based approach, Peel and McLachlan [7] assumed that the two group-conditional distributions were bivariate normal with

F. Greselin - Dip. Metodi Quantitativi per le Scienze Economiche e Aziendali - Università di Milano-Bicocca - Via degli Arcimboldi, 8 - 20126 Milano (Italy) - e-mail: francesca.greselin@unimib.it
S. Ingrassia, A. Punzo - Dip. Economia e Metodi Quantitativi - Università di Catania - Corso Italia, 55 - 95129 Catania (Italy) e-mail: [s.ingrassia,antonio.punzo]@unict.it

common covariance matrix, on the basis of Hawkins' test. As pointed out by the authors, this assumption produced a larger misallocation rate than the unconstrained model. Greselin and Ingrassia in [5] resolved this apparent contradiction, requiring only covariances matrices with the same set of (ordered) eigenvalues, i.e. ellipsoids of equal concentration with the same size and shape. In the present paper, Flury's proposal and the above remarks on patterned covariance matrices are jointly considered to provide a unified framework for a more informative scedasticity comparison among groups.

Now we introduce some basic notation and definition: supposing to deal with p variables measured on objects arising from $k \geq 2$ groups, let $\mathbf{x}_1^{(h)}, \ldots, \mathbf{x}_{n_h}^{(h)}$ denote n_h independent observations, for the hth group, drawn from a normal distribution with mean vector μ_h and covariance matrix $\boldsymbol{\Sigma}_h$, $h = 1, \ldots, k$. Let $\boldsymbol{\Sigma}_h = \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}_h \boldsymbol{\Gamma}'_h$ be the spectral decomposition of $\boldsymbol{\Sigma}_h$, where $\boldsymbol{\Lambda}_h = \text{diag}(\lambda_1^{(h)}, \ldots, \lambda_p^{(h)})$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_h$ sorted in non-increasing order and $\boldsymbol{\Gamma}_h$ is the $p \times p$ orthogonal matrix whose columns $\gamma_1^{(h)}, \ldots, \gamma_p^{(h)}$ are the normalized eigenvectors of $\boldsymbol{\Sigma}_h$ ordered according to their eigenvalues, $h = 1, \ldots, k$; here, and in what follows, the prime denotes the transpose of a matrix. On the other hand, k covariance matrices which share the same matrix of orthonormalized eigenvectors $\boldsymbol{\Gamma}_1 = \cdots = \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}$ have ellipsoids of equal concentration with the same *axis orientation* in p-space, i.e. they are congruent up to a suitable dilation/contraction (alteration in size) and/or "deformation" (alteration in shape). By using the Greek term "*trò pos*" (orientation), we call it "*homotroposcedasticity*".

When homometroscedasticity and homotroposcedasticity simultaneously hold, we have homoscedastic data. Moreover, an intermediate situation between heteroscedasticity and homoscedasticity, denoted as "*weak homoscedasticity*" can be devised, when only one of the two above conditions holds.

2 More informative multiple tests to compare "scedasticity"

Consider the null hypothesis of homometroscedasticity $H_0^\Lambda : \Lambda_1 = \cdots = \Lambda_k = \Lambda$ versus the alternative $H_1^\Lambda : \Lambda_h \neq \Lambda_l$ for some $h, l \in \{1, \ldots, k\}, h \neq l$. $H_0^\Gamma : \Gamma_1 = \cdots = \Gamma_k = \Gamma$ versus the alternative $H_1^\Gamma : \Gamma_h \neq \Gamma_l$ for some $h, l \in \{1, \ldots, k\}, h \neq l$.

A test of homoscedasticity can be re-expressed by a union-intersection (UI) test as $H_0^{\text{S}} : H_0^\Lambda \cap H_0^\Gamma$ versus $H_1^{\text{S}} : H_1^\Lambda \cup H_1^\Gamma$. Differently from a usual test of homoscedasticity, the present approach shows its profitability whenever the null hypothesis H_0^{S} is rejected: the component tests for H_0^Λ and H_0^Γ can discriminate the nature of the departure from H_0^{S}. On the other hand, a test of weak homoscedasticity may be formulated as an intersection-union (IU) test: $H_0^{\text{W}} : H_0^\Lambda \cup H_0^\Gamma$ versus $H_1^{\text{W}} : H_1^\Lambda \cap H_1^\Gamma$.

Testing homometroscedasticity. Let $\bar{\mathbf{x}}_h$ and $\mathbf{S}_h$ respectively be the sample mean vector and the unbiased sample covariance matrix in the hth group, $h = 1, \ldots, k$. Moreover, let G_h be the $p \times p$ orthogonal matrix whose columns are the normalized eigenvectors of S_h ordered by the non-increasing sequence of the eigenvalues of S_h, $h = 1,\ldots,k$. According to the *principal component (linear) transformation* $\mathbf{x}_i^{(h)} \rightarrow \mathbf{y}_i^{(h)} = \mathbf{G}'_h(\mathbf{x}_i^{(h)} - \bar{\mathbf{x}}_h)$, $i = 1,\ldots,n_h$, the data $\mathbf{y}_1^{(h)},\ldots,\mathbf{y}_{n_h}^{(h)}$, are uncorrelated, and their covariance matrix $\mathbf{L}_h$ is the diagonal matrix containing the non-negative eigenvalues of $\mathbf{S}_h$. Since the assumption of multivariate normality holds, these components are also independent and normally distributed. Based on these results, the null hypothesis H_0^{Λ} can be re-expressed as follows $H_0^{\Lambda} : \bigcap_{j=1}^{p} H_0^{\lambda_j}$, where $H_0^{\lambda_j} : \lambda_j^{(1)} = \cdots = \lambda_j^{(k)} = \lambda_j$, and λ_j is the unknown jth eigenvalue, common to the k groups. The problem can now be approached by a UI test, through p simpler tests of equality of variances in k groups.

Testing homotroposcedasticity. Consider the spectral decomposition of $\boldsymbol{\Sigma}_h$, i.e. $\boldsymbol{\Sigma}_h = \boldsymbol{\Gamma}_h \boldsymbol{\Lambda}_h \boldsymbol{\Gamma}'_h$. Under H_0^{Γ}, the matrix $\boldsymbol{\Gamma}$ simultaneously diagonalizes all covariance matrices and generates the eigenvalue matrices Λ_h. Thus H_0^{Γ} can be restated as $H_0^{\Gamma} : \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_h\boldsymbol{\Gamma} = \boldsymbol{\Lambda}_h, h = 1,\ldots,k$.

To the best of the authors' knowledge, a direct method to deal with the latter expression for H_0^{Γ} does not exist. However, under the assumption of multivariate normality, Flury derived the log-LR statistics for testing the weaker null hypothesis of common principal components $H_0^{\mathrm{CPC}} : \underset{\sim}{\boldsymbol{\Gamma}}'\boldsymbol{\Sigma}_h\underset{\sim}{\boldsymbol{\Gamma}} = \underset{\sim}{\boldsymbol{\Lambda}}_h$, $h = 1,\ldots,k$, where $\underset{\sim}{\boldsymbol{\Gamma}}$ is a $p \times p$ orthonormalized matrix that diagonalizes all covariance matrices, and $\underset{\sim}{\boldsymbol{\Lambda}}_h$ is one of the possible $p!$ diagonal matrices of eigenvalues in the hth group, $h = 1,\ldots,k$. Note that, in contrast with the latter formulation of H_0^{Γ}, no canonical ordering of the columns of $\underset{\sim}{\boldsymbol{\Gamma}}$ is specified here. In order to apply Flury's proposal, the *F-G algorithm* can estimate the sample counterpart $\underset{\sim}{\mathbf{G}}$ of $\underset{\sim}{\boldsymbol{\Gamma}}$. Under H_0^{CPC}, the following transformation $\mathbf{x}_i^{(h)} \rightarrow \underset{\sim}{\mathbf{y}}_i^{(h)} = \underset{\sim}{\mathbf{G}}'(\mathbf{x}_i^{(h)} - \bar{\mathbf{x}}_h)$ holds, where the data $\underset{\sim}{\mathbf{y}}_1^{(h)},\ldots,\ \underset{\sim}{\mathbf{y}}_{n_h}^{(h)}$ are uncorrelated with diagonal covariance matrix $\underset{\sim}{\mathbf{L}}_h$ (the sample counterpart of $\underset{\sim}{\Lambda}_h$), for $h = 1,\ldots,k$. From a geometrical point of view, under H_0^{Γ} the k ellipsoids of equal concentration have the same orientation in p-space while, under H_0^{CPC} they have only the same principal axes. For Hence, to test H_0^{Γ}, first we perform Flury's test H_0^{CPC} of equality between principal axes; afterwards, if H_0^{CPC} is accepted, perform a statistical test to evaluate H_0^{R}. For further details about the tests, see [6]. Finally we remark that in order to preserve the chosen level α for the overall null hypothesis, the significance levels in the individual tests have to be devised with some care. In any case, the implemented R routine requires only the overall significance level and derives the component ones.

3 An application to clustering

In mixture-model based clustering, the conceptual analysis on similarities between covariance matrices allows a more parsimonious parametrization, and the scedasticity tests can be employed for model assessment. The following example illustrates the real gain in using these multiple tests combining them to parsimonious parametrization in mixture-modeling.

Table 1 shows the results obtained by applying the `Mclust` procedures, followed by the normality and scedasticity tests on the obtained clusters on the Crab dataset. The package suggests three best models, but they significantly differ in covariance structure and/or number of groups. For $k = 2$, a homometroscedastic (EEV) and a heteroscedastic (VVV) model are proposed; further, the slight difference of 0.983% in BIC values is of little help for the choice. Now, considering the first model and performing Mardia's test on each cluster, the hypothesis of normality is accepted in both groups, while Box's test fails, the new tests assess equality of the eigenvalues and reject equality of the eigenvectors in the two groups. In the second model, even if normality holds in each cluster and the test for homoscedasticity fails (as we expected, for a VVV model), the homometroscedasticity test on the two clusters assesses that they have the same eigenvalues. This contradiction leads to discard the second model. With reference to the third model, Mardia's test rejects the null hypothesis of normality in one out of three groups and this allow us to discard the model. Hence the testing methodology selects the first model out of the three proposed by Mclust. As the true classification on the Crab is known, we can verify also that the first model has a significantly lower misclassification error, showing the real gain achieved by the testing methodology.

	`Mclust` results			Test results	
Model	number of groups	identifier	BIC	Normality	Scedasticity
1	2	EEV	-916.1354	in both groups	homometroscedasticity
2	2	VVV	-925.2317	in both groups	homometroscedasticity
3	3	EEV	-933.9190	only in 2 groups	homometroscedasticity

Table 1 Best `Mclust` models on the Crab dataset and tests results for model assessment.

References

1. E. Anderson. The irises of the Gaspe peninsula. *Bullettin of the American Iris Society*, 59:2–5, 1935.
2. J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.
3. N.A. Campbell and R.J. Mahon. A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology*, 22:417–425, 1974.
4. B.N. Flury. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons, 1988.
5. F. Greselin and S. Ingrassia. Weakly homoscedastic constraints for mixtures of *t* distributions. In Andreas Fink, Berthold Lausen, Wilfried Seidel, and Alfred Ultsch, editors, *Advances in Data Analysis, Data Handling and Business Intelligence*. Springer, 2009.
6. F. Greselin S. Ingrassia and A. Punzo. A more informative approach to compare scedasticity under the assumption of multivariate normality. *Rapporti di ricerca Dip. Metodi Quantitativi Sc. Econ. ed Aziendali*, Univ. Milano Bicocca, n°166, 2009, available at http://www.dimequant.unimib.it/_ricerca/pubblicazione.jsp?id=189.
7. D. Peel and G.J. McLachlan. Robust mixture modelling using the *t* distribution. *Statistics & Computing*, 10:339–348, 2000.

Printed by C.L.E.U.P. “Coop. Libraria Editrice Università di Padova”
Via G. Belzoni, 118/3 - Padova - Italy (Tel. +39 049 650261)
www.cleup.it