Les Cahiers du GERAD ISSN: 0711–2440

A Weighted Multivariate Sign Test for Cluster Correlated Data

D. Larocque, J. Nevalainen,
H. Oja

G–2005–74

September 2005

Les textes publiés dans la série des rapports de recherche HEC n'engagent que la responsabilité de leurs auteurs. La publication de ces rapports de recherche bénéficie d'une subvention du Fonds québécois de la recherche sur la nature et les technologies.

A Weighted Multivariate Sign Test for Cluster Correlated Data

Denis Larocque
GERAD and Department of Quantitative Methods
HEC Montréal
3000 chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7
denis.larocque@hec.ca

Jaakko Nevalainen
Department of Mathematics, Statistics and Philosophy
University of Tampere
Finland

Hannu Oja
Tampere School of Public Health
University of Tampere
Finland

September 2005

Les Cahiers du GERAD

G–2005–74

Abstract

We consider the multivariate location problem with cluster correlated data. A family of multivariate weighted sign tests are introduced for which observations from different clusters can receive different weights. Under weak assumptions, the test statistic is asymptotically distributed as a chi-squared random variable as the number of clusters goes to infinity. The asymptotic distribution of the test statistic is also given for a local alternative model under multivariate normality. Optimal weights maximizing Pitman asymptotic efficiency are provided. These weights depend on the cluster sizes and on the intracluster correlation. Several approaches for estimating these weights are presented. Using Pitman asymptotic efficiency, it is shown that appropriate weighting can increase substantially the efficiency compared to a test that gives the same weight to each cluster. A multivariate weighted t-test is also introduced. The finite sample performance of the weighted sign test is explored through a simulation study which shows that the proposed approach is very competitive.

Key Words: Multivariate location problem, Spatial sign test, Intraclass correlation, One-way random effect, Clustered observations, Affine-invariance.

Résumé

Dans cet article, nous proposons et étudions les propriétés d'une classe de tests du signe pondéré pour le problème de position multivarié avec des données corrélées en grappes.

Acknowledgments: The research work of Denis Larocque was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT). The research work of Jaakko Nevalainen and Hannu Oja was partially supported by research grants from Academy of Finland and by doctoral program SIIDA. Part of this work was done while Denis Larocque was visiting the University of Tampere. He wishes to thank Hannu Oja for inviting him.

1 Introduction

Analysts often face situations where the classical assumption of N independent observations is not reasonable. Multilaboratory studies, multiple measurements taken from the same individual or questionnaires put on several classes of students are examples where one cannot assume independence: outcomes from the same laboratory, individual or school class tend to be alike. It is rather well-known that if such within-cluster dependency is not carefully taken into account during the course of the analysis, p–values are likely to be too small thus inflating the Type I error rate of tests. A complete treatment of longitudinal and clustered data including numerous examples drawn from studies in the biomedical and health sciences can be found in Fitzmaurice, Laird and Ware (2004).

Extensions of nonparametric univariate tests to cluster correlated data have been proposed in the literature; Datta and Satten (2005), Rosner and Grove (1999) and Rosner, Glynn and Ting Lee (2003). Other approaches for treating clustered data include the use of generalized estimating equations (Williamson, Datta and Satten, 2003; Stoner and Leroux, 2002) and resampling methods (Hoffman, Sen and Weinberg, 2001).

In this paper our aim is to consider the multivariate location problem under minimal assumptions of the underlying model. Much less attention has been paid to this multivariate problem before. Recently, Larocque (2003) proposed an extension of the one-sample affine-invariant multivariate sign test to cluster correlated data. A parallel development for the estimation problem appears in Nevalainen, Larocque and Oja (2005). The purpose of this paper is to extend Larocque's test to a whole family of (cluster) weighted tests that keep the advantages of the original procedure which are affine-invariance, validity under very slight assumptions and computing ease. However, by incorporating cluster weights, the new tests can improve the power of the original test when cluster sizes are different.

The description and the asymptotic null distribution of the weighted sign test is given in Section 2. Optimal weights maximizing Pitman asymptotic efficiency under a general multivariate normal model are derived in Section 3. Different approaches to estimate the optimal weights are proposed in Section 4. In Section 5, a parallel development is made for a weighted multivariate t-test. Results of a simulation study are given in Section 6 followed by concluding remarks in Section 7. Technical details and some proofs are reported in the Appendix.

2 Description of the weighted sign statistic

Suppose that we have n clusters of respective size $m_1, \ldots, m_n$ for a total of $N = \sum_{j=1}^{n} m_j$ observations. Let $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})'$ be the p-vectors corresponding to the j^{th} observation of the i^{th} cluster. Furthermore, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ be a fixed location vector of interest. The multivariate normal one-way random effects is often used to treat cluster correlated data. This model can be written as

$$\mathbf{X}_{ij} = \boldsymbol{\mu} + \mathbf{a}_i + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i, \tag{1}$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are independent and identically distributed (iid) as multivariate normal random p-vectors with expectation $\mathbf{0}$ and positive semidefinite covariance matrix $\mathbf{\Sigma_a}$ and where the $\boldsymbol{\epsilon}_{ij}$'s are iid (and independent of the $\mathbf{a}_i$'s) as multivariate normal random p-vectors with expectation $\mathbf{0}$ and positive definite covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$. Clearly, for that model, $E[\mathbf{X}_{ij}] = \boldsymbol{\mu}$ and

$$Cov[\mathbf{X}_{ij}, \mathbf{X}_{kl}] = \begin{cases} \mathbf{\Sigma_a} + \mathbf{\Sigma}_{\boldsymbol{\epsilon}} & \text{if} \quad i = k, j = l \\ \mathbf{\Sigma_a} & \text{if} \quad i = k, j \neq l \\ \mathbf{0} & \text{if} \quad i \neq k \end{cases}$$

In this paper, we consider a more general multivariate location model for which (1) is a special case. This model can be written as

$$\mathbf{X}_{ij} = \boldsymbol{\mu} + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i, \tag{2}$$

where the $\boldsymbol{\epsilon}_{ij}$'s are angularly symmetric identically distributed continuous random p-vectors. By angularly symmetric, it is meant that, if $|| \cdot ||$ denote the Euclidean norm, $\boldsymbol{\epsilon}_{ij}/||\boldsymbol{\epsilon}_{ij}||$ and $-\boldsymbol{\epsilon}_{ij}/||\boldsymbol{\epsilon}_{ij}||$ are identically distributed. Assume also that $\boldsymbol{\epsilon}_{ij}$ and $\boldsymbol{\epsilon}_{kl}$ are independent if $i \neq k$ and possibly dependent if $i = k$. That is, the observations are possibly correlated within clusters. Furthermore, assume that $(\boldsymbol{\epsilon}_{ij_i}, \boldsymbol{\epsilon}_{ij'_i})$ and $(\boldsymbol{\epsilon}_{kj_k}, \boldsymbol{\epsilon}_{kj'_k})$ are identically distributed for any $i, k = 1, \ldots, n$ and where j_i, j'_i and j_k, j'_k are indices chosen in $\{1, \ldots, m_i\}$ and $\{1, \ldots, m_k\}$ respectively. Note that this assumption on bivariate distributions is weaker than the regular assumption that the $\boldsymbol{\epsilon}_{ij}$'s are exchangeable within clusters. Note also that the existence of the first two moments is not required in model (2) nor is the assumption that the errors are symmetrically distributed (only angular symmetry is needed).

Assuming model (2), we wish to confront the hypotheses

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{and} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \tag{3}$$

An affine-invariant multivariate sign test is introduced in Larocque (2003) where affine-invariance is achieved by using the so-called "Tyler's transformation matrix"; Tyler (1987). However, the symmetrized version of this transformation, proposed by Dümbgen (1998), will be used in this paper. The reason is that, unlike Tyler's transformation, Dümbgen's transformation does not need a separate location estimate to be valid under the alternative as Tyler's transformation do. Moreover, Dümbgen's shape matrix will also be used to compute canonical correlations in order to estimate cluster weights as explained in Section 4. The Appendix contains details about these transformations.

Let $\{j_1, \ldots, j_n\}$ be a vector consisting of n indices (one for each cluster) such that $j_k \in \{1, \ldots, m_k\}$, $k = 1, \ldots, n$. Define $\hat{\mathbf{A}}_D$ to be the Dümbgen's transformation matrix obtained from $\mathbf{X}_{1j_1}, \mathbf{X}_{2j_2}, \ldots, \mathbf{X}_{nj_n}$. In practice, the vector of indices $\{j_1, \ldots, j_n\}$ may be chosen at random among all such vectors which is the same as choosing one index at random for each cluster.

Define the transformed points to be

$$\mathbf{Y}_{ij} = \hat{\mathbf{A}}_D \mathbf{X}_{ij} \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i.$$

Moreover, define

$$\mathbf{U}_{ij} = \mathbf{Y}_{ij}/||\mathbf{Y}_{ij}|| \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i.$$

to be the "signs" (unit vectors) of the transformed observations that are sometimes called "standardized signs". Note that in the particular case of one-dimensional observations ($p = 1$), $\hat{\mathbf{A}}_D$ is equal to 1 by definition and $\mathbf{U}_{ij}$ is then $\mathrm{sign}(X_{ij})$.

The test statistic proposed in Larocque (2003) is

$$S_N = N\bar{\mathbf{U}}'\hat{\mathbf{\Sigma}}_{\mathbf{U}}^{-1}\bar{\mathbf{U}} \tag{4}$$

where

$$\bar{\mathbf{U}} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{U}_{ij} \tag{5}$$

is the overall unweighted average of the signs and where $\hat{\mathbf{\Sigma}}_{\mathbf{U}}$ is a consistent estimator of the asymptotic covariance matrix of $\sqrt{N}\bar{\mathbf{U}}$ under H_0 defined by

$$\hat{\mathbf{\Sigma}}_{\mathbf{U}} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{U}_i\mathbf{U}_i' \tag{6}$$

where

$$\mathbf{U}_i = \sum_{j=1}^{m_i}\mathbf{U}_{ij}.$$

To be precise, Larocque (2003) pre-transforms the data using Tyler's transformation matrix instead of $\hat{\mathbf{A}}_D$. The statistic S_N gives equal weight to each observation. Clearly, this would be the right way to weight the data points if they were independent within clusters since we would then be in the usual iid setting. But other weighting schemes might be preferable when cluster sizes are different and when observations are correlated within clusters. This possibility was mentioned in Larocque (2003) but was not pursued there.

We are now ready to define a generalized version of (4). Let $w_1, w_2, \ldots, w_n$ be a sequence of cluster weights satisfying $(1/N)\sum_{i=1}^{n} m_i w_i = 1$. The weighted multivariate sign test is a quadratic form based on the weighted average

$$\bar{\mathbf{U}}_w = \frac{1}{N}\sum_{i=1}^{n} w_i \sum_{j=1}^{m_i}\mathbf{U}_{ij}. \tag{7}$$

Let

$$\hat{\mathbf{\Sigma}}_{\mathbf{U}_w} = \frac{1}{N}\sum_{i=1}^{n} w_i^2 \mathbf{U}_i \mathbf{U}_i'. \tag{8}$$

The weighted multivariate sign test is defined by

$$S_w = N \bar{\mathbf{U}}_w' \hat{\mathbf{\Sigma}}_{\mathbf{U}_w}^{-1} \bar{\mathbf{U}}_w. \tag{9}$$

The choice $w_i \equiv 1$ gives equal weight to each individual observation and the resulting statistic is equivalent to the one defined by (4). At the other extreme, the choice $w_i = N/(nm_i)$ gives equal weight to each cluster. Since the original statistic S_N is affine-invariant, it is straightforward to see that S_w also has that property.

We conclude this section by giving the asymptotic (as the number of clusters goes to infinity) null distribution of S_w that can be used to apply the test in practice. In fact, this asymptotic distribution is used in the simulation study described in Section 6.

For the rest of the paper, we will assume that the two following limits exists and are finite.

$$c_{w1} = \lim_{n\to\infty}\left(\frac{1}{N}\sum_{i=1}^{n} m_i w_i^2\right) \quad \text{and} \quad c_{w2} = \lim_{n\to\infty}\left(\frac{1}{N}\sum_{i=1}^{n} m_i(m_i-1) w_i^2\right). \tag{10}$$

The next result is a direct generalization of Theorem 1 of Larocque (2003) and can be proven using similar arguments.

Theorem 2.1 *Assume model (2). Under H_0, as $n \to \infty$,*

$$S_w \xrightarrow{D} \chi_p^2.$$

3 Asymptotic efficiency and optimal weights

In this section, the asymptotic distribution of S_w is obtained under a local alternative multivariate normal model. Then, optimal weights are derived and the test using these optimal weights is compared to the unweighted sign test S_N using Pitman asymptotic efficiency.

Let $\boldsymbol{\mu}_N = \boldsymbol{\mu}/\sqrt{N}$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ $(\neq \mathbf{0})$ is a fixed p-vector. Without loss of generality, assume that $\boldsymbol{\mu}'\boldsymbol{\mu} = 1$. The sequence of local alternatives considered is

$$H_{1n} \quad : \quad \mathbf{X}_{ij} = \boldsymbol{\mu}_N + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i \tag{11}$$

where the $\boldsymbol{\epsilon}_{ij}$'s are multivariate normal random p-vectors with expectation $\mathbf{0}$ and positive definite covariance matrix $Var(\boldsymbol{\epsilon}_{ij}) = \mathbf{\Sigma}_{11}$. Moreover, observations from different clusters are independent and the covariance between two vectors from the same cluster $(j \neq j')$ is $Cov(\boldsymbol{\epsilon}_{ij}, \boldsymbol{\epsilon}_{ij'}) = \mathbf{\Sigma}_{12}$.

Consider the $2p$ vector $\mathbf{Z}$ formed by stacking any 2 vectors from the same cluster. Then the covariance matrix of $\mathbf{Z}$ is

$$\begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{11} \end{pmatrix}.$$

Let $\mathbf{V}$ and $\mathbf{\Lambda}$ be the orthogonal matrix and the diagonal matrix corresponding to the eigenvalue decomposition $\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1/2} = \mathbf{V\Lambda V}'$. The diagonal of $\mathbf{\Lambda}$ contains the canonical correlations $(\rho_1, \rho_2, \ldots, \rho_p)$ for two vectors from the same cluster. We will assume that they are placed in descending order $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p$.

Hence, if we define the standardized vectors $\mathbf{\Sigma}_{11}^{-1/2}\mathbf{X}_{ij} = \mathbf{X}_{ij}^* = (X_{ij1}^*, \ldots, X_{ijp}^*)'$, this model specifies that for $j \neq j'$, $Corr(X_{ijk}^*, X_{ij'k}^*) = \rho_k$, that is the correlation between the k^{th} components from two standardized observations from the same cluster is ρ_k. We will see shortly that the asymptotic distribution of the test statistic, under the sequence (11), depends on the covariance structure ($\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{12}$) only through the ρ_k's.

Define

$$F(\rho) = \frac{F(1/2, 1/2; p/2+1; \rho^2)}{F(1/2, 1/2; p/2+1; 1)} \tag{12}$$

where $F(a, b; c; d)$ is the hypergeometric function of Gauss; Abramowitz and Stegun (1970). Let

$$d = \sqrt{2}\frac{\Gamma((p+1)/2)}{\Gamma(p/2)}. \tag{13}$$

Note that $F(\rho)$ and d depend also on the dimension p but we use this simplified notation since no confusion is possible.

Theorem 3.1 *Under the sequence (11), as $n \to \infty$,*

$$S_w \xrightarrow{D} \chi_p^2(\delta_{S_w})$$

where

$$\delta_{S_w} = \frac{d^2}{p}\sum_{j=1}^{p}\frac{\mu_j^2}{c_{w1} + \rho_j F(\rho_j) c_{w2}}. \tag{14}$$

To maximize the power of the test, we must seek the weights that make the noncentrality parameter as large as possible. We call these the optimal weights for the normal model. But we clearly see from (14) that in general, the optimal weights will depend on the direction of the shift $\boldsymbol{\mu}$ which is unknown. There are many ways to get around that.

Firstly, assume that all the ρ's are the same, that is, $\rho_j = \rho$ for all j. Then the noncentrality parameter becomes

$$\delta_{S_w} = \boldsymbol{\mu}'\boldsymbol{\mu}\frac{d^2}{p}\frac{1}{c_{w1} + \rho F(\rho) c_{w2}} = \frac{d^2}{p}\frac{1}{c_{w1} + \rho F(\rho) c_{w2}}. \tag{15}$$

In that case, it is shown in the Appendix that the optimal weights, maximizing the noncentrality parameter, are

$$w_i^{(o1)} = \left(\frac{1}{N} \sum_{j=1}^{n} \frac{m_j}{1 + \rho F(\rho)(m_j - 1)} \right)^{-1} \frac{1}{1 + \rho F(\rho)(m_i - 1)}. \tag{16}$$

We thus see that the optimal weights are inversely proportional to the cluster sizes m_i and also decrease as the intracluster correlation ρ increases.

Secondly, we can try to maximize the power in the least favorable case. Recall that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p$. Thus $\rho_1 = \max(\rho_1, \rho_2, \ldots, \rho_p)$. Assume that $\rho_1 > \rho_2$, then ρ_1 is the unique maximum. The minimum value (over $\boldsymbol{\mu}$) of (14) is then attained when $\mu_1 = \pm 1$ and $\mu_j = 0$ for all other j. In that case, the noncentrality parameter becomes

$$\delta_{S_w} = \frac{d^2}{p} \frac{1}{c_{w1} + \rho_1 F(\rho_1) c_{w2}}$$

and the weights maximizing the power in this least favorable case are

$$w_i^{(o2)} = \left(\frac{1}{N} \sum_{j=1}^{n} \frac{m_j}{1 + \rho_1 F(\rho_1)(m_j - 1)} \right)^{-1} \frac{1}{1 + \rho_1 F(\rho_1)(m_i - 1)}. \tag{17}$$

In practice, unless we have a priori knowledge, the ρ's are unknown. Estimates are then needed to be able to use the weights (16) or (17). One possibility is to assume that all the ρ's are the same (or close to each other), estimate the common value and plug it in (16) to obtain the weights for the test statistic. Another possibility is to look for protection against the worst possible case and estimate $\max(\rho_1, \rho_2, \ldots, \rho_p)$. Then again, this estimate can be plugged into (17) to obtain the weights. But the important thing is that the estimate of the common ρ value (or the maximum ρ value) must be affine-invariant if we want the test statistic to remain affine-invariant. Note that the ρ's are indeed affine-invariant parameters. Hence, it is reasonable to use affine-invariant estimators of them. Possible estimators will be described in the next section.

Since the possibilities are endless when the ρ's are distinct, we will assume that $\rho_j = \rho$, for all j, for the rest of this section to study the behavior of the Pitman asymptotic efficiency. Scenarios with distinct ρ's will be explored through simulations in Section 6.

Define by $S_{w(o1)}$ and $S_{w(o2)}$ the sign tests that use the optimal weights (16) and (17) respectively.

By plugging the optimal weights (16) into (14), we obtain the noncentrality parameter with the optimal weights

$$\delta_{S_{w(o1)}} = \frac{d^2}{p} l_1 \tag{18}$$

where

$$l_1 = lim_{n\to\infty}\frac{1}{N}\sum_{i=1}^{n}\frac{m_i}{1+\rho F(\rho)(m_i-1)}.$$

On the other hand, if we give the same weight to each observation ($w_i \equiv 1$), then the noncentrality parameter of the unweighted sign test S_N is

$$\delta_{S_N} = \frac{d^2}{p}\frac{1}{(1+l\rho F(\rho))} \qquad (19)$$

where

$$l = lim_{n\to\infty}\frac{1}{N}\sum_{i=1}^{n} m_i(m_i-1);$$

see Larocque (2003). Since the Pitman asymptotic efficiency between two tests is simply the ratio of their noncentrality parameters, we have

$$\mathrm{ARE}(S_{w(o1)}, S_N) = l_1(1+l\rho F(\rho)).$$

In order to get an idea of the behavior of this ARE, assume that we have R different clusters sizes $m_1, m_2, \ldots, m_R$. Moreover, assume that, asymptotically, a proportion α_1 of the clusters are of size m_1, a proportion α_2 are of size m_2 and so on. Thus $\alpha_1+\alpha_2+\cdots\alpha_R = 1$. Then

$$\mathrm{ARE}(S_{w(o1)}, S_N) =$$

$$\sum_{r=1}^{R}\left(\frac{\alpha_r m_r}{1+\rho F(\rho)(m_r-1)}\right)\frac{1}{\sum_{r=1}^{R}\alpha_r m_r}\left(1+\left(\frac{\sum_{r=1}^{R}\alpha_r m_r(m_r-1)}{\sum_{r=1}^{R}\alpha_r m_r}\right)\rho F(\rho)\right).$$

Figure 1 presents the value of this ARE as a function of ρ for different configurations of cluster sizes $(m_1, \ldots, m_R)$ and respective proportions $(\alpha_1, \ldots \alpha_R)$ and for dimensions p =1, 3 and 10. Obviously, we can see that, as soon as $\rho > 0$, the weighted sign test with optimal weights is always more efficient then the unweighted sign test. The higher ρ is, the better $S_{w(o1)}$ is compared to S_N. Moreover, the higher the dimension is, the higher the ARE gets. Possible cluster sizes range between 1 and 5 in the first column and between 1 and 10 in the second column. The cluster size comes a binomial distribution in row 1, a uniform distribution in row 2 and an "extreme" distribution in row 3 where only the lowest and highest cluster size can appear each with probability 0.5. We can clearly see that the more dispersed the distribution of the cluster is, the higher the ARE is. This is not surprising because when all clusters are of the same size, then the two tests are equivalent.

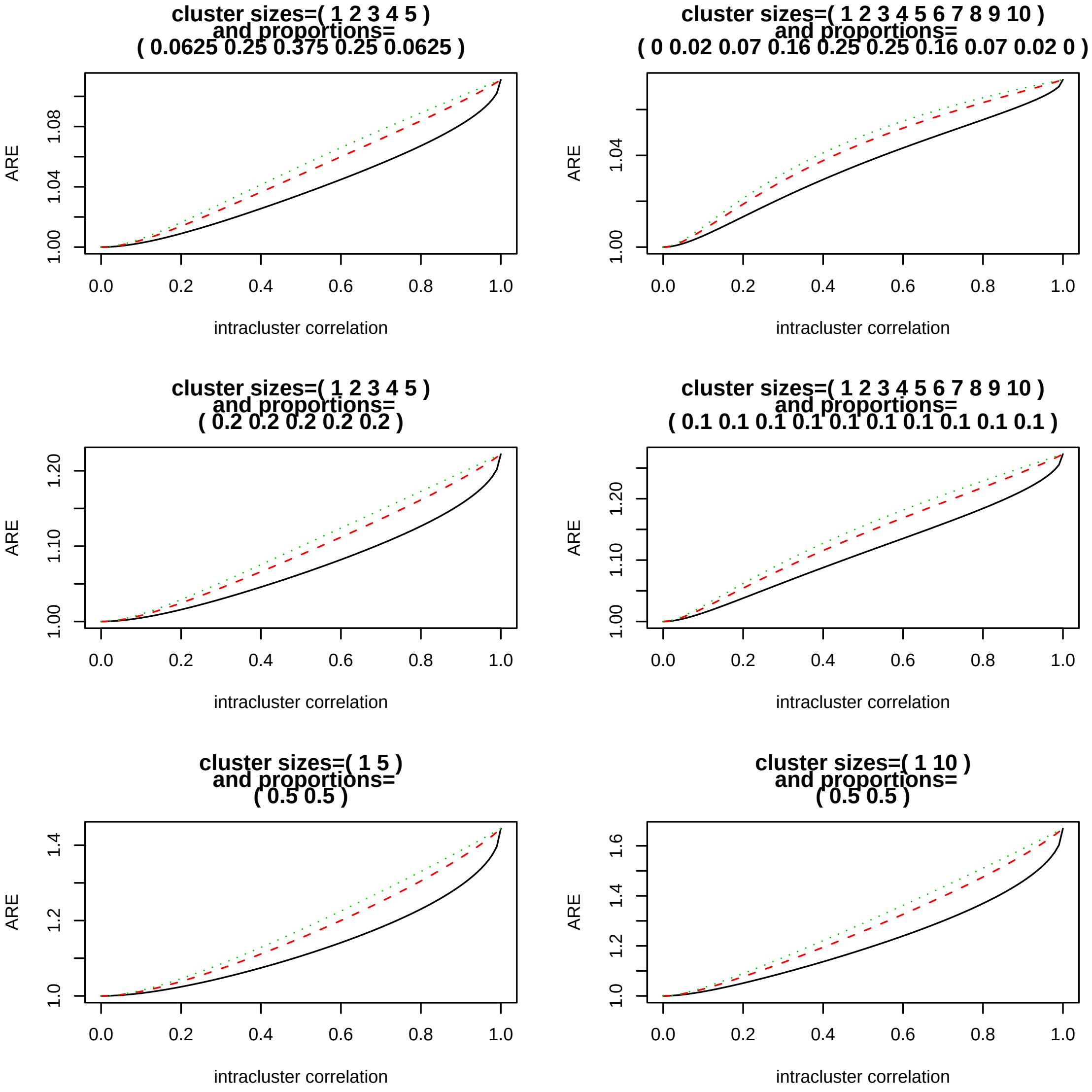

Figure 1: Asymptotic efficiency of the optimal weighted sign test $S_{w(o1)}$ relative to the unweighted sign test S_N. The solid line corresponds to $p = 1$, the dashed line to $p = 3$ and the dotted line to $p = 10$.

4 Estimating the weights

From the last section, we have that estimates of the ρ's are needed to compute either the weights (16) or (17). Three approaches are described in this section. The first two approaches are general and the third one assumes that the ρ's are all the same.

The first two approaches consist in finding the sample canonical correlations for two vectors from the same cluster using either a regular sample covariance matrix or a more robust shape matrix.

Let's begin with the simplest one that uses a sample covariance matrix. Let n_0 $(\leq n)$ be the number of clusters with more than one observation. For simplicity and without loss of generality, assume that these are the first n_0 clusters. For each of these clusters, choose 2 distinct observations at random. Let $\mathbf{X}_{ij}$ and $\mathbf{X}_{ij'}$ be the observations selected for cluster i, $i = 1, \ldots, n_0$. Let $\mathbf{Z}_i = (\mathbf{X}'_{ij}, \mathbf{X}'_{ij'})'$ be the $2p$ vector obtained by stacking the two observations for cluster i. We will compute the sample canonical correlations between the $\mathbf{X}_{ij}$'s and the $\mathbf{X}_{ij'}$'s. Let

$$\begin{pmatrix} \hat{\mathbf{\Sigma}}_{11} & \hat{\mathbf{\Sigma}}_{12} \\ \hat{\mathbf{\Sigma}}_{21} & \hat{\mathbf{\Sigma}}_{22} \end{pmatrix}$$

be the sample covariance matrix of $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_{n_0}$. Let $\hat{\mathbf{V}}$, $\hat{\mathbf{W}}$ and $\hat{\mathbf{\Lambda}}$ be the orthogonal matrices and the diagonal matrix corresponding to the decomposition $\hat{\mathbf{\Sigma}}_{11}^{-1/2}\hat{\mathbf{\Sigma}}_{12}\hat{\mathbf{\Sigma}}_{22}^{-1/2} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{W}}'$. The diagonal of $\hat{\mathbf{\Lambda}}$ contains the sample canonical correlations $(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_p)$ in descending order.

By using only two observations per cluster, we are not using all data points if some clusters are larger than two. We are thus proposing to repeat the steps above B times to obtain B estimates and then take their average. In what follows, $(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_p)$ will refer to this average. In the simulation study of Section 6, the canonical correlations are estimated with $B = 30$ repetitions.

Once we have our final estimate of the ρ's, there are different ways to utilize them. Firstly, we could take some sort of average of them to obtain a single value of ρ and then use the weights (16). If all the $\hat{\rho}$'s are strictly positive, we could take the geometric average

$$\hat{\rho} = \left(\prod_{j=1}^{n_0} \hat{\rho}_j \right)^{1/p} = (\det(\hat{\mathbf{\Lambda}}))^{1/p}.$$

The usual arithmetic average could also be used. Secondly, if we want protection against the worst possible case, we could take the estimate of the maximum, $\hat{\rho}_1$, and use the weights (17).

The second general approach to estimate the ρ's consist of replacing the regular covariance matrix above by a more robust estimate. We propose using Dümbgen's shape matrix described in the Appendix. Note that if the $\mathbf{Z}_i$'s are from an elliptical population, then using this shape (or any other) estimates the same population parameters as using the regular covariance matrix; Taskinen, Croux, Kankainen, Ollila and Oja (2005).

The third and last approach starts by assuming that all the ρ's are the same and is derived through the MANOVA table (see Ebel (1951) for the univariate case). Define

$$\bar{\mathbf{X}}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{X}_{ij} \quad \text{and} \quad \bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{X}_{ij}$$

as the observations average for cluster i and the overall average. The between clusters MS is then

$$\text{MSC} = \frac{1}{n-1}\sum_{i=1}^{n} m_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})'.$$

The within clusters MS is

$$\text{MSE} = \frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)'.$$

Let

$$c = \frac{N - \sum_{i=1}^{n} m_i^2/N}{n-1}.$$

The proposed estimator of ρ is defined through generalized variances and is given by

$$\hat{\rho} = \frac{|\text{MSC} - \text{MSE}|^{(1/p)}}{|\text{MSC} + (c-1)\text{MSE}|^{(1/p)}}.$$

This last approach has the advantage of being easy to implement and it uses directly all the observations.

It is straightforward to see that all those estimates of the ρ's are affine-invariant. Consequently, the statistic S_w that uses weights with those estimated ρ's remains affine-invariant.

5 A weighted multivariate t-test

A parallel development can easily be made for a multivariate test based on the weighted average of the observations; a weighted multivariate t-test. Let

$$\bar{\mathbf{X}}_w = \frac{1}{N}\sum_{i=1}^{n} w_i \sum_{j=1}^{m_i}\mathbf{X}_{ij} \tag{20}$$

denote a weighted average of the original data points. Let

$$\hat{\mathbf{\Sigma}}_{\mathbf{X}_w} = \frac{1}{N}\sum_{i=1}^{n} w_i^2 \mathbf{X}_i \mathbf{X}_i' \tag{21}$$

where $\mathbf{X}_i = \sum_{j=1}^{m_i} \mathbf{X}_{ij}$. The weighted multivariate t-test is defined by

$$T_w = N\bar{\mathbf{X}}_w' \hat{\mathbf{\Sigma}}_{\mathbf{X}_w}^{-1} \bar{\mathbf{X}}_w. \tag{22}$$

It is straightforward to show that under model (2), if $\boldsymbol{\mu} = E(\mathbf{X}_{ij})$, if second moments exists and if (10) holds, then under H_0, $T_w \xrightarrow{D} \chi_p^2$ as $n \to \infty$. Moreover, under the local model (11), the noncentrality parameter of the asymptotic χ_p^2 distribution of T_w is

$$\delta_{T_w} = \sum_{j=1}^{p} \frac{\mu_j^2}{c_{w1} + \rho_j c_{w2}}.$$

This noncentrality is very similar to the one of the sign test (14). As for the sign test, the optimal weights may depend on the direction of the shift. But if we assume that all the ρ's are the same, that is, $\rho_j = \rho$ for all j, then it becomes

$$\delta_{T_w} = \frac{1}{c_{w1} + \rho c_{w2}}.$$

If we maximize this function with respect to the weights, we find that the optimal weights are given by

$$w_i^{(t1)} = \left(\frac{1}{N}\sum_{j=1}^{n} \frac{m_j}{1+\rho(m_j-1)}\right)^{-1} \frac{1}{1+\rho(m_i-1)}.$$

We will call the t-test that uses these optimal weight $T_{w(o1)}$.

6 Simulation Study

In this section we will mainly compare the performance of the unweighted sign test S_N, the optimal sign test $S_{w(o1)}$ and the optimal t-test $T_{w(o1)}$in the case of finite samples. All methods described in Section 4 for estimating the weights were tried. Moreover the sign and t-test using the weights (17) defined through the maximum of the ρ's were also included and will be discussed briefly. Only a relevant subset of the results are presented and discussed here but the complete simulation results are available upon request.

Data points were generated using the model

$$\mathbf{X}_{ij} = \boldsymbol{\mu} + \mathbf{a}_i + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, n \quad ; \quad j = 1, \ldots, m_i$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are iid random p-vectors with expectation $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{a}} = \rho\mathbf{I}_p$ and where the $\boldsymbol{\epsilon}_{ij}$'s are iid (and independent of the $\mathbf{a}_i$'s) random p-vectors with expectation $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = (1-\rho)\mathbf{I}_p$.

- Two dimensions are used: p=1 and 3.

- Two distributions are used:
  1. Normal where both the $\mathbf{a}_i$'s and the $\boldsymbol{\epsilon}_{ij}$ are distributed as normal random vectors.
  2. $t_3 - t_3$ where both the $\mathbf{a}_i$'s and the $\boldsymbol{\epsilon}_{ij}$ have the multivariate t distribution with 3 degrees of freedom.
- Ten values of the intraclass correlation ρ are used: 0; .1; .2; .3; .4; .5; .6; .7; .8 and .9.
- Three design are used, each of them involving 60 clusters and 330 observations:
  1. Binomial for which we have 1 cluster of size 2, 4 clusters of size 3, 10 clusters of size 4, 15 clusters of size 5, 15 clusters of size 6, 10 clusters of size 7, 4 clusters of size 8 and 1 cluster of size 9.
  2. Uniform for which we have 6 clusters of size 1, 6 clusters of size 2, . . . , 6 clusters of size 9 and 6 clusters of size 10.
  3. Extreme for which we have 30 clusters of size 1 and 30 clusters of size 10.

The vector $\boldsymbol{\mu}$ was either set to $\mathbf{0}$ (null hypothesis) of to a value of the form c when p=1 or (c, c, c) when p=3 such that the power of the unweighted sign test was between .25 and .3. Note that (when $p = 3$) all the ρ's are identical under this model. Situations with different ρ's will be discussed later in this section.

All tests were performed at the 5% level. The proportions of rejection was calculated with 10000 replications using the critical point from the χ^2_p distribution.

For both the sign test and the t-test, the value of ρ was estimated using the three methods described in Section 4. That is, we have three different version of the optimal sign test and also three of the optimal t-test. But in the end, the results were very similar whatever estimation method was used. Consequently, we present only the results for the versions of the test that are most natural. The natural way of estimating ρ for the weighted sign test is by taking the arithmetic average of the canonical correlations obtained with Dümbgen's shape matrix. The natural way of estimating ρ for the weighted t-test is by taking the arithmetic average of the canonical correlations obtained with the regular covariance matrix.

First, let's have a look at the observed levels. Overall, considering all tests and all configurations, 1488 observed levels were obtained and each of them was estimated with 10000 replications. The minimum value obtained is 0.028, the maximum is 0.055, the average and median values are 0.046. Moreover, the middle 95% of the observed levels are comprised in the interval $[0.038, 0.052]$ and the middle 90% are in $[0.041, 0.051]$. Consequently, all tests maintained reasonably their prescribed level of 5% with perhaps a very slight tendency towards being conservative.

We can now move to the power comparisons. The results are presented in Figures 2 and 3 for $p = 1$ and 3 respectively. The benchmark is the unweighted sign test (the full straight line in all plots) and results are given in terms of difference in power (in %) compared to the benchmark. Results for the normal distribution are given in the first column of the figures and the second column contains the $t_3 - t_3$ distribution. The binomial, uniform and extreme designs appear in rows 1, 2 and 3.

Comparing firstly the two sign tests, we see that the optimal sign test is always more powerful except in a few cases when $\rho = 0$. The gain in power can be higher than 60% in some cases. The improvement is very similar for both the normal and $t_3 - t_3$ distributions. Even though the optimal weights are derived under a normal model, we see that their use are also beneficial for the $t_3 - t_3$ distribution. Also, the results coming from the ARE analysis are confirmed. That is, the more dispersed the clusters are, the better is $S_{w(o1)}$ compared to S_N. The cluster are the most dispersed for the extreme design (last rows), followed by the uniform design and then by the binomial design.

Comparing the optimal sign test and the optimal t-test, we see that the t-test is more powerful for the normal distribution and the opposite is true for the $t_3 - t_3$ distribution. Moreover, the difference in power between the two tests seems to be quite stable as ρ varies (the two curves are almost parallel in each plot). On one hand, the difference between the two tests seems to be constant for the $t_3 - t_3$ distribution when we move from $p = 1$ to $p = 3$. On the other hand, the difference between the two tests is smaller when $p = 3$ compared to when $p = 1$ for normal data. This fact was also noted in Larocque (2003) when comparing the unweighted versions of the sign and t-tests.

Even though these results are not reported here, we can mention that the power of the tests using the weights (17) by estimating the maximum of the ρ's is sometimes very similar and sometimes slightly less than the power of the corresponding test that uses the weights (16) by estimating the average of the ρ's. But this is not surprising since all the ρ's are equal in the cases considered so far. The real potential value of estimating the maximum of the ρ's is when they are in fact different and when the shift is in the direction of the component with the highest ρ. This is why we included such configurations in the simulation when $p = 3$. More precisely, we used the same two distributions and the same three designs. But this time, instead of having canonical correlations of the form $(\rho_1, \rho_2, \rho_3) = (\rho, \rho, \rho)$ for ρ ranging between 0 and 0.9 as in the scenarios reported so far, we used $(\rho_1, \rho_2, \rho_3) = (0.2, 0.5, 0.8)$. Moreover, we used four different directions for the shifts under H_1. Namely, the shifts are of the form (c, c, c), $(c, 0, 0)$, $(0, c, 0)$ and $(0, 0, c)$. A test using an estimate of the maximum of the ρ's should be at its best when the shift is in the direction $(0, 0, c)$. Some results for the weighted sign test are reported in Table 1. It gives the power of the optimal sign test that uses the weights (16) by estimating ρ with the arithmetic average of the sample canonical correlations obtained with Dümbgen's shape matrix and the power of the optimal sign test that uses the weights (17) by estimating the maximum of the ρ's with the maximum of the sample canonical correlations obtained with Dümbgen's shape matrix.

We can observe that the two tests have very similar powers most of the time. For shifts of the form (c, c, c) and $(c, 0, 0)$, $S_{w(o1)}$ is slightly better for the extreme design. The only

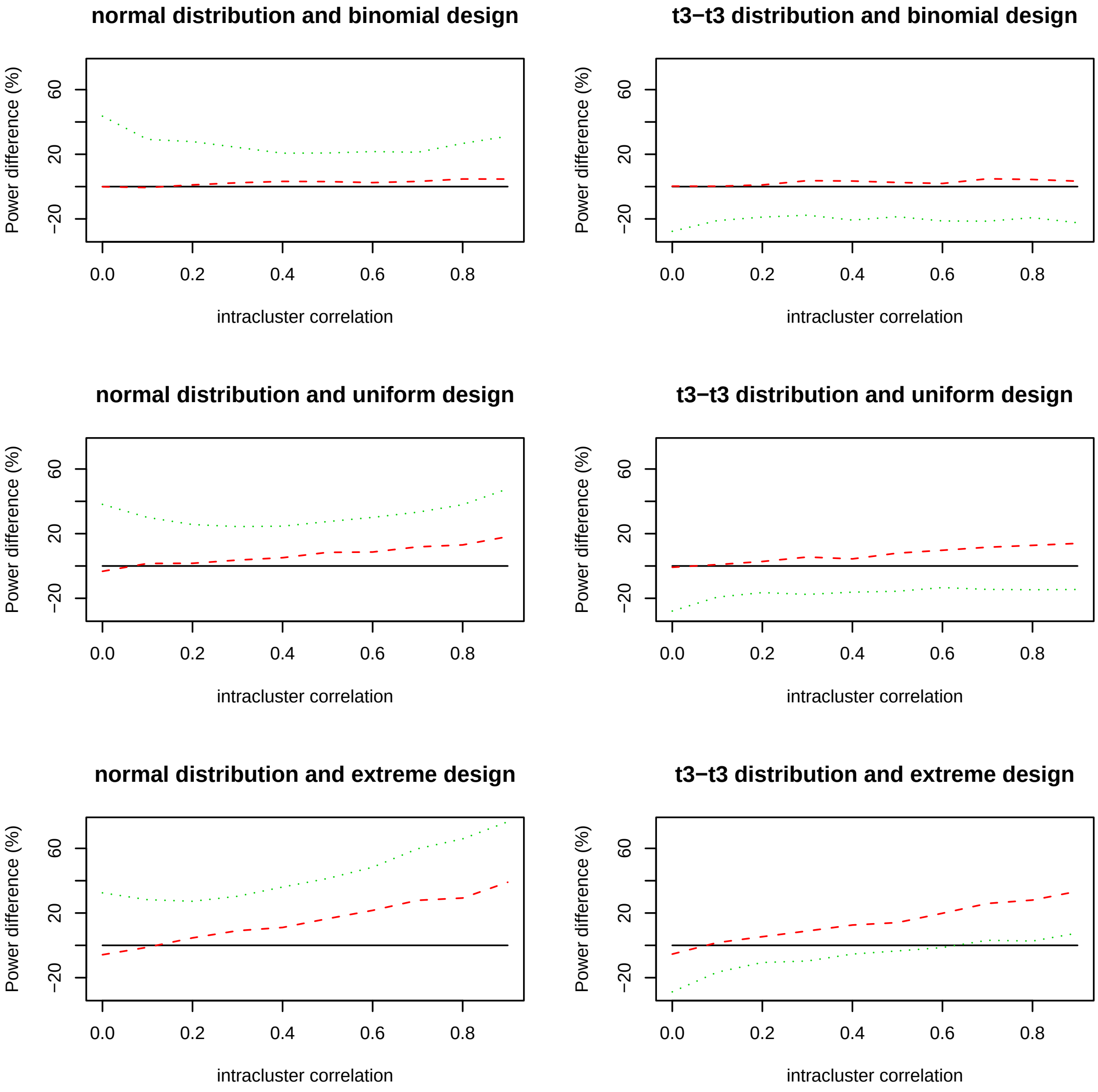

Figure 2: Percent difference in power between the optimal weighted sign test $S_{w(o1)}$ (dashed line) and optimal weighted t-test $T_{w(o1)}$ (dotted line) compared to the unweighted sign test S_N (full line) for $p = 1$.

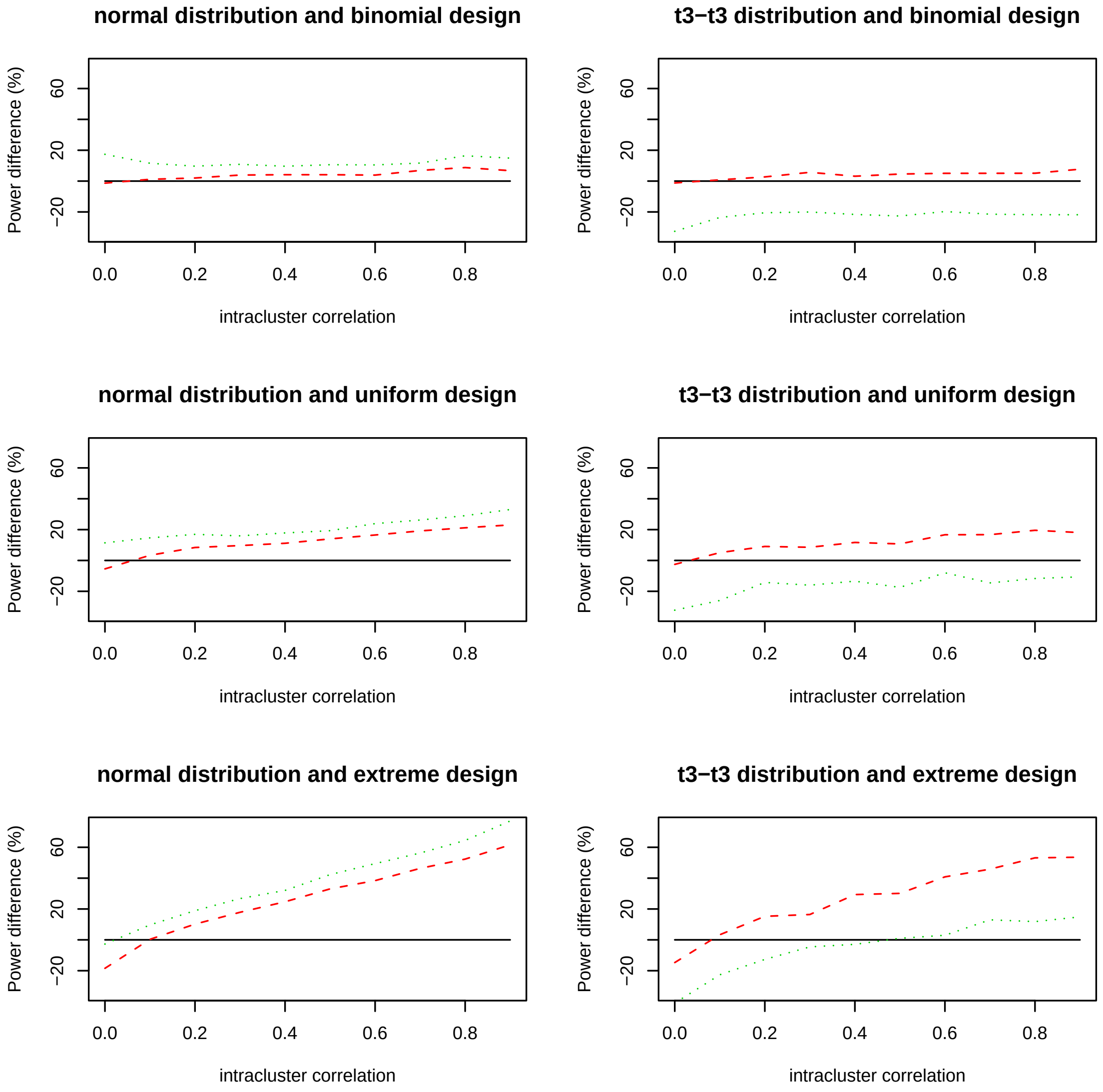

Figure 3: Percent difference in power between the optimal weighted sign test $S_{w(o1)}$ (dashed line) and optimal weighted t-test $T_{w(o1)}$ (dotted line) compared to the unweighted sign test S_N (full line) for $p = 3$.

Table 1: Observed powers when $(\rho_1, \rho_2, \rho_3) = (0.2, 0.5, 0.8)$

			Direction of shift			
Test	Design	Distribution	(c,c,c)	$(c,0,0)$	$(0,c,0)$	$(0,0,c)$
$S_{w(o1)}$	binomial	normal	0.309	0.284	0.306	0.297
$S_{w(o2)}$	binomial	normal	0.306	0.285	0.307	0.297
$S_{w(o1)}$	binomial	$t_3 - t_3$	0.282	0.334	0.274	0.328
$S_{w(o2)}$	binomial	$t_3 - t_3$	0.281	0.331	0.274	0.327
$S_{w(o1)}$	uniform	normal	0.351	0.293	0.292	0.380
$S_{w(o2)}$	uniform	normal	0.348	0.279	0.291	0.386
$S_{w(o1)}$	uniform	$t_3 - t_3$	0.306	0.309	0.339	0.321
$S_{w(o2)}$	uniform	$t_3 - t_3$	0.308	0.300	0.342	0.325
$S_{w(o1)}$	extreme	normal	0.340	0.297	0.391	0.421
$S_{w(o2)}$	extreme	normal	0.318	0.257	0.384	0.441
$S_{w(o1)}$	extreme	$t_3 - t_3$	0.347	0.301	0.340	0.412
$S_{w(o2)}$	extreme	$t_3 - t_3$	0.324	0.273	0.331	0.425

time that $S_{w(o1)}$ seems less powerful than $S_{w(o2)}$ if when the shift is of the form $(0, 0, c)$ for the extreme design and normal distribution where the observed powers are 0.421 and 0.441 respectively. All things being considered, it seems preferable to estimate the average of the ρ's instead of the maximum of the ρ's.

7 Concluding remarks

In this paper, we extended the multivariate affine-invariant sign test of Larocque (2003) to a whole family of weighted tests. The new tests keep all the advantages of the original test: affine-invariance, valid under very slight assumptions, easy to compute and implement. But by being able to incorporate cluster weights, the new family of tests can improve the power of the original test when cluster sizes are different. Optimal weights under a general multivariate normal model were obtained. Several ways of estimating the weights with the data were described. Following a parallel development, a multivariate weighted t-test was also proposed. The extent of how much more efficient the optimal sign test is compared to the sign test that gives the same weight to each cluster was explored by calculating Pitman efficiencies. We saw that the more dispersed the clusters are, the more efficient the optimal test is compared to the unweighted test. These results were confirmed in a simulation study. Moreover, the simulation showed that the optimal sign test is more powerful than the optimal t-test for the $t_3 - t_3$ distribution but the opposite is true for the normal distribution. But in that case, the difference between the two tests is smaller for $p = 3$ than for $p = 1$. The simulation also demonstrated that, at least in the case covered, it is better to use an estimate of the average of the ρ's instead of an estimate of the maximum of the ρ's. That's why our recommendation is to favor the optimal sign test by using the weights (16) by estimating ρ with the arithmetic average of the sample canonical correlations obtained with Dümbgen's shape matrix.

A Appendix

Tyler's and Dümbgen's transformations:

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n random p-vectors with location $\boldsymbol{\mu}$. Tyler (1987) shape matrix, $\hat{\mathbf{V}}_T$, is the positive definite symmetric matrix with trace$(\hat{\mathbf{V}}_T) = p$ such that, for any $\hat{\mathbf{A}}_T$ with $\hat{\mathbf{A}}_T' \hat{\mathbf{A}}_T = \hat{\mathbf{V}}_T^{-1}$,

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{\mathbf{A}}_T(\mathbf{X}_i - \boldsymbol{\mu})}{||\hat{\mathbf{A}}_T(\mathbf{X}_i - \boldsymbol{\mu})||} \right) \left(\frac{\hat{\mathbf{A}}_T(\mathbf{X}_i - \boldsymbol{\mu})}{||\hat{\mathbf{A}}_T(\mathbf{X}_i - \boldsymbol{\mu})||} \right)' = \frac{1}{p} \mathbf{I}_p.$$

This matrix exists and is unique when $n > p(p-1)$. $\hat{\mathbf{A}}_T$ is called Tyler's transformation matrix. For a given $\boldsymbol{\mu}$, a simple and fast algorithm to compute those matrices is given in Oja and Randles (2004). Tyler's transformation was first used in Randles (2000) to construct an affine-invariant version of the spatial sign test. In Randles (2000) and in Larocque (2003), the matrix $\mathbf{A}_T$ is computed by setting $\boldsymbol{\mu}$ to its value under the null hypothesis, that is $\mathbf{0}$. But we see that in general, the computation of $\mathbf{A}_T$ necessitates a separate location estimate.

A symmetrized version of Tyler's matrix that does not need a separate location estimator was proposed in Dümbgen (1998). Dümbgen's shape matrix, $\hat{\mathbf{V}}_D$, is the positive definite symmetric matrix with trace$(\hat{\mathbf{V}}_D) = p$ such that, for any $\hat{\mathbf{A}}_D$ with $\hat{\mathbf{A}}_D' \hat{\mathbf{A}}_D = \hat{\mathbf{V}}_D^{-1}$,

$$\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{\hat{\mathbf{A}}_D(\mathbf{X}_i - \mathbf{X}_j)}{||\hat{\mathbf{A}}_D(\mathbf{X}_i - \mathbf{X}_j)||} \right) \left(\frac{\hat{\mathbf{A}}_D(\mathbf{X}_i - \mathbf{X}_j)}{||\hat{\mathbf{A}}_D(\mathbf{X}_i - \mathbf{X}_j)||} \right)' = \frac{1}{p} \mathbf{I}_p.$$

Note that the algorithm to compute Tyler's matrices can be used to compute Dümbgen's matrices by replacing the $(\mathbf{X}_i - \boldsymbol{\mu})$'s by the differences $(\mathbf{X}_i - \mathbf{X}_j)$'s.

Shape matrices can be viewed as standardized scatter matrices that retain information about the form of the distribution but not about the scale. See Taskinen et al. (2005) for more details.

Proof of Theorem 3.1:

The proof is based upon LeCam's lemmas on contiguity.

Let

$$\boldsymbol{\Lambda}_1 = \text{diag}(\rho_1 F(\rho_1), \rho_2 F(\rho_2) \ldots, \rho_p F(\rho_1))$$

where F is defined by (12).

We begin with a lemma that will be used in the proof of Theorem 3.1. It is a slight generalization of lemma 1 of Larocque (2003) and can be proven using similar arguments.

Lemma A.1 *Let $(\mathbf{X}', \mathbf{Y}')'$ be distributed as a multivariate normal random $(p \times 2)$-vector and suppose that each of $\mathbf{X}$ and $\mathbf{Y}$ are distributed as a multivariate normal random p-vector with expectation $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$. Further assume that $Cov[\mathbf{X}, \mathbf{Y}] = \mathbf{\Lambda}$ where $\mathbf{\Lambda} = diag(\rho_1, \dots, \rho_p)$ with $-1 < \rho_i < 1$ for all i. Then*

$$E\left[\mathbf{X}\frac{\mathbf{X}'}{||\mathbf{X}||}\right] = \frac{d}{p}\mathbf{I}_p \quad , \quad E\left[\mathbf{X}\frac{\mathbf{Y}'}{||\mathbf{Y}||}\right] = \frac{d}{p}\mathbf{\Lambda} \quad \text{and} \quad E\left[\frac{\mathbf{X}}{||\mathbf{X}||}\frac{\mathbf{Y}'}{||\mathbf{Y}||}\right] = \frac{1}{p}\mathbf{\Lambda}_1$$

where d is defined by (13).

Since the test statistic is affine-invariant, we can assume without loss of generality, after multiplying each data points by $\mathbf{V}'\mathbf{\Sigma}_{11}^{-1/2}$, that $\mathbf{\Sigma}_{11} = \mathbf{I}_p$ and $\mathbf{\Sigma}_{12} = \mathbf{\Lambda}$.

Define $\mathbf{\Sigma}_{\rho,m} = \rho\mathbf{J}_m + (1-\rho)\mathbf{I}_m$ where $\mathbf{J}_m$ and $\mathbf{I}_m$ are the $m \times m$ matrices of ones and the identity matrix respectively. Then

$$\mathbf{\Sigma}_{\rho,m}^{-1} = \frac{1}{1-\rho}\left(\mathbf{I}_m - \frac{\rho}{1+(m-1)\rho}\mathbf{J}_m\right). \tag{23}$$

Let $\mathbf{Z}_i^*$ be the pm_i vector formed by stacking up all the observations in cluster i in such a way that all the first components appear first, then the second components and so on. That is $\mathbf{Z}_i^* = (X_{i11}, X_{i21}, \dots, X_{im_i1}, X_{i112}, X_{i22}, \dots, X_{im_i2}, \dots, X_{i1p}, X_{i2p}, \dots, X_{im_ip})'$. The covariance matrix of $\mathbf{Z}_i^*$ is $\mathbf{\Sigma}_i = \text{diag}(\mathbf{\Sigma}_{\rho_1,m_i}, \mathbf{\Sigma}_{\rho_2,m_i}, \dots, \mathbf{\Sigma}_{\rho_p,m_i})$. Let $\boldsymbol{\mu}_m^* = (\mu_1, \dots, \mu_1, \mu_2, \dots, \mu_2, \mu_p, \dots, \mu_p)'$ be the pm vector where each μ is repeated m times. Let

$$Z_{ijk} = \frac{X_{ijk}}{1+(m_i-1)\rho_k}$$

and $\mathbf{Z}_{ij} = (Z_{ij1}, \dots, Z_{ijp})'$. Following tedious but straightforward calculations using (23), we have that

$$\boldsymbol{\mu}_{m_i}^{*'}\mathbf{\Sigma}_i^{-1}\mathbf{Z}_i^* = \boldsymbol{\mu}'\mathbf{Z}_i$$

where $\mathbf{Z}_i = \sum_{j=1}^{m_i} \mathbf{Z}_{ij}$.

Define $\mathbf{C}_{i1} = \text{diag}(\frac{1}{(1+(m_i-1)\rho_1)}, \dots, \frac{1}{(1+(m_i-1)\rho_p)})$ and $\mathbf{C}_{i2} = \text{diag}(\frac{\rho_1}{(1+(m_i-1)\rho_1)}, \dots, \frac{\rho_p}{(1+(m_i-1)\rho_p)})$.

Let $Q = Q_n(\mathbf{Z}_1^*, \dots, \mathbf{Z}_n^*)$ denote the likelihood function under model (11). Let $P = P_n(\mathbf{Z}_1^*, \dots, \mathbf{Z}_n^*)$ denote the null likelihood function, that is, when $\boldsymbol{\mu} = \mathbf{0}$ in model (11). We have

$$\ln\left(\frac{Q}{P}\right) = \frac{1}{\sqrt{N}}\boldsymbol{\mu}'\sum_{i=1}^{n}\mathbf{Z}_i - \frac{1}{2}\frac{1}{N}\sum_{i=1}^{n} m_i\boldsymbol{\mu}'\mathbf{C}_{i1}\boldsymbol{\mu}.$$

Under H_0, $\hat{\mathbf{A}}_D$ converges in probability to $\mathbf{I}_p$ as $n \to \infty$. Hence, we have that, under H_0, the statistic S_w computed using $\mathbf{Y}_{ij} = \mathbf{X}_{ij}$ is asymptotically equivalent to S_w computed using $\mathbf{Y}_{ij} = \hat{\mathbf{A}}_D\mathbf{X}_{ij}$. Moreover, since the sequence (11) is contiguous to H_0, they are also asymptotically equivalent under that sequence. Thus, we can assume that $\mathbf{U}_{ij} = \mathbf{X}_{ij}/||\mathbf{X}_{ij}||\ \forall i, j$ for the rest of the proof.

Under H_0, $E(\sqrt{N}\bar{\mathbf{U}}_w) = \mathbf{0}$ and

$$V(\sqrt{N}\bar{\mathbf{U}}_w) = \frac{1}{N}\sum_i^n m_i w_i^2 \frac{\mathbf{I}_p}{p} + \frac{1}{N}\sum_i^n m_i(m_i-1)w_i^2\frac{\mathbf{\Lambda}_1}{p}$$

which converges to

$$\frac{1}{p}(\mathbf{I}_p c_{w1} + \mathbf{\Lambda}_1 c_{w2})$$

as n goes to ∞.

Next, using lemma A.1,

$$\begin{aligned} E_{H_0}\left[\ln\left(\frac{Q}{P}\right)\sqrt{N}\bar{\mathbf{U}}_w\right] &= \frac{\boldsymbol{\mu}'}{N}\sum_{i=1}^n\sum_{j=1}^{m_i}\sum_{k=1}^n\sum_{l=1}^{m_k} w_k E_{H_0}[\mathbf{Z}_{ij}\mathbf{U}'_{kl}] \\ &= \frac{\boldsymbol{\mu}'}{N}\frac{d}{p}\sum_{i=1}^n w_i m_i(\mathbf{C}_{i1} + (m_i-1)\mathbf{C}_{i2}) \\ &= \frac{\boldsymbol{\mu}'d}{p}. \end{aligned}$$

Hence, under the sequence (11),

$$\sqrt{N}\bar{\mathbf{U}}_w \xrightarrow{D} N_p\left(\boldsymbol{\mu}\frac{d}{p}, \frac{1}{p}(\mathbf{I}_p c_{w1} + \mathbf{\Lambda}_1 c_{w2})\right)$$

and Theorem 3.1 follows since

$$\left(\boldsymbol{\mu}\frac{d}{p}\right)'\left(\frac{1}{p}(\mathbf{I}_p c_{w1} + \mathbf{\Lambda}_1 c_{w2})\right)^{-1}\left(\boldsymbol{\mu}\frac{d}{p}\right) = \frac{d^2}{p}\sum_{j=1}^p \frac{\mu_j^2}{c_{w1} + \rho_j F(\rho_j)c_{w2}}.$$

Optimal weights:

It will be shown here that (16) are the optimal weights. For a finite n, the optimal weights are the solution to

$$\max_{w_1,\ldots,w_n}\left(\sum_{i=1}^n m_i w_i^2 + \rho\sum_{i=1}^n m_i(m_i-1)w_i^2\right)^{-1} \quad \text{subject to} \quad \sum_{i=1}^n m_i w_i = N.$$

This comes from the finite sample version of the noncentrality parameter (15). Setting $\mathbf{w} = (w_1, \ldots, w_n)'$, $\mathbf{m} = (m_1, \ldots, m_n)'$ and $\mathbf{A} = \mathrm{diag}(m_1(1+\rho(m_1-1)), \ldots, m_n(1+\rho(m_n-1)))$, this problem can be written as

$$\min_{\mathbf{w}}(\mathbf{w}'\mathbf{A}\mathbf{w}) \quad \text{subject to} \quad \mathbf{w}'\mathbf{m} = N.$$

This can be solved directly using the standard Lagrange multiplier rule and the solution is given by (16).

References

Abramowitz, M. and Stegun, I. A. (1970). *Handbook of Mathematical Functions*. Dover. New York.

Datta, S. and Satten, G. A. (2005). Rank-Sum Tests for Clustered Data. *Journal of the American Statistical Association* **100**, 908–915.

Dümbgen, L. (1998). On Tyler's M-Functional of Scatter in High Dimension. *Annals of the Institute of Statistical Mathematics* **50**, 471–491.

Ebel, R. L. (1951). Estimation of the Reliability of Ratings. *Psychometrika* **16**, 407–424.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-Cluster Resampling. *Biometrika* **88**, 1121–1134.

Larocque, D. (2003). An Affine-Invariant Multivariate Sign Test for Cluster Correlated Data. *Canadian Journal of Statistics* **31**, 437–455.

Nevalainen, J., Larocque, D. and Oja, H. (2005). On Multivariate Spatial Median for Clustered Data. Submitted.

Oja, H. and Randles, R. H. (2004). Multivariate Nonparametric Tests. *Statistical Science* **19**, 598–605.

Randles, R. H. (2000). A Simpler, Affine-Invariant Multivariate, Distribution-Free Sign Test. *Journal of the American Statistical Association*, **95**, 1263–1268.

Rosner, B. and Grove, D. (1999). Use of the Mann-Whitney U-Test for Clustered Data. *Statistics in Medicine* **18**, 1387–1400.

Rosner, B. Glynn, R. J. and Ting Lee, M.-L. (2003). Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large Sample Approach. *Biometrics* **59**, 1089–1098.

Stoner, J. A. and Leroux, B. G. (2002). Analysis of Clustered Data: A Combined Estimating Equations Approach. *Biometrika* **89**, 567–578.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E. and Oja, H. (2005). Canonical Analysis based on Scatter Matrices. To appear in *Journal of Multivariate Analysis*.

Tyler, D. E. (1987). A Distribution-Free M-Estimator of Multivariate Scatter. *Annals of Statistics*, **15**, 234–251.

Williamson, J., Datta, S., and Satten, G. A. (2003). Marginal Analyses of Clustered Data when Cluster Size is Informative. *Biometrics* **59**, 36–42.