6.5 Solution to the multivariate BFP

Overview

Anderson *et al.* (2017) described a general dissimilarity-based solution to the multivariate Behrens-Fisher problem. Their solution uses a statistic (called 'F_2' therein) that is a modification of the original PERMANOVA pseudo *F* statistic (called 'F_1' therein). This modification is a direct dissimilarity-based multivariate analogue to a solution to the univariate BFP proposed by Brown & Forsythe (1974). The PERMANOVA routine in PRIMER is the only known software implementation of this method that correctly accounts for heterogeneity in multivariate dispersions *not only* in the construction of the test-statistic itself, *but also* in the algorithm used for permutations to estimate the *p*-value.

What is described below are the details of the one-way case, but for more complex ANOVA designs with multiple factors, the end-user must specify wherein the heterogeneity lies that needs to be accounted for, and the construction of the correct F statistic and associated permutation algorithm required to achieve rigorous inference in the context of the full study design is not trivial. We re-iterate: the PERMANOVA routine in PRIMER is the only software we know of that will do all of this correctly.

PERMANOVA in a nutshell

For a description of PERMANOVA, please see the original articles (Anderson (2001), McArdle & Anderson (2001)) and also the rather more recent and more thorough encyclopedia entry provided by Anderson (2017). We shall describe the original PERMANOVA test-statistic here, in brief, following Anderson *et al.* (2017), with notation that facilitates the description of the modified test.

Suppose $\bf Y$ is an $N \times p$ matrix of multivariate row vectors ${\bf y}_ {ij}$ corresponding to sampling units, each of length p and each belonging to one of $i = 1, \ldots, a$ groups, with $j = 1, \ldots, n_{i}$ sampling units (rows) in the ith group and $N = \sum_{i=1}^a n_i$. Let $\bf D$ be an $N \times N$ symmetric matrix of dissimilarities $\lbrace d_{ij,i'j'} \rbrace$ calculated between every pair of sampling units.

Next, as in Gower (1966), let matrix $\bf A$ be comprised of elements $\lbrace a_{ij,i'j'}\rbrace$ = $\lbrace -0.5 \times d_{ij,i'j'}^2 \rbrace$, then define matrix $\bf G$ (a centred version of matrix $\bf A$) as: $$ {\bf G} = ({\bf I} - (1/N) {\bf J}_ {\scriptscriptstyle N}) {\bf A} ({\bf I} - (1/N) {\bf J}_ {\scriptscriptstyle N}) $$ where ${\bf J}_ {\scriptscriptstyle N}$ denotes an $N \times N$ matrix of 1s and $\bf I$ denotes an $N \times N$ identity matrix.

For the one-way case, let $\bf X$ be a $N \times r$ matrix of full rank $r = (a-1)$ containing orthogonal contrasts among the groups. We can construct a linear projection matrix for the design as: $$ {\bf H} = {\bf X} [{\bf X}'{\bf X}]^{-1} {\bf X}' $$ Then, the PERMANOVA pseudo F statistic for comparing the centroids among the a groups is: $$ F_1 = \frac { \text{tr} ({\bf HG})/(a-1) } { \text{tr} [{\bf (I-H) G}] /(N-a) } $$

where '$ \text{tr}(\cdot)$' denotes the trace (sum of diagonal elements) of a matrix. Note that if $p = 1$ and $\bf D$ is calculated using Euclidean distances, then $F_1 = F_0$, the classical univariate F ratio.

Test by permutation

We may calculate a *p*-value to test the null hypothesis of equality of centroids in the space of the chosen dissimilarity measure under the sole assumption that the sampling units (rows) are exchangeable among the a groups. First, we calculate an observed value of the test statistic, F_1, with the rows of the data (sampling units) in their original order. Then, we randomly permute (re-order) the $1,\ldots,N$ rows of matix $\bf Y$ to obtain a matrix of permuted data ${\bf Y}^\pi$, yet leaving the grouping structure fixed (i.e., the original ordering is retained in matrix $\bf X$ and hence also in the projection matrix $\bf H$). In other words, under a true null hypothesis, any ordering of the sampling units across the groups is equally likely *via* exchangeability. Indeed, exchangeability is the only assumption of the PERMANOVA test, which is distribution-free.

Re-calculation of F_1 replacing $\bf Y$ with ${\bf Y}^\pi$ yields F_1^π, a value of the test-statistic under permutation. Repeating this random permutation and re-calculation a large number of times (say, $n_\text{perm} = 9999$), yields a distribution of values of F_1^π from which we can empirically calculate a p-value as $P = \text{Pr}(F_1^\pi \ge F_1)$. Specifically, the *p*-value is calculated directly as:

$$ P = \frac{(\text{no. of } F_1^{\pi} \ge F_1) + 1}{ (n_\text{perm} + 1) } $$

This mirrors what we saw for the test by permutation for univariate ANOVA (and so many other permutation tests offered in PRIMER); namely, that the $+1$ in each of the numerator and denominator are there simply to acknowledge the inclusion of the original (genuine) ordering of the data as one of the possible 'random' outcomes we could have obtained - they would not be needed in the equation if *all* possible orderings were to be done systematically and exhaustively.

Modified PERMANOVA to account for heterogeneity

Anderson *et al.* (2017) suggested a modification to the PERMANOVA pseudo F statistic to account for heterogeneity, following directly from the univariate solution to the BFP proposed by Brown & Forsythe (1974) . Specifically, instead of F_1, we can use:

$$ F_2 = \frac { \text{tr} ({\bf HG}) } { \sum_{i=1}^a (1-n_i/N) \cdot V_i } $$

where V_i is the within-group dispersion for group i, defined as:

$$ V_i = \sum_{j=1}^{(n-1)} \sum_{j'=(j+1)}^n d_{ij,i'j'}^2 / [n_i(n_i-1)] $$

Some important things to note about this modified test-statistic are:

- The null hypothesis for F_2 is equality of centroids in the space of the chosen dissimilarity measure *given* potential differences in dispersions among the groups.
- The potential for heterogeneity is explicitly acknowledged in the modified test-statistic, F_2 by the calculation of separate individual dispersions (V_i) for each group.
- F_2 is equivalent to F_1 if sample sizes are equal across all groups. This is sensible, as PERMANOVA (like ANOVA) is very robust to heterogeneity of dispersions when the design is balanced.
- F_2, like F_1, is carefully constructed so that the numerator and denominator *have the same expectation* when the null hypothesis is true.
- If we are dealing with univariate data, so $p = 1$ response variable, and the entries in ${\bf D}$ are Euclidean distances, then V_i is the usual classical univariate unbiased measure of the sample variance (s^2) for group i.
- If PERMANOVA is run using F_2, then the *degrees of freedom* are also modified to reflect the Satterthwaite approximation given by Brown & Forsythe (1974) . This is to ensure compatibility between the PERMANOVA implementation and the solution provided by Brown & Forsythe (1974) for univariate cases, but in practice the *p*-value itself is always calculated in PERMANOVA using permutation algorithms (see below), so there is no direct consequence of this change in the degrees of freedom (which will remain constant under permutation) for the level of significance in the outcome.

Obtaining a *p*-value for the modified test

On the face of it, we would not expect to be able to do a permutation test for F_2, because how can we view the sampling units as *exchangeable* among the groups if we know already that the groups have different dispersions? Some alternative method, such as a separate-sample bootstrap, either with or without some kind of bias-adjustment (e.g., Efron & Tibshirani (1993) , Manly (2006)), would seem to be more appropriate to use here, at least conceptually. Simulation work by Anderson *et al.* (2017) demonstrated, however, that the modified test based on F_2 had better statistical behaviour when permutations were done to obtain the *p*-value, rather than bootstraps. To be specific, the Type I error was closer to the nominated significance level and the distribution of *p*-values under a true null hypothesis was more uniform (as is desirable) when tests were done using permutations. In contrast, the results obtained using bootstrapping methods were always much more conservative, and the degree of conservatism increased with increasing dimensionality, increasing degree of heterogeneity and increasing differences in sample sizes among groups. In addition, under all simulation scenarios, tests by permutation using F_2 had the greatest power, matching or exceeding that of F_1, compared to bootstrap alternatives.

Thus, PERMANOVA in PRIMER that is done using F_2 to account for heterogeneity calculates *p*-values using permutation algorithms rather than using bootstrapping methods.

Revision #103
Created 7 April 2025 22:52:53 by Marti

Updated 30 November 2025 01:59:47 by Marti