NeuroImage 123 (2015) 253–268

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Multi-level block permutation

Anderson M. Winkler [a,*], Matthew A. Webster [a], Diego Vidaurre [b], Thomas E. Nichols [a,c], Stephen M. Smith [a]

[a] Oxford Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK
[b] Oxford Centre for Human Brain Activity, University of Oxford, Oxford, UK
[c] Department of Statistics & Warwick Manufacturing Group, University of Warwick, Coventry, UK

ARTICLE INFO

Article history:
Received 8 December 2014
Accepted 24 May 2015
Available online 11 June 2015

Keywords:
Permutation inference
Multiple regression
General linear model
Repeated measurements

ABSTRACT

Under weak and reasonable assumptions, mainly that data are exchangeable under the null hypothesis, permutation tests can provide exact control of false positives and allow the use of various non-standard statistics. There are, however, various common examples in which global exchangeability can be violated, including paired tests, tests that involve repeated measurements, tests in which subjects are relatives (members of pedigrees) — any dataset with known dependence among observations. In these cases, some permutations, if performed, would create data that would not possess the original dependence structure, and thus, should not be used to construct the reference (null) distribution. To allow permutation inference in such cases, we test the null hypothesis using only a subset of all otherwise possible permutations, i.e., using only the rearrangements of the data that respect exchangeability, thus retaining the original joint distribution unaltered. In a previous study, we defined exchangeability for blocks of data, as opposed to each datum individually, then allowing permutations to happen within block, or the blocks as a whole to be permuted. Here we extend that notion to allow blocks to be nested, in a hierarchical, multi-level definition. We do not explicitly model the degree of dependence between observations, only the lack of independence; the dependence is implicitly accounted for by the hierarchy and by the permutation scheme. The strategy is compatible with heteroscedasticity and variance groups, and can be used with permutations, sign flippings, or both combined. We evaluate the method for various dependence structures, apply it to real data from the Human Connectome Project (HCP) as an example application, show that false positives can be avoided in such cases, and provide a software implementation of the proposed approach.

Introduction

In the context of hypothesis testing using the general linear model (GLM) (Scheffé, 1959; Searle, 1971), permutation tests can provide exact or approximately exact control of false positives, and allow the use of various non-standard statistics, all under weak and reasonable assumptions, mainly that the data are *exchangeable* under the null hypothesis, that is, that the joint distribution of the error terms remains unaltered after permutation. Permutation tests that compare, for instance, groups of subjects, are of great value for neuroimaging (Holmes et al., 1996; Nichols and Holmes, 2002), and in Winkler et al. (2014), extensions were presented to more broadly allow tests in the form of a GLM, and also to account for certain types of well structured non-independence between observations, which ordinarily would preclude the use of permutation methods. This was accomplished by redefining the basic exchangeable unit from each individual datum to blocks of data, i.e., rather than asserting exchangeability across all observations of a given experiment, blocks of exchangeable units are defined; these *exchangeability blocks* (EBS) can be rearranged as a whole (*whole-block exchangeability*), or the observations within block can be shuffled among themselves (*within-block exchangeability*), using either permutations, sign flippings, or permutations combined with sign flippings.

In the same work, the *G*-statistic, a generalisation over various commonly used statistics, including the *F*-statistic, was proposed. *G* is robust to known heteroscedasticity (i.e., the situation in which the variances are known to be not equal across all observations, which can be then classified into variance groups) and can be used with the GLM, ensuring that pivotality[1] is preserved, a crucial requisite for exact control over familywise error rate (FWER) using the distribution of the most extreme statistic (Westfall and Young, 1993), as needed in many neuroimaging studies. Indeed, the use of EBS allows for variances to be heterogeneous, provided that the groups of observations sharing the same variance (i.e., *variance groups*, VGS) (Woolrich et al., 2004) are compatible with the EBS; specifically, for within-block exchangeability the VGS must coincide with the blocks, and for whole-block exchangeability they must include one or more observations from each block in a consistent order.

---

* Corresponding author.
E-mail address: winkler@fmrib.ox.ac.uk (A.M. Winkler).
URL: http://www.fmrib.ox.ac.uk (A.M. Winkler).

[1] A *pivotal* statistic has a sampling distribution that does not depend on unknown parameters.

http://dx.doi.org/10.1016/j.neuroimage.2015.05.092

This arrangement, using a statistic that is robust to heteroscedasticity, the use of variance groups, and the imposition of restrictions on exchangeability through the use of EBS, allows inference on various designs that, otherwise, would be much more difficult to do non-parametrically. These designs include paired tests, longitudinal designs, and other common tests that involve repeated measurements. However, certain study designs, despite exhibiting well-structured dependence between observations, still cannot be accommodated in the above framework. This occurs when the overall covariance structure is known, but its exact magnitude is not. An example occurs when multiple measurements per subject are performed in more than one session, with more than one measurement per session: the measurements within session may be exchangeable, but not across sessions. Another example is for studies using siblings, such as designs using discordant sib-pairs (in which only one sibling is affected by a given disorder), or using twins: permutations that disrupt the constitution of any sibship cannot be performed, as this would violate exchangeability.

Studies such as these are relatively common, notably those that involve siblings. However, whereas in classical twin designs the central objective is to quantify the fraction of the variation in a measurement (trait) that can be explained by the familial relationship between subjects after potential confounds have been taken into account, a quantity known as *heritability*, here the concern is with a general linear model, and the objective is to test the influence of explanatory variables on the observed data. In other words, the interest lies on the relationship between the covariates and the main trait, while the non-independence between observations, which is a feature of interest in a heritability study, is here a form of nuisance that imposes restrictions on exchangeability for permutation inference for the GLM.

Rather than inadvertently breaking these restrictions, here we propose to test the null hypothesis using a subset of all otherwise possible permutations, only allowing the rearrangements that respect exchangeability, thus retaining original joint distribution unaltered.[2] As in our previous work, we treat observations or entire blocks of data as weakly exchangeable, but here we further extend the definition of EBS to allow more complex designs to be addressed. This is accomplished through the use of *multi-level exchangeability blocks*, in which levels consist of *nested* blocks; for each such block the state of within- or whole-block exchangeability can be specified. The blocks are defined hierarchically, based on information about the dependence within data, but not requiring the modelling of the actual dependency. Even though the possibility of using nested blocks was anticipated in Winkler et al. (2014) ("Whole-block and within-block can be mixed with each other in various levels of increasing complexity", page 386), nothing further was studied or presented at the time. Here we provide a comprehensive description of the approach, investigate its performance, its power, and present an applied example using the data structure of the ongoing Human Connectome Project (HCP). In the Appendix A, we present an implementation strategy.

Theory

Terminology

When contrasting the method described in this article with simple data rearrangement, various terms could be adopted: *single-level* vs. *multi-level* block shuffling, emphasising the levels of relationship between observations; *unrestricted* vs. *restricted*, emphasising the imposition of restrictions on how the data are allowed to be rearranged at each shuffling; *free* vs. *tree* shuffling, emphasising the tree-like structure of the relationships between observations that allow shuffling. All these terms have equivalent meaning in the context of this article, and are used interchangeably throughout. The generic terms *shuffling* and *rearrangement* are used when the distinction between permutations, sign flippings or permutations with sign flippings is not relevant.

Notation

We consider a GLM that can be expressed as $\mathbf{Y} = \mathbf{M}\boldsymbol{\psi} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the $N \times 1$ vector of observed data, $\mathbf{M}$ is the full-rank $N \times r$ design matrix that includes explanatory variables (i.e., effects of interest and possibly nuisance effects), $\boldsymbol{\psi}$ is the $r \times 1$ vector of r regression coefficients, and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of random errors. Estimates for the $\boldsymbol{\psi}$ can be computed by ordinary least squares, i.e., $\hat{\boldsymbol{\psi}} = \mathbf{M}^{+}\mathbf{Y}$, where the superscript $(^{+})$ denotes a pseudo-inverse. One generally wants to test the null hypothesis that a given combination (contrast) of the elements in $\boldsymbol{\psi}$ equals to zero, that is, $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi} = \mathbf{0}$, where $\mathbf{C}$ is a $r \times s$ full-rank matrix of s contrasts, $1 \le s \le r$. The commonly used F statistic can be computed as usual and used to test the null hypothesis. When $s = 1$, the Student's t statistic can be computed as $t = \mathsf{sign}\left(\hat{\boldsymbol{\psi}}\right)\sqrt{F}$. A p-value for the statistic is calculated by means of shuffling the data, the model, the residuals, or variants of these (Winkler et al., 2014, Table 2). In any of these cases, to allow rearrangements of the data, some assumptions need to be made: either of *exchangeable errors* (EE) or of *independent and symmetric errors* (ISE). The first allows permutations, the second sign flippings; if both are available for a given model, permutations and sign flippings can be performed together. These rearrangements are represented by permutation and/or sign flipping matrices $\mathbf{P}$, and the set of all such matrices allowed for a given design is denoted as $\mathcal{P}$.

At its simplest, the EBS for within- or whole-block exchangeability can be identified or represented by a set of indices $\{1,2\ldots,B\}$, one for each of the B blocks. A vector of size $N \times 1$, can be used to indicate to which EB each observation from $\mathbf{Y}$ belongs (Fig. 1, *left*); an extra flag is passed to the shuffling algorithm (such as the randomise algorithm) to indicate whether the rearrangements of the data should happen as within- or as whole-block. While this notation probably covers the majority of the most common study designs, it allows only within- *or* whole-block, but not *both* simultaneously; in other words, if in a study the observations can be permuted within block, and the blocks as a whole can also be permuted, such notation does not convey all possibilities for reorganising the data while preserving their joint distribution unaltered, and algorithms would perform fewer shufflings than those that are effectively allowed.

This can be addressed by extending the notation from a single column to a multi-column array, allowing nested EBS to be defined, such that blocks can contain sub-blocks, in a hierarchical fashion, and where each column represents a level; we use the leftward columns to indicate higher, and rightward to indicate lower levels. More columns alone, however, are not sufficient, because at each level, shufflings of observations or of sub-blocks can be allowed within-block, or the blocks at that level can be shuffled as a whole. Hence to discriminate between one type or the other, we use negative indices to indicate that the exchangeable units at the level immediately below should not be permuted, and positive indices indicate that shuffling of these units is allowed (Fig. 1, *right*). The exchangeable units can be sub-blocks, which can contain yet other sub-blocks, or observations if the next level immediately below is the last.

These two notations, i.e., using single- or multi-column indices, do not represent mathematical entities, and are not meant to be used for algebraic manipulation; rather, these notations are shorthand methods to represent structured relationships between observations. The covariance structure prevents unrestricted shuffling from being considered, but it often permits shufflings to happen in a certain orderly manner that preserves the joint distribution of the data. These notations are to be used by the algorithm that performs the test to construct the permutation and/or sign flipping matrices, which then can be used to effectively disarrange the model to construct the distribution of the statistic under the null hypothesis.

---

[2] Exchangeability with respect to a subset of all possible permutations is termed *weak exchangeability* (Good, 2005). For conciseness, we will use the solitary term "exchangeability", while making clear the subsets of permutations for which this is valid.