Tests of independence based on correlations

Robert Schlicht
Department of Forest Sciences, TU Dresden
Pienner Str. 8, 01737 Tharandt, Germany
robert.schlicht@tu-dresden.de

2026

Abstract

This article closes some gaps regarding the computational equivalence of several classical statistical procedures, which is largely known, but whose formal derivation remains incomplete and in some instances anecdotal. First, a result by Mardia and Kent (1991) on Rao's score test of independence based on independent identically distributed pairs of multivariate observations is extended to observations that include covariates and particular forms of dependence. Generalizing special cases observed by several authors, it is then shown that various classical statistical tests represent the same procedure, with differences only in the potential prior elementary transformation of the variables, in the level of conditioning on which the statistical model is presented and in the choice of the approximation of the distribution of the test statistic. This includes F-tests, tests in multivariate linear models based on Pillai's statistic, parametric and nonparametric tests of correlation, Kruskal–Wallis tests, standard approximate versions of Wilcoxon rank-sum and signed rank tests, χ^2-tests of independence, score tests in multinomial logit models, among others. Recognizing such relationships can help in teaching, in the theoretical analysis and in the implementation of those procedures.

1 Introduction

For data X with probability density functions proportional to $x \mapsto e^{s(x)^\top \theta}$, where θ is from an open subset of $\mathbb{R}^m$, it is well known that the test statistic of Rao's score test of the hypothesis that θ is in a given linear subspace of $\mathbb{R}^m$ is the squared Mahalanobis distance of $s(X)$ from $\mathbb{E}_\theta s(X)$ with respect to $\mathrm{Var}_\theta(s(X))$ evaluated at a value θ from that subspace such that $s(X) - \mathbb{E}_\theta s(X)$ is orthogonal to the subspace; see Section 2.1 for details.

Mardia and Kent (1991) analyze several instances of this test, and in particular, for n independent identically distributed pairs of multivariate observations, a hypothesis under which the two components of the pair are independent. They show that in some important cases the score statistic is then equal to n times the sum of the squares of the sample canonical correlations.

The purpose of this article is to extend this result to a more general linear model setting that includes covariates and particular forms of dependence of the observations, and to demonstrate how this provides an intuitive and powerful framework that transparently handles many classical statistical procedures. The resulting test covers not only common tests in correlation and regression analysis, including the usual F-tests and their multivariate extensions based on Pillai's statistic, but most other classical parametric and nonparametric tests traditionally taught in introductory statistics courses, perhaps after a prior elementary transformation of the involved variables. See Table 1 for an overview.

While the general derivation of the equivalent computations underlying all of those procedures is new, several special cases have been widely used and recognized by many authors over the decades since Pearson's investigation of correlation coefficients, which already contains rudiments of some of the computations found below (Pearson, 1896, p. 265). To name a few, this includes relationships of χ^2-tests to multivariate normal distributions (Pearson, 1900), of rank correlations to correlation coefficients (Spearman, 1904), of the Wilcoxon rank-sum test to an analysis of variance (Kruskal, 1952), of linear models, χ^2-tests of independence and linear discriminant analyses to canonical correlations (Knapp,

A	X	Y	X given?	
any	normal	normal	no	multivariate test for correlation (4.1)
any	any	normal	yes	tests in multivariate linear regression, F-test (4.2)
zero	$(1 \quad \cdots \quad 1)^\top$	normal	yes	one-sample t-test (4.2)
	indicator	normal	yes	one-way analysis of variance (4.2)
$\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$	any	indicator	yes	score test for constant in logit regression (4.3)
	indicator	indicator	no	χ^2-test of independence (4.4)
	indicator	ranks	yes	Kruskal–Wallis test (4.5)
	ranks	ranks	no	test based on Spearman rank correlations (4.6)
zero	abs. val., ranks	signs	no	Wilcoxon signed rank test (nonzero data) (4.7)

Table 1: Examples (classical parametric and nonparametric tests). Details are discussed in the subsection of Section 4 indicated in parentheses. The first three columns represent the choice of or transformations underlying A, X, Y.

1978), of a generalized form of χ^2-tests of independence to logit regression (Day and Byar, 1979), of score tests of independence to canonical correlations (Mardia and Kent, 1991) and of various tests in 2×2 contingency tables to each other (Martín Andrés et al., 1995).

A crucial observation that has apparently evaded earlier investigations and is implicit in Mardia and Kent (1991) in a special case only, is the fact that Pillai's statistic in multivariate linear regression and correlation analysis is, up to a factor, a score statistic. This property simplifies deriving many relationships and is the central tool in this article.

2 Preliminaries

2.1 Score test of linear hypotheses in exponential families

Consider data X that follow an exponential family having probability density functions $p_\theta(x) \propto e^{s(x)^\top \theta}$, $\theta \in \Theta$, with respect to some measure, where Θ is an open subset of $\mathbb{R}^m$ and s is an $\mathbb{R}^m$-valued measurable function on the set in which X has its values. Then $p_t(x)/p_\theta(x) = e^{s(x)^\top(t-\theta)}/\mathbb{E}_\theta e^{s(X)^\top(t-\theta)}$, so $\mathbb{E}_\theta e^{s(X)^\top(t-\theta)}$ is finite for all θ, t in Θ. Consequently, all moments of $s(X)$ and so the expectation $\mu_\theta = \mathbb{E}_\theta s(X)$ and covariance matrix $V_\theta = \mathrm{Var}_\theta(s(X))$ exist and are finite. These, of course, are standard facts about exponential families (see, e.g., Brown, 1986, Thm. 2.2 and Prop. 1.5), along with the statement that the score vector, or gradient of the log-likelihood function $\theta \mapsto \log p_\theta(X)$ has the form $S_\theta = s(X) - \mu_\theta$. Indeed, the statement we really need, the supergradient property

$$\log p_t(X) - (\log p_\theta(X) + S_\theta^\top(t-\theta)) = \mathbb{E}_\theta \log \tfrac{p_t(X)}{p_\theta(X)} \le \mathbb{E}_\theta \tfrac{p_t(X)}{p_\theta(X)} - 1 = 0$$

for all θ, t in Θ, is elementary, too, and just expresses the nonnegativity of the Kullback–Leibler information (Csiszár, 1975). This also shows that a θ from an affine subspace of $\mathbb{R}^m$, i.e., column space (optionally shifted) of a matrix G, maximizes the likelihood among all such parameter values if and only if S_θ is orthogonal to that subspace in the sense that $S_\theta^\top G = 0$; and any two such estimates, or more generally any θ, t from the subspace with $\mu_t^\top G = \mu_\theta^\top G$, define the same distribution of X because the expression on the left with θ, t exchanged differs in sign only and "$\le$" is an equality only if $p_t(X)/p_\theta(X) = 1$ with probability 1 under θ.

For given data X, suppose such a maximum likelihood estimate under the hypothesis H_0 that the parameter is in a given affine subspace of $\mathbb{R}^m$ exists, and write $\mathbb{E}$, μ, V etc. for the corresponding maximum likelihood estimates of $\mathbb{E}_\theta, \mu_\theta, V_\theta$ etc. Then the test statistic of Rao's *score test* (Rao, 1948) of H_0, or score statistic, is the squared length of US or, in the terminology used henceforth, of $S = s(X) - \mu$ *standardized* with respect to V by means of any U such that $UVU^\top$ is the identity matrix with the rank of V. This does not depend on the choice of U, which is unique up to multiplication by an orthogonal matrix from the left, and is a slightly more general form $S^\top U^\top U S$ of the usual squared Mahalanobis distance with respect to V since we do not require the Fisher information matrix V to possess an inverse $U^\top U$. Also, applying the standardization instead to US for any matrix U whose

row space includes the values of S clearly leads to the same standardized score vectors and hence to the same test statistic. So, in addition to being likelihood-based and thus not depending on the representation of the data, the score test is, by the chain rule, invariant under differentiable parameter transformations with a differentiable inverse. This, again, is immediate from the preceding derivation in the important case of a linear transformation $\theta \mapsto (U^\top)^{-1}\theta$ accompanying an affine transformation $s(x) \mapsto Us(x) + u$ with invertible U and any u, which preserves the exponential family structure $p_\theta(x) \propto e^{(Us(x)+u)^\top((U^\top)^{-1}\theta)}$ and the affine form of H_0.

It is probably worth pointing out that, as usual with tests, probabilistic properties in expressions like $\mathbb{E}_\theta...$ always refer to a copy of the experiment assuming θ were the true parameter value, even if we then insert estimates of θ based on data. To avoid misunderstandings, especially in conjunction with conditional expectations and parameter transformations, we therefore do not use a separate random variable representing a maximum likelihood estimator under H_0, and estimates may not even uniquely exist for all realizations of the data.

The standard approximation of the distribution of the test statistic under H_0, inspired by the large-sample argument for independent observations by Rao (1948), is a χ^2-distribution with a degrees of freedom parameter given by the dimension of the subspace of the elements v of the column space of V orthogonal to the subspace defining H_0, i.e., with $v^\top G = 0$, where G is as above. For every θ the column space of V_θ is the smallest linear subspace that S_θ has, with probability 1, its values in since, for every t, $S_\theta^\top t$ is zero with probability 1 if and only if its variance is zero. So, due to the strictly positive densities, it is here identical for all θ. Still, we avoid the popular approach of reducing the dimension m so that V becomes invertible, which would simplify the presentation in this section, but does not fit with our applications.

A practical limitation is that a maximum likelihood estimate under H_0 may fail to exist, so that the test statistic is not defined. To partially alleviate this, we here extend the definition to include limits of distributions of X under H_0 in the sense that $\mathbb{E}_\theta f(X) \to \mathbb{E}f(X)$ as $\mu_\theta^\top G \to \mu^\top G$ for a suitable class of functions f, where $\mu = \mathbb{E}s(X)$ etc. are the corresponding expectations. Thus, whenever such a limit with $S = s(X) - \mu$ satisfying $S^\top G = 0$ exists, we say $\mathbb{E}$ is an (extended) *maximum likelihood estimate under* H_0 and define the (extended) *score statistic* as the squared length of the S standardized with respect to $V = \mathrm{Var}(s(X))$. The drawback is that the column space of V, and hence the χ^2 approximation, can now depend on the data. Regarding the functions f, while technically we at least need to include $f(x) = s(x)$ and $f(x) = s(x)s(x)^\top$ to obtain convergence of first and second moments of $s(X)$ (or $Us(X)+u$ with U, u as earlier), we would typically assume strong forms of limits that include all $f(x) = g(s(x))e^{s(x)^\top t}$ with bounded continuous g and t in a neighborhood of zero. The latter implies the limit distributions are again part of exponential families of the same form, with respect to different measures, so we have an aggregate exponential family (Barndorff-Nielsen, 1978, Ch. 9.3; Brown, 1986, Ch. 6). In many important examples the existence of such continuous extensions is easy to verify.

An advantage of the score test is that the computation of the moments under H_0 often implies considerable simplifications. In particular, for a given random vector, standardization can be applied separately to component vectors if these are uncorrelated, i.e., the covariance matrix has a block diagonal form, and likewise to the factors of a Kronecker product if these are independent with expectation zero because the covariance matrix then also has a corresponding product form.

A number of authors (see Agresti, 2013, Ch. 4.5.5) recognized that the score statistic in generalized linear models often has a generalized Pearson form. The following examples are relevant in the context of our applications in Section 4.

Example 1 (Poisson GLMs). Suppose Y is a random vector in $\mathbb{R}^n$ with independent Poisson distributed components with expectations $\mathbb{E}_\theta Y_i = e^{a_i+(X\theta)_i}$, $i = 1, ..., n$, where $\theta \in \mathbb{R}^m$ is unknown, $a \in \mathbb{R}^n$ is known (often zero) and X is a known $n \times m$ matrix. Thus, Y has a probability density $p_\theta(y) \propto e^{s(y)^\top\theta}$, where $s(y) = X^\top y$, with respect to the weighted sum $\sum_{y\in\{0,1,...\}^n}(y_1! \cdots y_n!)^{-1}e^{a^\top y} \cdots$. Here we can always combine identical rows in X by replacing the corresponding Y_i (and the factors e^{a_i}) by their sum, which is again Poisson distributed. This results in the same standardized score vectors because the sufficient statistic $s(Y)$ and likelihood function are unaffected by this modification.

Example 2 (multinomial distributions and log-linear models). Continuing Example 1, consider a matrix

M with each row containing at most one element equal to 1 and all others 0, for example, $M = (1 \quad \cdots \quad 1)^\top$. Then densities of the same form as in Example 1, with an extra factor $\mathbf{1}_{M^\top y=N}$ in the weighted sum, also define the conditional distribution of Y given the sums $N = M^\top Y$ in the groups defined by the columns of M, namely, the multinomial distributions with parameters given by the elements N_j of N and the corresponding subvectors of $\mathbb{E}_\theta Y$ divided by $\mathbb{E}_\theta N_j$ (cf. Fisher, 1922, p. 89). Each of these distributions independently describes counts in N_j independent identically distributed trials, so both $\mathbb{E}_\theta(Y \mid N)$ and $\mathrm{Var}_\theta(Y \mid N)$ depend linearly on N. Suppose the column space of X includes the columns of M, so $X^\top \mathbb{E}_\theta(Y \mid N)$ also depends linearly on $s(Y) = X^\top Y$, or, alternatively, assume a particular θ with that property. We can then write $\binom{\mathbb{E}_\theta(s(Y)|N)}{s(Y)-\mathbb{E}_\theta(s(Y)|N)} = U_\theta s(Y)$ with some matrix U_θ and use $U_\theta s(Y)$ instead of $s(Y)$ in computing a standardized score vector at θ. From the definition of the conditional expectation we get

$$\mathrm{Var}_\theta(U_\theta s(Y)) = \begin{pmatrix} \mathrm{Var}_\theta(\mathbb{E}_\theta(s(Y)|N)) & 0 \\ 0 & \mathbb{E}_\theta \mathrm{Var}_\theta(s(Y)|N) \end{pmatrix}, \quad \mathrm{Var}_\theta(U_\theta s(Y) \mid N) = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Var}_\theta(s(Y)|N) \end{pmatrix}$$

in the Poisson and the conditional model. Restricted to the event $\{N = \mathbb{E}_\theta N\}$, the linearity in N implies $\mathbb{E}_\theta(s(Y) \mid N) = \mathbb{E}_\theta s(Y)$ and $\mathrm{Var}_\theta(s(Y) \mid N) = \mathbb{E}_\theta \mathrm{Var}_\theta(s(Y) \mid N)$. So, for given data Y, suppose θ is a parameter value such that the expectation $\mathbb{E}_\theta N$ in the Poisson model equals the observed N, and suppose this is the same realization of N at which we are considering the conditional model. Then the standardized score vector at θ in the Poisson model, which then only the lower part of $U_\theta s(Y)$ contributes to, coincides with the standardized score vector at θ for the same data Y in the conditional model. Since, moreover, S_θ is the same in both models, the condition for θ being a maximum likelihood estimate under the hypothesis H_0 that $X\theta$ belongs to a given linear subspace, is also equivalent. Thus, provided such θ exists (and yields $\mathbb{E}_\theta N = N$ as before, e.g., if the subspace contains the columns of M), we arrive at identical score statistics in the Poisson and the conditional model.

For example, if we take the identity matrix X in Examples 1 and 2, then $\mathrm{Var}_\theta(s(Y)) = \mathrm{Var}_\theta(Y)$ is the diagonal matrix with diagonal $\mathbb{E}_\theta Y$, and the standardized score vector at θ becomes Pearson's statistic (Pearson, 1900)

$$\sum_{i=1}^n \frac{(Y_i - \mathbb{E}_\theta Y_i)^2}{\mathbb{E}_\theta Y_i} = \sum_{i=1}^n \frac{(Y_i - \mathbb{E}_\theta(Y_i \mid N))^2}{\mathbb{E}_\theta(Y_i \mid N)}.$$

Without going into the details, we note that with the extended maximum likelihood estimates introduced above, it would be possible to extend these results to edge cases consisting of estimated distributions with $\mathbb{E}Y_i = 0$ for some i.

2.2 Canonical correlations

If $\tilde{\xi}$ and $\tilde{\eta}$ are standardized versions (in the sense of Section 2.1) of random vectors ξ in $\mathbb{R}^k$ and η in $\mathbb{R}^l$ with expectation zero and covariance matrices with ranks K ($\le k$) and L ($\le l$), the $K \times L$ covariance matrix $\mathbb{E}(\tilde{\xi}\tilde{\eta}^\top)$ is, with a proper choice of $\tilde{\xi}, \tilde{\eta}$ based on the singular value decomposition of that matrix, zero except for nonnegative elements on its main diagonal arranged in decreasing order; these are the canonical correlations of ξ and η introduced by Hotelling (1936, full-rank case).

What we here need is the empirical counterpart in a linear model setting $(X \quad Y) = A\,(\alpha \quad \beta) + (\delta \quad \varepsilon)$ with data matrices X, Y with k and l columns, known A, unknown α, β and error terms δ, ε. Let K and L be the ranks of $\mathcal{X} = X - P_A X$ and $\mathcal{Y} = Y - P_A Y$, with P_A representing the orthogonal projection on the column space of A. Again, with a proper choice of matrices $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ whose columns are orthonormal bases of the column spaces of $\mathcal{X}$ and $\mathcal{Y}$, the matrix $\tilde{\mathcal{X}}^\top \tilde{\mathcal{Y}}$ is zero except for its main diagonal elements $R_1 \ge \cdots \ge R_J \ge 0$, where $J = \min\{K, L\}$. As usual we will refer to those statistics as *canonical correlations* even without distributional assumptions on the error terms.

In the special case of A having identical nonzero rows, typically $A = (1 \quad \cdots \quad 1)^\top$, these are the sample canonical correlations based on simple centering most often presented in the literature for full column ranks $K = k$ and $L = l$ (Hotelling, 1936); if both ranks are 1, then X, Y are column vectors and R_1 is the absolute value of their correlation coefficient. Geometrically, the canonical correlations describe the relative positions of the column spaces of $\mathcal{X}$ and $\mathcal{Y}$ to each other in that they are the

radii of the ellipsoid obtained from orthogonal projection of the unit ball in either of the spaces on the other. If $K > 0$ and $L = l = 1$, then the (only) canonical correlation R_1 is the length of $\tilde{\mathcal{X}}^\top Y$ or $\tilde{\mathcal{X}}\tilde{\mathcal{X}}^\top Y = P_{\mathcal{X}} Y$ divided by the length of $\mathcal{Y}$, and R_1^2 has the familiar form of a quotient of sums of squares.

Suppose either δ or ε has independent rows with an identical multivariate normal distribution with mean zero, and just for clarity note that this implies the corresponding rank K or L equals, with probability 1, the rank of the covariance matrix of this distribution or the dimension r of the orthogonal complement of the column space of A, whichever is smaller, and so is independent of X or Y. Under the hypothesis of independence of X and Y, the joint distribution of $R_1^2, ..., R_J^2$, given K and L, is then known and in the case $r \geq K + L$ has a probability density function (Hsu, 1939; Anderson, 2003, Ch. 12–13; Anderson, 2007) given by

$$\rho(t_1, ..., t_J) \propto \prod_{j=1}^{J} t_j^{(|K-L|-1)/2} (1 - t_j)^{(r-K-L-1)/2} \prod_{j'>j} (t_j - t_{j'})$$

with respect to the Lebesgue measure on the simplex $\{t \in \mathbb{R}^J : 1 \geq t_1 \geq \cdots \geq t_J \geq 0\}$.

For $J = 1$ this reduces to the well-known case of a single beta distributed $R^2 = R_1^2$ or equivalently an F-distributed $\frac{R^2/(KL)}{(1-R^2)/(rJ-KL)}$, with the divisors in the numerator and denominator representing the degrees of freedom, or twice the parameters of the beta distribution. For general $J > 0$, following Pillai (1954, p. 99, p. 44), this is also used for $R^2 = (R_1^2 + \cdots + R_J^2)/J$ as an approximation.

3 Main result

After these preparations, we are ready to state and prove the main result, which extends the statement from Mardia and Kent (1991) to particular forms of dependence and non-identical distributions of the observations. We write $\langle A, B \rangle = \operatorname{tr}(A^\top B) = \sum_{i,j} A_{ij} B_{ij}$ for the standard inner product of matrices rearranged as vectors and $A_1, ..., A_n$ for the transposed rows of a matrix $A = (A_1 \quad \cdots \quad A_n)^\top$.

Model. *Suppose X and Y are $n \times k$ and $n \times l$ random matrices that jointly follow an exponential family*

$$p_\theta(x, y) \propto e^{\langle a, A^\top x \rangle + \langle b, A^\top y \rangle + \langle c, x^\top x \rangle + \langle d, y^\top y \rangle + \langle e, x^\top y \rangle} \tag{1}$$

with respect to some measure, where θ, from an open set Θ, consists of the elements of the parameter matrices a, b, c, d, e and A is a known matrix with n rows. We assume that for every realization $\mathbb{E}$ of $\mathbb{E}_\theta$ under $H_0 : e = 0$, the matrices $\mathcal{X} = X - P_A \mathbb{E}X$, $\mathcal{Y} = Y - P_A \mathbb{E}Y$ satisfy

$$\gamma \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}(\mathcal{X}_i \mathcal{X}_j^\top \otimes \mathcal{Y}_i \mathcal{Y}_j^\top) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}(\mathcal{X}_i \mathcal{X}_i^\top) \otimes \mathbb{E}(\mathcal{Y}_j \mathcal{Y}_j^\top) \tag{2}$$

with a known constant γ and are such that $\mathcal{X}^\top \mathcal{Y}$

(i) has expectation zero,

(ii) is uncorrelated to $A^\top \mathcal{X}$ and $\mathcal{X}^\top \mathcal{X}$,

(iii) is uncorrelated to $A^\top \mathcal{Y}$ and $\mathcal{Y}^\top \mathcal{Y}$.

Remarks. (a) The situation we are interested in is that X, Y (and hence $\mathcal{X}, \mathcal{Y}$) are independent under H_0; this is clearly equivalent to a corresponding product structure of the reference measure. With this assumption, conditions (i), (ii) and (i), (iii), respectively, follow if $\mathbb{E}Y$ or $\mathbb{E}X$ has columns in the column space of A because the product of two independent random variables, one of which has expectation zero, has expectation zero and is uncorrelated to expressions involving only the other one.

(b) Condition (2) states that the vector representation $\sum_{i=1}^{n} \mathcal{X}_i \otimes \mathcal{Y}_i$ of $\mathcal{X}^\top \mathcal{Y}$ has γ times a second-moment matrix equal to $\mathbb{E}(\mathcal{X}^\top \mathcal{X}) \otimes \mathbb{E}(\mathcal{Y}^\top \mathcal{Y})$. We only strictly need this condition and (i)–(iii) at a maximum likelihood estimate under H_0.

(c) Since $x^\top x$ and $y^\top y$ are symmetric, it is possible to work, with adjustments to s in the following proof, with symmetric c, d represented by (usually) fewer elements in θ.

Proposition. *Assuming the Model, if* $\mathbb{E}$ *represents a maximum likelihood estimate under* H_0 *(possibly in the limit sense from Section 2.1), then* $\mathcal{X} = X - P_A X$, $\mathcal{Y} = Y - P_A Y$, *and the score statistic is the squared length of the vector* $\sum_{i=1}^n \mathcal{X}_i \otimes \mathcal{Y}_i$ *of elements of* $\mathcal{X}^\top \mathcal{Y}$ *standardized with respect to its covariance matrix, which, multiplied by* γ, *equals* $\mathcal{X}^\top \mathcal{X} \otimes \mathcal{Y}^\top \mathcal{Y}$.

Proof. Let $s(x, y)$ be the vector of the elements of the matrices $A^\top x$, $A^\top y$, $x^\top x$, $y^\top y$, $x^\top y$ in the exponent of (1). For any fixed realization $\mathbb{E}$ of $\mathbb{E}_\theta$ under H_0, writing $P_A \mathbb{E} \begin{pmatrix} X & Y \end{pmatrix} = A \begin{pmatrix} \alpha & \beta \end{pmatrix}$ with some α, β and noting that $\mathcal{X}^\top \mathcal{X} = X^\top X - \alpha^\top (A^\top X) - (X^\top A)^\top \alpha + \alpha^\top A^\top A \alpha$ etc., we see that $s(\mathcal{X}, \mathcal{Y})$ is the result of an affine transformation applied to $s(X, Y)$, and we can substitute this on the right side of (1) as discussed in Section 2.1, while replacing a and b with shifted versions $a + \alpha(c + c^\top) + \beta e^\top$ and $b + \beta(d + d^\top) + \alpha e$. In particular, in view of (b), all assumptions on the covariance structure can be written in terms of the first and second moments of $s(X, Y)$ and so, by continuity, extend to limits $\mathbb{E}$ as introduced in Section 2.1 and the corresponding $\mathcal{X}, \mathcal{Y}$ if they hold for all non-limit realizations under H_0. Also, the substitution just presented remains valid for the limit case.

So we can compute standardized score vectors, possibly in the limit sense, from $s(\mathcal{X}, \mathcal{Y})$ minus its expectation, which we here do for the distribution represented by the same $\mathbb{E}$ we use for the substitution. With the assumptions on the covariance structure, the elements of $s(\mathcal{X}, \mathcal{Y})$ representing $\mathcal{X}^\top \mathcal{Y}$ are uncorrelated to the others and have expectation zero and γ times a covariance matrix equal to $\mathbb{E}(\mathcal{X}^\top \mathcal{X}) \otimes \mathbb{E}(\mathcal{Y}^\top \mathcal{Y})$. The uncorrelatedness, or block diagonal structure of the joint covariance matrix, means we can apply the standardization to $\mathcal{X}^\top \mathcal{Y}$ and the other components of the score vector separately.

Now, if $\mathbb{E}$ is a maximum likelihood estimate under H_0 for given data X, Y, we also know the expectations of those other components of $s(X, Y)$ equal their observed values, and the same follows for $s(\mathcal{X}, \mathcal{Y})$, which can be verified directly or by noting that H_0 is invariant under the above parameter transformation. So in that case only the standardized $\mathcal{X}^\top \mathcal{Y}$ contributes to the score statistic, and $\mathbb{E}$ also disappears in the definitions of $\mathcal{X}, \mathcal{Y}$. □

So the score statistic becomes γ times the sum of squared elements of $\mathcal{X}^\top \mathcal{Y}$ after standardization in the rows and columns with respect to $\mathcal{X}^\top \mathcal{X}$ and $\mathcal{Y}^\top \mathcal{Y}$, respectively, and the degrees of freedom parameter of the χ^2 approximation from Section 2.1 equals the rank of $\mathcal{X}^\top \mathcal{X} \otimes \mathcal{Y}^\top \mathcal{Y}$:

Corollary. *Assuming the Model, the score statistic of* H_0, *if defined, equals* γ *times Pillai's statistic* $\sum_{j=1}^J R_j^2$, *where* $R_1, ..., R_J$ *are the canonical correlations as defined in Section 2.2, and the approximation of its distribution under* H_0 *from Section 2.1 is a* χ^2*-distribution with* KL *degrees of freedom.*

Indeed, in Section 4 we consider tests based on approximating the distribution of $\sum_{j=1}^J R_j^2$ by either of the following: the distribution of J times a random variable having a beta distribution with parameters $KL/2$ and $(rJ - KL)/2$, following Pillai's approach to the normal case from Section 2.2, with possibly stochastic J *(beta approximation);* and a gamma distribution with shape parameter $KL/2$ and rate parameter $\gamma/2$ as in the Corollary *(gamma approximation).*

4 Special cases

Consider θ under H_0. In our applications we always assume X, Y are independent for θ and $\mathbb{E}_\theta Y$ has columns in the column space of A, i.e., $\mathcal{Y}$ has expectation zero, or equivalently $\mathcal{Y} = Y - \mathbb{E}_\theta Y$; so properties (i), (ii) are satisfied. We also assume $\mathrm{Cov}_\theta(Y_i, Y_j) = \begin{cases} \gamma V & \text{if } i = j \\ (\gamma - n) V & \text{if } i \neq j \end{cases}$ for all i, j and some matrix V, either with $\gamma = n$ or additionally assuming X such that $\sum_{j=1}^n \mathcal{X}_j = 0$; this implies (2) because the difference of the terms for given i on the right and left side of (2) is, by independence of $\mathcal{X}$ and $\mathcal{Y}$,

$$\mathbb{E}_\theta(\mathcal{X}_i \mathcal{X}_i^\top) \otimes n\gamma V - \gamma \sum_{j=1}^n \mathbb{E}_\theta(\mathcal{X}_i \mathcal{X}_j^\top) \otimes \begin{cases} \gamma V & \text{if } i = j \\ (\gamma - n) V & \text{if } i \neq j \end{cases} = \gamma \mathbb{E}_\theta(\mathcal{X}_i \textstyle\sum_{j=1}^n \mathcal{X}_j^\top) \otimes (n - \gamma) V = 0.$$

This assumption is satisfied with

- $\gamma = n$ if $Y_1, ..., Y_n$ are uncorrelated with identical covariance matrix ($= nV$),

- $\gamma = n - 1$ if $X_1 + \cdots + X_n$ and $Y_1 + \cdots + Y_n$ are known, $\mathrm{Cov}_\theta(Y_i, Y_j)$ is identical ($= -V$) for all pairs $i \neq j$, and either $(1 \quad \cdots \quad 1)^\top$ or the columns of $\mathbb{E}_\theta X$ are also in the column space of A; here $\mathrm{Var}_\theta(\sum_{i=1}^n Y_i) = 0$, and the last condition implies $(1 \quad \cdots \quad 1)\,(I - P_A)\mathbb{E}_\theta X = 0$, so the sum of rows of $\mathcal{X}$ equals that of $X - \mathbb{E}_\theta X$, which is zero.

To satisfy the remaining property (iii), we additionally require that one of the following conditions holds under H_0:

A. $\mathbb{E}_\theta X$ *also has columns in the column space of* A.

B. $A^\top Y$ *and* $Y^\top Y$ *are known.* Then $A^\top\mathcal{Y}$ and $\mathcal{Y}^\top\mathcal{Y}$ are independent of $\mathcal{X}^\top\mathcal{Y}$.

C. $Y_1, ..., Y_n$ *are uncorrelated with an identical covariance matrix and distributions that are symmetric about their expectations.* For any two columns of u, v of $\mathcal{Y}$ this implies $\mathrm{Cov}_\theta(A^\top u, \mathcal{X}^\top v) = A^\top\mathbb{E}_\theta(uv^\top)\mathbb{E}_\theta\mathcal{X} = A^\top\mathbb{E}_\theta(uv^\top)(I - P_A)\mathbb{E}_\theta X = 0$ since $\mathbb{E}_\theta(uv^\top)$ is a multiple of the identity matrix; this proves $A^\top\mathcal{Y}$ and $\mathcal{X}^\top\mathcal{Y}$ are uncorrelated. Similarly $\mathbb{E}_\theta((\mathcal{Y}^\top u)(\mathcal{X}^\top v)^\top) = \sum_{i,j}\mathbb{E}_\theta(\mathcal{Y}_i u_i v_j)\mathbb{E}_\theta\mathcal{X}_j = 0$ because the distribution of $\mathcal{Y}_i u_i v_j$ is symmetric about 0 for all i, j; this proves $\mathcal{Y}^\top\mathcal{Y}$ and $\mathcal{X}^\top\mathcal{Y}$ are uncorrelated.

D. $Y_1, ..., Y_n$ *are independent identically distributed, and* $(1 \quad \cdots \quad 1)^\top$ *is in the column space of* A. As in C we see that $A^\top\mathcal{Y}$ and $\mathcal{X}^\top\mathcal{Y}$ are uncorrelated, and here, in the sum $\sum_{i,j}\mathbb{E}_\theta(\mathcal{Y}_i u_i v_j)\mathbb{E}_\theta\mathcal{X}_j$, the terms with $i \neq j$ are zero because $\mathcal{Y}$ has independent rows and $\mathbb{E}_\theta v_j = 0$, and the terms with $i = j$ have the sum $\mathbb{E}_\theta\,(\mathcal{Y}_1 u_1 v_1 \quad \cdots \quad \mathcal{Y}_n u_n v_n)\,(I - P_A)\mathbb{E}_\theta X$, which is also zero because the transpose of the first expectation has identical rows and hence columns in the column space of A.

As we will now see, this covers several well-known procedures. Let us begin with two major applications in correlation and regression analysis, the second one of which can be seen as a conditional version, given X, of the first one:

4.1 Multivariate test for correlation

Consider the multivariate linear model $(X \quad Y) = A\,(\alpha \quad \beta) + (\delta \quad \varepsilon)$ with known A, unknown α, β and error terms $(\delta \quad \varepsilon)$ having independent $\mathcal{N}(0, \Sigma)$ rows with an invertible $\Sigma = -\left(\begin{smallmatrix} c+c^\top & e \\ e^\top & d+d^\top \end{smallmatrix}\right)^{-1}$ and $(\alpha \quad \beta) = (a \quad b)\,\Sigma$. The joint distribution, with parameter combination θ, has a probability density function of the form (1),

$$p_\theta(x, y) \propto e^{-\frac{1}{2}\mathrm{tr}\left(((x\ y) - A(\alpha\ \beta))\Sigma^{-1}((x\ y) - A(\alpha\ \beta))^\top\right)} \propto e^{\langle -\frac{1}{2}\Sigma^{-1}, (x\ y)^\top(x\ y)\rangle + \langle(\alpha\ \beta)\Sigma^{-1}, A^\top(x\ y)\rangle}$$

with respect to Lebesgue measure. The hypothesis $H_0 : e = 0$ is here equivalent to the independence of X and Y, and we are in case A with $\gamma = n$. The score test from the Corollary with the beta approximation, which is here exact if $J = 1$, is one the procedures (Pillai, 1954) commonly found in textbooks on multivariate statistics.

4.2 Linear regression: test of linear hypothesis

Consider the multivariate linear model $Y = A\alpha + X\beta + \varepsilon$ with known A, X, unknown α, β and error terms ε having independent $\mathcal{N}(0, \Sigma)$ rows with an invertible $\Sigma = -(d + d^\top)^{-1}$ and $(\alpha \quad \beta) = (b \quad e)\,\Sigma$. In this case

$$p_\theta(x, y) \propto e^{\langle a, A^\top x\rangle + \langle c, x^\top x\rangle} e^{-\frac{1}{2}\mathrm{tr}\left((y - A\alpha - x\beta)\Sigma^{-1}(y - A\alpha - x\beta)^\top\right)},$$

which again follows (1), provides a probability density function of (X, Y) with respect to the product of the point mass at the constant value of X, where only values at that particular x are relevant and a, c have no effect on the distribution, and the Lebesgue measure. The hypothesis $H_0 : e = 0$ is equivalent to $\beta = 0$, which means X can be removed from the regression function, and we are in case C with $\gamma = n$. Again, the test with the beta approximation, which is exact if $J = 1$, is a common procedure in multivariate linear regression, including multivariate analysis of variance models, with

the most popular alternative being Wilks' statistic $\prod_{j=1}^{J}(1-R_j^2)$, which is here known to lead to a test equivalent to the likelihood ratio test (Wilks, 1932). In the case $L = l = 1$ and nonzero K all of those tests coincide with the classical F-test.

The applications in Sections 4.1 and 4.2 can both be extended to singular Σ with the procedure from Section 2.1 and the strong form of limits discussed there. Also, in Sections 4.1 and 4.2, a model of the same form and with the same parameters, but with reduced n, is obtained by multiplying A, X, Y from the left by a matrix whose rows are an orthonormal basis of a subspace containing the orthogonal complement of the column space of A. This does not change the column spaces of $\mathcal{X}$ and $\mathcal{Y}$, so the canonical correlations and the shape parameters of the approximating beta and gamma distributions stay the same. The invariance we have for sufficient statistics does not apply, though, and the reduced $\gamma = n$ is directly reflected in a smaller score statistic, which is relevant in the gamma approximation. The computations in Section 4.1 in the original and maximally reduced forms correspond, respectively, to ML and REML variance estimation under H_0 for the rows of Y that the statistic $\mathcal{X}^\top\mathcal{Y}$ from the Proposition linearly depends on (Patterson and Thompson, 1971).

4.3 Multinomial logit regression: score test of constant function

Consider the matrix $Y = (\mathbf{1}_{\tilde{Y}_i=\tilde{y}_j})_{i,j}$ of dummy variables representing indicator-encoded independent $\tilde{Y}_i$, $i = 1, ..., n$, each of which has, with probabilities proportional to $e^{(A\alpha+X\beta)_{ij}}$, $j = 1, ..., l$, one of l distinct values $\tilde{y}_j$, where the proportionality constant may depend on i, and A, X are known and α, β unknown matrices. We are interested in $H_0 : \beta = 0$.

In the special case $A = (1 \quad \cdots \quad 1)^\top$, we can apply our test from Section 3: For $\alpha = b + (\text{diagonal of } d)$ and $\beta = e$, (1) supplies a probability density function of (X, Y) with respect to the sum of point masses at its possible realizations, restricted to which this density has the form $p_\theta(x, y) \propto e^{\operatorname{tr}(\alpha^\top A^\top y + \beta^\top x^\top y)}$ since x is fixed and $y^\top y$ is the diagonal matrix with diagonal elements $A^\top y$. The hypothesis $H_0 : e = 0$ is again equivalent to $\beta = 0$, which implies $\tilde{Y}_1, ..., \tilde{Y}_n$ are identically distributed and $\mathbb{E}_\theta Y$ has columns in the column space of A, and we have case D with $\gamma = n$. Best known is the classical logistic regression case $l = 2$ with binary response data and log-odds $\log \mathbb{E}_\theta Y_{i1} - \log \mathbb{E}_\theta Y_{i2}$ for the first value.

More generally, since the rows of Y have multinomial distributions and the elements of $A\alpha + X\beta$ depend linearly on those of α, β, the model could also be analyzed as a special case of Example 2 and hence by procedures for the score test in Poisson GLMs. To be specific, we would consider the model in which the elements of Y are independent Poisson distributed with expectations $\mathbb{E}_\theta Y_{ij} = e^{\lambda_i + (A\alpha+X\beta)_{ij}}$, $i = 1, ..., n$, $j = 1, ..., l$. Since we keep the matrix arrangement of the elements of Y, the vector of group sizes from Example 2, $N = Y(1 \quad \cdots \quad 1)^\top$, here represents the sums of elements in each row and is obtained by a multiplication from the right. In the special case $A = (1 \quad \cdots \quad 1)^\top$, under $H_0 : \beta = 0$, where $\mathbb{E}_\theta Y_{ij} = e^{\lambda_i} e^{\alpha_j}$ has a simple product form, a maximum likelihood estimate $\mathbb{E}$ in the sense of Section 2.1 is easily computed since $\mathbb{E}Y$ has to reproduce the observed row and column sums, and thus, for multinomial data with row sums equal to 1, has rows all equal to some vector, say η. We therefore obtain the same estimate $\mathbb{E}$ in a simpler model with λ removed and then also the same score statistic because for any fixed realizations of $\mathbb{E}, \eta$ under H_0 the elements of $(A \quad X)^\top \mathbb{E}(Y \mid N) = (A \quad X)^\top N\eta^\top$ depend linearly on those of $s(Y) = (A \quad X)^\top Y$, cf. Example 2. A practical approach is to replace λ with an expression of the form $X\lambda$ that additionally preserves the degrees of freedom computation.

4.4 χ^2-test of independence

Suppose $X = (\mathbf{1}_{\tilde{X}_i=\tilde{x}_j})_{i,j}$ and $Y = (\mathbf{1}_{\tilde{Y}_i=\tilde{y}_{j'}})_{i,j'}$ are both matrices of dummy variables representing independent $(\tilde{X}_i, \tilde{Y}_i)$, $i = 1, ..., n$, each of which has one of $k \times l$ distinct values $(\tilde{x}_j, \tilde{y}_{j'})$ with probabilities proportional to $e^{(A\alpha)_{ij} + (A\beta)_{ij'} + e_{jj'}}$, $j = 1, ..., k$, $j' = 1, ..., l$. We again consider $A = (1 \quad \cdots \quad 1)^\top$, where the last expression becomes $e^{\alpha_j + \beta_{j'} + e_{jj'}}$ and the rows (X_i, Y_i) are identically distributed. Applying Example 2 to the contingency table $Z = X^\top Y$, which is a sufficient statistic and multinomially distributed, we have a corresponding Poisson model that, as in 4.3, has a simple product form $\mathbb{E}_\theta Z_{jj'} = e^{\alpha_j} e^{\beta_{j'}}$ under $H_0 : e = 0$, where a maximum likelihood estimate $\mathbb{E}Z$ has to repreduce the observed row and column sums. This shows the score statistic of H_0 equals Pearson's statistic for the test

of independence (Fisher, 1922). On the other hand, as in 4.3, for $\alpha = a +$ (diagonal of c) and $\beta = b +$ (diagonal of d), (1) supplies a probability density function of (X, Y) with respect to the sum of point masses at its possible realizations, restricted to which $p_\theta(x, y) \propto e^{\langle A\alpha, x\rangle + \langle A\beta, y\rangle + \langle e, x^\top y\rangle}$. The hypothesis H_0 implies $(\tilde{X}_1, ..., \tilde{X}_n)$ and $(\tilde{Y}_1, ..., \tilde{Y}_n)$ are independent, and we are in case A with $\gamma = n$.

4.5 Random permutations, Kruskal–Wallis and Wilcoxon rank-sum tests

In this and the remaining applications 4.6–4.7, we consider hypotheses H_0 consisting of a single distribution of (X, Y) only. We can always, as we here do, take this as the reference measure for the exponential family, that is, the distribution for a, b, c, d, e all zero. So the only effect of the exponential family assumption (1) is the specification of the alternatives. We will see that several classical nonparametric tests are computationally equivalent to the score tests from Section 3 after transforming the data into ranks and conditioning on the structure of ties. This is a formal equivalence because the particular alternatives investigated in our test are not precisely those that make most sense on the level of the untransformed data, where they could depend on the sample size, for example.

We begin with a slightly more general test of a uniformly distributed random permutation. Let $A = (1 \quad \cdots \quad 1)^\top$, and suppose x and y are known matrices with n rows. Let Π be a random permutation of $\{1, ..., n\}$ such that the corresponding permutation matrix $(\mathbf{1}_{\Pi(i)=j})_{i,j}$ has a probability density $q_\theta(\pi) \propto e^{\langle e, x^\top \pi y\rangle}$ with respect to the sum of point masses at its possible realizations. So Π has a uniform distribution under $H_0 : e = 0$. Let $X = x$ and Y the result of applying the permutation matrix to y. These are clearly sufficient statistics, thus yielding the same score statistic of H_0, and they have a joint density following the exponential family (1), where the parameters a, b, c, d have no effect since $A^\top X, A^\top Y, X^\top X, Y^\top Y$ are all known. Under H_0, the rows $y_{\Pi(1)}, ..., y_{\Pi(n)}$ of Y are identically distributed, so $\mathbb{E}_\theta Y$ has columns in the column space of A, and we are in case B with $\gamma = n - 1$ and can handle this with our test.

Now assume $X = x$ is a known matrix of dummy variables defining K non-empty groups of observations, and consider the case $l = 1$ and $y_1, ..., y_n$ not all equal, so $J = 1$. By the sums-of-squares computation of R_1^2, the score statistic then becomes $\gamma \sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2 / \sum_{i=1}^n (Y_i - \bar{Y})^2$, where $\bar{Y}$ is the mean of all $Y_1, ..., Y_n$ and $\bar{Y}_i$ is the mean in the group that observation i belongs to. The denominator is known and hence equals its expectation, which under H_0, where $\mathbb{E}_\theta Y_i = \bar{Y}$, is $\sum_{i=1}^n \mathrm{Var}_\theta(Y_i)$. For example, suppose $Y_1, ..., Y_n$ present ranks of data $\tilde{Y}_1, ..., \tilde{Y}_n$ on an ordinal scale, with mean ranks assigned in case of ties. Then, given the structure of the ties (i.e., which ranks occur and how often) and assuming $\tilde{Y}_1, ..., \tilde{Y}_n$ are independent identically distributed under the hypothesis to be tested, we can apply our test. If there are no ties, each Y_i has the uniform distribution on the integers $1, ..., n$, so the test statistic coincides with that of the Kruskal–Wallis test as derived by Kruskal (1952, p. 526), and this remains true if there are ties because the correction factor derived by Kruskal (1952, p. 538) from the second-moment behavior is precisely the one that affects our denominator $\sum_{i=1}^n \mathrm{Var}_\theta(Y_i)$. The χ^2-distribution with $K - 1$ degrees of freedom assumed in that test corresponds to our gamma approximation.

As discussed by Kruskal and Wallis (1952) and Kruskal (1952), the test includes, in the special case of two groups, the Wilcoxon rank-sum or Mann–Whitney test (Mann and Whitney, 1947), although in that case often either the exact distribution of the test statistic is used or an additional continuity correction is applied.

4.6 Test based on Spearman rank correlations

Let $A = (1 \quad \cdots \quad 1)^\top$, and suppose the rows of both X and Y are obtained by uniformly distributed random permutations, independent under $H_0 : e = 0$, from known values $x_1, ..., x_n$ and $y_1, ..., y_n$, respectively, as in 4.5. Then $A^\top X, A^\top Y$, $X^\top X, Y^\top Y$ are all known, so again the parameters a, b, c, d have no effect. Because the rows are identically distributed, both $\mathbb{E}_\theta X$ and $\mathbb{E}_\theta Y$ always have columns in the column space of A, and we have case A with $\gamma = n - 1$. In the case $k = l = 1$ and neither $x_1, ..., x_n$ nor $y_1, ..., y_n$ all equal, the test statistic becomes γ times the squared sample correlation coefficient of X and Y. As in 4.5 we can consider the situation where the rows present ranks of independent identically distributed data $\tilde{X}_1, ..., \tilde{X}_n$ and $\tilde{Y}_1, ..., \tilde{Y}_n$, respectively, on ordinal scales, with mean ranks

assigned in case of ties. Given the structure of the ties and assuming independence of $(\tilde{X}_1, ..., \tilde{X}_n)$ and $(\tilde{Y}_1, ..., \tilde{Y}_n)$ under the hypothesis being tested, we are in the situation described. If neither the rows of X nor those of Y are all equal, the test statistic is $n-1$ times the squared sample correlation coefficient of the ranks. With the beta approximation, i.e., the formal application of the test from Section 4.1 to the ranked data, this becomes the usual approximate procedure for testing independence based on Spearman's rank correlation coefficient (Spearman, 1904).

4.7 Wilcoxon signed rank test

Suppose A is zero, the rows of X are obtained by a uniformly distributed random permutation from known values $x_1, ..., x_n$ as in 4.5 and 4.6, and those of Y have values $+1$ and -1 and, under H_0 : $e = 0$, take both signs with equal probability and are independent and also independent of X. Since $A^\top X, A^\top Y$, $X^\top X, Y^\top Y$ are then all known, the parameters a, b, c, d of the exponential family (1) again have no effect on the distribution. Under H_0, the values $Y_1, ..., Y_n$ are independent identically distributed with expectation zero, and we have case C with $\gamma = n$, as in 4.2. By the Proposition, in the case $k = 1$ $(= l)$ and neither $x_1, ..., x_n$ nor $Y_1, ..., Y_n$ all equal, the score statistic is the square of $\sum_{i=1}^n X_i Y_i$ standardized with respect to its variance under H_0. For example, suppose the rows of X and Y are, respectively, the ranks of the absolute values, with mean ranks assigned in case of ties, and the signs of independent identically distributed data in $\mathbb{R} \setminus \{0\}$. Given the structure of the ties and assuming a distribution symmetric about 0 under the hypothesis to be tested, we are in the situation just described. Since $\sum_{i=1}^n X_i Y_i$ is, up to a linear transformation, the test statistic of the Wilcoxon signed rank test (Wilcoxon, 1945), the usual large-sample normal approximation of this statistic is therefore equivalent to assuming a χ^2-distribution with $KL = 1$ degree of freedom for the score statistic, which is our gamma approximation.

5 Discussion

Building upon earlier observations by Mardia and Kent (1991) and others, we showed that several classical statistical tests and a number of advanced procedures are, on a computational level, special cases of the score test in a particular exponential family. These procedures, which despite their limitations and known deficiencies remain some of the most widely used in many branches of sciences, can in this respect be seen as applications of the multivariate extension of late 19th century correlation analysis.

Recognizing the equivalences can help in several areas. In teaching, the exposition can become simpler if the question is not about which test to apply, but which prior transformation (e.g., indicator variables or ranking) data should be subjected to before a uniform procedure is applied. The theoretical analysis can benefit from the presentation of the procedures in a unified framework. The implementation in software can make efficient use of identical underlying computations, as the various special computations devised at a time when calculations had to be done by hand no longer provide much advantage in a computer; a reference implementation is available in the R package provided at `https://cran.r-project.org/package=cctest`.

While in our applications it is always correct to say that we are testing a strong form of independence of X and Y in that we assume independence is implied by H_0, there are further arguments that justify the informal interpretation as a *test of independence.* In some cases (4.1, 4.4 with $n > 0$), H_0 or perhaps the slightly broader hypothesis that X, Y have the same joint distribution as for some parameter combination under H_0, is in fact equivalent to independence. In the cases with given X (4.2, 4.3, 4.5), we can consider the same model for different choices of X while keeping the reference measure for Y fixed, cf. Remark (a); then H_0 implies the distribution of Y, which could be interpreted as a conditional distribution, is functionally independent of X. In the cases where X or Y are obtained by a transformation (4.3–4.7), a similar reasoning applies to the underlying hypothesis.

Pillai's statistic can be computed as the sum of squared elements (or squared Frobenius norm) of $\tilde{\mathcal{X}}^\top \tilde{\mathcal{Y}}$ for any $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ whose columns are orthonormal bases of the column spaces of $\mathcal{X}$ and $\mathcal{Y}$. So the computationally expensive singular value decomposition from Section 2.2 is not strictly needed, although the diagonal form clarifies the relationship to other test statistics (see 4.2) and to exploratory

techniques such as linear discriminant analysis and correspondence analysis. In view of the fact that we arrive, at least for full ranks K, L, at a simple rational function of the data for the test statistic, the rather limited setting – the dual role of A in the Model and the normal or identical distribution under H_0 in the applications – is not surprising.

The results presented in this article apply to the computational level of the test statistics and they provide no justification of the tests at all. Indeed, the limited assumptions regarding the reference measure mean examples in which the test from Section 3 performs poorly can be easily constructed. The classical building block of a theoretical justification, forming a significant part of the literature on the procedures referred to in Section 4, are, of course, asymptotic results on the large sample behavior. As these rarely make guarantees for a given sample size, Monte Carlo analysis of a procedure in the specific setting is an essential tool, too.

References

A. Agresti. *Categorical data analysis*. John Wiley & Sons, 3rd edition, 2013.

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Interscience, 3rd edition, 2003.

T. W. Anderson. Multiple discoveries: Distribution of roots of determinantal equations. *J. Statist. Plann. Inference*, 137(11):3240–3248, 2007. https://doi.org/10.1016/j.jspi.2007.03.008.

O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons, Ltd., 1978. https://doi.org/10.1002/9781118857281.

L. D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, 1986. https://doi.org/10.1214/lnms/1215466757.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3(1):146–158, 1975. https://doi.org/10.1214/aop/1176996454.

N. E. Day and D. P. Byar. Testing hypotheses in case-control studies: Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics*, 35(3):623–630, 1979. https://doi.org/10.2307/2530253.

R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. Roy. Statist. Soc.*, 85(1):87–94, 1922. https://doi.org/10.1111/j.2397-2335.1922.tb00768.x.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. https://doi.org/10.1093/biomet/28.3-4.321.

P. L. Hsu. On the distribution of roots of certain determinantal equations. *Ann. Eugenics*, 9(3): 250–258, 1939. https://doi.org/10.1111/j.1469-1809.1939.tb02212.x.

T. R. Knapp. Canonical correlation analysis: A general parametric significance-testing system. *Psychol. Bull.*, 85(2):410–416, 1978. https://doi.org/10.1037/0033-2909.85.2.410.

W. H. Kruskal. A nonparametric test for the several sample problem. *Ann. Math. Statistics*, 23(4): 525–540, 1952. https://doi.org/10.1214/aoms/1177729332.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.*, 47(260):583–621, 1952. https://doi.org/10.1080/01621459.1952.10483441.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18(1):50–60, 1947. https://doi.org/10.1214/aoms/1177730491.

K. V. Mardia and J. T. Kent. Rao score tests for goodness of fit and independence. *Biometrika*, 78 (2):355–363, 1991. https://doi.org/10.1093/biomet/78.2.355.

A. Martín Andrés, I. Herranz Tejedor, and A. Silva Mato. The Wilcoxon, Spearman, Fisher, χ^2-, Student and Pearson tests and 2×2 tables. *The Statistician*, 44(4):441–450, 1995. https://doi.org/10.2307/2348893.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. https://doi.org/10.1093/biomet/58.3.545.

K. Pearson. Mathematical contributions to the theory of evolution — III. Regression, heredity, and panmixia. *Philos. Trans. Roy. Soc. London Ser. A*, 187:253–318, 1896. https://doi.org/10.1098/rsta.1896.0007.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. (5)*, 50(302):157–175, 1900. https://doi.org/10.1080/14786440009463897.

K. C. S. Pillai. *On some distribution problems in multivariate analysis*, volume 88 of *Institute of Statistics mimeo series*. North Carolina State University, Dept. of Statistics, 1954.

C. R. Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philos. Soc.*, 44(1):50–57, 1948. https://doi.org/10.1017/S0305004100023987.

C. Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1): 72–101, 1904. https://doi.org/10.2307/1412159.

F. Wilcoxon. Individual comparisons by ranking methods. *Biom. Bull.*, 1(6):80–83, 1945. https://doi.org/10.2307/3001968.

S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24(3-4):471–494, 1932. https://doi.org/10.1093/biomet/24.3-4.471.