Les Cahiers du GERAD ISSN: 0711–2440

Multivariate variance components tests for multilevel data

D. Larocque,
J. Nevalainen, H. Oja

G–2018–58

August 2018

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2018
– Bibliothèque et Archives Canada, 2018

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2018
– Library and Archives Canada, 2018

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Multivariate variance components tests for multilevel data

Denis Larocque [a]

Jaakko Nevalainen [b]

Hannu Oja [c]

[a] *GERAD & Department of Decision Sciences, HEC Montréal, Montréal (Québec), Canada, H3T 2A7*

[b] *School of Health Sciences, University of Tampere, 33014 Tampere, Finland*

[c] *University of Turku, Department of Mathematics and Statistics, 20014 Turku, Finland*

denis.larocque@hec.ca
jaakko.nevalainen@uta.fi
hannu.oja@utu.fi

August 2018
Les Cahiers du GERAD
G–2018–58

Abstract: We consider the multivariate linear model for multilevel data where units are nested within a hierarchy of clusters. We propose permutation procedures to test for variance components at any given level. The tests are moment based and require no distributional assumptions except finite second moments. We introduce the R package `mvctm`, which implements the tests. It can perform tests based on the original observations and score–based tests using ranks and signs. A simulation study shows that the new tests maintain the desired type I error. It also compares their power. The results suggest that the tests based on the original observations and the rank-based tests are very competitive. With univariate data, the former one is even more powerful than a likelihood ratio test based on a mixture of chi-squared distributions. The proposed tests are illustrated using the PISA data.

Keywords: Multilevel data, permutation test, multivariate sign, multivariate ranks, variance components, R

Acknowledgments: This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by HEC Montreal.

1 Introduction

Cluster correlated data (or just clustered data) are common in practice; Song (2007), Fitzmaurice et al. (2011). Typical examples are children nested within classrooms, patients nested within physicians, and family members nested within households. The classrooms, physicians and households are the clusters in these examples. Typically, the responses are assumed independent if they are in different clusters but possibly dependent if they are in the same cluster. It is also common to have more levels of clustering. The term multilevel (or hierarchical) data is usually used in this situation; Kreft and De Leeuw (1998), Raudenbush and Bryk (2001), Goldstein (2010), Scott et al. (2013). Clustered data can then be seen as the special case of 2-level data. Children, nested within classrooms, which are in turn nested within schools would be an example of 3-level data. The school is the level 1 cluster while the classroom is the level 2 cluster. Likewise, patients, nested within physicians (level 2 cluster), nested within clinics (level 1 cluster) would be another example.

In a regression setting, it is well-known that when intra-cluster correlation is present, it needs to carefully be taken into account in order to have valid inferences for the regression parameters; Fitzmaurice et al. (2011). Adding random intercepts (variance components) at the different cluster levels is one way to model the intra-cluster correlation. With a continuous response, this can be achieved with linear mixed models; Verbeke and Molenberghs (2009). However, in this paper, the focus is not on the regression part (or fixed effects part) of the model, but rather on the covariance part. Testing whether or not there is heterogeneity at a given level is very often of interest. For instance, in the children example, we may ask: is there heterogeneity among classrooms or heterogeneity among schools? This typically amounts to test for variance components, i.e., to test that the variance of a random effect is 0.

Variance components testing have been studied by many authors in various settings, e.g. Ofversten (1993), Stram and Lee (1994), Christensen (1996), Berkhof and Snijders (2001), Zhang and Lin (2008), (Nobre et al., 2013). However, in this paper we focus on permutation tests.

The principal reference for this work is Fitzmaurice et al. (2007). They proposed permutation tests for variance components in multilevel generalized linear mixed models, based on the likelihood ratio test (LRT). It is well-known that when we test that a variance parameter is 0 with the LRT in a linear mixed model, the null distribution of the test statistic is not the typical chi-squared but rather a mixture of chi-squared distributions; see Section 6.3.4 in Verbeke and Molenberghs (2009) and Section 7.5 in Fitzmaurice et al. (2011). This comes from the fact that the parameter lies on the boundary of the parameter space. For 2-level data, the basic idea of Fitzmaurice et al. (2007) is to i) compute the LRT statistic to test that the variance of the random effect is 0 with the original data, ii) randomly permute the cluster indices while holding fixed the number of observations within a cluster, and compute the same statistic with the permuted data, iii) repeat the permutation process a large number of times, and, iv) the permutation test p-value is the proportion of time that the LRT statistic of the permuted data is greater than or equal to the one of the original data. Obviously, this procedure can also be used for linear mixed models. The simulation study in Fitzmaurice et al. (2007), with 2-level data, shows that the permutation test is more powerful than the test based on the mixture of chi-squared distributions. In their concluding remarks, they discuss that the procedure can be used for data with more than two levels. However, no detailed description on how to proceed nor any empirical study are provided.

Samuh et al. (2012) proposed to perform the permutation test using a different statistic than the one used in Fitzmaurice et al. (2007). They work under the simple linear regression model with 2-level data. Their idea is to use the usual F-test statistic from ANOVA after removing the effect of the covariate. Their simulation study shows that the power of this procedure is similar to the one of the Fitzmaurice et al. (2007) test.

In Fitzmaurice et al. (2007) and Samuh et al. (2012), only a single random intercept is present in the model. Lee and Braun (2012) studied the more general linear mixed model, with multiple random effects. Hence in addition to a random intercept, random effects for some of the covariates can also be present. They proposed two permutation tests. The first one is based on the best linear unbiased predictions (BLUPs) and can be used to test any single random effect. The second one is based on the restricted LRT under normality and can be used to test multiple random effects.

Drikvandi et al. (2013) also considered the general linear mixed model with possibly multiple random effects. They propose a permutation test using a statistic based on the variance least square estimator. Their method can be used to test any subset of the variance components.

Zeng et al. (2015) propose a permutation test for multiple variance components in a generalized linear mixed models based on the LRT. They use the penalized quasi-likelihood algorithm to compute the LRT statistic. All random effects are assumed to have the same variance so their method is effectively testing that a single variance parameter is equal to 0. Their simulation study shows that their test has higher power than the score test and the tests based on mixtures under a wide range of scenarios.

In this paper, we are extending the Fitzmaurice et al. (2007) idea in different directions. Even though Fitzmaurice et al. (2007) discuss the possibility to extend their approach to data with more than two levels, all the methods presented above are developed and studied mainly for 2-level data. Moreover, they are limited to univariate responses. This papers aims to extend the basic idea in those two directions. We are proposing practical permutation tests for the multivariate multilevel linear model. The proposed procedures allows to test a single (possibly multivariate) random effect at any level in the hierarchy. The tests are readily available in the R (R Core Team, 2017) package `mvctm` (Larocque, 2017), available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=mvctm`.

The rest of the paper is organized as follows: Section Models, proposed tests and package `mvctm` describes the model, the proposed tests and the `mvctm` package, Section Simulation study gives the results from a simulation study, Section Data example provides a real data example and Section Discussion and concluding remarks concludes. Three appendices provide technical details, more details about the `mvctm` package and more details about the simulation study.

2 Models, proposed tests and package `mvctm`

2.1 Model and hypotheses

We are interested in the multivariate multilevel linear model. However, to fix ideas, we consider first the univariate 2-level model without covariates but only an intercept μ. The usual notation for this model is:

$$Y_{ij} = \mu + a_i + \epsilon_{ij}\, , i = 1, \dots, n_1, j = 1, \dots, m_i,$$

where Y_{ij} is the response variable for observation j in cluster i, a_i is the random effect of cluster i and the ϵ's are individual random errors. However, to avoid multiple indices and facilitate the description of the test statistic, we will use a different notation. Moreover, since we are working in the general multivariate case, we will also use the notation $A \geq (>)$ to indicate that a matrix A is semidefinite (definite) positive. Also, $A = 0$ will denote a matrix of zeros.

Assume we have K-level data. Level 1 is the outer level. Level 2 is nested within level 1 and so on up to level $K-1$ which is the deepest level (nested within level $K-2$). Assume that at level k ($k = 1, \dots, K-1$), we have n_k clusters indexed from 1 to n_k (no two clusters at the same level have the same index). We have a total of N observations. We will use the notation $l(k, i)$ to indicate the level k ($k = 1, \dots, K-1$) cluster index of observation i ($i = 1, \dots, N$). The multivariate multilevel linear model is written as:

$$Y_i = \beta X_i + \sum_{k=1}^{K-1} a_{kl(k,i)} + \epsilon_i, \quad i = 1, \dots, N.$$

In details, Y_i is the $p \times 1$ ($p \geq 1$) vector of responses for observation i, X_i is the $q \times 1$ vector of covariates, β is the $p \times q$ matrix of coefficients, and ϵ_i is the $p \times 1$ individual error vector. We assume that the ϵ's are independent and identically distributed (iid) with $E[\epsilon_i] = 0$ and covariance matrix $V[\epsilon_i] = \Sigma > 0$. For a given k, $a_{k1}, \dots, a_{kn_k}$ are the random effects at level k. We assume that they are iid with $E[a_{kj}] = 0$ and covariance matrix $V[a_{kj}] = \Sigma_k \geq 0$. Furthermore, the random effects from different levels are independent and independent of the ϵ's. Note that the standard assumptions that the random effects and individual errors are normally distributed are not required. Conditionally on the X_i's (which we will not repeat to simplify the presentation), the covariance structure of the Y_i's is $V[Y_i] = \sum_{k=1}^{K-1} \Sigma_k + \Sigma$ and

$$Cov[Y_i, Y_j] = \sum_{k=0}^{m(i,j)} \Sigma_k$$

where $m(i, j)$ is the maximum value of k such that $l(k, i) = l(k, j)$. That is, $m(i, j)$ is the deepest level that observations i and j have in common. Here we use the convention that $m(i, j) = 0$ when $l(1, i) \neq l(1, j)$ and define $\Sigma_0 = 0$. Hence, $Cov[Y_i, Y_j] = 0$ when observations i and j belong to different level 1 clusters. Note that when $\Sigma_1 = \ldots = \Sigma_{K-1} = 0$, the model reduces to the ordinary multivariate regression model.

For K-level data and for $t \in \{1, \ldots, K-1\}$, define the following hypothesis:

$$H_{0t} : \Sigma_t = 0 \quad \text{vs} \quad H_{1t} : \text{at least one eigenvalue of } \Sigma_t \text{ is positive.}$$

We will refer to it as "testing level t with K-level data". To fix ideas, and since the package presented in this paper can handle up to 4-level data, we explicitly write the covariance structure and hypotheses of interests in theses cases. For 2-level data, we have

$$Cov[Y_i, Y_j] = \begin{cases} 0 & \text{if} \quad m(i,j) = 0 \\ \Sigma_1 & \text{if} \quad m(i,j) = 1. \end{cases}$$

The only hypothesis of interest is then

$$H_{01} : \Sigma_1 = 0 \quad \text{vs} \quad H_{11} : \text{at least one eigenvalue of } \Sigma_1 \text{ is positive.}$$

For 3-level data, we have

$$Cov[Y_i, Y_j] = \begin{cases} 0 & \text{if} \quad m(i,j) = 0 \\ \Sigma_1 & \text{if} \quad m(i,j) = 1 \\ \Sigma_1 + \Sigma_2 & \text{if} \quad m(i,j) = 2. \end{cases}$$

Testing level 1, as above, is still of interest, as is the following hypothesis:

$$H_{02} : \Sigma_2 = 0 \quad \text{vs} \quad H_{12} : \text{at least one eigenvalue of } \Sigma_2 \text{ is positive.}$$

For 4-level data, we have

$$Cov[Y_i, Y_j] = \begin{cases} 0 & \text{if} \quad m(i,j) = 0 \\ \Sigma_1 & \text{if} \quad m(i,j) = 1 \\ \Sigma_1 + \Sigma_2 & \text{if} \quad m(i,j) = 2 \\ \Sigma_1 + \Sigma_2 + \Sigma_3 & \text{if} \quad m(i,j) = 3. \end{cases}$$

Three hypotheses are of interests, the same two as for 3-level data and a third one:

$$H_{03} : \Sigma_3 = 0 \quad \text{vs} \quad H_{13} : \text{at least one eigenvalue of } \Sigma_3 \text{ is positive.}$$

The `mvctm` package can perform tests for the above six testing problems, one for 2-level data, two for 3-level data and three for 4-level data.

2.2 Permutation principle

Assume that we have a statistic S such that greater values indicate "more dependence" in the data. The specific statistic we have in mind will be presented in the next subsection. But here, we want to describe the generic permutation testing method. The basic idea behind the permutation test procedure to test level t with K-level data is the following.

Generic permutation procedure

1. Compute S with the original data. Call it S_o.
2. For $b = 1, \ldots, B$. Randomly reassign the observations within level t with a valid permutation (see below). Compute S with the permuted data. Call it S_b.
3. The p-value of the test is the proportion of time that S_b is greater than or equal to S_o.

Valid permutation

To test level t with K-level data, a permutation is valid if it possesses the following characteristics:

1. For $k = 1, \ldots, K-1$ ($K \neq t$), and $i, j = 1, \ldots, N$, $I[l(k,i) = l(k,j)]$ remains constant before and after the permutation, where I is the indicator function.
2. The number of $t+1$ clusters (or the number of observations if $t+1 = K$) within each level t cluster remains constant before and after the permutation.

Basically, a valid permutation reassigns the tested level memberships by keeping intact the configuration of the other levels. In order to fix the permutation idea more clearly, we use some graphical examples. Starting with 2-level data, the only testing problem is to test for level 1. Figure 1 gives the original data, two valid permutations, and one invalid one. The numbers inside the clusters are the observation numbers. The original data have three clusters, one of size two, one of size three and one of size four. We see that the number of observations within each level 1 cluster remains constant for a valid permutation. The bottom right plot shows an invalid permutation because there are three clusters of size three, which is not the same as the original configuration.

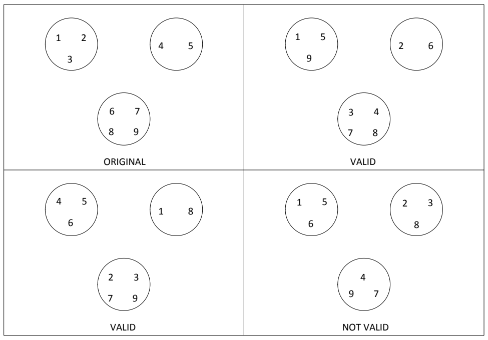

Figure 1: Examples of valid and invalid permutations when testing level 1 with 2-level data.

With 3-level data, we look first at the case where we want to test level 2, the deepest level. Figure 2 gives the original data. The configuration is two level 1 clusters, one with two level 2 clusters and one with three level 2 clusters. Again, two valid permutations, and an invalid one are depicted. We see, for a valid permutation, that a level 2 cluster stays inside its original level 1 cluster, and that the number of level 2 clusters within each level 1 cluster remain constant. The permutation in the bottom right is invalid because some observations jumped to another level 1 cluster (namely observations 4, 5, 6, 11).

Still with 3-level data, we now look at the case where we want to test level 1, the outer level, with Figure 3. We use the same configuration as in the last figure. We see, for a valid permutation, that the level 2 clusters are not broken and that the number of level 2 clusters within each level 1 cluster remains constant. It is the full level 2 clusters that are permuted. The bottom right permutation is invalid because i) some level 2 clusters are broken and ii) the number of level 2 cluster inside each level 1 cluster did not remain constant. Note that any one of these two facts is sufficient to say that the permutation is invalid.

Finally, the idea is the same with 4-level data. We only illustrate, in Figure 4, the middle case where we want to test level 2. We see, for a valid permutation, that the level 3 clusters are not broken, that the observations remain in the same level 1 cluster, and that the number of level 3 clusters, within each level 2 cluster, remains constant. The permutation in the bottom right is invalid because i) some observations have switched level 1 cluster, ii) the level 3 clusters have been broken (observations 13 and 15 were not together originally), and iii) the number of level 2 clusters, within each level 1 cluster did not remain constant. Again, any one of these facts is sufficient to say that the permutation is invalid.

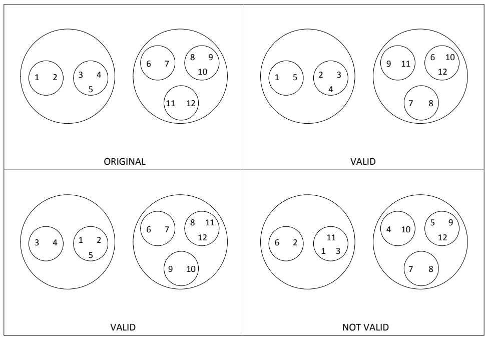

Figure 2: Examples of valid and invalid permutations when testing level 2 with 3-level data.

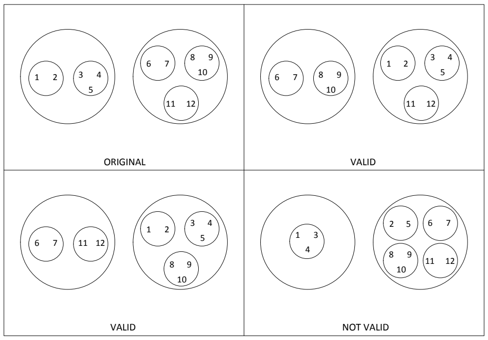

Figure 3: Examples of valid and invalid permutations when testing level 1 with 3-level data.

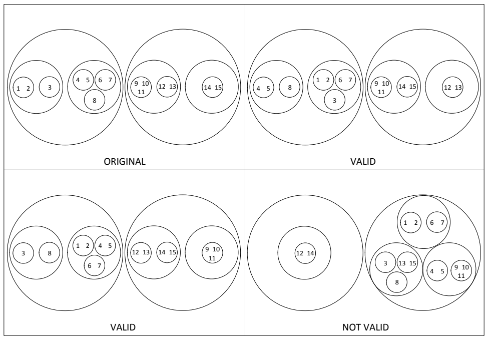

Figure 4: Examples of valid and invalid permutations when testing level 2 with 4-level data.

2.3 Description of the basic statistic

The test is based on a statistic used to evaluate the dependence between observations up to a given level of clustering. We will denote by T_i, the transformation of the original responses used in the calculation of the test statistic. In its simplest form, T_i will be the centered version of Y_i, after removing the effect of the covariates. Namely, $T_i = Y_i - \hat{\beta}X_i$ where $\hat{\beta}$ is a suitable estimate of β. But other transformations (based on multivariate signs and ranks) are also possible with the `mvctm` package and are discussed below. Let $I_t = \{i, j = 1, \dots, N : i \neq j, m(i,j) = t\}$, i.e., the set of pairs of index such that t is their largest level in common. The general form of the statistic used to evaluate the dependency up to level t is:

$$S_t = \left(\sum_{I_t} w_{ij}\right)^{-1} \sum_{I_t} w_{ij} T_i T_j' \tag{1}$$

where the w's are weights chosen by the analyst. The test statistic used in the permutation procedure is

$$\lambda_M(S_t) = \text{largest eigenvalue of } S_t.$$

The fact that S_t has a closed-form expression is an advantage for a permutation test since it has to be evaluated many times. Hence, problems of convergence of statistics based the maximization of a likelihood function are avoided. Moreover, the effect of the covariates is removed only once with the original data and not for each permuted sample, which also speeds up the computations.

The weights w_{ij} can be used to balance the importance given to clusters as opposed to single observations. Three obvious choices are:

$$\begin{gathered} 1 \text{ (pair)} \\ (m(i)+m(j))^{-1} \text{ (observation)} \\ (n_t(i))^{-1} \text{ (cluster)} \end{gathered}$$

where $m(i)$ is the number of times that i appears in I_t, and where $n_t(i)$ is the number of pairs, from the level t cluster to which i belongs, that appears in I_t. The "pair" weights give a weight of one to each pair. Hence, observations in larger level t clusters will have more weights. At the other extreme, the "cluster" weights give the same weight to each level t cluster. As a compromise between these two, the "observation" weights give an equal weight to each individual observation. Note that with fully balanced data, i.e., when the number of level $k+1$ clusters is the same inside each level k clusters for all k, and the number of observations is the same inside each level K cluster, these three weight functions are the same. The three weights functions are available in the `mvctm` package.

As mentioned, S_t estimates the dependency up to level t. To fix ideas, assume that β is known and that $T_i = Y_i - \beta X_i$, then $E[S_t] = \sum_{k=0}^{t} \Sigma_k$. The idea behind the test is that, for a permuted sample, some pairs will have an expectation of $E[T_i T_j'] = \sum_{k=0}^{t} \Sigma_k$ and some others will have $E[T_i T_j'] = \sum_{k=0}^{t-1} \Sigma_k$. Hence, under H_1, a permuted sample will show "less" dependency, as measured by $\lambda_M(S_t)$.

In principle, many different scores could be used for T_i in (1). The `mvctm` package allows the use of multivariate ranks and signs. For multivariate data, many versions of signs and ranks are available. Spatial ranks and signs are used in `mvctm`; Oja (2010), Nevalainen et al. (2010), Nordhausen and Oja (2018). Some details are provided in Appendix 1. Here we described only the univariate case. With the rank scores, once the effect of the covariates is removed, the residuals are simply replaced by their ranks. Hence, $T_i = r(Y_i - \hat{\beta} X_i)$, where $r(z_i)$ is the rank of z_i among $z_1, \ldots, z_N$. It is a Spearman-type measure of dependency. With the sign scores, once the effect of the covariates is removed, the residuals are simply replaced by their signs. Hence, $T_i = \text{sign}(Y_i - \hat{\beta} X_i)$, giving a Kendall-type measure of dependency. A complete description of the options available in the package `mvctm` is given next.

2.4 The mvctm package

The R package `mvctm` can perform the permutation tests described above. Tests for 2, 3 and 4-level data are available. Only the function `mvctm` is needed to apply the tests. The function has the following parameters and default values:

```
mvctm(fixed, cluster, data, leveltested, method="ls", npermut=1000, weight="observation",
affequiv=TRUE)
```

Using the defaults for the other, only four arguments are required.

Required arguments

- fixed: An object of class "formula" describing the fixed effects part of the model using the variables in the data frame data.
- cluster: A vector giving the name of the variables in the data frame data to specify the clustering configuration. The order is important. For 2-level data it is a vector of dimension 1 specifying the level 1 cluster. For 3-level data, it is a vector of dimension 2. The first element specifies the level 1 (outer) cluster and the second one specifies the level 2 (inner) cluster. For 4-level data, it is a vector of dimension 3. The first element specifies the level 1 (outer) cluster, the second one specifies the level 2 (middle) cluster, and the last one specifies the level 3 (inner) cluster.
- data: A data frame containing the data.
- leveltested: An integer giving the level to be tested. It must be 1 for 2-level data, 1 or 2 for 3-level data, and 1, 2 or 3 for 4-level data. It corresponds to the element in cluster.

The other arguments of the function provide other options and allow the user more control.

Other arguments

- method: The scores to be used. The four choices "ls", "mixed", "rank" and "sign" are available. The default is "ls". The choice "mixed" is only available for a univariate response.
- npermut: The number of permutation used to perform the test. The default is 1000.
- weight: The weight function to be used. The three choices "pair", "observation" and "cluster" are available. The default is "observation".
- affequiv: Whether or not we want to use the affine-equivariant version of the tests. This is only relevant when $p > 1$ and method="rank" or "sign". See Appendix 1 for details. The default is TRUE.

The function mvctm returns a list with the following two elements:

1. The p-value of the test.
2. The value of the test statistic for the original data, $\lambda_M(S_t)$.

Examples of calls to the function mvctm are given in Section Data example. Here we give more details about the method argument. Basically, method="ls", estimates the fixed effects part of the model (β) by ordinary least-squares. Then the test is performed on the residuals from this fit. Hence, the clustering structure is not used to estimate β. While this approach is valid, it may not be the most efficient one to estimate β. This is why with method="mixed", only available with a univariate response, β is estimated with a linear mixed model. Then the test is performed on the marginal (population) residuals from this fit. It is important to use the marginal residuals because we want to keep the correlation structure intact for the test. With method="rank", a rank-based method is used to estimate β. Then the test is performed on the ranks of the residuals from this fit. Finally, with method="sign", a sign-based method is used to estimate β. Then the test is performed on the signs of the residuals from this fit. Specific details about the R functions and packages used are given in Appendix 2.

As we just saw, the package mvctm has a few built-in ways to remove the fixed effects, transform the residuals, and then perform the test on the transformed residuals. But it is very flexible because it is possible to remove the fixed effects by any other means (outside mvctm) , compute the residuals, and then perform the permutation test with mvctm using these residuals. As an illustration, suppose we have 2-level data with a response y and two covariates x, z in a data frame name "dataschool", with a variable "school" to represent the clustering structure. Assume that, instead of using the available methods ("ls", "mixed", "rank" or "sign"), the analyst wishes to use an M-estimator to remove the fixed effects. This can be achieved easily like this:

```
R> library("MASS")

R> dataschool[,"mresid"]=rlm(y~x+z,data=dataschool)$residuals

R> mvctm(fixed=mresid~0,cluster=c("school"),data=dataschool,leveltested=1)
```

The second line adds the residuals from an M-estimate fit in the data frame. It comes from the function rlm in the package MASS; Venables and Ripley (2002). The third line calls mvctm. By specifying mresid$\sim$0, there is no covariates nor intercept. Hence, the test is performed using directly the response, here the residuals, without any centering or transformation. Thus, the package mvctm provides a platform to perform the permutation tests with any centering and/or transformation of the responses. Moreover, the utility function permcluster allows the user to get permuted data sets. This way, other test statistics can be used to perform the permutation test. The function has the following parameters: permcluster(cluster, data, leveltested). The arguments are the same as for the mvctm function. The type of permutation used depends on the number of levels and which level is tested.

3 Simulation study

3.1 Description of the study

A simulation study was performed to investigate the performance of the tests in the mvctm package. Here is the description of the study.

1. Two dimensions are used: $p = 1$ and 3.
2. Two distributions for the random effects and the error terms are used: normal and t_3 (t distribution with 3 degrees of freedom).
3. Six testing problems are investigated: testing Σ_1 with 2-level data, Σ_1 and Σ_2 with 3-level data, and Σ_1, Σ_2 and Σ_3 with 4-level data.
4. Two hypotheses are used: H_0 and H_1.

When we cross these four elements, we get 48 different scenarios ($2 \times 2 \times 6 \times 2$). They were all included in the study. Data were generated according to the model:

$$Y_i = \mu + \sum_{k=1}^{K-1} a_{kl(k,i)} + \epsilon_i, \quad i = 1, \ldots, N.$$

Hence, no covariates are present. Each individual a_{ij} and ϵ_i has the specified distribution (normal or t_3). Moreover, we use $\mu = 0$ to generate the data, without loss of generality. However, the intercept is still estimated by the testing procedure since it's unrealistic to assume that we know beforehand that $\mu = 0$.

The values of the Σ's that are used are given in Table 12 in Appendix 3. With 2-level data, a sample is formed by 20 clusters of three observations for a total of $N = 60$ observations.

For 3-level data, a sample is formed by ten level 1 clusters. Inside each of them there were three level 2 clusters containing three observations each. This is illustrated in the left part of Figure 5 and this gives a total sample size of $N = 90$.

For 4-level data, the right part of Figure 5 depicts the configuration of a single level 1 cluster. It is repeated ten times for a total sample size of $N = 120$.

Since the cluster configurations are balanced, method="ls" and method="mixed" are equivalent. Hence three methods are compared, namely, method="ls"("mixed"), method="rank" and method="sign". We use the "observation" weights throughout. But once again, since the cluster configurations are balanced, the three available weight functions ("observation", "pair" and "cluster") are equivalent. The affine-equivariant versions of the tests are used throughout.

In addition, with univariate data, a standard LRT based on a mixture of chi-squared distributions is used as a benchmark. The restricted maximum likelihood (REML) version of the test is used since it performs slightly better than the ML version for maintaining the nominal level; see Section 6.3.4 of Verbeke and Molenberghs (2009). It was obtained with PROC GLIMMIX in SAS; Inc. (2003).

For each scenario, 1000 samples are used. Each permutation test is performed with 1000 permutations. All tests are performed at the 5% level.

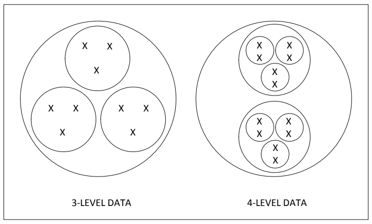

Figure 5: Cluster configurations used in the simulation for 3 and 4-level data.

3.2 Results

The results of the simulations are presented in Tables 1, 2 and 3. The first one is for 2-level data, the second one for 3-level data and the last one for 4-level data. In each table, the upper (lower) part presents the results under H_0 (H_1). Looking at the results under H_0, we can see that all tests maintain the prescribed level (5%) reasonably well except for the sign test with univariate data, which is liberal in many cases. This is probably due to the fact that the test statistic S_t does not have a rich enough distribution for the permutation approach to work. Indeed, $T_i T_j'$ in (1) can only take the values -1 or 1 with univariate data. With multivariate data ($p > 1$), the spatial signs distribution is a lot richer. Consequently, we see that the sign test is able to maintain its level with three-variate data. We can also notice that the LRT has a tendency towards being conservative. This was also noted in the simulation study in Fitzmaurice et al. (2007).

Table 1: Simulation results with 2-level data.

			Under H_0			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.052	0.080	0.054	0.050
1	1	t_3	0.041	0.076	0.044	0.034
3	1	normal	0.05	0.042	0.044	-
3	1	t_3	0.042	0.041	0.044	-
			Under H_1			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.493	0.361	0.455	0.481
1	1	t_3	0.532	0.484	0.574	0.503
3	1	normal	0.571	0.477	0.547	-
3	1	t_3	0.549	0.597	0.689	-

The proportion of time that the null hypothesis is rejected over the 1000 repetitions is presented. The permutation tests were performed with 1000 permutations.

The power of the tests can be compared by looking at the lower parts of the tables. Since the ls/mixed version of the permutation tests uses the original observations (not the signs or the ranks), it can be seen as the most direct competitor to the LRT. It is striking that either these two tests are very close, or the permutation test is more powerful. A similar finding was obtained in Fitzmaurice et al. (2007). Their permutation test, based on the LRT statistic, was more powerful than the mixture of chi-squared distributions test for both a linear model and a logistic regression model with 2-level data. Here we see that the permutation test continues to be better when more than two levels are present, regardless of which level is tested. Part of this may be explained by the fact that the permutation test tends to be closer to the prescribed level while the LRT is slightly conservative.

Table 2: Simulation results with 3-level data.

			Under H_0			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.046	0.052	0.046	0.044
1	1	t_3	0.042	0.047	0.043	0.033
1	2	normal	0.050	0.094	0.059	0.042
1	2	t_3	0.051	0.091	0.049	0.036
3	1	normal	0.056	0.066	0.056	-
3	1	t_3	0.039	0.045	0.046	-
3	2	normal	0.042	0.056	0.046	-
3	2	t_3	0.034	0.052	0.039	-
			Under H_1			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.449	0.376	0.440	0.444
1	1	t_3	0.446	0.412	0.485	0.413
1	2	normal	0.564	0.368	0.501	0.544
1	2	t_3	0.552	0.425	0.546	0.512
3	1	normal	0.535	0.480	0.525	-
3	1	t_3	0.499	0.534	0.555	-
3	2	normal	0.509	0.383	0.481	-
3	2	t_3	0.518	0.443	0.556	-

The proportion of time that the null hypothesis is rejected over the 1000 repetitions is presented. The permutation tests were performed with 1000 permutations.

Table 3: Simulation results with 4-level data.

			Under H_0			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.065	0.062	0.064	0.066
1	1	t_3	0.040	0.051	0.045	0.030
1	2	normal	0.048	0.048	0.046	0.038
1	2	t_3	0.063	0.075	0.062	0.032
1	3	normal	0.049	0.120	0.047	0.050
1	3	t_3	0.035	0.121	0.049	0.035
3	1	normal	0.055	0.056	0.053	-
3	1	t_3	0.060	0.054	0.052	-
3	2	normal	0.054	0.051	0.053	-
3	2	t_3	0.047	0.057	0.051	-
3	3	normal	0.048	0.046	0.045	-
3	3	t_3	0.040	0.043	0.043	-
			Under H_1			
	Level		Permutation Test			LRT
p	tested	Distribution	ls/mixed	sign	rank	mixture of χ^2
1	1	normal	0.550	0.434	0.517	0.552
1	1	t_3	0.488	0.402	0.493	0.481
1	2	normal	0.426	0.260	0.387	0.397
1	2	t_3	0.393	0.339	0.402	0.342
1	3	normal	0.506	0.345	0.431	0.494
1	3	t_3	0.489	0.431	0.493	0.465
3	1	normal	0.480	0.447	0.487	-
3	1	t_3	0.498	0.470	0.523	-
3	2	normal	0.442	0.373	0.434	-
3	2	t_3	0.456	0.402	0.482	-
3	3	normal	0.434	0.300	0.402	-
3	3	t_3	0.431	0.399	0.452	-

The proportion of time that the null hypothesis is rejected over the 1000 repetitions is presented. The permutation tests were performed with 1000 permutations.

In all scenarios, the rank test is more powerful than the sign test. In all scenarios but one, the ls/mixed test is more powerful than the rank test for the normal distribution. The opposite is true for the t_3 distribution. In all scenarios but one, the rank test is more powerful than the ls/mixed test in these cases. However, both tests are fairly competitive over all scenarios. Perhaps the only exception is when $p = 3$ with the t_3 distribution and 2-level data, where a larger difference is seen. In that case, the power of the rank test is 0.689 as opposed to 0.549 for the ls/mixed test. Even the sign test (0.597) is more powerful than the ls/mixed test in this scenario.

4 Data example

For this example, we use the 2012 data from the Programme for International Student Assessment (PISA). PISA is a worldwide study on children of age 15, measuring their performance in mathematics, reading, and science; PISA (2014). It is organized by the Organisation for Economic Co-operation and Development (OECD). Only the data from Canada are used in this example. The scores on the three dimension are used, namely mathematics, reading and science. Each on them is computed by averaging the available corresponding items. The total number of items is 109, 44 and 53 for mathematics, reading and science, respectively. A minimum of eight items is required to compute the score, otherwise we set the observation to missing. The sample size is 8161. Basic summary measures along with the correlation matrix are provided in Table 4. The three responses are highly correlated with correlations ranging between 0.604 and 0.675.

Table 4: Basic statistics and correlation matrix for the PISA data for Canada (N=8161).

	Mathematic	Reading	Science
Mean	0.527	0.653	0.592
Std	0.247	0.205	0.217
		Correlation	
	Mathematic	Reading	Science
Mathematic	1	0.604	0.649
Reading	-	1	0.675
Science	-	-	1

For the sake of illustration, multivariate analysis using jointly the three responses and univariate analysis on the separate responses are performed. The students are nested within schools (875) which are nested within strata (46). Moreover, since education is within provincial jurisdiction in Canada, we are also considering the ten provinces in the analysis. In one set of analysis, the province is considered as another level of clustering. In another set of analysis, the province is modeled as a fixed effect. Hence, when the province is considered as a level of clustering, we have a 4-level data set (schools, nested within strata, nested within provinces), with no covariates (only an intercept). When we model the province as a fixed effect, we have a 3-level data set (schools, nested within strata), with a categorical covariate taking ten values.

For the multivariate analysis, only the permutation test with method="ls" is used. For the univariate analysis, the permutation test with method="mixed" is used, along with the same standard LRT from the simulation study of the last section. Namely, this is the test based on a mixture of chi-squared distributions from PROC GLIMMIX. We are taking a nominal level of 0.05 for all tests performed.

Table 5 presents the p-values for the permutation tests applied jointly to the three responses (multivariate analysis). For example, the call to perform the test for the province variance component in the 4-level model is:

```
R> mvctm(fixed=cbind(mathematic,reading,science)~1, cluster=c("province","strata","school"),
data=canpisa, leveltested=1)
```

The data frame canpisa contains the data in this example. In Table 5, the variance components are significant for the strata and school levels in both models. The variance component for the province level is not significant in the 4-level model.

Table 5: Multivariate variance components tests of the PISA data for Canada (N=8161).

Level tested	3-level	4-level
Province	-	0.109
Strata	0.012	0.015
School	0.000	0.000

The table reports the p-values of the tests performed with 1000 permutations. An intercept is included in the fixed part of the model for the 4-level model, and a fixed province effect is included in the 3-level model.

We then performed separate tests on the three responses. However, with such a large sample size, the question about statistical significance versus effect size is relevant. This is why we also report the individual estimates of the variance components, their standard errors, and the corresponding intra-class correlations (ICC). They are taken from the REML fit from PROC GLIMMIX. Tables 6 and 7 report the results for the 4-level model, and Tables 8 and 9 report them for the 3-level model.

For example, the call to test the school level in the 4-level model with the mathematic score as the response is:

```
R> mvctm(fixed=mathematic∼1, cluster=c("province","strata","school"), data=canpisa, leveltested=3,
method="mixed")
```

The call to test the school level in the 3-level model with the reading score as the response is:

```
R> mvctm(fixed=reading∼prov1+prov2+prov3+prov4+prov5+prov6+prov7+prov8+prov9,
cluster=c("strata","school"), data=canpisa, leveltested=2, method="mixed")
```

In the last call, “prov1” to “prov9” are dummy variables to model the 10 provinces.

Looking at the 4-level models first (Tables 6 and 7), we see that the two variance components tests agree with one exception. Both tests are significant for the variance components at the strata and school levels for each of the three responses (Table 6). The variance component at the province level is only found significant by the permutation test (p-value=0.019) for the mathematic score. From Table 7, the correlation (ICC) between two observations of the mathematic score from the same province but from different strata (and thus schools) is 0.00509, which is quite small. It is obtained as 0.000311/(0.000311+0.001153+0.005136+0.05454). The large sample size explains why the permutation test detects such a small effect. The correlations (all significant) between two observations from the same stratum but different schools vary between 0.0239 and 0.0357 across the three responses. For example, the one for the mathematic score is obtained as (0.000311+0.001153)/(0.000311+0.001153+0.005136+0.05454). Finally, the correlations (all significant) between two observations from the same school vary between 0.108 and 0.118.

Table 6: Univariate variance components tests for the 4-level models (PISA data for Canada, N=8161).

	p-value for the variance component					
	Variable					
Level	Mathematic		Reading		Science	
tested	perm.	mixt. χ^2	perm.	mixt. χ^2	perm.	mixt. χ^2
Province	0.019	0.2701	0.282	0.2695	1.000	0.3105
Strata	0.014	0.0006	0.001	<.0001	0.000	<.0001
School	0.000	<.0001	0.000	<.0001	0.000	<.0001

The table reports the p-values of the tests. The permutation tests were performed with 1000 permutations. Only an intercept is included in the fixed part of the model.

Table 7: Univariate variance components estimates for the 4-level models (PISA data for Canada, N=8161).

	Variable					
	Mathematic		Reading		Science	
Level	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)
Province	0.000311	0.000581	0.00022	0.000403	0.000215	0.000482
	ICC=0.00509		ICC=0.00520		ICC=0.00453	
Strata	0.001153	0.000699	0.000943	0.000496	0.00148	0.000663
	ICC=0.0239		ICC=0.0275		ICC=0.0357	
School	0.005136	0.000568	0.003811	0.000407	0.003881	0.000439
	ICC=0.108		ICC=0.118		ICC=0.118	
Residuals	0.05454	-	0.03731	-	0.04185	-

Only an intercept is included in the fixed part of the model.

Moving to the 3-level models (Tables 8 and 9), we see that the two tests are significant for the variance components at the strata and school levels for each of the three responses (Table 8). In all these models, there is a fixed effect for the province. We see from the bottom of Table 9 that the province fixed effect is only significant for the mathematic score (p-value=0.0083). It is interesting to note that in the preceding 4-level model for mathematic, the variance component at the province level was found significant by the permutation test but not by the LRT. The ICC (Table 9) are similar to what they were in the 4-level models, except that now they must be interpreted as the correlation between residuals after the province effect is removed.

Table 8: Univariate variance components tests for the 3-level models (PISA data for Canada, N=8161).

	p-value for the variance component Variable					
Level	Mathematic		Reading		Science	
tested	perm.	mixt. χ^2	perm.	mixt. χ^2	perm.	mixt. χ^2
Strata	0.019	0.0057	0.001	$<.0001$	0.000	$<.0001$
School	0.000	$<.0001$	0.000	$<.0001$	0.000	$<.0001$

The table reports the p-values of the tests. The permutation tests were performed with 1000 permutations. The province is modeled as a fixed effect.

Table 9: Univariate variance components estimates for the 3-level models (PISA data for Canada, N=8161).

	Variable					
	Mathematic		Reading		Science	
Level	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)
Strata	0.000662	0.000514	0.000752	0.00041	0.001334	0.000614
	ICC=0.0110		ICC=0.0180		ICC=0.028	
School	0.005188	0.000572	0.003816	0.000408	0.003891	0.00044
	ICC=0.097		ICC=0.109		ICC=0.111	
Residuals	0.05455	-	0.03732	-	0.04185	-
p-value to test the province fixed effect	0.0083		0.0700		0.1551	

The province is modeled as a fixed effect.

Since the sample size is rather large, it seems interesting to compare the methods further by using subsamples, in order to investigate their power by varying the sample size. To achieve this, we focus on the mathematic score in the 3-level model (strata-school) with a province fixed effect. For each sample size n =225, 400, 625, 900, 1225 , 1600, 2025, and 2500, we obtained 200 random samples from the original data of size $N = 8161$. We then performed the permutation test (again with method="mixed") and the LRT on all these samples (1600 samples in all) to test the variance component at the school level. We then obtained the proportion of times that H_0 is rejected for each sample size, over the 200 samples. These proportions are reported in Table 10 and plotted in Figure 6. We see that the permutation test is more powerful than the LRT for the smaller sample sizes. It is striking that the permutation test is 56% more powerful than the LRT when $n = 225$. Then the gap between the two tests becomes small for sample sizes of 1225 and over. Both tests have a power of 1 when we reach $n = 2500$. This goes in accordance with the findings from the simulation study where, either the two tests had a similar power, or the permutation test was more powerful.

Table 10: Proportions of rejections for testing the school level in a 3-level (strata-school) model with province as fixed effect, as a function of the sample size.

Sample size	Permutation (method="mixed")	LRT mixture of χ^2
225	0.125	0.080
400	0.160	0.120
625	0.360	0.295
900	0.455	0.400
1225	0.710	0.690
1600	0.890	0.900
2025	0.955	0.970
2500	1.000	1.000

The response is the mathematic score.

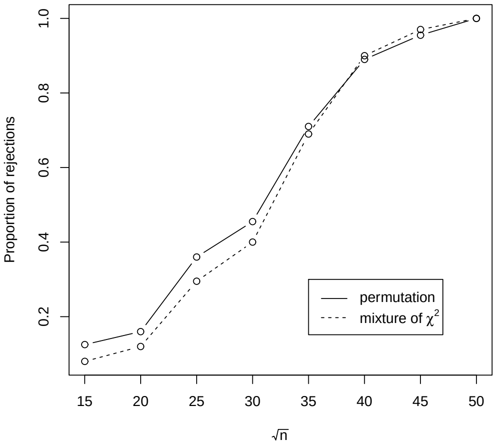

Figure 6: Proportions of rejections for testing the school level in a 3-level (strata-school) model with province as fixed effect, as a function of the sample size. The response is the mathematic score.

5 Discussion and concluding remarks

The goal of this paper is to propose a practical solution, based on the permutation approach, to test for random components with multilevel data and possibly multivariate responses. The R package `mvctm` implements several different methods. Some are based on the original observations, and others are based on the ranks or signs (the spatial version of them for multivariate data) of the observations. The test statistic is moment based and requires no distributional assumptions. Furthermore, it has a closed-formed expression which is convenient for the many computations required for a permutation test. The current version of the package can test any level for up to 4-level data, which should be sufficient for most practical applications.

A simulation study, covering all the implemented testing problems shows that the proposed approach works well. Indeed, for univariate data, the permutation test based on the original observations is preferable to a LRT using a mixture of chi-squared distributions, confirming the results obtained in Fitzmaurice et al. (2007). Moreover, the test based on the ranks of the observations is also very competitive, especially, with heavier-tailed distributions like the t_3. However, the sign test did not perform well in the simulations. First of all, it did not maintain the prescribed level with univariate data. Secondly, it was less powerful than the rank test in all scenarios. One might reasonably think that the sign test would do better when the distribution has heavier tails than the t_3. We ran an additional simulation to investigate that. With 2-level data, $p = 3$, and the same cluster configuration as the one used in the scenarios reported in Table 1, we ran two additional scenarios with the t_1 (Cauchy) distribution. The results are reported in Table 11. We see that the three tests maintain their level but the rank test is still more powerful than the sign test. The ls/mixed test is the least powerful, far behind the other two. Consequently, based on the limited empirical results, we can only recommend the ls, mixed and rank versions of the tests at the moment.

Table 11: Additional simulation with 2-level data.

		Under H_0			
	Level			Test	
p	tested	Distribution	ls/mixed	sign	rank
3	1	t_1	0.053	0.054	0.058
		Under H_1			
	Level			Test	
p	tested	Distribution	ls/mixed	sign	rank
3	1	t_1	0.441	0.758	0.847

The proportion of time that the null hypothesis is rejected over the 1000 repetitions is presented. The permutation tests were performed with 1000 permutations.

The package `mvctm` provides a flexible structure allowing the removal of the covariates effects using any method judged adequate, prior to performing the permutation test. This flexibility allows to use many other methods like robust ones for example.

The aim of the simulation was to show the validity of the proposed tests for all possible testing problems with 2, 3 and 4-level data. However, many more aspects would deserve a closer examination. First of all, all cluster designs used in the simulation were balanced, making the choice of the weight function irrelevant. Further simulations could investigate the merits of the three implemented weight functions for unbalanced designs. Secondly, the rank test was more powerful than the sign test in all the scenarios considered. Investigating the conditions (if any) that would make the sign test worthwhile could be interesting. Finally, we used a closed-form moment based estimator to measure the dependency in the data, which requires no distributional assumptions except finite second moments. But other statistics could be used within the permutation framework. The work mention in the Introduction (Fitzmaurice et al. (2007), Samuh et al. (2012), Lee and Braun (2012), Drikvandi et al. (2013), Zeng et al. (2015)) all use different statistics in the univariate case. This aspect would deserve further investigation, especially for multivariate data. The function permcluster in `mvctm` generates permuted data sets for all the testing problems considered in this article and thus could be useful for that purpose.

Appendix 1: Spatial signs, spatial ranks and affine-equivariance

A brief description of the notion of spatial signs and ranks is given here. More details can be found in Oja (2010). For a $p \times 1$ vector y, define the function

$$U(y) = \begin{cases} 0 & \text{if} \quad y = 0 \\ \frac{y}{||y||} & \text{if} \quad y \neq 0. \end{cases}$$

Let $Y_1, \ldots, Y_N$ be $p \times 1$ vectors. The spatial sign of Y_i is $U(Y_i)$. For univariate data, it is simply $\text{sign}(y_i)$. The spatial rank of Y_i, among $Y_1, \ldots, Y_N$, is

$$R(Y_i) = \frac{1}{N} \sum_{j=1}^{N} U(Y_i - Y_j).$$

For univariate data, $R(Y_i)$ is the centered rank of Y_i.

A natural property of a scatter statistic S, like S_t in (1), is affine-equivariance, which means that

$$S(AY_1, \ldots, AY_N) = AS(Y_1, \ldots, Y_N)A^\top,$$

for any non-singular matrix A. S_t has this property if computed with the original observations $Y_1, \ldots, Y_N$, but not if computed with the spatial ranks $R(Y_1), \ldots, R(Y_N)$ or the spatial signs $U(Y_1), \ldots, U(Y_N)$ (note that the transformation by A is applied to the original observations before the spatial ranks or signs are computed). In order to achieve affine-equivariance, a transformation-retransformation method can be used. Basically, this method applies a data-driven transformation to the spatial ranks (or signs), computes the statistic of interest (S_t for instance), and transforms the result back into the original coordinate system. By carefully selecting the transformation, affine-equivariance can be achieve. The `MNM` package, Nordhausen and Oja (2011), contains many functions to perform multivariate analysis using spatial ranks and signs. We are using this package in `mvctm`. Appendix 2 mentions which particular functions are used. Importantly, the data-driven transformations needed to achieve affine-equivariance are available in these functions from `MNM`. We refer to Oja (2010) and Nordhausen and Oja (2011) for technical details.

Appendix 2: Details about the argument method of the mvctm function

As described in Sections 2.3 and 2.4, the permutation test is performed on the (maybe transformed) residuals after removing the effect of the covariates. The way it is performed depends on the method argument. When the model has no intercept and no covariate, the raw data are used to perform the test. This may be appropriate if the data have been already centered using another method. For all other cases, the specific values T_i used to perform the permutation test are detailed below.

- When method="ls": β is estimated by ordinary least-squares and $T_i = Y_i - \hat{\beta}X_i$. The lm function is used.
- When method="mixed": β is estimated by a linear mixed model with random intercepts to account for the clustering structure, and $T_i = Y_i - \hat{\beta}X_i$. The function lme in the package `nlme` is used; Pinheiro et al. (2017).
- When method="rank":
  1. When there is only an intercept in the model. If $p = 1$, $T_i = r(Y_i) - \text{avg}(r(Y_i))$. If $p > 1$, the spatial ranks with inner standardization are used and obtained from the function mv.1sample.est in the `MNM` package. See Appendix 1.
  2. When other covariates are also present in the model. If $p = 1$, $T_i = r(\hat{e}_{ri}) - \text{avg}(r(\hat{e}_{ri}))$, where the $\hat{e}_{ri}$'s are the residuals after a rank estimation of β using the rfit function in the `Rfit` package; Kloke and McKean (2012). If $p > 1$, the T_i's are the residuals from a spatial rank fit using the function mv.l1lm in the `MNM` package. see Appendix 1.
- When method="sign":
  1. When there is only an intercept in the model. If $p = 1$, $T_i = \text{sign}(Y_i - \tilde{Y})$, where $\tilde{Y}$ is the median of $Y_1, \ldots, Y_N$. If $p > 1$, the spatial signs with inner standardization are used and obtained from the function mv.1sample.est in the `MNM` package. See Appendix 1.
  2. When other covariates are also present in the model. If $p = 1$, $T_i = \text{sign}(\hat{e}_{si})$, where the $\hat{e}_{si}$'s are the residuals after a median regression estimation of β using the rq function in the `quantreg` package; Koenker (2017). If $p > 1$, the T_i's are the residuals from a spatial sign fit using the function mv.l1lm in the `MNM` package. see Appendix 1.

Appendix 3: Parameters used in the simulation study

Table 12 gives the values of Σ, Σ_1, Σ_2 and Σ_3 used in the simulation study.

Table 12: Values used for the Σ's in the simulation study. A "0" means a matrix of 0 when $p = 3$.

Data	Level tested	p	Hypothesis	Σ	Σ_1	Σ_2	Σ_3
2-level	1	1	H_0	1	0	NA	NA
	1	1	H_1	1	0.3	NA	NA
2-level	1	3	H_0	D	0	NA	NA
	1	3	H_1	D	$0.25 \times D$	NA	NA
3-level	1	1	H_0	1	0	1	NA
	1	1	H_1	1	0.65	1	NA
	2	1	H_0	1	1	0	NA
	2	1	H_1	1	1	0,3	NA
3-level	1	3	H_0	D	0	D	NA
	1	3	H_1	D	$0.45 \times D$	D	NA
	2	3	H_0	D	D	0	NA
	2	3	H_1	D	D	$0.25 \times D$	NA
4-level	1	1	H_0	1	0	1	1
	1	1	H_1	1	1.7	1	1
	2	1	H_0	1	1	0	1
	2	1	H_1	1	1	0.5	1
	3	1	H_0	1	1	1	0
	3	1	H_1	1	1	1	0.3
4-level	1	3	H_0	D	0	D	D
	1	3	H_1	D	$1.1 \times D$	D	D
	2	3	H_0	D	D	0	D
	2	3	H_1	D	D	$0.5 \times D$	D
	3	3	H_0	D	D	D	0
	3	3	H_1	D	D	D	$0.25 \times D$

The matrix D is $\begin{pmatrix} 1 & .2 & .3 \\ .2 & 1 & -.1 \\ .3 & -.1 & 1 \end{pmatrix}$

References

Johannes Berkhof and Tom AB Snijders. Variance component testing in multilevel models. Journal of Educational and Behavioral Statistics, 26(2):133–152, 2001.

Ronald Christensen. Exact tests for variance components. Biometrics, pages 309–314, 1996.

Reza Drikvandi, Geert Verbeke, Ahmad Khodadadi, and Vahid Partovi Nia. Testing multiple variance components in linear mixed-effects models. Biostatistics, 14(1):144–159, 2013.

Garrett M Fitzmaurice, Stuart R Lipsitz, and Joseph G Ibrahim. A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics, 63(3):942–946, 2007.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. Applied longitudinal analysis, 2nd edition. Wiley, 2011.

Harvey Goldstein. Multilevel statistical models, 4th edition. Wiley, 2010.

SAS Institute Inc. SAS/STAT Software, Version 9.3. Cary, NC, 2003. http://www.sas.com/.

John D. Kloke and Joseph W. McKean. Rfit: Rank-based estimation for linear models,. The R Journal, 4(2):57–64, 2012.

Roger Koenker. quantreg: Quantile Regression, 2017. https://CRAN.R-project.org/package=quantreg. R package version 5.34.

Ita G G Kreft and Jan De Leeuw. Introducing multilevel modeling. Sage, 1998.

Denis Larocque. mvctm: Multivariate Variance Components Tests for Multilevel Data, 2017. https://CRAN.R-project.org/package=mvctm. R package version 1.1.

Oliver E Lee and Thomas M Braun. Permutation tests for random effects in linear mixed models. Biometrics, 68(2): 486–493, 2012.

Jaakko Nevalainen, Denis Larocque, Hannu Oja, and Ilkka Pörsti. Nonparametric analysis of clustered multivariate data. Journal of the American Statistical Association, 105(490):864–872, 2010.

Juvêncio S Nobre, Julio M Singer, and Pranab K Sen. U-tests for variance components in linear mixed models. Test, 22(4):580–605, 2013.

Klaus Nordhausen and Hannu Oja. Multivariate l_1 methods: The package mnm. Journal of Statistical Software, 43(5):1–28, 2011. http://www.jstatsoft.org/v43/i05/.

Klaus Nordhausen and Hannu Oja. Robust nonparametric inference. Annual Review of Statistics and Its Application, 5:473–500, 2018.

J Ofversten. Exact tests for variance components in unbalanced mixed linear models. Biometrics, pages 45–57, 1993.

Hannu Oja. Multivariate nonparametric methods with R. Springer, 2010.

Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. nlme: Linear and Nonlinear Mixed Effects Models, 2017. https://CRAN.R-project.org/package=nlme. R package version 3.1–131.

PISA, 2014. http://www.oecd.org/pisa/.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. https://www.R-project.org/.

Stephen W Raudenbush and Anthony S Bryk. Hierarchical linear models: Applications and data analysis methods. Sage, 2001.

Monjed H Samuh, Leonardo Grilli, Carla Rampichini, Luigi Salmaso, and Nicola Lunardon. The use of permutation tests for variance components in linear mixed models. Communications in Statistics-Theory and Methods, 41 (16-17):3020–3029, 2012.

Marc A Scott, Jeffrey S Simonoff, and Brian D Marx, editors. The SAGE handbook of multilevel modeling. SAGE Publications, 2013.

Peter X-K Song. Correlated data analysis. Springer, 2007.

Daniel O Stram and Jae Won Lee. Variance components testing in the longitudinal mixed effects model. Biometrics, pages 1171–1177, 1994.

W. N. Venables and B. D. Ripley. Modern applied statistics with S, 4th edition. Springer, 2002. http://www.stats.ox.ac.uk/pub/MASS4.

Geert Verbeke and Geert Molenberghs. Linear mixed models for longitudinal data. Springer, 2009.

Ping Zeng, Yang Zhao, Hongliang Li, Ting Wang, and Feng Chen. Permutation-based variance component test in generalized linear mixed model with application to multilocus genetic association study. BMC Medical Research Methodology, 15(1):37, 2015.

Daowen Zhang and Xihong Lin. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In Random effect and latent variable model selection, pages 19–36. Springer, 2008.