Permutation inference for canonical correlation analysis

Anderson M. Winkler [a,*], Olivier Renaud [b], Stephen M. Smith [c], Thomas E. Nichols [d]

[a] National Institute of Mental Health (NIMH), National Institutes of Health (NIH), Bethesda, MD, USA
[b] Methodology and Data Analysis, Department of Psychology, University of Geneva, Switzerland
[c] Wellcome Trust Centre for Integrative Neuroimaging (WIN-FMRIB), University of Oxford, Oxford, UK
[d] Big Data Institute, University of Oxford, Oxford, UK

ARTICLE INFO

ABSTRACT

Canonical correlation analysis (CCA) has become a key tool for population neuroimaging, allowing investigation of associations between many imaging and non-imaging measurements. As age, sex and other variables are often a source of variability not of direct interest, previous work has used CCA on residuals from a model that removes these effects, then proceeded directly to permutation inference. We show that a simple permutation test, as typically used to identify significant modes of shared variation on such data adjusted for nuisance variables, produces inflated error rates. The reason is that residualisation introduces dependencies among the observations that violate the exchangeability assumption. Even in the absence of nuisance variables, however, a simple permutation test for CCA also leads to excess error rates for all canonical correlations other than the first. The reason is that a simple permutation scheme does not ignore the variability already explained by previous canonical variables. Here we propose solutions for both problems: in the case of nuisance variables, we show that transforming the residuals to a lower dimensional basis where exchangeability holds results in a valid permutation test; for more general cases, with or without nuisance variables, we propose estimating the canonical correlations in a stepwise manner, removing at each iteration the variance already explained, while dealing with different number of variables in both sides. We also discuss how to address the multiplicity of tests, proposing an admissible test that is not conservative, and provide a complete algorithm for permutation inference for CCA.

1. Introduction

Canonical correlation analysis (CCA) (Jordan, 1875; Hotelling, 1936) is a multivariate method that aims at reducing the correlation structure between two sets of variables to the simplest possible form (hence the name "canonical'') through linear transformations of the variables within each set. Put simply, given two sets of variables, the method seeks linear mixtures within each set, such that each resulting mixture from one set is maximally correlated with a corresponding mixture from the other set, but uncorrelated with all other mixtures in either set.

From a peak use through from the late 1970's until mid-1980's, the method has recently regained popularity, presumably thanks to its ability to uncover latent, common factors underlying association between multiple measurements obtained, something relevant in recent research, particularly in the field of brain imaging, that uses high dimensional phenotyping and investigates between-subject variability across multiple domains. This is in contrast to initial studies that introduced CCA to the field (Friston et al., 1995, 1996; Worsley, 1997; Friman et al., 2001,

2002, 2003) for investigation of signal variation in functional magnetic resonance imaging (fMRI) time series. For example, Smith et al. (2015) used CCA to identify underlying factors associating brain connectivity features to various demographic, psychometric, and lifestyle measures; Rosa et al. (2015) used a sparse CCA method to investigate differences in brain perfusion after administration of two distinct antipsychotic drugs; Miller et al. (2016) used CCA to identify associations between imaging and non-imaging variables in the UK Biobank; Drysdale et al. (2017) used CCA to investigate associations between brain connectivity and clinical assessments, and found two canonical variables that would allow classification of participants into distinct categories (but see Dinga et al., 2019); Kernbach et al. (2018) used CCA to identify connectivity patterns in the default mode network associated with patterns of connectivity elsewhere in the brain; Bijsterbosch et al. (2018), Xia et al. (2018), and Mihalik et al. (2019) likewise used CCA to identify associations between functional connectivity and various indices of behaviour and psychopathology, whereas Sui et al. (2018) used a combination of multivariate methods, including CCA, to investigate brain networks associated with composite

* Corresponding author.
   E-mail address: anderson.winkler@nih.gov (A.M. Winkler).

cognitive scores; Li et al. (2019) used CCA to investigate, among subjects, the topography of the global fMRI signal and its relationship with a number of cognitive and behavioral measurements; Ing et al. (2019) used CCA to identify symptom groups that were correlated with brain regions assessed through a diverse set of imaging modalities; Alnæs et al. (2020) used CCA to investigate the association between imaging measurements and cognitive, behavioral, psychosocial and socioeconomic indices; Clemens et al. (2020) used a combination of pattern classification algorithms and CCA to study imaging and behavioral correlates of the subjective perception of oneself belonging to a particular gender. In most of these between-subject, group level studies, putative nuisance variables or confounds were regressed out from the data before proceeding to inference, and all of them used some form of permutation test to assess the significance of the results. In a recent review, Wang et al. (2020) described a permutation procedure for CCA as "a random shuffling of the rows (…) of the two variable sets".

Permutation tests are well known and widely used. Among their many benefits, these tests lead to valid inferences while requiring assumptions that are commonly satisfied in between-subject analyses, such that of exchangeability of observations under the null hypothesis. However, here we show that simple implementations of permutation inference for CCA are inadequate on four different grounds. First, simple, uncorrected permutation p-values are not guaranteed to be monotonically related to the canonical correlations, leading to inadmissible results; for the same reason, multiple testing correction using false discovery rate is also inadmissible. Second, except for the highest canonical correlation, a simple one-step estimation of all others without considering the variability already explained by previous canonical variables in relation to each of them also leads to inflated per comparison error rates and thus, invalid results. Third, regressing out nuisance variables without consideration to the introduction of dependencies among observations caused by residualisation leads to an invalid test, with excess false positives. Fourth, multiple testing correction using the distribution of the maximum test statistic leads to conservative results, except for the highest canonical correlation.

In this paper we explain and discuss in detail each of these problems, and offer solutions that address each of them. In particular, we propose a stepwise estimation method for the canonical correlations and canonical variables that remains valid even when the number of variables is not the same for both sides of CCA. We propose a method that transforms residualised data to a lower dimensional basis where exchangeability — as required for the validity of permutation tests — holds. We also propose that inference that considers multiple canonical correlations should use a closed testing procedure that is more powerful than the usual correction method used in permutation tests that use the distribution of the maximum statistic; the procedure also ensures a monotonic relationship between p-values and canonical correlations. Finally, we provide a complete, general algorithm for valid inferences for CCA.

2. Theory

2.1. Notation and general aspects

Thorough definition and derivation of CCA can be found in many classical textbooks on multivariate analysis (e.g., Kendall, 1975; Mardia et al., 1979; Brillinger, 1981; Muirhead, 1982; Seber, 1984; Krzanowski, 1988; Anderson, 2003); the reader is referred to these for a comprehensive overview. Here we present concisely and only allude to the distinction between population (ρ) and sample (r) canonical correlations where strictly needed. Let $\mathbf{Y}_{N \times P}$ and $\mathbf{X}_{N \times Q}$ be each one a collection of, respectively, P and Q variables observed from N subjects, $N \geq P + Q$. Without loss of generality, assume that $P \leq Q$, that the columns of $\mathbf{Y}$ and $\mathbf{X}$ are mean-centered, that these matrices are of full rank, and define $E(\mathbf{Y}'\mathbf{Y}) = \mathbf{\Sigma_{YY}}$, $E(\mathbf{X}'\mathbf{X}) = \mathbf{\Sigma_{XX}}$, and $E(\mathbf{Y}'\mathbf{X}) = \mathbf{\Sigma_{YX}} = \mathbf{\Sigma'_{XY}}$. The goal of CCA is to identify *canonical coefficients* or *canonical weights* $\mathbf{A}_{P \times K} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ and $\mathbf{B}_{Q \times K} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$, $K = \min(P,Q)$, such that the pairs $(\mathbf{u}_k, \mathbf{v}_k)$ of

canonical variables, defined as:

$$
\begin{aligned}
[\mathbf{u}_1, \ldots, \mathbf{u}_K] &= \mathbf{U}_{N \times K} = \mathbf{YA} \\
[\mathbf{v}_1, \ldots, \mathbf{v}_K] &= \mathbf{V}_{N \times K} = \mathbf{XB}
\end{aligned}
\tag{1}
$$

have correlations r_k that are maximal, under the constraint that $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$. Estimation of $\mathbf{A}$ and $\mathbf{B}$ amounts to finding the K solutions to:

$$
\begin{bmatrix} -r_k \mathbf{\Sigma_{YY}} & \mathbf{\Sigma_{YX}} \\ \mathbf{\Sigma_{XY}} & -r_k \mathbf{\Sigma_{XX}} \end{bmatrix} \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} = 0
\tag{2}
$$

where the unknowns are r_k, $\mathbf{a}_k$ and $\mathbf{b}_k$; r_k are the sample *canonical correlations*, i.e., the correlations between the estimated canonical variables $\mathbf{u}_k$ and $\mathbf{v}_k$. The coefficients $\mathbf{a}_k$ are eigenvectors of $\mathbf{\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}$, whereas $\mathbf{b}_k$ are eigenvectors of $\mathbf{\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}}$; the respective eigenvalues (for either $\mathbf{a}_k$ or $\mathbf{b}_k$, as these eigenvalues are the same) are r_k^2. For convenience, we call *canonical component* the ensemble formed by the k-th canonical correlation, its corresponding pair of canonical variables, and associated pair of canonical coefficients; canonical variables may also be termed *modes of variation*.

The typical method for estimation involves an iterative procedure that finds one r_k and $\mathbf{a}_k$ at a time, with $\mathbf{b}_k$ computed as a function of these. However, the method proposed by Bjorck and Golub (1973) is more concise and numerically more stable; it is described in the Appendix (Algorithm 3). The canonical correlations are then produced in descending order, $r_1 \geq \ldots \geq r_k \geq \ldots \geq r_K \geq 0$; this positiveness of all canonical correlations is a consequence of these values being explicitly maximised during estimation; reversal of the sign of the coefficients $\mathbf{a}_k$ can always be accompanied by the reversal of the sign of the corresponding coefficients $\mathbf{b}_k$ in the other side (and of $\mathbf{u}_k$ and $\mathbf{v}_k$), to no net effect on r_k. That is, the signs of the canonical variables and coefficients are indeterminate, and any solution is arbitrary; nothing can be concluded about the specific direction of effects with CCA.

2.2. Parametric inference

The distribution of the sample canonical correlations r_k is intractable, even under assumptions of normality and independence among subjects, and is a function of the population correlations ρ_k (Constantine, 1963; James, 1964). This difficulty led to the development of a rich asymptotic theory (Fisher, 1939; Hsu, 1941; Lawley, 1959; Fujikoshi, 1977; Glynn and Muirhead, 1978). However, these approximations have been shown to be extremely sensitive to departures from normality, or require additional terms that further complicate their use (Muirhead and Waternaux, 1980); Brillinger (1981) recommended resampling methods to estimate parameters used by normal approximations, which otherwise can be biased (Anderson, 2003). These difficulties pose challenges for inference. Even though some computationally efficient algorithms have been proposed (Koev and Edelman, 2006), these tests continue to be rarely used.

Instead, a test based on whether a certain number of correlations are equal to zero has been proposed. The null hypothesis is $\mathscr{H}_k^0 : \rho_k = \rho_{k+1} = \ldots = \rho_K = 0$, i.e., $\mathscr{H}_k^0 : \wedge_{i=k}^K \rho_i = 0$, for $1 \leq k \leq K$, that is, the null is that $K - k + 1$ population canonical correlations (the smaller ones) are zero (Bartlett, 1938, 1947; Marriott, 1952; Lawley, 1959; Fujikoshi, 1974), versus the alternative that at least one is not, i.e., $\mathscr{H}_k^1 : \vee_{i=k}^K \rho_i > 0$. The test is based on the statistic proposed by Wilks (1935), as:

$$
\lambda_k = -\left(N - C - \frac{P + Q + 3}{2}\right) \ln\left(\prod_{i=k}^K \left(1 - r_i^2\right)\right)
\tag{3}
$$

where the constant $C = 0$ if there are no nuisance variables (Section 2.6). Under the null hypothesis, λ_k follows an approximate χ^2 distribution with degrees of freedom $\nu = (P - k + 1)(Q - k + 1)$ if each of the columns of $\mathbf{Y}$ and $\mathbf{X}$ have values that are independent and identically distributed following a normal distribution (but see Glynn and Muirhead, 1978, for a

different expression). Unfortunately, this test is sensitive to departures from normality, particularly in the presence of outliers, and its use has been questioned (Seber, 1984; Harris, 1976).

Another test statistic is based on Roy (1953)[1] and is simply:

$$\theta_k = r_k^2 \tag{4}$$

The critical values for the corresponding parametric distribution at a given test level α can be found in the charts provided by Heck (1960), using as parameters $s = \min(P,Q) - k + 1$, $m = (|P - Q| - 1)/2$, and $n = (N - C - P - Q - 2)/2$ (Lee, 1978), where the constant $C = 0$ if there are no nuisance variables (Section 2.6), or in tables provided by Kres (1975); more recent approximations for normally distributed data can be found in Chiani (2016) and Johnstone and Nadler (2017). Some approximations, however, are considered conservative (Harris, 1976, 2013). Note that, while Roy (1953) proposed the use of the largest value as test statistic, which would then be $r_{k=1}^2$, any given null $\mathcal{H}_k^0$ at position k must have already considered the previous canonical components, from 1 until $k - 1$, such that the maximum statistic, after the previous canonical correlations have been removed from the model, is always the current one. A similar reasoning holds for the smallest canonical correlations in the test proposed by Wilks (1935). This feature is exploited in the stepwise rejective procedure proposed in Section 2.5.

2.3. Permutation inference

The above problems can be eschewed with the use of resampling methods, such as permutation. An intuitive (but inadequate) permutation test for CCA could be constructed by randomly permuting the rows of **Y** or **X**. For each shuffling of the data, indicated by $j = \{1, ..., J\}$, a new set of canonical correlations $(r_k)_j^*$ and associated statistics $(\lambda_k)_j^*$ would be computed. A p-value would be obtained as $p_k = \frac{1}{J}\sum_{j=1}^{J} I[(\lambda_k)_1 \geq (\lambda_k)_j^*]$, where $I[\cdot]$ is the indicator (Kronecker) function, which evaluates as 1 if the condition inside the brackets is true, or 0 otherwise, and the index $j = 1$ corresponds to the unpermuted data (i.e., no permutation, with the data in their original ordering).

Such a naïve procedure, however, would ignore the fact that, this resampling scheme matches the first null hypothesis $\mathcal{H}_1^0$, but not the subsequent ones. For a given canonical correlation at position k being tested, $k > 1$, one must generate a permutation that satisfies the corresponding null $\mathcal{H}_k^0$, but not necessarily $\{\mathcal{H}_1^0, ..., \mathcal{H}_{k-1}^0\}$. Otherwise, latent effects represented by the corresponding earlier canonical variables $[\mathbf{u}_1, ..., \mathbf{u}_{k-1}]$ and $[\mathbf{v}_1, ..., \mathbf{v}_{k-1}]$ would, in the procedure above, remain in the **Y** and **X** at the time these are permuted. However, the variance associated with these earlier canonical variables would have already been explained through the rejection of their respective null hypothesis up to $\mathcal{H}_{k-1}^0$. This variance is now a nuisance for the positions from k (inclusive) onward. It contains information that are not pertinent to position k and subsequent ones, and that therefore should not be used to build the null distribution. That is variance should not be re-used in the shufflings that lead to a given $(r_k)_j^*$ or subsequent correlations.

Fortunately, CCA is invariant to linear transformations that mix the variables in **Y** or in **X**. Since the canonical variables are themselves linear transformations of these input variables (Equation (1)), a second CCA using **U** and **V** in place of the initial **Y** and **X** leads to the same solutions. Yet, unless $P = Q$, **V** will not span the same space as **X**. In principle, this would be inconsequential as far as the canonical variables are concerned. However, ignoring the variability in **X** not contained in **V** would again affect the p-values should **U** and **V** be used in a permutation test, as the

permuted data would not be representative of the original (unpermuted) that led to these initial canonical variables. To mitigate the problem, include into the matrix of canonical coefficients their orthogonal complement, i.e., compute $\tilde{\mathbf{V}} = \mathbf{X}[\mathbf{B}, \text{null}(\mathbf{B}')]$, then use $\tilde{\mathbf{V}}$ instead of **V** as a replacement for **X**. In this paper we adopted the convention that $P \leq Q$, but the same procedure works in reverse and, algorithmically, it might as well be simpler to compute also a $\tilde{\mathbf{U}} = \mathbf{Y}[\mathbf{A}, \text{null}(\mathbf{A}')]$ and use it instead of **U** as a replacement for **Y**. If $P \leq Q$, then $\tilde{\mathbf{U}} = \mathbf{U}$.

While these transformations do not change in any way the canonical components, they allow the construction of an improved algorithm that addresses the issue of variability already explained by canonical variables of lower rank (i.e., the ones with order indices smaller than that of a given one). It consists of running an initial CCA using **Y** and **X** to obtain $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$, then subject these to a second CCA and permutation testing while, crucially, at each permutation, iteratively repeating CCA K times, each using not the whole $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$, but only from the k-th component onwards, i.e., $[\tilde{\mathbf{u}}_k, ..., \tilde{\mathbf{u}}_P]$ and $[\tilde{\mathbf{v}}_k, ..., \tilde{\mathbf{v}}_Q]$ for the test about the k-th canonical correlation. Of note, a test level α does not need to be specified at the time in which the above iterative (stepwise) procedure is performed; instead, and in combination with the multiple testing procedure described below, adjusted p-values are are computed, which then are used to accept or reject the null for the k-th correlation. Algorithm 1 (Section 2.8) shows the procedure in detail (the algorithm contains other details that are discussed in the next sections).

A number of further aspects need be considered in permutation tests: the number of possible reorderings of the data, the need for permutations that break the association between the variables being tested, the random selection of permutations from the permutation set when not all possible permutations can be used, the choice of the test statistic, how to correct for the multiplicity of tests, the number of permutations to allow narrow confidence intervals around p-values, among others. These topics have been discussed in Winkler et al. (2014, 2016) and references therein and will not be repeated here. However, for CCA, some aspects deserve special treatment and are considered below.

2.4. Choice of the statistic

Asymptotically, using Wilks' statistic λ_k or Roy's θ_k are expected to lead to the same conclusion since all correlations are sorted in descending order: if $r_k = 0$, then all subsequent ones must be zero; likewise, if $r_k > 0$, then clearly at least one correlation between k and K is larger than zero, which has to include r_k itself. Moreover, permutation under the null is justifiable in the complete absence of association between the two sets, which implies that, under the null $\mathcal{H}_k^0$, all correlations $r_k, r_{k+1}, ..., r_K$ are equal to zero. With finite data, however, one statistic can be more powerful than the other in different settings; their relative performance is studied in Sections 3 and 4.

Computationally, Wilks' requires more operations to be performed compared to Roy's statistic. Since the relationship between r_k and θ_k is monotonic, the two are permutationally equivalent, and using r_k alone is sufficient, which makes Roy's the absolute fastest. However, even for Wilks', the amount of computation required is negligible compared to the overall number of operations needed for estimation of the canonical coefficients, such that in practice, the choice between the two should be in terms of power.

In either case, while inference refers to the respective null hypothesis at position k, it is not to be understood as inference on the index k. Rather, the null is merely conditional on the nulls for all previous correlations from 1 to $k - 1$ having been rejected. Rejecting the null implies that the correlation observed at position k is too high under the null hypothesis of no association between the two variable sets after all previous (from 1 to $k - 1$) canonical variables have been sequentially removed, as described in Section 2.3. In Algorithm 1 (Section 2.8) this is done in line 29, that uses as inputs to CCA the precomputed canonical variables only from position k onwards, as opposed to all of them.

---

[1] Roy (1953) proposed two distinct but related test statistics; these are both known as "Roy's largest root''. Here we use the one that is interpreted as a coefficient of determination, and not the other that is interpreted as an *F*-statistic. See Kuhfeld (1986) for a complete discussion.

2.5. Multiplicity of tests

For either of these two test statistics, the ordering of the canonical correlations from larger (farther from zero) to smaller (closer to zero) imply that rejection of the null hypothesis at each k must happen sequentially, starting from $k = 1$, using the respective test statistic and associated p-value until the null $\mathscr{H}_k^0$, for some $k = \{1, ..., K\}$, is not rejected at a predefined test level α. Then, at that position k, the procedure stops, and the null is retained from that position (inclusive) onward until the final index K.

The ordering of the canonical correlations brings additional consequences. First, because rejection of $\mathscr{H}_k^0$ implies rejection of all joint (intersection) hypotheses that include $\mathscr{H}_k^0$, that is $\mathscr{H}_1^0, ..., \mathscr{H}_{k-1}^0$, such sequentially rejective procedure is also a closed testing procedure (CTP), which controls the amount of any type I error across all tests, i.e., the *familywise error rate* (FWER) in the strong sense (Marcus et al., 1976; Hochberg and Tamhane, 1987). Thus, there is no need for further correction for multiple testing. Another way of stating the same is that the test for a given r_k, $k > 1$, has been "protected" by the test at the position $k = 1$ at the level α. Adjusted p-values (in the FWER sense) can be computed as $[p_k]_{\text{FWER}} = \max(p_1, ..., p_k)$, that is, the FWER-adjusted p-value for the canonical component k is the cumulative maximum p-value up to position k. Such adjusted p-values can be considered significant if their value is below the desired test level α.

The second consequence is that FWER adjustment of p-values using the distribution of the maximum statistic (not to be confused with the cumulative maximum described in the above paragraph) will be conservative for all canonical components except the first. The reason is that the maximum statistic is always the distribution of the first, which is stochastically dominant over all others. The distribution of the maximum is usually sought as a shortcut to a CTP when the condition of subset pivotality holds (Westfall and Young, 1993), as that reduces the computational burden from 2^K tests to only K tests. Interestingly, the ordering of the canonical correlations from largest to smallest leads to a CTP that does not use the distribution of the maximum, and yet requires only K tests, regardless of whether subset pivotality holds.

A third consequence is that using permutation p-values outside the above sequentially rejective procedure that controls FWER is not appropriate since these simple, uncorrected p-values are not guaranteed to be monotonically related to the canonical correlations r_k. Attempts to use these uncorrected p-values outside a CTP would lead to paradoxical results whereby higher, stronger canonical correlations might not be considered significant, yet later ones, smaller, weaker, could be so; that is, it would make the test *inadmissible* (Lehmann and Romano, 2005, p. 232). For the same reason, such simple p-values should not be subjected to correction using false discovery rate (FDR; Benjamini and Hochberg, 1995), because the ordering of p-values for FDR, from smallest to largest, is not guaranteed to match the ordering of the canonical correlations, leading similarly to an inadmissible test.

2.6. Nuisance variables

Few authors discussed nuisance variables or confounds in canonical correlation analysis, e.g., Roy (1957); Rao (1969); Timm and Carlson (1976); Lee (1978); Sato et al. (2010). Let the $\mathbf{Z}$ be a $N \times R$ matrix of nuisance variables, including an intercept. *Partial* CCA consists of considering $\mathbf{Z}$ nuisance for both $\mathbf{Y}$ and $\mathbf{X}$. This is distinct from *part* CCA, which consists of considering $\mathbf{Z}$ a nuisance for either $\mathbf{Y}$ or $\mathbf{X}$, but not both. Finally, *bipartial* CCA consists of considering $\mathbf{Z}$ a nuisance for $\mathbf{Y}$, while considering another set of variables $\mathbf{W}$, of size $N \times S$, a nuisance for $\mathbf{X}$. In all three cases, such nuisance variables can be regressed out from the respective set of variables of interest, then the respective residuals subjected to CCA (Table 1). In the parametric case, inference can proceed

Table 1

Taxonomy of canonical correlation analysis with respect to nuisance variables. In all cases, the aim is to find linear combinations of variables in the left and in the right sets, such that each combination from one set is maximally correlated with a corresponding combination from the other, but uncorrelated with all other combinations from either set.

Name	Left set	Right set
CCA ("full", no nuisance)	$\mathbf{Y}$	$\mathbf{X}$
Partial CCA	$\mathbf{R_Z Y}$	$\mathbf{R_Z X}$
Part CCA	$\mathbf{R_Z Y}$	$\mathbf{X}$
Bipartial CCA	$\mathbf{R_Z Y}$	$\mathbf{R_W X}$

$\mathbf{R_Z}$ is a residual forming matrix that considers the nuisance variables in $\mathbf{Z}$, and is computed as $\mathbf{I}_{N \times N} - \mathbf{Z Z^+}$, where the symbol $^+$ represents a pseudo-inverse. $\mathbf{R_W}$ is computed similarly, considering the nuisance variables in $\mathbf{W}$. The choice which set is on left or right side is arbitrary.

using the distribution of λ_k or θ_k (Equations (3) and (4)) using $C = R$ for partial or part, and $C = \max(R, S)$ for bipartial CCA (Timm and Carlson, 1976; Lee, 1978).

Permutation inference, however, requires further considerations, otherwise, as shown in Section 4, results will be invalid. Consider first the case without nuisance variables. Let $\mathbf{M} = [\mathbf{Y}, \mathbf{X}]_{N \times (P+Q)}$ be the horizontal concatenation of the two sets of variables whose association is being investigated. Both $\mathbf{Y}$ and $\mathbf{X}$ occupy an N-dimensional space and, therefore, so does $\mathbf{M}$. A random permutation of the rows of either of the two sets of variables will not affect their dimensionalities. For example $\mathbf{M}^* = [\mathbf{PY}, \mathbf{X}]$ continues to occupy the same N-dimensional space as $\mathbf{M}$.

However, residualisation changes this scenario. Let $\mathbf{R_Z} = \mathbf{I} - \mathbf{Z Z^+}$ be the residual forming matrix associated with the nuisance variables $\mathbf{Z}$, with the symbol $^+$ representing the Moore–Penrose pseudo-inverse. $\mathbf{R_Z}$ has the following interesting properties: $\mathbf{R_Z} = \mathbf{R_Z}'$ (symmetry) and $\mathbf{R_Z R_Z} = \mathbf{R_Z}$ (idempotency), both of which will be exploited later. In partial CCA, $\mathbf{Z}$ can be regressed out from $\mathbf{Y}$ and $\mathbf{X}$ by computing $\tilde{\mathbf{Y}} = \mathbf{R_Z Y}$ and $\tilde{\mathbf{X}} = \mathbf{R_Z X}$. Let $\tilde{\mathbf{M}} = [\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}]$ be the concatenation of the residualised sets $\mathbf{Y}$ and $\mathbf{X}$ with respect to $\mathbf{Z}$. While $\mathbf{Y}$ occupies an N-dimensional space, $\tilde{\mathbf{Y}}$ occupies a smaller one; its dimensions are, at most, of a size given by the rank of $\mathbf{R_Z}$, which is $N - R$ assuming $\mathbf{Y}$ and $\mathbf{Z}$ are of full rank. The same holds for $\mathbf{X}$ and $\tilde{\mathbf{X}}$ and, therefore, for $\mathbf{M}$ and $\tilde{\mathbf{M}}$.

Permutation affects these relations: while $\mathbf{M}^*$ still occupies a space of N dimensions as the unpermuted $\mathbf{M}$, $\tilde{\mathbf{M}}^*$, differently than $\tilde{\mathbf{M}}$, may now occupy a space with dimensions anywhere between $N - R$ and N, depending on a given random permutation. With fewer effective observations determined by this lower space after residualisation, and the same number of variables, the sample canonical correlations in the unpermuted case are stochastically larger than in the permuted, which in turn leads to an excess of spuriously small p-values. For not occupying the same space as the original, the permuted data are no longer a similar realisation of the unpermuted, thus violating exchangeability, and specifically causing the distribution of the test statistics to be unduly shifted to the left.

Here the following solution is proposed: using the results from Huh and Jhun (2001), let $\mathbf{Q_Z}$ be a $N \times N'$ semi-orthogonal basis (Abadir and Magnus, 2005, p. 84) for the column space of $\mathbf{R_Z}$ constructed via, e.g., spectral or Schur decomposition, such that $\mathbf{Q_Z Q_Z}' = \mathbf{R_Z}$, $\mathbf{Q_Z}' \mathbf{Q_Z} = \mathbf{I}_{N' \times N'}$, where $N' = N - R$, and $\mathbf{Q_Z}' = \mathbf{Q_Z^+}$. Then CCA on $\widetilde{\widetilde{\mathbf{Y}}} = \mathbf{Q_Z}'\mathbf{Y}$ and $\widetilde{\widetilde{\mathbf{X}}} = \mathbf{Q_Z}'\mathbf{X}$ leads to the same solutions as on $\tilde{\mathbf{Y}} = \mathbf{R_Z Y}$ and $\tilde{\mathbf{X}} = \mathbf{R_Z X}$. The reason is that, from Section 2.1, $E((\mathbf{R_Z Y})'(\mathbf{R_Z Y})) = E(\mathbf{Y}'\mathbf{R_Z}'\mathbf{R_Z Y}) = E(\mathbf{Y}'\mathbf{R_Z Y})$, which is the same as $E((\mathbf{Q_Z}'\mathbf{Y})'(\mathbf{Q_Z}'\mathbf{Y})) = E(\mathbf{Y}'\mathbf{Q_Z Q_Z}'\mathbf{Y}) = E(\mathbf{Y}'\mathbf{R_Z Y})$, since, as discussed earlier, $\mathbf{R_Z}$ is symmetric and idempotent, and $\mathbf{Q_Z Q_Z}' = \mathbf{R_Z}$. In a similar manner, $E((\mathbf{Q_Z}'\mathbf{X})'(\mathbf{Q_Z}'\mathbf{X})) = E(\mathbf{X}'\mathbf{R_Z X})$,

and likewise, $E((\mathbf{Q_Z}'\mathbf{X})'(\mathbf{Q_Z}'\mathbf{Y})) = E(\mathbf{X}'\mathbf{R_Z}\mathbf{Y})$. While pre-multiplication by $\mathbf{Q_Z}'$ does not affect the CCA results,[2] it changes the dependence structure among the rows of the data: $\widetilde{\mathbf{M}} = [\widetilde{\mathbf{Y}}, \widetilde{\mathbf{X}}]$ occupies an N'-dimensional space, and so does $\widetilde{\mathbf{M}}^* = [\mathbf{P}\widetilde{\mathbf{Y}}, \widetilde{\mathbf{X}}]$, for a permutation matrix $\mathbf{P}$ of size N' $\times N$', such that exchangeability holds, thus allowing a valid permutation test.

The treatment of partial CCA, as described above, can be seen as a particular case of bipartial CCA in which $\mathbf{W} = \mathbf{Z}$, that is, the set of nuisance variables in both sides is the same. Of course, for bipartial CCA proper, this equality does not necessarily hold, and the two sets may be different in different ways: $\mathbf{Z}$ may be entirely orthogonal to $\mathbf{W}$, or some or all variables from one set may be fully represented in the other, either directly (e.g., some of the variables present in both sets), or as linear combinations of one set in the other, or it may be that these two sets are simply not orthogonal. The direct strategy of computing $\mathbf{R_W} = \mathbf{I} - \mathbf{W}\mathbf{W}^+$ and its respective semi-orthogonal matrix $\mathbf{Q_W}$ leads to difficulties because, unless $R = S$, the products $\widetilde{\mathbf{Y}} = \mathbf{Q_Z}'\mathbf{Y}$ and $\widetilde{\mathbf{X}} = \mathbf{Q_W}'\mathbf{X}$ will not have the same number of rows: $\widetilde{\mathbf{Y}}$ has N' $= N - R$, whereas $\widetilde{\mathbf{X}}$ has N'' $= N - S$ rows, thus preventing the computation of CCA.

A more general solution, that accommodates bipartial and, therefore, is a generalisation for all cases of nuisance variables in CCA, consists of randomly permuting rows of $\widetilde{\widetilde{\mathbf{Y}}}$ and/or $\widetilde{\widetilde{\mathbf{X}}}$ using, respectively, permutation matrices $\mathbf{P_Y}$ and $\mathbf{P_X}$ of respective sizes N' and N'', therefore permuting in the lower dimensional space where $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{X}}$ are exchangeable, then, crucially, reestablishing the original number N of rows using the property that the transpose of a semi-orthogonal matrix is the same as its inverse ($\mathbf{Q}' = \mathbf{Q}^+$), to only then perform CCA. Therefore, CCA is computed using $\mathbf{Q_Z}\mathbf{P_Y}\mathbf{Q_Z}'\mathbf{Y}$ and $\mathbf{Q_W}\mathbf{P_X}\mathbf{Q_W}'\mathbf{X}$. Left and right sides will continue to have rank N' and N'' respectively, will have already been permuted, and will both have N rows. The procedure is fully symmetric in that, when the permutation matrices $\mathbf{P_Y}$ and $\mathbf{P_X}$ are both identity matrices (of sizes N' and N'', respectively), which is equivalent to no permutation, the expressions for each side reduce to the residualised data $\mathbf{R_Z}\mathbf{Y}$ and $\mathbf{R_W}\mathbf{X}$. The concatenation $[\mathbf{Q_Z}\mathbf{P_Y}\mathbf{Q_Z}'\mathbf{Y}, \mathbf{Q_W}\mathbf{P_X}\mathbf{Q_W}'\mathbf{X}]$ has the same rank as that of $[\mathbf{R_Z}\mathbf{Y}, \mathbf{R_W}\mathbf{X}]$, thus addressing the above problem of the unpermuted test statistic having a different and stochastically dominant distribution over that of the permuted data. Table 2 summarises the proposed solution for all cases, including part CCA.

2.7. Restricted exchangeability

The above method uses the Huh–Jhun semi-orthogonal matrix applied to CCA and leads to a valid permutation test provided that there are no dependencies among the rows of $\mathbf{M}$. That is, the method takes into account dependencies introduced by the regression of $\mathbf{Z}$ and/or $\mathbf{W}$ out from $\mathbf{Y}$ and/or $\mathbf{X}$, but not dependencies that might already exist in the data, and which generally preclude direct use of permutation tests. However, structured dependencies, such as those that may exist, for instance, in studies that involve repeated measurements, or for those in which participants do not constitute independent observations, e.g., sib-pairs, as in the Human Connectome Project (HCP; Van Essen et al., 2012), can be treated by allowing only those permutations that respect such dependency structure (Winkler et al., 2015). Unfortunately, the Huh–Jhun semi-orthogonal matrix does not respect such structure, blurring information from observations across blocks, and preventing the definition of a meaningful mapping from the N original observations that

Table 2
Proposed permutation method for the various cases of CCA, with respect to nuisance variables.

Name	Left set	Right set
CCA ("full", no nuisance)	$\mathbf{P_Y}\mathbf{Y}$	$\mathbf{X}$
Partial CCA	$\mathbf{P_Y}\mathbf{Q_Z}'\mathbf{Y}$	$\mathbf{Q_Z}'\mathbf{X}$
Part CCA	$\mathbf{Q_Z}\mathbf{P_Y}\mathbf{Q_Z}'\mathbf{Y}$	$\mathbf{P_X}\mathbf{X}$
Bipartial CCA	$\mathbf{Q_Z}\mathbf{P_Y}\mathbf{Q_Z}'\mathbf{Y}$	$\mathbf{Q_W}\mathbf{P_X}\mathbf{Q_W}'\mathbf{X}$

$\mathbf{Q_Z}$ is a $N \times N$' semi-orthogonal basis for the column space of $\mathbf{R_Z}$, such that $\mathbf{Q_Z}\mathbf{Q}'_{\mathbf{Z}} = \mathbf{R_Z}$, $\mathbf{Q}'_{\mathbf{Z}}\mathbf{Q_Z} = \mathbf{I}_{N' \times N'}$, where N' $= N - R$, and $\mathbf{Q}'_{\mathbf{Z}} = \mathbf{Q}^+_{\mathbf{Z}}$. $\mathbf{Q_W}$ is a $N \times N$'' similarly defined matrix for the column space of $\mathbf{R_W}$, N'' $= N - S$. The bipartial CCA case generalizes all others: for "full" CCA, $\mathbf{R_W} = \mathbf{R_Z} = \mathbf{I}_{N \times N}$, and so, $\mathbf{Q_W} = \mathbf{Q_Z} = \mathbf{I}_{N \times N}$; for partial CCA, $\mathbf{W} = \mathbf{Z}$; for part CCA $\mathbf{R_W} = \mathbf{I}_{N \times N}$, and so, $\mathbf{Q_W} = \mathbf{I}_{N \times N}$. For full and partial, pre-multiplication by $\mathbf{Q_Z}$ can be omitted since $\mathbf{Q_Z}'\mathbf{Q_Z} = \mathbf{I}_{N \times N}$, such that results do not change. Once these simplifications are considered, the general bipartial CCA case reduces to the other three as shown in the Table. Full and partial have matching number of rows in both sides, such that only one side needs be permuted; part and bipartial, however, have at the time of the permutation a different number of rows in each side, such that both can be permuted separately through the use of suitably sized permutation matrices $\mathbf{P_Y}$ and $\mathbf{P_X}$; $\mathbf{P_Y}$ is size $N \times N$ for full CCA, and N' $\times N$' for the three other cases; $\mathbf{P_X}$ is size $N \times N$ for full and for part CCA, and N'' $\times N$'' for the two other cases.
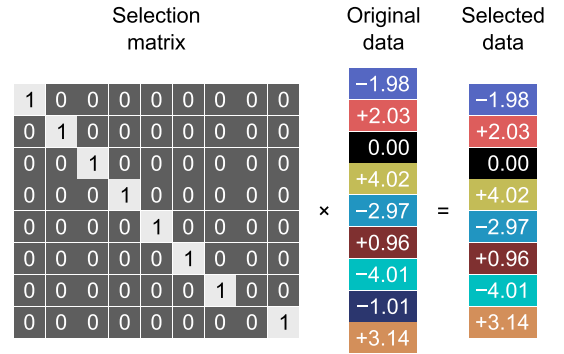

Fig. 1. A selection matrix is an identity matrix from which some specific rows have been removed. Pre-multiplication by a selection matrix deletes specific rows (those that correspond to columns that are all zero in the selection matrix).

define the block structure to the N' or N'' observations that are ultimately permuted.[3]

Such mapping, whereby each one of the N' and N'' rows of, respectively, $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{X}}$ corresponds uniquely to one of the N rows of the original data $\mathbf{Y}$ and $\mathbf{X}$, can be obtained using a different method, due to Theil (1965, 1968), and reviewed in detail by Magnus and Sinha (2005). Consider first the case of $\mathbf{Z}$. In the Theil method, that here is adapted for CCA, $\mathbf{Q_Z} = \mathbf{R_Z}\mathbf{S}'(\mathbf{S}\mathbf{R_Z}\mathbf{S}')^{-1/2}$, where the exponent $-1/2$ represents the positive definite matrix square root, and $\mathbf{S}$ is a $\overline{N} \times N$ selection matrix, $\overline{N} = \min(N', N'')$, that is, an identity matrix from which some $\max(R, S)$ rows have been removed. Pre-multiplication of a matrix by a selection matrix deletes specific rows, i.e., the ones that correspond to columns that are all zero in the selection matrix (Fig. 1). The $\mathbf{Q_Z}'\mathbf{Y}$ thusly computed are the best linear unbiased residuals with scalar covariance (BLUS), in that they are unbiased estimates of $\mathbf{S}\varepsilon_{\mathbf{Y}}$, where $\varepsilon_{\mathbf{Y}}$ are the (unknown) true errors after the nuisance effects of $\mathbf{Z}$ have been removed from $\mathbf{Y}$; $\mathbf{S}'\varepsilon_{\mathbf{Y}}$ contains the variance of interest, which may be shared among linear combinations of variables in both sides; it is an estimate of

---

[2] As originally proposed, in the context of the general linear model (GLM), Huh and Jhun (2001) use $\widetilde{\mathbf{Y}} = \mathbf{Q_Z}'\mathbf{R_Z}\mathbf{Y}$ and $\widetilde{\mathbf{X}} = \mathbf{Q_Z}'\mathbf{R_Z}\mathbf{X}$. These are equivalent to simply $\widetilde{\mathbf{Y}} = \mathbf{Q_Z}'\mathbf{Y}$ and $\widetilde{\mathbf{X}} = \mathbf{Q_Z}'\mathbf{X}$ as proposed here: since $\mathbf{R_Z} = \mathbf{Q_Z}\mathbf{Q_Z}'$, $\mathbf{Q_Z}'\mathbf{R_Z} = \mathbf{Q_Z}'\mathbf{Q_Z}\mathbf{Q_Z}' = \mathbf{I}_{N' \times N'}\mathbf{Q_Z}' = \mathbf{Q_Z}'$. This simplification holds true also for the GLM (not discussed in this article).

[3] There is an exception: if $\mathbf{Z}$ has a block diagonal structure and the observations encompassed by such blocks coincide with the exchangeability blocks, then an algorithm that uses Huh–Jhun and block permutation can be constructed.

Table 3

The semi-orthogonal matrix $\mathbf{Q}$ ($\mathbf{Q_Z}$ and/or $\mathbf{Q_W}$, subscripts dropped), discussed in Sections 2.6 and 2.7, is not unique. Two principled methods to obtain it are below.

Method	Matrix
Theil (1965)	$\mathbf{Q} = \mathbf{RS'}(\mathbf{SRS'})^{-1/2}$
Huh and Jhun (2001)	$\mathbf{QEQ'} = \mathbf{R}$ (via SVD or Schur)

$\mathbf{R}$ is the residual-forming matrix ($\mathbf{R_Z}$ or $\mathbf{R_W}$, for the respective set of nuisance variables, subscript dropped); since $\mathbf{R}$ is idempotent, all its eigenvalues (the diagonal elements of $\mathbf{E}$) are equal to 0 or 1. In the Theil method, $\mathbf{S}$ is a $N' = N \times (N-R)$ (for $\mathbf{Z}$) or $N'' = N \times (N-S)$ (for $\mathbf{W}$) selection matrix; the matrix square root (in the exponent $-1/2$) is the positive definite solution. In the Huh–Jhun method, after Schur or SVD factorisations of $\mathbf{R}$ are computed, and the R or S columns of $\mathbf{Q}$ that have corresponding zero eigenvalues in the diagonal of $\mathbf{E}$ are removed, such that $\mathbf{Q}$ computed from the factorisation is reduced from size $N \times N$ to $N \times N'$ or to $N \times N''$. At the end of these computations (see Algorithm 2, "semiortho'', in the Appendix), for both methods, $\mathbf{Q'Q} = \mathbf{I}$, $\mathbf{QQ'} = \mathbf{R}$, and $\mathbf{Q'} = \mathbf{Q^+}$. Both methods aim at obtaining residuals with a scalar covariance matrix $\sigma^2 \mathbf{I}$. Theil explicitly seeks BLUS residuals. However, strictly, $\mathbf{S}$ does not need to be a selection matrix: choose $\mathbf{S}$ to be $\mathbf{QR}$ (not to be confused with QR decomposition) using $\mathbf{Q}$ computed with the Huh–Jhun approach. Then, following Magnus and Sinha (2005, Theorem 2, p. 42), it can be shown that Huh–Jhun also provides BLUS residuals.

that is subjected to CCA and statistical testing. For partial CCA, $\mathbf{Q_Z}$ is the same for both sides; for bipartial CCA, similar computations hold for the other side, i.e., $\mathbf{Q_W} = \mathbf{R_W S'}(\mathbf{SR_W S'})^{-1/2}$. Table 3 summarises the two methods.

To construct a permutation procedure for CCA that respects the block structure, the Theil method can be used to compute $\mathbf{Q}$ instead of the Huh–Jhun approach. Choose $\max(R, S)$ observations to be removed from both sides (for partial CCA, $R = S$ since $\mathbf{W} = \mathbf{Z}$). Construct the selection matrix $\mathbf{S}$ of size $\overline{N} \times N$, define the exchangeability blocks based on $\overline{N}$ observations, compute $\mathbf{Q_Z}$ and $\mathbf{Q_W}$ using the same $\mathbf{S}$ for both (for part CCA, use the same strategy as for bipartial, replacing $\mathbf{R_W}$ for $\mathbf{I}$), residualise (in the BLUS sense) the input variables by computing $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{X}}$. These have the same number of rows, and the dependencies among these rows is the same for both sides; thus, only one side needs to be subjected to random permutations that respect such existing dependencies. Optionally, after permutation, the number N of observations may be reestablished by pre-multiplication by $\mathbf{Q_Z}$ and $\mathbf{Q_W}$. Finally, CCA is performed, with observation to the aspects discussed in Sections 2.3 and 2.5. A detailed algorithm is presented in Section 2.8.

It remains to be decided how to select the $\max(R, S)$ observations to be dropped. In principle, any set could be considered for removal, provided that the removed rows of $\mathbf{Z}$ or $\mathbf{W}$ form a full rank matrix. Some informed choices, however, could be more powerful than others. One of the conclusions from Winkler et al. (2015) is that the complexity of the dependence structure and the ensuing restrictions on exchangeability leads to reductions in power. Thus, natural candidates for removal are observations that, once removed, cause the overall dependence structure to be simpler. For example, it is sometimes the case that some observations are so uniquely related to all others that there are no other observations like them in the sample. These observations, therefore, cannot be permuted with any other, or perhaps with only a few. Their contribution to hypothesis testing in the permutation case is minimal, and their removal is less likely to affect a decision on rejection of the null hypothesis. Consider for example a design that has many monozygotic, dizygotic, and non-twin pairs of subjects, and that in the sample, there happens to be a single pair of half-siblings. It is well known that, for heritable traits, genetic resemblance depends on the kinship among individuals; half-siblings are expected to have a different degree of statistical dependency among each other compared to each one of the other

types of sibships in this sample. Thus, in there being just one such pair, it would be reasonable to prioritise it for exclusion, while keeping others.

2.8. General algorithm

A set of steps for permutation inference for CCA is described in Algorithm 1. In it, input variables $\mathbf{Y}$ and $\mathbf{X}$ will have been mean-centered before the algorithm begins, or an intercept will have been included as nuisance variable in both $\mathbf{Z}$ and $\mathbf{W}$. $\mathscr{P} = \{(\mathbf{P_Y}, \mathbf{P_X})_j\}$ is a set containing pairs of permutation matrices indexed by $j = \{1, ..., J\}$. In this set, the first permutation is always "no permutation'', i.e., $(\mathbf{P_Y}, \mathbf{P_X})_{j=1} = (\mathbf{I}_{N' \times N'}, \mathbf{I}_{N'' \times N''})$, such that $(\lambda_k)_{j=1}^* = \lambda_k$, for all k. For the cases in which only one side of CCA needs be permuted (Table 2), or for the cases in which $R = S$, or when there are dependencies among the data such that the Theil method is used to construct $\mathbf{Q}$ (Table 3), then $(\mathbf{P_X})_j$ can be set as $\mathbf{I}$ for all j. Details on how $\mathscr{P}$ is defined in observance to the null hypothesis and respecting structured dependencies among the data have been discussed in Winkler et al. (2014, 2015). In the algorithm, P can be larger, equal, or smaller than Q. Optional input arguments are the matrices with nuisance variables $\mathbf{Z}$ and $\mathbf{W}$, and the selection matrix $\mathbf{S}$. If $\mathbf{Z}$ is supplied but not $\mathbf{W}$, then the algorithm performs part or partial CCA, depending on the Boolean argument PARTIAL; if both $\mathbf{Z}$ and $\mathbf{W}$ are supplied, the algorithm performs bipartial CCA; if neither is supplied, then "full'' CCA is performed. If $\mathbf{S}$ is supplied, then the BLUS residuals based on Theil are used; otherwise, Huh–Jhun residuals are used. For either of these two cases, the semi-orthogonal matrix $\mathbf{Q}$ is computed using a separate, ancillary function named "semiortho'', described in the Appendix.

An initial CCA using residualised data is done in line 19; this uses another ancillary function, named "cca'', and also described in the Appendix; this function returns three results: the canonical coefficients $\mathbf{A}$ and $\mathbf{B}$, and the canonical correlations r_k. The canonical coefficients are used to compute the canonical variables $\mathbf{U}$ and $\mathbf{V}$, augmented by their orthogonal complement needed to ensure that they span the same space as the variables subjected to this initial CCA; the canonical correlations are ignored at this point and not stored (hence the placeholder "_''). A counter c_k for each canonical component is initialised as 0.

The core part of the algorithm are the two loops that run over the permutations in $\mathscr{P}$ and the K canonical components (between lines 25 and 35). At each permutation j, CCA is executed K times. In each, the columns of $\mathbf{U}$ and $\mathbf{V}$ that precede the current k are removed, such that their respective variances are not allowed to influence the canonical correlations at position k. At each permutation, the K canonical correlations are obtained (the third output from the function "cca'') and used to compute the associated test statistic. As shown, Wilks' statistic, λ_k, is used, simplified by the removal of the constant term, which does not affect permutation p-values. For numerical stability, sum of logarithms is favoured over the logarithm of a product (compare line 30 with Equation (3)). For inference using Roy's statistic, replace the condition $(\lambda_k)_j^* \geq \lambda_k$ for $(r_k)_j^* \geq r_k$ in line 31; this modification alone is sufficient as θ_k is permutationally equivalent to r_k. In that case, computations indicated in line 30 are no longer needed and can be removed to save computational time.

Whenever the statistic for the correlation at position k in a given permutation is higher or equal than that for the unpermuted data, the counter c_k is incremented (line 32). After the loop, the counter is converted into a p-value for each k. These simple, uncorrected p-values, however, are not useful. Instead, FWER-adjusted p-values are computed under closure using the cumulative maximum, i.e., the p-value for r_k is the largest (least significant) uncorrected p-value up to position k. The algorithm returns then these adjusted p-values, which can be compared to a predefined test level α to establish significance. Note that α itself is never used in the algorithm.

Algorithm 1. Permutation inference for CCA.

Require: $\mathbf{Y}_{N\times P}, \mathbf{X}_{N\times Q}, \mathcal{P}$. **Optional:** $\mathbf{Z}_{N\times R}, \mathbf{W}_{N\times S}, \mathbf{S}$, PARTIAL.   ▷ Inputs.

1:   $K \leftarrow \min(P, Q)$ ▷ Number of canonical components.
2:   **if** exist($\mathbf{Z}$) **then**   ▷ If left-side nuisance were defined.
3:      $\mathbf{R_Z} \leftarrow \mathbf{I} - \mathbf{ZZ}^+$ ▷ Residual forming matrix due to $\mathbf{Z}$.
4:      $[\mathbf{Q_Z}]_{N\times N'} \leftarrow \text{semiortho}(\mathbf{R_Z}, \mathbf{S})$ ▷ $\mathbf{Q_Z}$ via H–J/Theil; $N' = N - R$.
5:   **else**
6:      $[\mathbf{Q_Z}]_{N\times N'} \leftarrow \mathbf{I}_{N\times N}$ ▷ $\mathbf{Q_Z}$ is identity; $N' = N$.
7:   **end if**
8:   **if** $\neg\,\text{exist}(\mathbf{W}) \wedge$ PARTIAL **then**   ▷ If $\mathbf{W}$ not given, and this is partial CCA.
9:      $\mathbf{W} \leftarrow \mathbf{Z}$ ▷ Re-use $\mathbf{Z}$ as $\mathbf{W}$.
10:   **end if**   ▷ (otherwise, it is bipartial or part CCA.)
11:   **if** exist($\mathbf{W}$) **then**   ▷ If right-side nuisance were defined.
12:      $\mathbf{R_W} \leftarrow \mathbf{I} - \mathbf{WW}^+$ ▷ Residual forming matrix due to $\mathbf{W}$.
13:      $[\mathbf{Q_W}]_{N\times N''} \leftarrow \text{semiortho}(\mathbf{R_W}, \mathbf{S})$ ▷ $\mathbf{Q_W}$ via H–J/Theil; $N'' = N - S$.
14:   **else**
15:      $[\mathbf{Q_W}]_{N\times N''} \leftarrow \mathbf{I}_{N\times N}$ ▷ $\mathbf{Q_W}$ is identity; $N'' = N$.
16:   **end if**
17:   $\mathbf{Y} \leftarrow \mathbf{Q_Z'Y}$ ▷ Residualised $\mathbf{Y}$, exchangeable.
18:   $\mathbf{X} \leftarrow \mathbf{Q_W'X}$ ▷ Residualised $\mathbf{X}$, exchangeable.
19:   $\mathbf{A}_{P\times K}, \mathbf{B}_{Q\times K}, \_ \leftarrow \text{cca}(\mathbf{Y}, \mathbf{X}, R, S)$ ▷ Initial CCA (see Algorithm 3).
20:   $\mathbf{U}_{N\times P} \leftarrow \mathbf{Y}[\mathbf{A}, \text{null}(\mathbf{A}')]_{P\times P}$ ▷ Canonical variables from residualised $\mathbf{Y}$.
21:   $\mathbf{V}_{N\times Q} \leftarrow \mathbf{X}[\mathbf{B}, \text{null}(\mathbf{B}')]_{Q\times Q}$ ▷ Canonical variables from residualised $\mathbf{X}$.
22:   **for** $k \in \{1, \ldots, K\}$ **do**   ▷ For each component.
23:      $c_k \leftarrow 0$ ▷ Initialise a counter.
24:   **end for**
25:   **for all** $(\mathbf{P_Y}, \mathbf{P_X})_j \in \mathcal{P}$ **do**   ▷ For each permutation.
26:      **for** $k \in \{1, \ldots, K\}$ **do**   ▷ For each component.
27:        $[\mathbf{u}_k, \ldots, \mathbf{u}_P]_j^* \leftarrow \mathbf{Q_Z}(\mathbf{P_Y})_j[\mathbf{u}_k, \ldots, \mathbf{u}_P]$ ▷ Permute left side.
28:        $[\mathbf{v}_k, \ldots, \mathbf{v}_Q]_j^* \leftarrow \mathbf{Q_W}(\mathbf{P_X})_j[\mathbf{v}_k, \ldots, \mathbf{v}_Q]$ ▷ Permute right side.
29:        $\_, \_, [r_{(1)}, \ldots]_j^* \leftarrow \text{cca}([\mathbf{u}_k, \ldots]_j^*, [\mathbf{v}_k, \ldots]_j^*, R, S)$ ▷ Main CCA.
30:        $(\lambda_k)_j^* \leftarrow \sum_{i=1}^{K-k+1} \ln(1 - (r_{(i)}^2)_j^*)$ ▷ Wilks' test statistic.
31:        **if** $(\lambda_k)_j^* \geqslant (\lambda_k)_{j=1}^*$ **then**   ▷ If statistic after permutation is larger.
32:          $c_k \leftarrow c_k + 1$ ▷ Increment the counter.
33:        **end if**
34:      **end for**
35:   **end for**
36:   **for** $k \in \{1, \ldots, K\}$ **do**   ▷ For each canonical component.
37:      $[p_k]_{\text{unc}} \leftarrow c_k/J$ ▷ Uncorrected p-value.
38:      $[p_k]_{\text{FWER}} \leftarrow \max([p_1, \ldots, p_k]_{\text{unc}})$ ▷ FWER-corrected p-value (closure).
39:   **end for**
40:   **return** $[p_1, \ldots, p_K]_{\text{FWER}}$ ▷ Return the FWER-corrected p-values.

As presented, the algorithm does not cover dimensionality reduction or any penalty to enforce sparse solutions for CCA. Dimensionality reduction using methods such as principal component analysis (PCA), if included, would be performed after residualisation, but before CCA. Thus, in the algorithm, PCA, if executed, would be done between lines 18 and 19. As for the many forms of sparse or penalised CCA (Nielsen, 2002; Waaijenborg and Zwinderman, 2007; Wiesel et al., 2008; Parkhomenko et al., 2007, 2009; Witten et al., 2009; Soneson et al., 2010; Hardoon and Shawe-Taylor, 2011; Gao et al., 2017; Ma and Li, 2018; Tan et al., 2018),

in principle these can be incorporated into the algorithm through the replacement of the classical CCA in lines 19 and 29 for one of these methods.

3. Evaluation methods

In this section we describe the synthetic data and methods used to investigate error rates and power under the different choices for the various aspects presented in Section 2 at each stage of a permutation test

Table 4

A valid permutation test for CCA needs to consider several computational and statistical aspects, which can be seen as steps in a procedure; each of them may be addressed using different strategies, some of which are studied below. The recommendations for the different options for each case (column "Use") are based on theoretical grounds shown in the sections as indicated in the table, and verified empirically using synthetic data (Section 4).

Step	Possible strategies studied	Use	Theory	Scenarios
Estimation of the canonical components	(a) All in a single step.	✖	Section 2.3.	I–VI.
	(b) Stepwise; variance already explained removed.	✔		
Inclusion of the complement of the canonical coefficients	(a) Null space not included.	✖	Section 2.3.	I–VI.
	(b) Null space included.	✔		
Correction for multiple testing	(a) Uncorrected, simple p-values, $[p_k]_{unc}$.	✖	Section 2.5.	I–VI, XVIII.
	(b) Corrected, cumulative maximum, $[p_k]_{clo}$.	✔		
	(c) Corrected, distribution of the maximum, $[p_k]_{max}$.	●		
Treatment of nuisance variables	(a) Simple residualisation ($\mathbf{Q} = \mathbf{I}$).	✖	Sections 2.6 and 2.7.	VII–XVIII.
	(b) Huh–Jhun method.	✔		
	(c) Theil method.	✔		
Choice of the test statistic	(a) Wilks' λ_k.	✔	Sections 2.2 and 2.4.	XVII, XVIII.
	(b) Roy's largest root, θ_k.	✔		

✔ Can or should be used.
● Can but should not be used.
✖ Cannot or should not be used.

for CCA, providing empirical evidence for the approach proposed. An overview of these aspects and choices at each stage is shown in Table 4. For each case, we use a series of simulation scenarios: each consists of a set of synthetic variables constructed using random values drawn from a normal or a non-normal (kurtotic or binary) probability distribution, sometimes with or without dimensionality reduction using principal components analysis (PCA; Hotelling, 1933; Jolliffe, 2002), sometimes with or without signal, and sometimes with or without nuisance variables. We also consider cases with large sample sizes and large number of variables. An overview of these scenarios (there are twenty of them) is in Table 5.

We start by investigating aspects related to the estimation of the canonical components at each permutation. Specifically, we consider (a) a one-step estimation of all canonical components, from 1 to K, versus (b) sequential estimation that removes, for the k-th canonical component in a given permutation, the variance already explained by the previous ones, as described in Section 2.3. With respect to the inclusion of the complement of the canonical coefficients, we consider (a) without the inclusion of the null space of the canonical coefficients, versus (b) with its inclusion so as to ensure that all variance from the original data not

explained in the initial CCA is considered in the estimation at every permutation, as described in Section 2.3. With respect to multiple testing, we consider the following strategies: (a) simple, uncorrected p-values, $[p_k]_{unc}$, (b) corrected under closure, $[p_k]_{clo}$, and (c) corrected using the distribution of the maximum statistic $[p_k]_{max}$; both $[p_k]_{clo}$ and $[p_k]_{max}$ offer FWER control, as discussed in Section 2.5. Keeping the same notation, we define scenarios I–VI consisting of $N = 100$ observations, with $P = 16$ variables on the left side ($\mathbf{Y}$) of CCA and $Q = 20$ variables on the right side ($\mathbf{X}$) (the procedure is symmetric; the choice of sides is arbitrary and does not affect results); for these six scenarios, data are drawn from one of three possible distributions: a normal distribution with zero mean and unit variance, a Student's t distribution with variable degrees of freedom $\nu = \{2, 4, 6, 8, 10\}$ (kurtotic), or a Bernoulli distribution with parameter $q = 0.20$ (binary). Analyses with and without dimensionality reduction to 10 variables using PCA are considered. The number of permutations used to compute p-values was set as $J = 2000$, with 2000 realisations (repetitions), thus allowing the computation of error rates.

We then turn our attention to aspects related to nuisance and residualisation discussed in Sections 2.6 and 2.7. We consider (a) simple residualisation, (b) residualisation using the Huh–Jhun method, and (c)

Table 5

Simulation scenarios.

Scenarios		N	P	Q	R	S	#(PCA)	Distribution	Signals	#Perms.	#Reps.
Without nuisance	I	100	16	20	0	0	–	normal	–	2000	2000
	II	100	16	20	0	0	10	normal	–	2000	2000
	III	100	16	20	0	0	–	kurtotic	–	2000	2000
	IV	100	16	20	0	0	10	kurtotic	–	2000	2000
	V	100	16	20	0	0	–	binary	–	2000	2000
	VI	100	16	20	0	0	10	binary	–	2000	2000
Partial CCA	VII	100	16	20	15	R	–	normal	–	2000	2000
	VIII	100	16	20	15	R	10	normal	–	2000	2000
	IX	100	16	20	15	R	–	kurtotic	–	2000	2000
	X	100	16	20	15	R	10	kurtotic	–	2000	2000
	XI	100	16	20	15	R	–	binary	–	2000	2000
	XII	100	16	20	15	R	10	binary	–	2000	2000
Bipartial CCA	XIII	100	16	20	15	15	–	normal	–	2000	2000
	XIV	100	16	20	15	15	10	normal	–	2000	2000
Larger samples	XV	*	16	20	20	R	–	normal	–	1000	1000
	XVI	*	16	20	20	R	10	normal	–	1000	1000
With signal	XVII	100	16	20	0	0	–	normal	sparse	2000	2000
	XVIII	100	16	20	0	0	–	normal	dense	2000	2000

* For scenarios XV and XVI, sample size varied, $N = \{100, 200, ..., 900, 1000\}$. In the table, R and S refer to the number of nuisance variables other than the intercept, which is always included (so the number of nuisance variables in left and right sides for all the simulation scenarios was always, respectively $R+1$ and $S+1$). For partial CCA, the number of nuisance variables on one side is always the same as in the other, i.e., $S = R$, but that does not have to be for bipartial CCA, even though here the same size was used. The case with larger samples was used for investigation of partial CCA.

residualisation using the Theil method. For this purpose, scenarios VII–XVII are constructed similarly as I–VI, except that a third set $\mathbf{Z}$ of $R = 15$ variables is used as nuisance for partial CCA, whereas two other scenarios, XIII and XIV, a fourth set of variables $\mathbf{W}$ of $S = 15$ variables is used as nuisance for bipartial CCA.

The impact of ignoring, in samples substantially larger than the number of variables, the dependencies introduced by the residualisation of both sides of CCA is studied with scenarios XV and XVI, which consider samples progressively larger, $N = \{100, 200, \ldots, 900, 1000\}$, while keeping the other parameters similar as in scenarios VII and VIII. Finally, we briefly investigate power and the choice of the test statistic: we consider (a) Wilks' statistic (λ_k), as well as (b) Roy's largest root (θ_k), as discussed in Sections 2.2 and 2.4. We define scenarios XVII and XVIII similarly as I, this time including a strong, true signal in one canonical component, thus named "sparse'', or multiple, weaker signals shared across multiple (half of the smaller set, thus, "dense''). For all scenarios, an intercept is always included as nuisance variable in both sides such that the actual number of nuisance variables is $R + 1$ and $S + 1$ for each side, respectively. To report confidence intervals (95%), the Wilson (1927) method is used.

4. Results

In the results below, Sections 4.1 and 4.2 establish empirically that with an estimation method (i) that includes the null space of the canonical coefficients, (ii) that finds the canonical correlations in an iterative manner, and (iii) that after computing p-values through a closed testing procedure, the error rates are controlled. The subsequent results, from Section 4.3 onwards, consider only this valid approach.

4.1. Estimation strategies

Not including the complement of the canonical coefficients (null space) caused error rates to be dramatically inflated, well above the ex-

pected test level $\alpha = 0.05$ (5%), regardless of whether the estimation used the single step or the stepwise procedure, and regardless of any of the multiple testing correction strategies discussed; these results are shown in Table 6.

Even when the null space of the canonical coefficients was included, a single step procedure was never satisfactory. To understand this, consider the following consequence of the theory presented in Sections 2.2 and 2.5: for a valid, exact test in CCA, the expected error rate for each $\mathscr{H}_k^0$, i.e., the *per comparison error rate* (PCER; Hochberg and Tamhane, 1987) is α for $k = 1$, but for $k = 2$ it is $\alpha \times \alpha$, since the null can only be rejected if the previous one has also been declared significant at α. More generally, the PCER for a valid test is α^k for the k-th test, i.e., for the k-th canonical correlation. If the test level is set at 5%, then the PCER is 5% for $k = 1$, 0.25% for $k = 2$, 0.0125% for $k = 3$, and so forth. Error rates above this expectation render the test invalid; below render it conservative. In the simulations, a single step procedure never led to an exact test, with or without consideration to multiple testing, as shown in Table 6.

4.2. Multiple testing

As with the PCER, it is worth mentioning what the expected FWER for a valid, exact test is. That expectation is the test level itself, i.e., α. Any higher error rate renders a test invalid; lower error rate renders it conservative, though valid. Table 6 shows the FWER for the three different correction methods considered.

If the null space was not included, since the PCER was not controlled, the FWER could not be controlled either (first two columns of the table). If the null space of the canonical coefficients was included (last two columns), even though the single step estimation controlled the PCER, the FWER was not controlled for the simple, uncorrected p-values (third column, upper panel), which is not surprising. It should be emphasised, however, that these simple p-values have another problem: they are not guaranteed to be monotonically related to the respective canonical

Table 6

Observed per comparison error rate (%) and 95% confidence intervals for the first 6 canonical correlations in scenario I, assessed using the Wilks' statistic and three different multiple testing correction methods; the observed familywise error rate (FWER) for each case is also shown. Valid methods should have a FWER close to the nominal 5%, and PCER close to the nominal $(5\%)^k$; see Sections 4.1 and 4.2 for details.

	Null space not included		Null space included	
	Single step	Stepwise	Single step	Stepwise
(a) Uncorrected, simple p-values, $[p_k]_{\text{unc}}$				
$k = 1$	91.35 (90.04–92.50)	91.35 (90.04–92.50)	4.70 (3.86–5.72)	4.70 (3.86–5.72)
$k = 2$	93.50 (92.33–94.50)	60.40 (58.24–62.52)	4.60 (3.77–5.61)	0.25 (0.11–0.58)
$k = 3$	94.70 (93.63–95.60)	26.70 (24.81–28.68)	4.60 (3.77–5.61)	0.00 (0.00–0.19)
$k = 4$	95.55 (94.56–96.37)	7.25 (6.19–8.47)	4.85 (3.99–5.88)	0.00 (0.00–0.19)
$k = 5$	96.10 (95.16–96.86)	1.45 (1.01–2.07)	4.40 (3.59–5.39)	0.00 (0.00–0.19)
$k = 6$	96.75 (95.88–97.44)	0.25 (0.11–0.58)	4.30 (3.50–5.28)	0.00 (0.00–0.19)
FWER	99.90 (99.64–99.97)	91.35 (90.04–92.50)	18.30 (16.67–20.05)	4.70 (3.86–5.72)
(b) Corrected, cumulative maximum, $[p_k]_{\text{clo}}$				
$k = 1$	91.35 (90.04–92.50)	91.35 (90.04–92.50)	4.70 (3.86–5.72)	4.70 (3.86–5.72)
$k = 2$	90.35 (88.98–91.57)	60.40 (58.24–62.52)	3.40 (2.69–4.29)	0.25 (0.11–0.58)
$k = 3$	89.80 (88.40–91.05)	26.70 (24.81–28.68)	2.75 (2.12–3.56)	0.00 (0.00–0.19)
$k = 4$	89.55 (88.13–90.82)	7.25 (6.19–8.47)	2.40 (1.81–3.17)	0.00 (0.00–0.19)
$k = 5$	89.30 (87.87–90.58)	1.45 (1.01–2.07)	1.75 (1.26–2.42)	0.00 (0.00–0.19)
$k = 6$	88.85 (87.40–90.16)	0.25 (0.11–0.58)	1.45 (1.01–2.07)	0.00 (0.00–0.19)
FWER	91.35 (90.04–92.50)	91.35 (90.04–92.50)	4.70 (3.86–5.72)	4.70 (3.86–5.72)
(c) Corrected, distribution of the maximum, $[p_k]_{\text{max}}$				
$k = 1$	91.35 (90.04–92.50)	91.35 (90.04–92.50)	4.70 (3.86–5.72)	4.70 (3.86–5.72)
$k = 2$	7.95 (6.84–9.22)	10.95 (9.66–12.39)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 3$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 4$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 5$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 6$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
FWER	91.35 (90.04–92.50)	91.35 (90.04–92.50)	4.70 (3.86–5.72)	4.70 (3.86–5.72)

Using the Roy's statistic led to similar results as with Wilks (not shown). Dimensionality reduction with PCA led to similar results for the case in which the null space is included (not shown). For the case in which the null space is not included, results are not comparable with the ones above because, after PCA, $P = Q$ in the simulations, such that there is no null space to be considered as the matrices with canonical coefficients in both sides are then square.

correlations, such that it is possible that, using these p-values, the null hypothesis could be rejected for some canonical correlation, but retained for another that happens to be larger than the former. The use of such uncorrected, simple p-values, therefore, constitutes a test that is inadmissible. The problem with lack of monotonicity with uncorrected p-values is less severe if estimation is done in a stepwise manner (fourth column, upper panel), but is nonetheless still present, as shown in Fig. 2, and has potential to lead to an excess FWER, even though that did not occur in these simulations.

For the other two correction methods, when the null space of the canonical coefficients was included in the estimation process, FWER was controlled (third and fourth columns of Table 6, middle and lower panels), but there are particularities. Using the distribution of the maximum (lower panel) led to very conservative PCER, for both single step

or stepwise estimation, whereas correction with closure led to invalid PCER for single step estimation (third column, middle panel).

The only configuration that led to exact (neither conservative or invalid) control over PCER and FWER, and a monotonic relationship between canonical correlations and associated p-values, is the one in which a stepwise estimation was performed, with the null space of the canonical coefficients included, and with correction using a closed testing procedure (fourth column, middle panel of Table 6). Moreover, the FWER, when controlled using the cumulative maximum or the distribution of the maximum statistic, is guaranteed to match the PCER for $k = 1$: in the former case, any further rejection of the null is conditional on the first one having been rejected; in the latter, the distribution of the maximum coincides with the distribution of the first as the canonical correlations are ranked from largest to smallest.
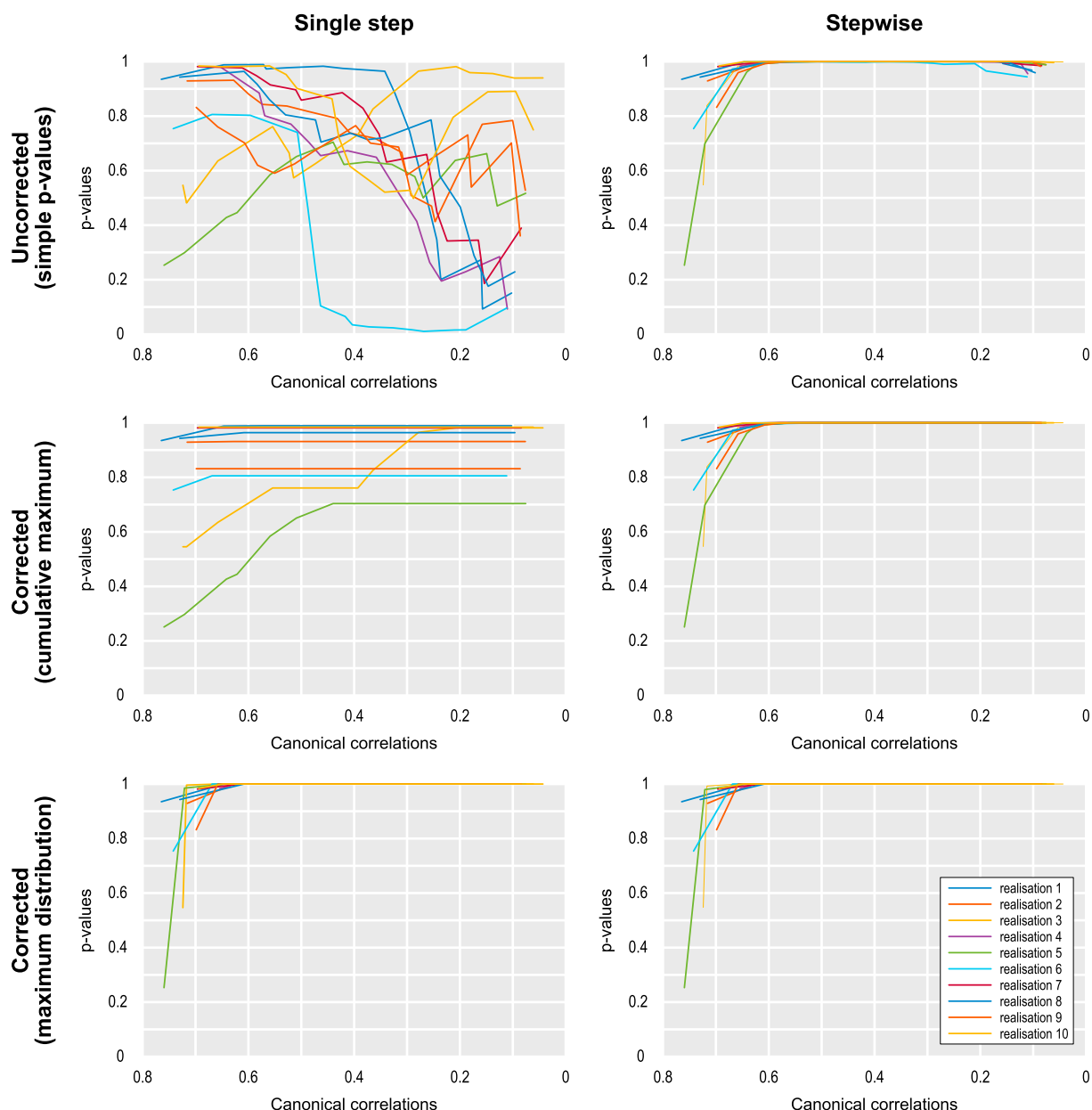

Fig. 2. Relationship between canonical correlations (horizontal axes) and associated p-values (vertical axes) for 10 realisations of scenario ɪ, considering two estimation methods (single step and stepwise) and three multiple testing correction methods (uncorrected, corrected using the cumulative maximum, and corrected using the distribution of the maximum statistic). The figure complements Table 6 by showing example realisations that average to the error rates shown in the table for the cases in which the null space is included. For simple, uncorrected p-values, the test is inadmissible; for corrected using the distribution of the maximum statistic, the test is overly conservative; single step does not control the familywise error rate.

Table 7

Observed per comparison error rate (PCER, %) and 95% confidence intervals for the first 6 canonical correlations in scenarios VII (partial CCA) and XIII (bipartial CCA), for the three different methods considered for treatment of nuisance variables.

	Simple residuals	Huh–Jhun	Theil
(a) Partial CCA			
$k = 1$ (FWER)	83.85 (82.17–85.40)	5.10 (4.22–6.15)	4.85 (3.99–5.88)
$k = 2$	44.15 (41.99–46.34)	0.30 (0.14–0.65)	0.35 (0.17–0.72)
$k = 3$	12.75 (11.36–14.28)	0.05 (0.01–0.28)	0.00 (0.00–0.19)
$k = 4$	1.75 (1.26–2.42)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 5$	0.20 (0.08–0.51)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 6$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
(b) Bipartial CCA			
$k = 1$ (FWER)	5.55 (4.63–6.64)	5.20 (4.31–6.26)	4.45 (3.63–5.44)
$k = 2$	0.10 (0.03–0.36)	0.30 (0.14–0.65)	0.20 (0.08–0.51)
$k = 3$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 4$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 5$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)
$k = 6$	0.00 (0.00–0.19)	0.00 (0.00–0.19)	0.00 (0.00–0.19)

Estimation included the null space of the canonical coefficients and a stepwise procedure, assessed using the Wilks' statistic, and corrected using a closed testing procedure (CTP). The CTP guarantees that the familywise error rate (FWER) matches the PCER for the first canonical correlation (i.e., for $k = 1$). Using the Roy's statistic led to similar results as with Wilks'; likewise, dimensionality reduction with PCA led to similar results (not shown).

Table 8

Observed per comparison error rate (PCER, %) and 95% confidence intervals for the first canonical correlation in scenarios I–VI (without nuisance) and VI–XII (partial CCA, with Huh–Jhun), considering different distributions for the data.

Distribution		Without nuisance	Partial CCA
Normal		4.70 (3.86–5.72)	5.15 (4.26–6.21)
Student	$\nu = 2$	3.95 (3.18–4.90)	14.70 (13.22–16.32)
	$\nu = 4$	5.45 (4.54–6.53)	5.40 (4.49–6.48)
	$\nu = 6$	4.15 (3.36–5.12)	5.40 (4.49–6.48)
	$\nu = 8$	4.70 (3.86–5.72)	5.00 (4.13–6.04)
	$\nu = 10$	3.85 (3.09–4.79)	5.10 (4.22–6.15)
Bernoulli	$q = 0.2$	5.30 (4.40–6.37)	5.30 (4.40–6.37)

ν: Degrees of freedom of the Student's t distribution used to simulate data; q: Parameter of the Bernoulli distribution used to simulate data. Estimation used the null space of the canonical coefficients and a stepwise procedure, assessed using the Wilks' statistic, and corrected using a closed testing procedure (CTP). The CTP guarantees that the familywise error rate (FWER) matches the PCER for the first canonical correlation (i.e., for $k = 1$). Using the Roy's statistic led to similar results as with Wilks'; using Theil led to similar results as Huh–Jhun; likewise, dimensionality reduction with PCA led to similar results (not shown).

4.3. Nuisance variables

For partial CCA, simple residualisation, even using the above procedure (stepwise estimation, null space included, correction via closure), resulted in the error rates being dramatically inflated, as shown in Table 7. The Huh–Jhun and the Theil residualisation methods, in contrast, resulted in the error rates being controlled at the nominal level, with no excess false positives. For bipartial CCA, the problem did not happen in the simulation settings: simple residualisation of both sides by entirely different sets of variables did not cause the error rates to be inflated; yet, using Huh-Jhun or Theil also produced nominal error rates, suggesting that these could be used in any configuration of nuisance variables, regardless of whether those in one side are not independent from those in the other.

4.4. Non-normality

Without nuisance variables and with kurtotic data simulated using a Student's t distribution with a small number of degrees of freedom, $\nu = \{2, 4, 6, 8, 10\}$, as well as with binary data simulated using a Bernoulli

distribution with parameter $q = 0.20$, error rates were controlled nominally, as shown in Table 8. In partial CCA, however, even using the Huh–Jhun method, highly kurtotic data led to excess error rates. In particular, for the simulated data using a Student's t distribution with degrees of freedom of only $\nu = 2$, the observed error rate was 14.7%, for a test level of 5%; using the Theil method led to also inflated error rate in this case, with 10.7% (95% confidence interval: 8.28–13.72, not shown in the table). For $\nu \geq 4$, error rates were controlled at the nominal level, for both Huh–Jhun (Table 8) and Theil (not shown).

4.5. Large samples

Increasing the sample size while keeping the number of variables fixed progressively reduced the amount of errors for the simple residualisation method to treat nuisance variables, as shown in Table 9; the trend was similar with or without dimensionality reduction using PCA. The reduction in the error rate as the sample size increased did not affect Huh–Jhun or Theil methods, for which error rates were already controlled even with a relatively smaller sample compared to the number of variables.

Without PCA: $P = 20$, $Q = 16$. With PCA: $P = Q = 10$. Estimation included the null space of the canonical coefficients and a stepwise procedure, assessed using the Wilks' statistic, and corrected using a closed testing procedure (CTP). The CTP guarantees that the familywise error rate (FWER) matches the PCER for the first canonical correlation (i.e., for $k = 1$). Using the Roy's statistic led to similar results as with Wilks'. The confidence intervals are wider than for other tables because the number of realisations (and also of permutations) was smaller (Table 5).

4.6. Dimensionality reduction

Dimensionality reduction with PCA did not affect error rates (PCER and FWER) with respect to single step vs. stepwise estimation of canonical coefficients, nor correction for multiple testing, nor method for addressing nuisance variables. That is, these results (not shown) were indistinguishable from those obtained without PCA (shown above). Moreover, as the simulations used the same number of principal components for both sides of CCA, including or not the null space could not have affected results, as $P = Q$ after dimensionality reduction. Using PCA did yield higher power to detect effects, for both Wilks' and Roy's test statistics (Table 10, next item). This apparent extra power can be attributed to the smaller number of variables after PCA, as the principal components that were retained contained most of the simulated signal, which, given the reduced dimensionality of the set of data, could then be detected with higher likelihood.

4.7. Choice of the statistic

The results above, that consider solely the error rates, and are based on results with the Wilks' statistic (λ_k), are essentially the same for Roy's largest root (θ_k; results not shown). That is, results regarding the estimation strategies, multiple testing, nuisance variables, non-normality, behaviour with large samples, and dimensionality reduction with PCA, are virtually the same for Wilks' and Roy's statistics. In the presence of synthetic signal, however, the two test statistics diverged. Table 10 shows that, with signal spread across multiple canonical components (i.e., "dense"), Wilks' is substantially more powerful than Roy's statistic. With signal concentrated in just one (the first) canonical variable (i.e., "sparse"), the trend reverses, and Roy's become more powerful than Wilks'.

5. Discussion

5.1. Permutation tests

Compared to univariate, multivariate tests pose the problem of

Table 9
Observed familywise error rate (FWER, %, that matches the PCER for $k = 1$) and 95% confidence intervals for scenarios XV and XVI, used to investigate effect of different sample sizes, for the three different methods for dealing with nuisance variables.

N	Without PCA			With PCA		
	Simple residuals	Huh–Jhun	Theil	Simple residuals	Huh–Jhun	Theil
100	96.60 (95.29–97.56)	4.80 (3.64–6.31)	5.00 (3.81–6.53)	59.20 (56.12–62.21)	5.00 (3.81–6.53)	5.10 (3.90–6.64)
200	42.20 (39.17–45.29)	5.50 (4.25–7.09)	5.40 (4.16–6.98)	22.30 (19.83–24.98)	5.00 (3.81–6.53)	4.50 (3.38–5.97)
300	25.10 (22.51–27.88)	5.00 (3.81–6.53)	5.40 (4.16–6.98)	12.20 (10.31–14.37)	5.40 (4.16–6.98)	5.10 (3.90–6.64)
400	18.00 (15.74–20.50)	4.30 (3.21–5.74)	5.10 (3.90–6.64)	11.80 (9.95–13.95)	4.50 (3.38–5.97)	5.70 (4.43–7.31)
500	11.50 (9.67–13.63)	6.10 (4.78–7.76)	4.10 (3.04–5.51)	9.50 (7.83–11.48)	6.80 (5.40–8.53)	5.10 (3.90–6.64)
600	10.80 (9.02–12.88)	5.20 (3.99–6.76)	5.00 (3.81–6.53)	9.30 (7.65–11.26)	4.70 (3.55–6.19)	4.70 (3.55–6.19)
700	11.20 (9.39–13.31)	4.20 (3.12–5.63)	4.40 (3.29–5.86)	7.20 (5.76–8.97)	5.70 (4.43–7.31)	5.50 (4.25–7.09)
800	10.10 (8.38–12.12)	5.50 (4.25–7.09)	4.20 (3.12–5.63)	7.00 (5.58–8.75)	4.80 (3.64–6.31)	5.90 (4.60–7.54)
900	8.40 (6.84–10.28)	5.00 (3.81–6.53)	4.00 (2.95–5.40)	7.70 (6.20–9.52)	5.20 (3.99–6.76)	5.20 (3.99–6.76)
1000	7.70 (6.20–9.52)	4.30 (3.21–5.74)	5.90 (4.60–7.54)	5.00 (3.81–6.53)	6.20 (4.87–7.87)	4.70 (3.55–6.19)

Table 10
Observed power (%) and 95% confidence intervals for the first canonical correlation in scenarios XVII and XVIII, that included a synthetic signal added to either one or half ("sparse" or "dense", respectively) of the initial variables, thus captured by only one (sparse case) or multiple (dense case) canonical correlations.

Signals		Without PCA		With PCA	
		Wilks (λ)	Roy (θ)	Wilks (λ)	Roy (θ)
Sparse	$k = 1$	42.10 (39.95–44.28)	57.90 (55.72–60.05)	81.55 (79.79–83.19)	94.75 (93.68–95.64)
Dense	$k = 1$	83.05 (81.34–84.63)	37.05 (34.96–39.19)	95.80 (94.83–96.59)	70.70 (68.67–72.65)
	$k = 2$	42.30 (40.15–44.48)	4.75 (3.90–5.77)	72.05 (70.04–73.97)	24.75 (22.91–26.69)
	$k = 3$	12.75 (11.36–14.28)	0.25 (0.11–0.58)	31.95 (29.94–34.03)	4.30 (3.50–5.28)
	$k = 4$	1.95 (1.43–2.65)	0.00 (0.00–0.19)	6.55 (5.55–7.72)	0.50 (0.27–0.92)
	$k = 5$	0.10 (0.03–0.36)	0.00 (0.00–0.19)	1.00 (0.65–1.54)	0.00 (0.00–0.19)
	$k = 6$	0.05 (0.01–0.28)	0.00 (0.00–0.19)	0.05 (0.01–0.28)	0.00 (0.00–0.19)

Estimation used the null space of the canonical coefficients and a stepwise procedure, assessed using the Wilks' statistic, and corrected using a closed testing procedure (CTP). The CTP guarantees that the familywise error rate (FWER) matches the PCER for the first canonical correlation (i.e., for $k = 1$).

establishing the distributional form for more complicated test statistics; in the parametric case, inference is marred by a set of difficulties: the assumption that all observations are independent and identically distributed following normal theory, the extremely complicated formulas for the density of the canonical correlations, which further depend on the (unknown) population canonical correlations, the sensitivity of asymptotic approximations to departures from assumptions, bias in estimations of parameters, and the validity of these approximations only for particular cases.

Permutation tests address these difficulties in different ways, and their advantages are well known (Ludbrook and Dudley, 1998; Nichols and Holmes, 2002; Good, 2005; Pesarin and Salmaso, 2012): no underlying distributions need be assumed, non-independence and even heteroscedastic variances can be accommodated, non-random samples can be used, and a wide variety of test statistics are allowed. Moreover, all information needed to build the null distribution lie within the data, as opposed to in some idealised population.

These many benefits extend to inference for multivariate methods. In the case of CCA, one benefit is immediately obvious: the complicated formulas and charts for the distribution of the canonical correlations can be bypassed completely, thus with no need to appeal to distributional assumptions. In effect, as shown in Section 4.4, even with all variables not following a normal distribution, error rates were still controlled at the nominal level. It should be noted, however, that extremely kurtotic data, such as that generated with a Student's t distribution with extremely low degrees of freedom, caused results to be invalid in the presence of nuisance variables, even with the Huh–Jhun or Theil methods. Such data, however, are rare (recall that with 2 degrees of freedom, the Student's t distribution has infinite variance); most applications of CCA investigate datasets that have variables with data that have diverse distributional properties.

Yet, although in the univariate case, algorithms for permutation inference tend to be relatively straightforward to implement and do lead to valid results, for CCA, the theory presented in the previous sections and

the results with synthetic data show that a simple permutation algorithm that does not consider aspects such as a stepwise estimation of the canonical correlations, nor the inclusion of the null space of canonical coefficients when the two sets of variables do not have the same size, or that does not accommodate specific treatment for nuisance variables, or addresses multiplicity respecting the ordering of the canonical correlations, leads to invalid results.

5.2. Estimation and multiple testing correction

Results from Sections 4.1 and 4.2 show that the estimation method that leads to exact, valid results (neither conservative or invalid) is the one that estimates one canonical correlation one at a time, in a stepwise, iterative manner, that includes the null space of the canonical coefficients when the sets of variables have different sizes (i.e., when $P \neq Q$), and that computes adjusted p-values using a closed testing procedure. All alternative approaches led to either invalid or conservative results when considering PCER or the FWER.

It should be emphasised, however, that there are cases in which the naïve permutation method, described at the beginning of Section 2.3 remains valid. The method is valid whenever only the first ($k = 1$) canonical component is of interest, and there are no nuisance variables or, if there are nuisance variables, those in the left and right side (**Z** and **W**) are completely orthogonal (thus, excluding partial CCA). Even though the naïve method was not explicitly tested, it is equivalent to the single step method with the null space included, which in the simulations led to an error rate of 4.70% at test level 0.05 (Table 6). The reason why it remains valid is that, if interest is only in the first canonical component, there is no need to perform an initial CCA to allow stepwise removal of previous (before the current k components) components. Moreover, there is no multiple testing to be considered.

The last column of Table 6 may suggest that uncorrected p-values (upper panel) and a CTP (middle) are equivalent for stepwise estimation. They are not, and their differences are manifest in two ways, both

previously discussed: first, uncorrected p-values are not monotonically related to the canonical correlations (Fig. 2), and second, FWER has potential to be higher than the PCER for $k = 1$, even though that did not happen in the simulations.

5.3. Inference in the presence of nuisance variables

It is sometimes the case that known, spurious variability needs to be taken into account. For example, variables such as age and sex are often considered confounds. Merely regressing out such nuisance variables from all other variables that are subjected to CCA, then proceeding to a simple permutation test, leads to inflated error rates and an invalid test, as expected from Section 2.6, and evidenced by the results in Section 4.3. The dependencies among observations introduced through the residualisation renders the data no longer exchangeable.

This inflated error rate, even after multiple testing correction, is the probably the most striking finding of the current study, as the results can be dramatically affected, particularly if the number of nuisance variables is relatively large compared to the sample size, as shown in Section 4.5. Transformations that make residuals exchangeable, through the use of a lower dimensional basis where exchangeability holds, namely, the Huh–Jhun and Theil methods, mitigate the problem, as evidenced by the theory and through the simulations.

Even though both methods led to similarly controlled error rates, they are not equivalent: Huh–Jhun always leads to same canonical components as they would have been obtained from the residualised data, whereas the Theil method can allow for multiple, different solutions depending on the choice of the selection matrix **S**. Theil (1965) suggested that the choice of the observations to be dropped should consider power; here we suggest that the choice of **S** can be based on restrictions on exchangeability: if all original data are freely exchangeable, the Huh–Jhun method is a preferable choice in that it does not require an additional argument that affect the results; however, it does require a Schur or singular value decomposition of the residual-forming matrix, which is a rank-deficient matrix, such that numerical stability should also be a factor for consideration.

For bipartial CCA, while error rates were controlled even in the simple residualisation case, it should be noted that **Z** and **W** were generated independently in the simulations, such that they were expected to be orthogonal. With real data, possible overlap among columns or linear combinations of columns between **Z** and **W** create a case that would lie between the two extremes of partial and bipartial CCA. In such case, and given the results for partial CCA, error rates are not expected to be controlled with simple residualisation. Huh–Jhun and Theil, being able to deal with the most extreme case of dependencies between **Z** and **W** (that is, when the two are the same, which defines partial CCA), constitute a general solution to all cases.

5.4. Relationship with the GLM

The dangers of residualising both dependent and independent variables in the general linear model (GLM) with respect to nuisance variables, then proceeding to a permutation test, as proposed originally by Kennedy (1995) are well known (Anderson and Robinson, 2001). It is not a complete surprise, therefore, that permutation inference for CCA would lead to invalid results in similar settings. The original Huh and Jhun (2001) method (see also Kherad-Pajouh and Renaud, 2010) was proposed for the GLM as a way to address shortcomings of the Kennedy method in accommodating nuisance variables. Both Kennedy and Huh–Jhun were evaluated by Winkler et al. (2014): among the methods that can be considered for permutation inference in the GLM, Huh–Jhun is the only that cannot be used directly with exchangeability blocks, as the reduction to a lower dimensional space does not respect the block structure. The solution proposed here for permutation inference for CCA in the presence of exchangeability blocks, which uses the Theil method, is expected to solve the same problem also for the GLM, i.e., as a replacement

for Huh–Jhun in cases where the data have a block dependence structure, as it does for freely exchangeable data (Ridgway, 2009).

As in the univariate case, permutation tests in the presence of nuisance variables are approximate. Their exactness is in the sense that, under the null hypothesis, the probability of finding a p-value smaller or equal to the test level is the test level itself. Such tests are not perfectly exact as the true relationship between the nuisance variables and the variables of interest are not known and needs to be estimated. Even in the absence of nuisance variables, however, permutation tests that use only a fraction of the total number of possible permutations are also approximate, for not covering the whole permutation space (the number of potential permutations tends to be very large, and grows very rapidly with increases in sample size). The same holds for other resampling methods that do not use all possible rearrangements of the data. Regardless of the reason why the tests are approximate, results are known to converge asymptotically to the true p-values.

5.5. Choice of the statistic

Among the two test statistics considered, Wilks' (λ_k) tends to be more powerful than Roy's (θ_k) for effects that span multiple canonical components; the converse holds for signals concentrated in only a few of the canonical components, i.e., when many of the canonical variables are zero; in these cases, Roy's tend to be more powerful than Wilks', as shown in Section 4.7. The respective formulas (Equations (3) and (4)) give insight on why that is the case: Roy's statistic is invariant to canonical correlations other than the first (largest), whereas Wilks' pool information across all correlations; past simulations, reviewed by Johnstone and Nadler (2017), corroborate to the finding.

The use of these two statistics for any canonical correlation other than the first (i.e., for $k > 1$) is possible in the proposed iterative procedure because, for the current position k being tested, all the variance associated with the previous canonical components at positions $\{1, ..., k-1\}$ will have already been removed from the model (Sections 2.3 and 2.8), such that the largest canonical correlation (Roy's statistic) is the current one being tested; for Wilks', the procedure holds because these earlier canonical correlations are not marked as zero; instead, they are ignored altogether when the statistic is computed, as if the previous canonical components have never existed.

Wilks' lambda and Roy's largest root are not the only possible statistics that can be considered for CCA, and permutation tests allow the use of yet others. Some, such as Hotelling–Lawley and Pillai–Bartlett, were considered by Friederichs and Hense (2003). Using simulations and Monte Carlo results, the authors found that parametric distributions of these classical multivariate statistics were accurate, and could be obtained quickly at low computational cost; it should be noted, however, that the study used normally distributed simulated data, in which case parametric assumptions are known to hold.

5.6. Relationship with previous studies

While a number of studies have used permutation tests with CCA, not many investigated the performance of these tests. Nandy and Cordes (2003) proposed a non-parametric strategy for inference with CCA for the investigation of task-based fMRI time series: the method uses a resting-state (no task-related activity) dataset to build the null distribution; as resampling time series can be challenging due to temporal autocorrelation, the null distribution uses multiple voxels selected far apart from each other so as to also avoid issues with spatial correlation. The approach differs from the one presented here in that it uses subject-level time series (as opposed to between-subject analyses), is specific to brain imaging (the proposed method is general) and does a resampling method that shares similarities with, yet is not the same as permutation. Eklund et al. (2011) specifically used permutation tests for CCA with fMRI time series whitened with a combination of methods to allow permutation; the authors demonstrated that permutation tests for

both the GLM and CCA could be greatly accelerated through the use of graphics processing units (GPUs).

Kazi-Aoual et al. (1995) proposed analytical formulas for the first three moments of the permutation distribution of Pillai's trace for CCA; these moments can be used to fit a Pearson (1895) type III distribution, from which p-values can be obtained. Legendre et al. (2011) studied parametric and permutation tests for redundancy analysis (RDA; Rao, 1964) and for canonical correspondence analysis (referred to also by the acronym "CCA"; ter Braak, 1986); the authors found that a simultaneous test of all canonical eigenvalues for the respective axes (eigenvectors of predicted response variables in a linear model) in RDA, despite simple, is not valid, whereas a marginal test on each eigenvalue, as well as a "forward" test in which previously tested canonical axes are added to a matrix of nuisance variables, performs well, even if conservatively for axes other than the first. Yoo et al. (2012) investigated the relationship between CCA and regression, proposing the use of permutations and studying cases without nuisance variables; in the method, for the k-th canonical correlation, variance not already explained by canonical variables $\{1, \ldots, k\}$ in one of the sides is permuted, whereas the other variables remain fixed. Turgeon et al. (2018) considered using a small number of permutations for CCA, recording of the empirical distribution function, then using it to estimate the parameters of a Tracy–Widom distribution (Tracy and Widom, 1996; Johnstone, 2008) for cases in which the number of observations is smaller than the number of variables in either $\mathbf{Y}$ or $\mathbf{X}$; the distribution is then used to obtain p-values; data are assumed to follow a normal distribution, and inference is for the largest canonical correlation.

Permutation tests for the method of partial least squares (PLS; Tucker, 1958; Wold et al., 1983; McIntosh et al., 1996; McIntosh and Lobaugh, 2004) have been considered. For example, Chin and Dibbern (2010) and Sarstedt et al. (2011) used a permutation test to investigate how differences in the strength of association between variables (magnitude of estimates) further differed between two or more groups. These would be equivalent to, in the context of CCA, testing whether canonical correlations obtained across different groups would differ. Le Floch et al. (2012) investigated strategies for dimensionality reduction and regularisation for imaging and genetic data, whereas Grellmann et al. (2015) compared the performance of variants of CCA and PLS for similar problems. Both studies used direct permutation of the data, and were mostly focused on the relative performance of the different methods, offering no specific treatment of nuisance variables or the other aspects considered here, and which concern validity. Monteiro et al. (2016) investigated a strategy for sparse PLS and sparse CCA in which data are split into training and hold out, and inference uses permutation of the training data, with coefficients applied to test data, in which measurements of association are computed.

The current paper therefore fills a substantial knowledge gap, whereby not many studies considered at all the validity of permutation inference with CCA, but those that did approach the topic were not sufficiently general; none covered the topics discussed here. Moreover, in principle, the method as proposed can be used with subject-level fMRI or other timeseries data provided that whitening has been successful in removing temporal dependencies. Additionally, given the conceptual similarity between PLS and CCA, it is possible that permutation inference for PLS would require similar strategies as described in Section 2, particularly in the presence of nuisance variables and re-use of variance already explained. Whether that is the case, it is a question that remains open for future investigation.

5.7. Recommendations

Given the above results, the main recommendations for permutation inference for CCA can be summarised as follows:

- When studying a given k-th canonical variable or canonical correlation, $1 < k \leq K$, remove the effects of the previous ones, i.e., the variance from one set that has already been explained by the other, as represented by the earlier canonical variables. These effects are surely significant (regardless of the test level), otherwise the current canonical variable or correlation would not be under consideration. Ignoring the earlier ones causes the error rates be inflated (empirical evidence provided in Section 4.1).

- For sets of variables with different sizes (i.e., $P \leq Q$), ensure that the variability not represented by the canonical variables produced at the first permutation is considered in all and every permutation. That is, include the null space of the canonical coefficients when computing the variables subjected to permutation. Not including the null space leads to excess false positives (empirical evidence provided in Section 4.1).

- Do not use simple p-values for inference, and make sure that a closed testing procedure is used. Using simple, uncorrected p-values has two negative consequences: (i) both the PCER and FWER are inflated, and (ii) since simple p-values are not guaranteed to be monotonically related to the canonical correlations, the resulting test is inadmissible (empirical evidence provided in Section 4.2).

- For the same reason, do not use FDR to correct for multiple testing after using simple p-values: while the p-values themselves satisfy the requirements of FDR, they lead to an inadmissible test even after correction, leading to non-sensical results whereby a stronger canonical correlation may be less significant than a weaker one (empirical evidence provided in Section 4.2).

- While valid, inference using the distribution of the maximum statistic across canonical correlations leads do conservative results, except for the first canonical correlation (empirical evidence provided in Section 4.2).

- If regressing out nuisance variables from both sets of variables subjected to CCA, make sure that the residuals are transformed to be exchangeable, e.g., with the Huh–Jhun or Theil methods, then permuted accordingly. Failure to observe this recommendation leads to excess false positives, particularly when the number of nuisance variables is a large fraction of the sample size (empirical evidence provided in Sections 4.3 and 4.5).

All these recommendations are integrated into Algorithm 1.

6. Conclusion

As evidenced by the theory and simulations in the previous sections, a simple permutation procedure leads to invalid results: (i) simple p-values are not admissible for inference in CCA, lead to excess PCER and FWER, and cannot be corrected using generic methods based on p-values such as FDR; (ii) ignoring the variability already explained by previous canonical variables leads to inflated error rates for all canonical correlations except for the first; (iii) regression of the same set of nuisance variables from both sides of CCA without further consideration leads to inflated error rates; and (iv) the classical method for multiple testing correction, that uses the distribution of the maximum statistic, leads to conservative results. The use of a stepwise estimation procedure, transformation of the residuals to a lower dimensional basis where exchangeability holds, and correction for multiple testing via closure, ensures the validity of permutation inference for CCA.

Appendix A. Ancillary functions

Algorithm 1 requires two relevant ancillary functions: one to compute the semi-orthogonal matrix $\mathbf{Q}$, and another to conduct the CCA proper and obtain the canonical coefficients $\mathbf{A}$ and $\mathbf{B}$; these two functions are described in pseudo-code in Algorithms 2 and 3. The "semiortho'' function takes as input a residual-forming matrix $\mathbf{R}$ and, optionally, a selection matrix $\mathbf{S}$. If $\mathbf{S}$ is supplied, it computes $\mathbf{Q}$ using the Theil method; otherwise, it uses the Huh–Jhun method (Table 3).

As shown, "semiortho'' uses Schur decomposition for Huh–Jhun, but that decomposition can be replaced by singular value decomposition (SVD) or QR decomposition. Another possibility consists of never using $\mathbf{R}$ directly, computing instead an orthogonal basis for the null space of $\mathbf{Z}$ (not shown; it would require taking $\mathbf{Z}$ as an input argument). All these are expected to produce the same results. However, as the residual-forming matrix is rank deficient and idempotent, all its eigenvalues are identical to 0 or 1. Thus, considerations about numerical stability and float point arithmetic (Moler, 2004), as well as speed, should determine the best choice for a particular programming language or computing architecture.

Algorithm 2. The "**semiortho**'' function, used in Algorithm 1.

Require: $\mathbf{R}_{N \times N}$. **Optional:** $\mathbf{S}_{N \times N'}$.   ▷ Inputs.
1: **if** $\mathrm{exist}(\mathbf{S})$ **then**   ▷ If a selection matrix was supplied.
2:    $\mathbf{Q}_{N \times N'} \leftarrow \mathbf{RS}'(\mathbf{SRS}')^{-1/2}$ ▷ Use Theil.
3: **else**   ▷ Otherwise, use Huh–Jhun.
4:    $\mathbf{Q}_{N \times N}, \mathbf{E} \leftarrow \mathrm{schur}(\mathbf{R})$ ▷ Schur decomposition.
5:    $\mathbf{Q}_{N \times N'} \leftarrow [\,\mathbf{q}_n \,|\, e_n > 0\,]$ ▷ Keep columns $\mathbf{q}_n$ with positive eigenvalues e_n.
6: **end if**
7: **return** $\mathbf{Q}$ ▷ Return the semi-orthogonal matrix.

Algorithm 3. The "**cca**'' function, used in Algorithm 1.

Require: $\mathbf{Y}_{N \times P}, \mathbf{X}_{N \times Q}, R, S$.   ▷ Inputs.
1: $K \leftarrow \min(P, Q)$ ▷ Number of canonical components.
2: $\mathbf{Q_Y}, \mathbf{R_Y}, \mathbf{T'_Y} \leftarrow \mathrm{qr}(\mathbf{Y})$;   ▷ QR decomposition.
3: $\mathbf{Q_X}, \mathbf{R_X}, \mathbf{T'_X} \leftarrow \mathrm{qr}(\mathbf{X})$;   ▷ QR decomposition.
4: $[\mathbf{L}_{P \times P}, \mathbf{D}, \mathbf{M}'_{Q \times Q}] \leftarrow \mathrm{svd}(\mathbf{Q'_Y Q_X})$;   ▷ Singular value decomposition.
5: $[r_1, \ldots, r_K] \leftarrow [d_{11}, \ldots, d_{KK}]$ ▷ Canonical correlations (diagonal of $\mathbf{D}$).
6: $\mathbf{L}_{P \times K} \leftarrow [\,\mathbf{l}_k, k = \{1, \ldots, K\}\,]$ ▷ Retain the first K columns of $\mathbf{L}$.
7: $\mathbf{M}_{Q \times K} \leftarrow [\,\mathbf{m}_k, k = \{1, \ldots, K\}\,]$ ▷ Retain the first K columns of $\mathbf{M}$.
8: $\mathbf{A}_{P \times K} \leftarrow \mathbf{T_Y R_Y^{-1} L}\sqrt{N - R}$ ▷ Canonical coefficients, left side.
9: $\mathbf{B}_{Q \times K} \leftarrow \mathbf{T_X R_X^{-1} M}\sqrt{N - S}$ ▷ Canonical coefficients, right side.
10: **return** $\mathbf{A}, \mathbf{B}, [r_1, \ldots, r_K]$ ▷ Return coefficients and correlations.

The "cca'' function takes as main inputs the sets of variables $\mathbf{Y}$ and $\mathbf{X}$. These will have been mean-centered and possibly residualised outside the function, such that no further mean-centering or residualisation is performed; if mean-centering was performed, then, at a minimum, the other two arguments are $R = S = 1$; if other variables were regressed out, as in part, partial, or bipartial CCA, then R and S are supplied with their corresponding values, minus 1 to account for the mean-centering. The algorithm uses the method described by Bjorck and Golub (1973), and is based on results of Olkin (1951) and Golub (1969); additional details can be found in Seber (1984). Inside this function, variables $\mathbf{Q}$ and $\mathbf{R}$ (subscripts omitted) refer to the factors of a QR factorization, hence with a different meaning than the similarly named matrices used elsewhere this paper. In the algorithm, $\mathbf{Y}$ and $\mathbf{X}$ are subjected to QR decomposition with pivoting (hence the matrices $\mathbf{T}$, subscripts omitted), using a numerically stable Householder transformation (Golub and Van Loan, 2013). The inner product $\mathbf{Q'_Y Q_X}$ of the orthogonal matrices from QR is subjected to singular value decomposition; the diagonal elements of $\mathbf{D}$ are the canonical correlations (line 5). The remaining computations are for the canonical coefficients: these are obtained via back substitution by solving the triangular sets of equations $\mathbf{L} = \mathbf{R_Y^{-1} A}$ and $\mathbf{M} = \mathbf{R_X^{-1} B}$. The permutation matrices $\mathbf{T_Y}$ and $\mathbf{T_X}$ are used for reordering. The constant factors in the square roots are normalising scalars to ensure unit variance for the canonical variables $\mathbf{U}$ and $\mathbf{V}$ (not returned by the algorithm, but computable as $\mathbf{U} = \mathbf{YA}$ and $\mathbf{V} = \mathbf{XB}$, Equation (1)); if an intercept was explicitly included in $\mathbf{Z}$ and $\mathbf{W}$ (for bipartial), these constant factors are as shown; if instead the data were mean-centered, further subtract 1 before taking the square root. Regardless, omission of these constant terms do not affect the canonical correlations.

Source code

Code related to this paper is available at https://github.com/andersonwinkler/PermCCA.

References

Abadir, K.M., Magnus, J.R., 2005. Matrix Algebra. Cambridge University Press, Cambridge.

Alnæs, D., Kaufmann, T., Marquand, A.F., Smith, S.M., Westlye, L.T., 2020. Patterns of sociocognitive stratification and perinatal risk in the child brain. Proc. Natl. Acad. Sci. Unit. States Am. 117 (22), 12419–12427.

Anderson, T.W., 2003. An Introduction to Multivariate Analysis, 3 ed. John Wiley & Sons, Hoboken, New Jersey.

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Austr. New Zealand J. Stat. Stat. 43, 75–88.

Bartlett, M.S., 1938. Further aspects of the theory of multiple regression. Math. Proc. Camb. Phil. Soc. 34, 33–40.

Bartlett, M.S., 1947. Multivariate analysis. J. Roy. Stat. Soc. Suppl. 9, 176–197.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300.

Bijsterbosch, J.D., Woolrich, M.W., Glasser, M.F., Robinson, E.C., Beckmann, C.F., Harrison, S.J., Smith, S.M., 2018. The relationship between spatial configuration and functional connectivity of brain regions. eLife 7, 27.

Bjorck, A., Golub, G.H., 1973. Numerical methods for computing angles between linear subspaces. Math. Comput. 27, 579–579.

Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden Day, San Francisco.

Chiani, M., 2016. Distribution of the largest root of a matrix for Roy's test in multivariate analysis of variance. J. Multivariate Anal. 143, 467–471.

Chin, W.W., Dibbern, J., 2010. An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services between Germany and the USA. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 171–193.

Clemens, B., Derntl, B., Smith, E., Junger, J., Neulen, J., Mingoia, G., Schneider, F., Abel, T., Bzdok, D., Habel, U., 2020. Predictive pattern classification can distinguish gender identity subtypes from behavior and brain imaging. Cerebr. Cortex.

Constantine, A.G., 1963. Some non-central distribution problems in multivariate analysis. Ann. Math. Stat. 34, 1270–1285.

Dinga, R., Schmaal, L., Penninx, B.W., van Tol, M.J., Veltman, D.J., van Velzen, L., Mennes, M., van der Wee, N.J., Marquand, A.F., 2019. Evaluating the evidence for biotypes of depression: methodological replication and extension of. Neuroimage: Clin. 22, 101796.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B., Dubin, M.J., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23, 28–38.

Eklund, A., Andersson, M., Knutsson, H., 2011. Fast random permutation tests enable objective evaluation of methods for single-subject FMRI analysis. Int. J. Biomed. Imag. 2011, 627947.

Fisher, R.A., 1939. The sampling distribution of some statistics obtained from non-linear equations. Ann. Eugenics 9, 238–249.

Friederichs, P., Hense, A., 2003. Statistical inference in canonical correlation analyses exemplified by the influence of north atlantic SST on European climate. J. Clim. 16, 522–534.

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., Knutsson, H., 2001. Detection of neural activity in functional MRI using canonical correlation analysis. Magn. Reson. Med. 45, 323–330.

Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2002. Exploratory fMRI analysis by autocorrelation maximization. Neuroimage 16, 454–464.

Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2003. Adaptive analysis of fMRI data. Neuroimage 19, 837–845.

Friston, K.J., Frith, C.D., Frackowiak, R.S., Turner, R., 1995. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2, 166–172.

Friston, K., Holmes, A.P., Frith, C.D., 1996. A multivariate analysis of PET activation studies. Hum. Brain Mapp. 4, 140–151.

Fujikoshi, Y., 1974. The likelihood ratio tests for the dimensionality of regression coefficients. J. Multivariate Anal. 4, 327–340.

Fujikoshi, Y., 1977. Asymptotic expansions of the distributions of the latent roots in MANOVA and the canonical correlations. J. Multivariate Anal. 7, 386–396.

Gao, C., Ma, Z., Zhou, H.H., 2017. Sparse CCA: adaptive estimation and computational barriers. Ann. Stat. 45, 2074–2101.

Glynn, W.J., Muirhead, R.J., 1978. Inference in canonical correlation analysis. J. Multivariate Anal. 8, 468–478.

Golub, G.H., 1969. Matrix decompositions and statistical calculations. In: Statistical Computation. Elsevier, pp. 365–397.

Golub, G.H., Van Loan, C.F., 2013. Matrix Computations, 4 ed. Johns Hopkins University Press, Baltimore.

Good, P., 2005. Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3 ed. Springer, New York.

Grellmann, C., Bitzer, S., Neumann, J., Westlye, L.T., Andreassen, O.A., Villringer, A., Horstmann, A., 2015. Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. Neuroimage 107, 289–310.

Hardoon, D.R., Shawe-Taylor, J., 2011. Sparse canonical correlation analysis. Mach. Learn. 83, 331–353.

Harris, R.J., 1976. The invalidity of partitioned-U tests in canonical correlation and multivariate analysis of variance. Multivariate Behav. Res. 11, 353–365.

Harris, R.J., 2013. A Primer of Multivariate Statistics, 3 ed. Taylor and Francis, New York.

Heck, D.L., 1960. Charts of some upper percentage points of the distribution of the largest characteristic root. Ann. Math. Stat. 31, 625–642.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. John Wiley & Sons, Inc.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Hsu, P.L., 1941. On the limiting distribution of the canonical correlations. Biometrika 32, 38–45.

Huh, M., Jhun, M., 2001. Random permutation testing in multiple linear regression. Commun. Stat. Theor. Methods 30, 2023–2032.

Ing, A., Sämann, P.G., Chu, C., Tay, N., Biondo, F., Robert, G., Jia, T., Wolfers, T., Desrivières, S., Banaschewski, T., Bokde, A.L.W., Bromberg, U., Büchel, C., Conrod, P., Fadai, T., Flor, H., Frouin, V., Garavan, H., Spechler, P.A., Gowland, P., Grimmer, Y., Heinz, A., Ittermann, B., Kappel, V., Martinot, J.L., Meyer-Lindenberg, A., Millenet, S., Nees, F., van Noort, B., Orfanos, D.P., Martinot, M.L.P., Penttilä, J., Poustka, L., Quinlan, E.B., Smolka, M.N., Stringaris, A., Struve, M., Veer, I.M., Walter, H., Whelan, R., Andreassen, O.A., Agartz, I., Lemaitre, H., Barker, E.D., Ashburner, J., Binder, E., Buitelaar, J., Marquand, A., Robbins, T.W., Schumann, G., 2019. Identification of neurobehavioural symptom groups based on shared brain mechanisms. Nat. Human Behav. 3, 1306–1318.

James, A.T., 1964. Distributions of matrix variates and latent roots derived from normal samples. Ann. Math. Stat. 35, 475–501.

Johnstone, I.M., 2008. Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy–Widom limits and rates of convergence. Ann. Stat. 36.

Johnstone, I.M., Nadler, B., 2017. Roys largest root test under rank-one alternatives. Biometrika 104, 181–193.

Jolliffe, I.T., 2002. Principal Component Analysis, 2 ed. Springer.

Jordan, C., 1875. Essai sur la géométrie à n dimensions. Bull. Soc. Math. Fr. 2, 103–174.

Kazi-Aoual, F., Hitier, S., Sabatier, R., Lebreton, J.D., 1995. Refined approximations to permutation tests for multivariate inference. Comput. Stat. Data Anal. 20, 643–656.

Kendall, M.G., 1975. Multivariate Analysis. Griffin, London.

Kennedy, P., 1995. Randomization tests in econometrics. J. Bus. Econ. Stat. 13, 85–94.

Kernbach, J.M., Yeo, B.T.T., Smallwood, J., Margulies, D.S., Thiebaut de Schotten, M., Walter, H., Sabuncu, M.R., Holmes, A.J., Gramfort, A., Varoquaux, G., Thirion, B., Bzdok, D., 2018. Subspecialization within default mode nodes characterized in 10,000 UK Biobank participants. Proc. Natl. Acad. Sci. Unit. States Am. 115, 12295–12300.

Kherad-Pajouh, S., Renaud, O., 2010. An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. Comput. Stat. Data Anal. 54, 1881–1893.

Koev, P., Edelman, A., 2006. The efficient evaluation of the hypergeometric function of a matrix argument. Math. Comput. 75, 833–847.

Kres, H., 1975. Statistische Tafeln zur multivariaten Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg.

Krzanowski, W.J., 1988. Principles of Multivariate Analysis: A User's Perspective. Clarendon Press, Oxford.

Kuhfeld, W.F., 1986. A note on Roy's largest root. Psychometrika 51, 479–481.

Lawley, D.N., 1959. Tests of significance in canonical analysis. Biometrika 46, 59–66.

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B., Duchesnay, E., 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. Neuroimage 63, 11–24.

Lee, S.Y., 1978. Generalizations of the partial, part and bipartial canonical correlation analysis. Psychometrika 43, 427–431.

Legendre, P., Oksanen, J., ter Braak, C.J.F., 2011. Testing the significance of canonical axes in redundancy analysis: test of canonical axes in RDA. Meth. Ecol. Evol. 2, 269–277.

Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses, 3 ed. Springer-Verlag, New York.

Li, J., Bolt, T., Bzdok, D., Nomi, J.S., Yeo, B.T.T., Spreng, R.N., Uddin, L.Q., 2019. Topography and behavioral relevance of the global signal in the human brain. Sci. Rep. 9, 14286.

Ludbrook, J., Dudley, H., 1998. Why permutation tests are superior to t and F tests in biomedical research. Am. Statistician 52, 127–132.

Ma, Z., Li, X., 2018. Subspace Perspective on Canonical Correlation Analysis: Dimension Reduction and Minimax Rates. arXiv:1605.03662 [math, Stat] ArXiv: 1605.03662.

Magnus, J., Sinha, A., 2005. On Theil's errors. Econom. J. 8, 39–54.

Marcus, R., Peritz, E., Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Marriott, F.H.C., 1952. Tests of significance in canonical analysis. Biometrika 39, 58–64.

McIntosh, A.R., Lobaugh, N.J., 2004. Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23, 250–263.

McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3, 143–157.

Mihalik, A., Ferreira, F.S., Rosa, M.J., Moutoussis, M., Ziegler, G., Monteiro, J.M., Portugal, L., Adams, R.A., Romero-Garcia, R., Vértes, P.E., Kitzbichler, M.G., Váša, F., Vaghi, M.M., Bullmore, E.T., Fonagy, P., Goodyer, I.M., Jones, P.B., Dolan, R., Mourão-Miranda, J., 2019. Brain-behaviour modes of covariation in healthy and clinically depressed young people. Sci. Rep. 9, 11536.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain

imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536.

Moler, C.B., 2004. Numerical Computing with MATLAB. SIAM, Philadelphia.

Monteiro, J.M., Rao, A., Shawe-Taylor, J., Mourão-Miranda, J., 2016. A multiple hold-out framework for sparse partial least squares. J. Neurosci. Methods 271, 182–194.

Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. John Wiley & Sons, Hoboken, New Jersey.

Muirhead, R.J., Waternaux, C.M., 1980. Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. Biometrika 67, 31–43.

Nandy, R.R., Cordes, D., 2003. Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. Magn. Reson. Med. 50, 354–365.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Nielsen, F.r., 2002. Neuroinformatics in Functional Neuroimaging. Ph.D. thesis. Technical University of Denmark.

Olkin, I., 1951. On Distribution Problems in Multivariate Analysis. Ph.D. thesis. University of North Carolina at Chapel Hill, Chapel Hill, NC.

Parkhomenko, E., Tritchler, D., Beyene, J., 2007. Genome-wide sparse canonical correlation of gene expression with genotypes. BMC Proc. 1, S119.

Parkhomenko, E., Tritchler, D., Beyene, J., 2009. Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. 8, 1–34.

Pearson, K., 1895. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos. Trans. R. Soc. London, Ser. A 186, 343–414.

Pesarin, F., Salmaso, L., 2012. A review and some new results on permutation testing for multivariate problems. Stat. Comput. 22, 639–646.

Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhya 26, 329–358.

Rao, B.R., 1969. Partial canonical correlations. Trab. Estad. Invest. Oper. 20, 211–219.

Ridgway, G.R., 2009. Statistical Analysis for Longitudinal MR Imaging of Dementia. Ph.D. thesis. University College London.

Rosa, M.J., Mehta, M.A., Pich, E.M., Risterucci, C., Zelaya, F., Reinders, A.A.T.S., Williams, S.C.R., Dazzan, P., Doyle, O.M., Marquand, A.F., 2015. Estimating multivariate similarity between neuroimaging datasets with sparse canonical correlation analysis: an application to perfusion imaging. Front. Neurosci. 9.

Roy, S.N., 1953. On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Stat. 24, 220–238.

Roy, S.N., 1957. Some Aspects of Multivariate Analysis, 1 ed. John Wiley & Sons, New York, NY.

Sarstedt, M., Henseler, J., Ringle, C.M., 2011. Multigroup Analysis in partial least squares (PLS) path modeling: alternative methods and empirical results. In: Sarstedt, M., Schwaiger, M., Taylor, C.R. (Eds.), Advances in International Marketing, vol. 22. Emerald Group Publishing Limited, pp. 195–218.

Sato, J.R., Fujita, A., Cardoso, E.F., Thomaz, C.E., Brammer, M.J., Amaro, E., 2010. Analyzing the connectivity between regions of interest: an approach based on cluster Granger causality for fMRI data analysis. Neuroimage 52, 1444–1455.

Seber, G.A.F., 1984. Multivariate Observations. John Wiley and Sons, New York.

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E.J., Glasser, M.F., Ugurbil, K., Barch, D.M., Van Essen, D.C., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci. 18, 1565–1567.

Soneson, C., Lilljebjörn, H., Fioretos, T., Fontes, M., 2010. Integrative Analysis of Gene Expression and Copy Number Alterations Using Canonical Correlation Analysis, vol. 11, p. 191.

Sui, J., Qi, S., van Erp, T.G.M., Bustillo, J., Jiang, R., Lin, D., Turner, J.A., Damaraju, E., Mayer, A.R., Cui, Y., Fu, Z., Du, Y., Chen, J., Potkin, S.G., Preda, A., Mathalon, D.H., Ford, J.M., Voyvodic, J., Mueller, B.A., Belger, A., McEwen, S.C., O'Leary, D.S.,

McMahon, A., Jiang, T., Calhoun, V.D., 2018. Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion. Nat. Commun. 9, 3028.

Tan, K.M., Wang, Z., Liu, H., Zhang, T., 2018. Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow. J. Roy. Stat. Soc. B 80, 1057–1086.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.

Theil, H., 1965. The analysis of disturbances in regression analysis. J. Am. Stat. Assoc. 60, 1067–1079.

Theil, H., 1968. A simplification of the BLUS procedure for analyzing regression disturbances. J. Am. Stat. Assoc. 63, 11.

Timm, N.H., Carlson, J.E., 1976. Part and bipartial canonical correlation analysis. Psychometrika 41, 159–176.

Tracy, C.A., Widom, H., 1996. On orthogonal and symplectic matrix ensembles. Commun. Math. Phys. 177, 727–754.

Tucker, L.R., 1958. An inter-battery method of factor analysis. Psychometrika 23, 111–136.

Turgeon, M., Greenwood, C.M., Labbe, A., 2018. A Tracy-Widom Empirical Estimator for Valid P-Values with High-Dimensional Datasets. arXiv , 1811, 07356.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage 62, 2222–2231.

Waaijenborg, S., Zwinderman, A.H., 2007. Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers. BMC Proc. 1, S122.

Wang, H.T., Smallwood, J., Mourao-Miranda, J., Xia, C.H., Satterthwaite, T.D., Bassett, D.S., Bzdok, D., 2020. Finding the needle in a high-dimensional haystack: canonical correlation analysis for neuroscientists. Neuroimage 216, 116745.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing. John Wiley & Sons, New York, NY.

Wiesel, A., Kliger, M., Hero III, A.O., 2008. A Greedy Approach to Sparse Canonical Correlation Analysis. arXiv:0801.2748 [stat].

Wilks, S.S., 1935. On the independence of k sets of normally distributed statistical variables. Econometrica 3, 309.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 22, 209–212.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.

Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., Smith, S.M., 2015. Multi-level block permutation. Neuroimage 123, 253–268.

Winkler, A.M., Ridgway, G.R., Douaud, G., Nichols, T.E., Smith, S.M., 2016. Faster permutation inference in brain imaging. Neuroimage 141, 502–516.

Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Kågström, B., Ruhe, A. (Eds.), Matrix Pencils, ume 973. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 286–293.

Worsley, K.J., 1997. An overview and some new developments in the statistical analysis of PET and fMRI data. Hum. Brain Mapp. 5, 254–258.

Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkins, M.E., Cook, P.A., Garcáa de la Garza, A., Vandekar, S.N., Cui, Z., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C., Gur, R.E., Shinohara, R.T., Bassett, D.S., Satterthwaite, T.D., 2018. Linked dimensions of psychopathology and connectivity in functional brain networks. Nat. Commun. 9, 3003.

Yoo, J.K., Kim, H.Y., Um, H.Y., 2012. Canonical correlation: permutation tests and regression. Commun. Stat. Appl. Method 19, 471–478.