Tests basés sur les rangs pour le modèle à Composantes Principales Communes

Marc Hallin, Davy Paindaveine, Thomas Verdebout

Tests basés sur les rangs pour le modèle à Composantes Principales Communes

Marc Hallin, Davy Paindaveine & Thomas Verdebout

E.C.A.R.E.S., Institut de Recherche en Statistique, and Département de Mathématique
Université Libre de Bruxelles, Brussels, Belgium

Les composantes principales sont l'un des plus anciens et des plus populaires outils de l'Analyse Multivariée puisqu'elles furent introduites par Pearson (1901) et redécouvertes un peu plus tard par Hotelling (1933). Abordant des problèmes à plusieurs échantillons, Flury introduit en 1984 le modèle à Composantes Principales Communes (CPC). Ce modèle a, depuis lors, trouvé de nombreuses applications, surtout en Biologie (Airoldi and Hoffmann (1984), Flury and Riedl (1988)). Sous l'hypothèse de CPC, m populations possèdent des matrices de covariance $\mathbf{\Sigma}_{i;\text{Cov}}$, $i = 1, \ldots, m$ qui ont les mêmes vecteurs propres avec des valeurs propres possiblement différentes. Plus précisément, ces matrices de covariance admettent les décompositions spectrales $\mathbf{\Sigma}_{i;\text{Cov}} = \boldsymbol{\beta}\mathbf{\Lambda}_{\mathbf{\Sigma}_{i;\text{Cov}}}\boldsymbol{\beta}'$ pour un certain m-uple de matrices diagonales positives $\mathbf{\Lambda}_{\mathbf{\Sigma}_{i;\text{Cov}}}$, $i = 1, \ldots, m$, et une certaine matrice orthogonale $\boldsymbol{\beta}$ caractérisant les vecteurs propres communs ou, de manière équivalente, les composantes principales communes.

Avant de considérer une analyse statistique basée sur ce modèle, il semble naturel de vérifier s'il est compatible avec les données traitées. Flury (1984) développe un test du rapport de vraisemblance gaussien pour l'hypothèse nulle $\mathcal{H}_0$ de composantes principales communes. Ce test est basé sur la loi asymptotique (sous $\mathcal{H}_0$) de $-2\log\Lambda$ où, notant $\mathbf{S}_i^{(n)}$, $i = 1, \ldots, m$ les matrices de covariance empiriques calculées à partir de m échantillons mutuellement indépendants d'observations à valeurs dans $\mathbb{R}^k$ et par $\hat{\boldsymbol{\beta}}$ l'estimateur maximum de vraisemblance contraint de $\boldsymbol{\beta}$,

$$\Lambda := \prod_{i=1}^{m} \left(\frac{\left|\hat{\boldsymbol{\beta}}'\mathbf{S}_i^{(n)}\hat{\boldsymbol{\beta}}\right|}{\left|\text{diag}(\hat{\boldsymbol{\beta}}'\mathbf{S}_i^{(n)}\hat{\boldsymbol{\beta}})\right|} \right)^{n_i/2}$$

(nous notons $\text{diag}(\mathbf{A})$ la matrice diagonale ayant les mêmes éléments diagonaux que la matrice carrée $\mathbf{A}$). L'interprétation de ce test est relativement claire: sous $\mathcal{H}_0$, $\hat{\boldsymbol{\beta}}'\mathbf{S}_i^{(n)}\hat{\boldsymbol{\beta}}$ doit être approximativement diagonale, donc $|\hat{\boldsymbol{\beta}}'\mathbf{S}_i^{(n)}\hat{\boldsymbol{\beta}}|$ et $|\text{diag}(\hat{\boldsymbol{\beta}}'\mathbf{S}_i^{(n)}\hat{\boldsymbol{\beta}})|$ approximativement égaux, auquel cas, Λ est proche de 1. La distribution asymptotique de $-2\log\Lambda$ est une conséquence directe de résultats classiques de Wilks (1938), résultats qui néanmoins restent valables sous des hypothèses gaussiennes seulement.

Hallin et al.(2008a) introduisent un test pseudo-gaussien $\phi^{(n)}_{\text{HPV}}$ pour la même hypothèse nulle $\mathcal{H}_0$ de CPC. Ce test reste valide sous des familles de lois elliptiques possèdant des paramètres d'applatissemment différents. Il requiert néanmoins que les lois elliptiques

sous-jacentes possèdent des moments d'ordre quatre finis. Asymptotiquement équivalent au test de Flury sous des densités gaussiennes, il conserve les propriétés d'optimalité de ce dernier. En dehors du cas gaussien, ses résultats en terme de puissance sont beaucoup moins bons. Nous introduisons des tests de rangs signés pour le même problème de test. Nos tests ne requièrent aucune hypothèse de moment sur les lois elliptiques sous-jacentes pour conserver leurs validités. Ils atteignent les bornes d'efficacité semiparamétrique sous un m-uple de densités elliptiques spécifiées. Dans le cas homogène (toutes les densités elliptiques sont les mêmes), la version gaussienne de nos tests de rangs signés domine uniformément, au sens de Pitman, le test pseudo-gaussien optimal. Les résultats sont obtenus en utilisant la théorie asymptotique de Le Cam adaptée au contexte d'expériences statistiques dites courbées.

Principal components—arguably, the oldest and most popular tool of multivariate analysis—were originally introduced by Pearson (1901), then rediscovered by Hotelling (1933), in a one-sample setup. Multisample principal component problems only came much later. In 1984, Flury introduced the Common Principal Components (CPC) model which since then has been used in a number of applications, mainly in a biometric context (see e.g. Airoldi and Hoffmann (1984), Flury and Riedl (1988)). Under such a model, m k-dimensional populations, with covariance matrices $\mathbf{\Sigma}_{i;\mathrm{Cov}}$, $i = 1, \ldots, m$, are assumed to share, with possibly different eigenvalues, the same principal components: namely, these covariance matrices factorize into $\mathbf{\Sigma}_{i;\mathrm{Cov}} = \boldsymbol{\beta}\mathbf{\Lambda}_{\mathbf{\Sigma}_{i;\mathrm{Cov}}}\boldsymbol{\beta}'$ for some m-tuple of positive diagonal matrices $\mathbf{\Lambda}_{\mathbf{\Sigma}_{i;\mathrm{Cov}}}$, $i = 1, \ldots, m$, and some orthogonal matrix $\boldsymbol{\beta}$—the matrix of *common eigenvectors*, characterizing the *common principal components*.

Before considering a statistical analysis based on such model, however, it is natural to check whether the CPC assumption is compatible with the data under study. Flury (1984) therefore developed a Gaussian likelihood ratio test for the null hypothesis $\mathcal{H}_0$ of common principal components. This test is based on the asymptotically chi-square null distribution of $-2\log\Lambda$ where, denoting by $\mathbf{S}_i^{(n)}$, $i = 1, \ldots, m$ the empirical covariance matrices computed from m mutually independent samples of k-dimensional independent observations and by $\hat{\boldsymbol{\beta}}$ the constrained maximum likelihood estimator of $\boldsymbol{\beta}$,

$$\Lambda := \prod_{i=1}^{m} \left(\frac{\left|\hat{\boldsymbol{\beta}}' \mathbf{S}_i^{(n)} \hat{\boldsymbol{\beta}}\right|}{\left|\mathrm{diag}(\hat{\boldsymbol{\beta}}' \mathbf{S}_i^{(n)} \hat{\boldsymbol{\beta}})\right|} \right)^{n_i/2}$$

(we write $\mathrm{diag}(\mathbf{A})$ for the diagonal matrix having the same diagonal elements as a squared matrix $\mathbf{A}$). The intuition behind this test is clear: under $\mathcal{H}_0$, $\hat{\boldsymbol{\beta}}' \mathbf{S}_i^{(n)} \hat{\boldsymbol{\beta}}$ should be nearly diagonal, hence $|\hat{\boldsymbol{\beta}}' \mathbf{S}_i^{(n)} \hat{\boldsymbol{\beta}}|$ and $|\mathrm{diag}(\hat{\boldsymbol{\beta}}' \mathbf{S}_i^{(n)} \hat{\boldsymbol{\beta}})|$ approximately equal, in which case Λ is close to one. The asymptotically chi-square distribution of $-2\log\Lambda$ follows from the classical asymptotic result of Wilks (1938) which however is valid under Gaussian assumptions only.

Hallin et al.(2008a) introduce a pseudo-Gaussian test $\phi^{(n)}_{\mathrm{HPV}}$ for the same hypothesis $\mathcal{H}_0$ of CPC which however remains valid under possibly *heterokurtic* elliptical families but still requires, as its competitors, finite moments of order four. This test, which retains the optimality properties of Flury's LRT under Gaussian densities, however appears to suffer a severe lack of power under non-Gaussian distributions. We introduce signed-rank tests for $\mathcal{H}_0$. Contrary to all existing methods, our tests do not require any moment assumptions; they reach semiparametric efficiency bounds at correctly specified (possibly heterogeneous) m-tuples of densities, but their powers does not deteriorate in case densities are misspecified. In the homogeneous case, the normal-score version of our signed-rank tests uniformly dominate, in the Pitman sense, the optimal pseudo-Gaussian test. The results are obtained via a nonstandard application of Le Cam's LAN methodology in the context of curved statistical experiments.

Keywords: elliptical distributions, common principal components, rank-based tests.

Mots clés: lois elliptiques, composantes principales communes, tests basés sur les rangs.

Bibliographie

[1] Airoldi, J.P., and R.S. Hoffmann (1984). Age variation in voles (*Microtus californicus* and *Microtus ochrogaster*) and its significance for systematic studies. *Occasional Papers of the Museum of the Natural History*, University of Kansas, Lawrence, 111, 1-45.

[2] Boente, G., and L. Orellana (2001). A robust approach to common principal components. *In: Fernholz, L.T., Morgenthaler, S., Stahel, W. (Eds.), Statistics in Genetics and in the Environmental Sciences*. Birkhäuser Verlag AG, Basel, Switzerland, 117145

[3]Flury, B. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79, 892-898.

[4] Flury, B. (1986). Asymptotic theory for common principal components analysis. *Annals of Statistics*, 14, 418-430.

[5] Flury, B., and H. Riedwyl (1988). *Multivariate Statistics: a practical approach*. Chapman and Hall, New York.

[6] Hallin, M., and D. Paindaveine (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Annals of Statistics*, 34, 2707-2756.

[7] Hallin, M., and D. Paindaveine (2008a). A general method for constructing pseudo-Gaussian tests (with M. Hallin). *Journal of the Japan Statistical Society*, 38, 27-40.

[8] Hallin, M., D. Paindaveine, and T. Verdebout (2009a). Pseudo-Gaussian tests for common principal components. Submitted.

[9] Hallin, M., D. Paindaveine, and T. Verdebout (2009b). Optimal rank-based testing for principal components. Manuscript in preparation.

[10] Le Cam, L., and G. L. Yang (2000). *Asymptotics in Statistics*, 2nd edition. Springer-Verlag, New York.

[11] Paindaveine, D. (2008). A canonical definition of shape. *Statistics and Probability Letters*, 78, 2240-2247.

[12]Shapiro, A., and M.W. Browne (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, 82, 1092-1097.

[13] Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60-62.