Biometrika (2022), **109**, 4, *pp.* 1067–1083
doi: 10.1093/biomet/asab059
Advance Access publication 18 November 2021

Significance testing for canonical correlation analysis in high dimensions

By IAN W. McKEAGUE

Department of Biostatistics, Columbia University,
Room R639, 722 West 168th Street, New York, New York 10032, U.S.A.
im2131@columbia.edu

and XIN ZHANG

Department of Statistics, Florida State University,
214 OSB, 117 N. Woodward Avenue, Tallahassee, Florida 32306, U.S.A.
xzhang8@fsu.edu

Summary

We consider the problem of testing for the presence of linear relationships between large sets of random variables based on a postselection inference approach to canonical correlation analysis. The challenge is to adjust for the selection of subsets of variables having linear combinations with maximal sample correlation. To this end, we construct a stabilized one-step estimator of the Euclidean norm of the canonical correlations maximized over subsets of variables of prespecified cardinality. This estimator is shown to be consistent for its target parameter and asymptotically normal, provided the dimensions of the variables do not grow too quickly with sample size. We also develop a greedy search algorithm to accurately compute the estimator, leading to a computationally tractable omnibus test for the global null hypothesis that there are no linear relationships between any subsets of variables having the prespecified cardinality. We further develop a confidence interval that takes the variable selection into account.

Some key words: Efficient one-step estimator; Greedy search algorithm; Large-scale testing; Postselection inference.

1. Introduction

When exploring the relationships between two sets of variables measured on the same set of objects, canonical correlation analysis, CCA (Hotelling, 1936), sequentially extracts linear combinations with maximal correlation. Specifically, with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ as two random vectors, the first step of CCA targets the parameter

$$\rho = \max_{\alpha \in \mathbb{R}^q,\, \beta \in \mathbb{R}^p} \operatorname{corr}(\alpha^{\mathrm{T}} Y, \beta^{\mathrm{T}} X) \quad \text{subject to} \quad \operatorname{var}(\alpha^{\mathrm{T}} Y) = 1 = \operatorname{var}(\beta^{\mathrm{T}} X). \tag{1}$$

Subsequent steps of CCA repeat this process subject to the constraint that the next linear combinations of X and of Y are uncorrelated with earlier ones, giving a decreasing sequence of correlation coefficients. We are interested in testing whether the maximal canonical correlation coefficient $\rho \neq 0$ versus the null hypothesis $\rho = 0$ in the high-dimensional setting in which p and q grow with sample size n. This is equivalent to testing whether all of the canonical correlation coefficients vanish, or whether their sum of squares τ^2, known as the Pillai (1955) trace, vanishes.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Biometrika Trust.
All rights reserved. For permissions, please email: journals.permissions@oup.com

Over the last dozen years, numerous sparse CCA methods (e.g., Witten et al., 2009; Hardoon & Shawe-Taylor, 2011; Gao et al., 2017; Mai & Zhang, 2019; Qadar & Seghouane, 2019; Shu et al., 2020) have been developed as extensions of classical CCA by adapting regularization approaches from regression, e.g., lasso (Tibshirani, 1996), elastic net (Zou & Hastie, 2005) and soft thresholding.

Sparse CCA methods have been widely applied to high-dimensional omics data to detect associations between gene expression and DNA copy number/polymorphisms/methylation, with the aim of revealing networks of coexpressed and coregulated genes (Waaijenborg & Zwinderman, 2007; Waaijenborg et al., 2008; Parkhomenko et al., 2009; Naylor et al., 2010; Wang et al., 2015). A problem with the indiscriminate use of such methods, however, is selection bias, arising when the effects of variable selection on subsequent statistical analyses are ignored, i.e., failure to take into account double dipping of the data when assessing evidence of association.

Devising valid tests for associations in high-dimensional sparse CCA, along with confidence interval estimation for the strength of the association, poses a challenging postselection inference problem. Nevertheless, some progress on this problem has been made. Yang & Pan (2015) proposed the sum of sample canonical correlation coefficients as a test statistic, and established a valid calibration under the sparsity assumption that the number of nonzero canonical correlations is finite and fixed, with the dimensions p and q proportional to sample size. Their approach comes at the cost of assuming that X and Y are jointly Gaussian, and thus fully independent under the null; similar results for the maximal sample canonical correlation coefficient are developed by Bao et al. (2019). Zheng et al. (2019) developed a test for the presence of correlations among arbitrary components of a given high-dimensional random vector, for both sparse and dense alternatives, but their approach also requires an independent components structure. Under the Gaussian assumption, the test of canonical correlation is equivalent to the test of independence, which is an extensively studied research topic in recent years (e.g., Zhu et al., 2017; Bodnar et al., 2019; Shi et al., 2021), but having a different goal from the present article.

In this paper, we provide valid postselection inference for a new version of sparse CCA in high-dimensional settings. We obtain a computationally tractable and asymptotically valid confidence interval for $\tau_{\max}$, where $\tau_{\max}^2$ is the maximum of the Pillai trace over all subvectors of X and Y having prespecified dimensions s_x and s_y, respectively. The method is fully nonparametric in the sense that no distributional assumptions or sparsity assumptions are required. Rather than adopting a penalization approach or making a sparsity assumption on the number of nonzero canonical correlations to regularize the problem, we use the sparsity levels $s_x \ll p$ and $s_y \ll q$ for regularization, and also for controlling the computational cost of searching over large collections of subvectors. We introduce a test statistic $\hat{\tau}_{\max}$ constructed as a stabilized and efficient one-step estimator of $\tau_{\max}$. Then, assuming that p and q do not grow too quickly with sample size, specifically that $n^{-1/2}\log(p+q) \to 0$, we show that a studentized version of $\hat{\tau}_{\max}$, after centring by $\tau_{\max}$, converges weakly to the standard normal. This leads to a practical way of calibrating a formal omnibus test for the global null hypothesis of $\tau_{\max} = 0$ that there are no linear relationships between any subsets of variables having the prespecified cardinality, along with an asymptotically valid Wald-type confidence interval for $\tau_{\max}$.

The proposed approach applies to any choice of prespecified sparsity levels s_x and s_y, which do not need to be the same as the true number of active variables in the population CCA, although they should be sufficiently large to capture the key associations. The test procedure and confidence interval for the target parameter $\tau_{\max}$ are asymptotically valid for any prespecified sparsity levels, and work well provided the sample cross-covariance matrices between subvectors of X and Y having dimensions s_x and s_y are sufficiently accurate.

Our approach is related to the type of postselection inference procedure for marginal screening developed by McKeague & Qian (2015), which applies to the one-dimensional response case, $q = 1$, and sparsity levels $s_x = s_y = 1$. To extend this approach to the sparse CCA setting, in which both p and q can be large, requires a trade-off between computational tractability and statistical power. The calibration used by McKeague & Qian (2015) is a double-bootstrap technique, which is computationally expensive. To obtain a fast calibration method for sparse CCA, we adapt the sample-splitting stabilization technique of Luedtke & van der Laan (2018) from univariate screening to the multivariate canonical correlation analysis setting. This provides calibration using a standard normal limit.

Adapting the approach of Luedtke & van der Laan (2018) to the present setting is challenging. New concentration results for the canonical gradient of the root-Pillai trace and its second-order remainder terms are needed. Also, nonidentifiable local parameters appear in the canonical gradient at the null hypothesis of interest, and standard calibration methods such as bootstrap for the sample version of the target parameter fail due to nonregularity. The stabilization approach controls the nonregularity at the null and furnishes the standard normal limit. Furthermore, to control the computational complexity of searching through large collections of subvectors of X and Y when computing $\hat{\tau}_{\max}$, we develop a greedy search algorithm that is more accurate and computationally more efficient than that of Wiesel et al. (2008). This is mainly because we are able to maximize and update exact increments in the Pillai trace, whereas in Wiesel et al. (2008) the maximization is carried out on lower bounds of the increments.

2. INFERENCE FOR THE MAXIMAL PILLAI TRACE

2.1. *Preliminaries*

Let $\Sigma_X > 0$ and $\Sigma_Y > 0$ denote the covariance matrices of X and Y, with cross-covariance matrix Σ_{XY} and standardized cross-covariance matrix $\Lambda_{XY} \equiv \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2} \in \mathbb{R}^{p\times q}$, also known as the coherence matrix. The sample counterparts are denoted by S_X, S_Y, S_{XY} and C_{XY}, respectively. The coherence matrix Λ_{XY} has $\min(p, q)$ singular values; when listed in decreasing order, they coincide with the canonical correlation coefficients, and ρ defined in (1) is the largest. A closely related parameter in multivariate analysis of variance is the Pillai trace τ^2 (Pillai, 1955), defined as the sum of squares of the canonical correlation coefficients, or, equivalently,

$$\tau^2 = \|\Lambda_{XY}\|_F^2 = \mathrm{tr}(\Lambda_{XY}\Lambda_{XY}^{\mathrm{T}}) = \mathrm{tr}\{H(H+E)^{-1}\},$$

where $\|\cdot\|_F$ is the Frobenius norm, and $H = \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ and $E = \Sigma_Y - H$ are population versions of covariance matrices in a linear model for predicting Y from X. Specifically, H is the covariance matrix of the least-squares-predicted outcome in the linear model $Y = A + BX + \varepsilon$, where $\mathrm{cov}(\varepsilon) = E$ and ε is uncorrelated with X.

Clearly, the null hypotheses $\rho = 0$ and $\tau = 0$ are equivalent, but the root-Pillai trace τ, the positive square root of τ^2, which contains information of all nonzero canonical correlations, is a more informative target parameter than the leading canonical correlation ρ, although the two would coincide if there is only a single nonzero canonical correlation coefficient. Moreover, because estimating the maximal values of τ or ρ, subject to sparsity constraints, needs repeated evaluation and updating of the estimates, the choice of τ provides considerable computational savings over ρ, as the latter would require updating the entire eigendecomposition at each step.

Our approach is to develop asymptotic distribution results for a regularized empirical version of this target parameter when the dimensions p and q grow with sample size n. Specifically, given sparsity levels s_x and s_y for X and Y, respectively, we are interested in selecting index sets $\mathcal{K} \subset \{1, \ldots, p\}$ and $\mathcal{J} \subset \{1, \ldots, q\}$ with cardinality $|\mathcal{K}| \leqslant s_x$ and $|\mathcal{J}| \leqslant s_y$ that maximize the Pillai trace, or, equivalently, the root-Pillai trace, of their corresponding subvectors $Y_{\mathcal{J}}$ and $X_{\mathcal{K}}$. The sparsity levels s_x and s_y are prespecified and fixed, e.g., $(s_x, s_y) = (1, 2)$.

More specifically, given independent observations $O_i = (X_i^{\mathrm{T}}, Y_i^{\mathrm{T}})^{\mathrm{T}}$ $(i = 1, \ldots, n)$, drawn from a distribution P on $\mathbb{R}^{p+q}$, we target the parameter

$$\tau_{\max} \equiv \max_{d \in \mathcal{D}_n} \Psi^d(P), \tag{2}$$

where

$$\mathcal{D}_n = \{(\mathcal{J}, \mathcal{K}) \mid |\mathcal{K}| = s_x \leqslant p,\ |\mathcal{J}| = s_y \leqslant q,\ \mathcal{K} \subseteq \{1, \ldots, p\},\ \mathcal{J} \subseteq \{1, \ldots, q\}\}$$

and $\Psi^d(P) = \|\Lambda_{X_{\mathcal{K}} Y_{\mathcal{J}}}\|_F$. There may be no unique maximizer $d \in \mathcal{D}_n$, and $\tau_{\max} \leqslant \tau$, with equality when $s_x = p$, $s_y = q$. The subscript n in $\mathcal{D}_n$ indicates that the dimensions $p = p_n$ and $q = q_n$ are allowed to increase with n. Although our focus is on $\tau_{\max}$, the estimation and inference procedures, as well as the theoretical results, can be extended straightforwardly to the maximal Pillai trace $\tau^2_{\max}$. See § 4.3 for further discussion and analysis.

The numbers of active variables $(s_x^\star, s_y^\star)$ are the smallest values of the sparsity levels (s_x, s_y) for which $\tau_{\max} = \tau$. Note that $s_x^\star$ and $s_y^\star$ can be as large as p and q, respectively, and as small as the number of nonzero canonical correlation coefficients for X and Y. The rank of Λ_{XY} is denoted by K in the sequel. The nonregularity arises for various reasons, including the fact that multiple elements of $\mathcal{D}_n$ may achieve the same maximum in (2), e.g., when $\tau_{\max} = 0$. This may occur, for example, if the prespecified sparsity levels are larger than the true sparsity levels $(s_x^\star, s_y^\star)$, but as we see later in this section, the sample root-Pillai trace is nonregular at $\tau_{\max} = 0$ even when d is fixed. Moreover, though the numbers of active variables $(s_x^\star, s_y^\star)$ are unique by definition, nonregularity still occurs because the sets of active variables may not be unique.

2.2. *Stabilized one-step estimator*

In this section we develop the stabilized one-step estimator for the target parameter $\tau_{\max}$ in terms of the canonical gradient $D^d(P)$ of the functional $\Psi^d(P)$ for a fixed $d \in \mathcal{D}_n$. The canonical gradient is derived later in § 3.2, and will be estimated by plugging-in empirical distributions P_j of the first j observations in place of P. We consider subsamples consisting of the first j observations for $j = \ell_n, \ldots, n-1$, where $\{\ell_n\}$ is some positive integer sequence such that both ℓ_n and $n - \ell_n$ tend to infinity. The following procedure is a version of the construction of the stabilized one-step estimator in Luedtke & van der Laan (2018).

For each $j = \ell_n, \ldots, n-1$, compute the following quantities.

Step 1. The selected subsets of variables $d_{nj} = (\hat{\mathcal{K}}, \hat{\mathcal{J}})$, given by

$$d_{nj} \equiv \operatorname*{arg\,max}_{d \in \mathcal{D}_n} \Psi^d(P_j) = \operatorname*{arg\,max}_{|\mathcal{K}| = s_x,\, |\mathcal{J}| = s_y} \|C_{X_{\mathcal{K}} Y_{\mathcal{J}}}(P_j)\|_F. \tag{3}$$

Step 2. The corresponding maximum $\Psi^{d_{nj}}(P_j) = \|C_{X_{\hat{\mathcal{K}}} Y_{\hat{\mathcal{J}}}}(P_j)\|_F$ and $\hat{D}_j(O_{j+1}) \equiv D^{d_{nj}}(P_j)(O_{j+1})$ using the canonical gradient given later by (5) and (6) with $P = P_j$.

Step 3. An estimate of the variance of $\hat{D}_j(O_{j+1})$:

$$\hat{\sigma}_j^2 = \frac{1}{j}\sum_{i=1}^{j}\left\{\hat{D}_j(O_i) - \frac{1}{j}\sum_{m=1}^{j}\hat{D}_j(O_m)\right\}^2.$$

Step 4. Weights $w_j = \overline{\sigma}_n/\hat{\sigma}_j$, where

$$\overline{\sigma}_n = \left(\frac{1}{n-\ell_n}\sum_{j=\ell_n}^{n-1}\hat{\sigma}_j^{-1}\right)^{-1}$$

is the harmonic mean.

The stabilized one-step estimator for the target parameter $\tau_{\max}$ is then given by

$$\hat{\tau}_{\max} = \frac{1}{n-\ell_n}\sum_{j=\ell_n}^{n-1} w_j\{\Psi^{d_{nj}}(P_j) + \hat{D}_j(O_{j+1})\},$$

and an asymptotic $100(1-\alpha)\%$ Wald-type confidence interval for $\tau_{\max}$ is

$$[\mathrm{LB}_n, \mathrm{UB}_n] = \left[\hat{\tau}_{\max} - z_{\alpha/2}\frac{\overline{\sigma}_n}{(n-\ell_n)^{1/2}}, \hat{\tau}_{\max} + z_{\alpha/2}\frac{\overline{\sigma}_n}{(n-\ell_n)^{1/2}}\right],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal. For an α-level test, we reject the null hypothesis $\tau_{\max} = 0$ if the lower bound of the $100(1-2\alpha)\%$ confidence interval exceeds 0.

The estimator $\hat{\tau}_{\max}$ is a weighted version of the online one-step estimator in van der Laan & Lendle (2014), where in our case $\Psi^{d_{nj}}(P_j)$ is improved using its estimated canonical gradient evaluated at a new observation O_{j+1}. The updates for a sample stream $j = \ell_n, \ldots, n-1$ take advantage of the recursive properties of the algorithm, which are partly due to the choice of harmonic mean $\overline{\sigma}_n$ and allow considerable speed-up in the computation, see the Supplementary Material. When the sample size n is large, we follow the suggestion of Luedtke & van der Laan (2018) to speed up the $(n-\ell_n)$ updates by restricting the sample stream over $j = \ell_n, \ldots, n-1$ to only involve increments in j of size $C \geqslant 2$. The asymptotic properties of the stabilized one-step estimator are not affected by C. In our experience, the results are insensitive to the choice of C, provided n is sufficiently large relative to C. We fixed $C = 20$ and $\ell_n = \lceil n/2 \rceil$ in our numerical studies. While in simulations, the data can be treated as a sample stream in any order, in real datasets the ordering in the samples may not be arbitrary. In that case, we recommend randomly ordering the data 10 times and then combining the proposed confidence intervals by averaging; see § 5.

2.3. *Greedy search for the maximal Pillai trace*

When computing the stabilized one-step estimator, the computationally costly part is the optimization in (3). To obtain d_{nj}, we need to search over subsets $\mathcal{K}$ of size s_x and, similarly, over subsets $\mathcal{J}$ of size s_y. This means a search over $\binom{p}{s_x}\binom{q}{s_y}$ possible combinations, which is too expensive to compute when p and q are large. In some applications, there may be a neighbourhood structure that can be exploited to reduce computational expense. For example, restricting to neighbourhoods of the form $\mathcal{K} = \{1, \ldots, s_x\}, \{2, \ldots, s_x+1\}, \ldots$ gives $p - s_x + 1$ possible subsets in total. Nevertheless, in general, there is a need to speed up the first step of the computation of the stabilized one-step estimator. To that end, we introduce the scalable greedy search in Algorithm 1 to approximately maximize the Pillai trace $\|C_{X_{\mathcal{K}}Y_{\mathcal{J}}}\|_F^2$ over $\mathcal{K}$ and $\mathcal{J}$.

Algorithm 1. Greedy search.

1. Initialize $\mathcal{J}=\{j\}$ and $\mathcal{K}=\{k\}$, where (j,k) maximizes $\|C_{Y_jX_k}\|_F^2=\widehat{\text{corr}}^2(Y_j,X_k)$.
2. Select over $j\notin\mathcal{J}$ and $k\notin\mathcal{K}$.
    - a. If $|\mathcal{J}|<s_y$ and $|\mathcal{K}|<s_x$, find $j\notin\mathcal{J}$ and $k\notin\mathcal{K}$ that maximizes $\delta_j^{\mathcal{J},\mathcal{K}}\equiv\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2$ and $\gamma_k^{\mathcal{J},\mathcal{K}}\equiv\|C_{Y_{\mathcal{J}}R_{k|\mathcal{K}}}\|_F^2$, respectively. Then update $\mathcal{J}\to\mathcal{J}\cup j$ if $\max_{j\notin\mathcal{J}}\delta_j^{\mathcal{J},\mathcal{K}}>\max_{k\notin\mathcal{K}}\gamma_k^{\mathcal{J},\mathcal{K}}$; otherwise, update $\mathcal{K}\to\mathcal{K}\cup k$.
    - b. If $|\mathcal{J}|<s_y$ and $|\mathcal{K}|=s_x$, update $\mathcal{J}\to\mathcal{J}\cup j$, where $j\notin\mathcal{J}$ maximizes $\delta_j^{\mathcal{J},\mathcal{K}}$.
    - c. If $|\mathcal{J}|=s_y$ and $|\mathcal{K}|<s_x$, update $\mathcal{K}\to\mathcal{K}\cup k$, where $k\notin\mathcal{K}$ maximizes $\gamma_k^{\mathcal{J},\mathcal{K}}$.
3. Update the Pillai trace based on the increment given in Lemma 1 below.
4. Repeat steps 2 and 3 until $|\mathcal{J}|=s_y$ and $|\mathcal{K}|=s_x$.
5. Output: $\hat{\mathcal{J}}$, $\hat{\mathcal{K}}$ and $\|C_{Y_{\hat{\mathcal{J}}}X_{\hat{\mathcal{K}}}}\|_F^2$ or $\|C_{Y_{\hat{\mathcal{J}}}X_{\hat{\mathcal{K}}}}\|_F$.

This algorithm is much more efficient than a full combinatorial search. For all $j\notin\mathcal{J}$ and $k\notin\mathcal{K}$, we consider the increments in the Pillai trace $\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2$ by replacing $\mathcal{J}$ with $\mathcal{J}\cup\{j\}$ and replacing $\mathcal{K}$ with $\mathcal{K}\cup\{k\}$. Let $E_{j|\mathcal{J}}=Y_j-\mathrm{E}(Y_j)-\Sigma_{Y_jY_{\mathcal{J}}}\Sigma_{Y_{\mathcal{J}}}^{-1}\{Y_{\mathcal{J}}-\mathrm{E}(Y_{\mathcal{J}})\}$ be the residual of Y_j regressed on $Y_{\mathcal{J}}$, and, similarly, let $R_{k|\mathcal{K}}$ be the residual of X_k regressed on $X_{\mathcal{K}}$. The sample versions of $E_{j|\mathcal{J}}$ and $R_{k|\mathcal{K}}$ are obtained using ordinary least squares, and then substituted into the calculations of $C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}$ and $C_{R_{k|\mathcal{K}}Y_{\mathcal{J}}}$.

This relies on the following result, which gives the increment in the Pillai trace when including an additional variable in either X or Y, or both, and allows us to implement the greedy search via forward stepwise selection.

LEMMA 1. *Assume that* $S_{Y_{\mathcal{J}}}>0$, $S_{X_{\mathcal{K}}}>0$ *and* $n>\max(s_x,s_y)+1$. *Then*

$$\begin{aligned}\|C_{Y_{\mathcal{J}\cup\{j\}}X_{\mathcal{K}}}\|_F^2&=\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2+\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2,\\ \|C_{Y_{\mathcal{J}}X_{\mathcal{K}\cup\{k\}}}\|_F^2&=\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2+\|C_{Y_{\mathcal{J}}R_{k|\mathcal{K}}}\|_F^2,\\ \|C_{Y_{\mathcal{J}\cup\{j\}}X_{\mathcal{K}\cup\{k\}}}\|_F^2&=\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2+\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2+\|C_{Y_{\mathcal{J}}R_{k|\mathcal{K}}}\|_F^2+\|C_{E_{j|\mathcal{J}}R_{k|\mathcal{K}}}\|_F^2.\end{aligned}\tag{4}$$

An alternative version of Algorithm 1 involves maximizing over increments of the root-Pillai trace, which by (4) can be expressed in terms of increments of the Pillai trace as

$$\begin{aligned}&\|C_{Y_{\mathcal{J}\cup\{j\}}X_{\mathcal{K}}}\|_F-\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F\\ &\quad=\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2/\{(\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2+\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2)^{1/2}+\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F\}.\end{aligned}$$

The above expression is an increasing function of $\|C_{E_{j|\mathcal{J}}X_{\mathcal{K}}}\|_F^2$, so the same index j must maximize both of these increments and the alternative version of Algorithm 1 is therefore equivalent.

This greedy search algorithm is related to that proposed by Wiesel et al. (2008), which was designed for sparse maximization of the sample version of the leading canonical correlation coefficient ρ. However, we have two important advantages. First, due to Lemma 1, we are able to

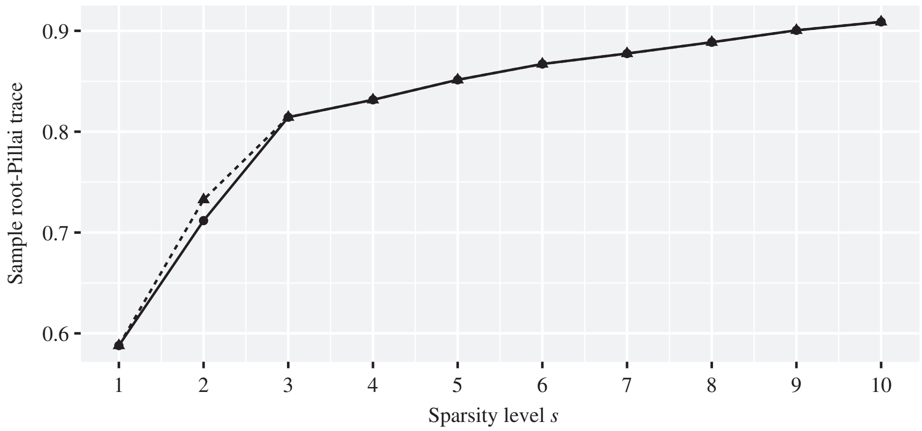

Fig. 1. Results based on single samples generated under Model A1 with $n = 500$, $s_x^\star = s_y^\star = 3$, $\tau_{\max} = 0.8$. Values of $\hat{\tau}_{\text{samp}}$ from a full search (dotted line) and a greedy search (solid line) as the sparsity level $s = s_x = s_y$ varies from 1–10 for $p = q = 10$.

maximize and update exact increments in the Pillai trace, namely $\|C_{E_{j|\mathcal{J}} X_{\mathcal{K}}}\|_F^2$ and $\|C_{Y_{\mathcal{J}} R_{k|\mathcal{K}}}\|_F^2$, whereas in Wiesel et al. (2008) the exact increments in ρ are not available and the maximization is carried out on lower bounds of the increments. Second, to obtain the maximal canonical correlation, the CCA directions α and β also need to be updated at each step of including an additional variable, while the update for the Pillai trace is automatically obtained by the equations in Lemma 1 using linear regression residuals. Therefore, our approach is both more accurate and computationally more efficient than the greedy search algorithm of Wiesel et al. (2008).

Figure 1 gives the results from a toy example showing that the proposed Algorithm 1 for finding the maximal sample root-Pillai trace under varying sparsity constraints provides almost perfect agreement with the full search. The data were generated from Model A1 used in the simulation study, in § 4, with $p = q = 10$, the true numbers of active variables $s_x^\star = s_y^\star = 3$, $\tau_{\max} = 0.8$, the true number of nonzero canonical correlations $K = 1$, and $n = 500$. The result of the greedy search agrees with the full search at all sparsity levels except at $s = s_x = s_y = 2$.

Algorithm 1 can naturally be modified so as not to require prespecified sparsity levels. In step 2, either j or k is added, whichever gives the larger increment in the Pillai trace. The algorithm can then be terminated in step 4 when the increment is smaller than some given tolerance, say 0.05 or 0.01. In the Supplementary Material we propose a graphical tool that is analogous to the scree plot used in principal component analysis and factor analysis, providing intuition and graphical diagnostics for how sparse the true model might be; to gain some further insight into the performance of the proposed greedy search algorithm, we also show that the root-Pillai trace comes close to satisfying the submodular property (Nemhauser & Wolsey, 1978; Krause & Golovin, 2014; Khim et al., 2016). Methods for estimating the number of nonzero canonical correlation coefficients K have been extensively studied in the signal processing literature (e.g., Song et al., 2016; Seghouane & Shokouhi, 2019), and these can be used to provide a lower bound on the choice of s_x and s_y.

3. THEORETICAL RESULTS

3.1. *Basic definitions*

We need some general concepts from semiparametric efficiency theory. Suppose that we observe a general random vector $O \sim P$. Let $L_0^2(P)$ denote the Hilbert space of P-square integrable functions with mean zero. Consider a smooth one-dimensional family of probability measures

$\{P_t, t \in [0,1]\}$ with $P_0 = P$ and having score function $k \in L_0^2(P)$ at $t = 0$. The tangent space $T(P)$ is the $L_0^2(P)$ closure of the linear span of all such score functions k. For example, if nothing is known about P then $P_t(do) = \{1 + tk(o)\}P(do)$ is such a submodel for any bounded function k with mean zero, provided t is sufficiently small, so $T(P)$ is seen to be the whole of $L_0^2(P)$ in this case. Let $\psi(P)$ be a real parameter that is pathwise differentiable at P: there exists $g \in L_0^2(P)$ such that $\lim_{t\to 0}\{\psi(P_t) - \psi(P)\}/t = \langle g, k\rangle$ for any smooth submodel $\{P_t\}$ with score function k, where $\langle \cdot, \cdot \rangle$ is the inner product in $L_0^2(P)$. The function g is called a gradient, or influence function, for ψ; the projection IF_ψ of any gradient into the tangent space $T(P)$ is unique and is known as the canonical gradient, or efficient influence function. The supremum of the Cramér–Rao bounds for all submodels is given by the second moment of $\text{IF}_\psi(O)$. Furthermore, the influence function as derived using von Mises calculus (van der Vaart, 2000, Ch. 20) of any regular and asymptotically linear estimator must be a gradient (Pfanzagl, 1990, Proposition 2.3).

A one-step estimator is an empirical bias correction of a naïve plug-in estimator in the direction of a gradient of the parameter of interest (Pfanzagl, 1982); when this gradient is the canonical gradient, then this results in an efficient estimator under some regularity conditions. Given an initial estimator $\hat{P}$ of P and any gradient $D(\hat{P})$ of the parameter ψ evaluated at $\hat{P}$, we have $\psi(\hat{P}) - \psi(P) = -PD(\hat{P}) + \mathrm{Rem}_\psi(\hat{P}, P)$, where $\mathrm{Rem}_\psi(\hat{P}, P)$ is negligible if $\hat{P}$ is close to P in an appropriate sense. Here Pf denotes the expectation under P of a random real-valued function f, ignoring the randomness in f. As $D(P)$ has mean zero under P, we expect that $PD(\hat{P})$ is close to zero if D is continuous in its argument and $\hat{P}$ is close to P. However, the rate of convergence of $PD(\hat{P})$ to zero as the sample size grows may be slower than $n^{-1/2}$. The one-step estimator aims to improve $\psi(\hat{P})$, and achieve $n^{1/2}$ consistency and asymptotic normality by adding an empirical mean $\mathbb{P}_n D(\hat{P})$ of its deviation from $\psi(P)$. The one-step estimator $\hat{\psi} \equiv \psi(\hat{P}) + \mathbb{P}_n D(\hat{P})$ then satisfies the expansion $\hat{\psi} - \psi(P) = (\mathbb{P}_n - P)D(\hat{P}) + \mathrm{Rem}_\psi(\hat{P}, P)$. Under an empirical process and the $L^2(P)$-consistency condition on $D(\hat{P})$, the leading term on the right is asymptotically equivalent to $(\mathbb{P}_n - P)D(P)$, which converges in distribution to a mean-zero Gaussian limit with consistently estimable covariance. To minimize the variance of the Gaussian limit, $D(\hat{P})$ can be taken as the canonical gradient of ψ at $\hat{P}$.

3.2. *Canonical gradient*

We now use von Mises calculus to derive the canonical gradient $D^d(P)(o)$ of the functional $\Psi^d(P)$ for a fixed $d \in \mathcal{D}_n$. This canonical gradient can be found in terms of the influence function of its square $\Phi^d(P) = \{\Psi^d(P)\}^2$, and using the fact that the tangent space is the whole of $L_0^2(P)$ in this nonparametric setting. Let $P_\epsilon = (1-\epsilon)P + \epsilon\delta_o$, where $\epsilon \in [0,1]$ and δ_o is the Dirac measure at the point $o = (x^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$.

THEOREM 1 (Canonical gradient). *When* $\Psi^d(P) > 0$, *we have*

$$D^d(P)(o) = \left.\frac{d\Psi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \frac{1}{2\Psi^d(P)} \left.\frac{d\Phi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0}, \tag{5}$$

where

$$\begin{aligned} \left.\frac{d\Phi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = &-\{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\}^{\mathrm{T}} \Sigma_{X_{\mathcal{K}}}^{-1} \Sigma_{X_{\mathcal{K}} Y_{\mathcal{J}}} \Sigma_{Y_{\mathcal{J}}}^{-1} \Sigma_{Y_{\mathcal{J}} X_{\mathcal{K}}} \Sigma_{X_{\mathcal{K}}}^{-1} \{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\} \\ &- \{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\}^{\mathrm{T}} \Sigma_{Y_{\mathcal{J}}}^{-1} \Sigma_{Y_{\mathcal{J}} X_{\mathcal{K}}} \Sigma_{X_{\mathcal{K}}}^{-1} \Sigma_{X_{\mathcal{K}} Y_{\mathcal{J}}} \Sigma_{Y_{\mathcal{J}}}^{-1} \{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\} \\ &+ 2\{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\}^{\mathrm{T}} \Sigma_{Y_{\mathcal{J}}}^{-1} \Sigma_{Y_{\mathcal{J}} X_{\mathcal{K}}} \Sigma_{X_{\mathcal{K}}}^{-1} \{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\}. \end{aligned} \tag{6}$$

Moreover, $\mathrm{E}_P\{D^d(P)(O)\} = 0$, *so the influence function belongs to the tangent space* $L_0^2(P)$ *and is thus efficient.*

A continuous extension of $D^d(P)(o)$ to the case $\Psi^d(P) = 0$ is obtained as follows. The matrix-valued parameter $\psi(P) = \Lambda \equiv \Lambda_{Y_{\mathcal{J}} X_{\mathcal{K}}}$ is pathwise differentiable, so when $\psi(P) = 0$, there exists a matrix G, which we can take as the efficient influence function, of the same dimensions as Λ and having entries in $L_0^2(P)$ such that $\psi(P_t)/t \to \langle G, k\rangle$ as $t \to 0$ for any smooth one-dimensional parametric submodel $\{P_t, t \in [0,1]\}$ with score function $k \in L_0^2(P)$ at $t = 0$. Here the inner product notation in $\langle G, k\rangle$ is understood to be applied entrywise to G.

Corollary 1. *When* $\Psi^d(P) = 0$, *the canonical gradient is given by*

$$D^d(P)(o) \equiv \lim_{t\to 0} D^d(P_t)(o) = \{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\}^{\mathrm{T}} \Sigma_{Y_{\mathcal{J}}}^{-1/2} L \Sigma_{X_{\mathcal{K}}}^{-1/2} \{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\}, \tag{7}$$

where $\Lambda_t \equiv \psi(P_t)$ *is arranged so that it does not vanish at any* t *apart from* $t = 0$, *and* $\Lambda_t/\|\Lambda_t\|_F \to \langle G, k\rangle/\|\langle G, k\rangle\|_F \equiv L$ *in the Frobenius norm as* $t \to 0$.

For univariate X and Y, the functional $P \mapsto \mathrm{corr}_P(X, Y)$ has canonical gradient

$$\frac{\{x - \mathrm{E}_P(X)\}\{y - \mathrm{E}_P(Y)\}}{\{\mathrm{var}(X)\mathrm{var}(Y)\}^{1/2}} - \frac{\mathrm{corr}(X,Y)}{2\mathrm{var}(X)}\{x - \mathrm{E}_P(X)\}^2 - \frac{\mathrm{corr}(X,Y)}{2\mathrm{var}(Y)}\{y - \mathrm{E}_P(Y)\}^2,$$

a result due to Colin Mallows (Devlin et al., 1975). When $\mathrm{corr}(X, Y) = 0$, the last two terms above drop out, and the expression agrees with the canonical gradient of $\Psi^d(P)$ in (7), since $L = 1$ in this case. In the multivariate case, the entries of the matrix L are nuisance parameters that are absent in the univariate case.

The nuisance parameters in L vary with d and the score function k, indicating the presence of nonregularity in the root-Pillai trace at zero, as the underlying k is not identifiable and plays the role of a local parameter. When target parameters take values on the boundary of their parameter space, zero is on the boundary in our case, nonregularity is known to cause unstable asymptotics, such as inconsistency of the bootstrap, even in the simple example of a population mean restricted to be nonnegative (Andrews, 2000). That is, dependence of a canonical gradient or efficient influence function on an arbitrary score function implies unstable behaviour of the estimator, especially in small samples. This form of nonregularity is present in dimensions $p \geqslant 2$ and $q \geqslant 2$, even without selection of $d \in \mathcal{D}_n$, but not in the case of univariate X and Y since the parameter space for the correlation coefficient is $(-1, 1)$, which has no boundary.

This boundary type of nonregularity is distinct from the postselection type of nonregularity noted by McKeague & Qian (2015, § 2) in the case $p \geqslant 2$ and $q = 1$, in which the asymptotic distribution of the maximal absolute sample correlation is discontinuous at $\tau_{\max} = 0$. This type of nonregularity occurs in the present setting with the sample estimator of $\tau_{\max}$, given by

$$\hat{\tau}_{\mathrm{samp}} = \max_{|\mathcal{K}|\leqslant s_x,\, |\mathcal{J}|\leqslant s_y} \|C_{Y_{\mathcal{J}} X_{\mathcal{K}}}\|_F = \max_{|\mathcal{K}|= s_x,\, |\mathcal{J}|= s_y} \|C_{Y_{\mathcal{J}} X_{\mathcal{K}}}\|_F,$$

where the second equality is a direct consequence of Lemma 1. It is challenging to use the estimator $\hat{\tau}_{\mathrm{samp}}$ as a test statistic for the global null hypothesis that $\tau_{\max} = 0$ because of the discontinuity in its asymptotic distribution at the null, but the estimator $\hat{\tau}_{\max}$ avoids this difficulty.

Curiously, the boundary type of nonregularity does not arise with the Pillai trace itself, since its canonical gradient (6) does not depend on any score function k; an intuitive explanation is that, by squaring the root-Pillai trace, the nonregularity is smoothed out at zero. However, this squaring has the effect of causing severe bias in the sampling distribution of the stabilized one-step estimator of $\tau_{\max}^2$, especially when $\tau_{\max}$ is small and in small samples; see Figs. 2 and 3 in § 4.3. This problem does not arise with $\hat{\tau}_{\max}$; hence, we focus on the root-Pillai trace.

Many authors have studied hypothesis testing problems in which a nuisance parameter is only identifiable under the alternative (e.g., Davies, 1977, 1987, 2002; Hansen, 1996). Here we encounter the situation where nuisance parameters appear only in the null, so calibration of the test statistic may potentially depend on L. Leeb & Pötscher (2017) have studied a postselection calibration method that uses estimates of such nuisance parameters, but, as we will see, our approach leads to an asymptotically pivotal estimator of $\tau_{\max}$ without the need to estimate L.

3.3. *Asymptotic properties of the stabilized one-step estimator*

We first lay out some technical assumptions similar to Luedtke & van der Laan (2018). For convenience, each component of X and Y is assumed to take values in $[-1, 1]$. We also assume that the canonical gradient of $\Psi^d(P)$ satisfies

$$\inf_{n \geqslant 2} \min_{d \in \mathcal{D}_n} \operatorname{var}_P\{D^d(P)(O)\} \geqslant \gamma \tag{8}$$

for some constant $\gamma > 0$. This moment condition requires that the quadratic forms in (5) and (7) have bounded variances, and is imposed to ensure a nondegenerate asymptotic distribution for the one-step estimator, as needed to form nontrivial confidence intervals for $\tau_{\max}$. To ensure that the canonical gradient is uniformly bounded for all d, we also assume that, for some $\delta > 0$,

$$\sup_{d \in \mathcal{D}_n} \max\{\|\Sigma_{X_{\mathcal{K}}}^{-1}\|, \|\Sigma_{Y_{\mathcal{J}}}^{-1}\|\} < \delta^{-1}, \tag{9}$$

where $\|M\|$ denotes the operator norm, or largest singular value, of a matrix M. This condition does not require the full invertibility of Σ_X and Σ_Y, and only means that the smallest eigenvalue of any $\Sigma_{X_{\mathcal{K}}}$ or $\Sigma_{Y_{\mathcal{J}}}$ considered by our procedure is bounded away from zero. We treat δ, γ and $s = s_x = s_y$ as fixed, and thus omit the dependence on δ, γ and s in the asymptotic statements. On the other hand, we allow both dimensions p and q to grow with the sample size n. When $p = p_n \to \infty$ and $q = q_n \to \infty$, it suffices to assume that $n^{-1/2} \log(p + q) \to 0$. More generally, define $\beta_n^2 = n^{-1/2} \log \max\{n, p, q\}$, and let, for some $\epsilon \in (0, 2)$,

$$\ell_n = \max[\{\log \max(n, p, q)\}^{1+\epsilon}, n \exp(-\beta_n^{-2+\epsilon})].$$

In particular, the above choice of ℓ_n satisfies

$$\frac{\log \max(n, p, q)}{\ell_n} \to 0, \qquad \beta_n^2 \log \frac{n}{\ell_n} \to 0, \qquad \limsup_{n \to \infty} \frac{\ell_n}{n} < 1. \tag{10}$$

For the estimation procedure described in § 2.2, we then have the following result on the lower bound of the confidence interval.

Theorem 2 (Tightness of the lower bound). *Under conditions* (8), (9) *and* (10), *for any sequence* $t_n \to \infty$, $\Psi_n(P) < \mathrm{LB}_n + t_n n^{-1/4} \beta_n$ *with probability approaching* 1.

Theorem 2 establishes the validity and tightness of the lower bound of the confidence interval for $\tau_{\max}$. This result immediately implies the asymptotic validity of our testing procedure for $H_0 : \tau_{\max} = 0$ versus $H_a : \tau_{\max} > 0$. To establish the upper bound, we further assume the following margin condition: for some sequence $t_n \to \infty$, there exists a sequence of nonempty subsets $\mathcal{D}_n^\star \subseteq \mathcal{D}_n$ such that, for all n,

$$\begin{gathered}\sup_{d_1,d_2\in\mathcal{D}_n^\star} \{\Psi^{d_1}(P) - \Psi^{d_2}(P)\} = o(n^{-1/2}), \\ \inf_{d\in\mathcal{D}_n^\star} \Psi^d(P) - \sup_{d\in\mathcal{D}_n\setminus\mathcal{D}_n^\star} \Psi^d(P) \geqslant t_n n^{-1/2}\beta_n.\end{gathered} \tag{11}$$

THEOREM 3 (Validity of the upper bound). *Under the same conditions as in Theorem 2, if we further assume* (11) *or* $\Psi_n(P) = 0$ *for all* n, *then* $\text{LB}_n \leqslant \Psi_n(P) \leqslant \text{UB}_n$ *with probability approaching* $1 - \alpha$.

These theorems, as well as their technical assumptions, are generalizations of Theorems 2 and 3 of Luedtke & van der Laan (2018), and specialize to their results when $s_x = 1$ and $q = s_y = 1$ in connection with the univariate maximal correlation setting of McKeague & Qian (2015). The extension to the general multivariate analysis of variance setting, such as the maximal Pillai trace, is highly nontrivial because of extra challenges that arise when analysing the canonical gradient, given by (5) and (6), and specifically in bounding its second-order remainder term, given in the Supplementary Material. Intuitively, the margin condition (11) allows the nonuniqueness of approximate maximizers of the root-Pillai trace, provided they are well separated from the root-Pillai trace of any other combination of variables.

4. SIMULATION STUDY

4.1. *Simulation set-up*

The sample size is fixed at $n = 500$, while we vary the dimensions of X and Y from $p = q = 30$ to $p = q = 5000$. We generated independent and identically distributed samples $(X_i^{\mathrm{T}}, Y_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+q}$ $(i = 1, \ldots, n)$, from a joint normal distribution with mean zero and covariance specified by

$$(\Sigma_X)_{jl} = (\Sigma_Y)_{jl} = \begin{cases} 0.5^{|j-l|}, & j, l \leqslant 100, \\ I(j = l), & \text{otherwise}, \end{cases} \qquad \Sigma_{XY} = \Sigma_X \left(\sum_{k=1}^{K} \rho_k \beta_k \alpha_k^{\mathrm{T}} \right) \Sigma_Y. \tag{12}$$

The above structured Σ_{XY} is commonly used in the sparse CCA literature (e.g., Mai & Zhang, 2019), where K is the number of nonzero CCA coefficients, $\rho_k > 0$ is the kth canonical correlation, and α_k and β_k are the corresponding sparse CCA directions that satisfy all the length, orthogonality and sparsity constraints. Also, the maximal canonical correlation coefficient $\rho = \rho_1$. The number K is irrelevant in our estimation as we did not use that information. Under this simulation setting, the covariance matrices Σ_X, Σ_Y and Σ_{XY} are not sparse, while the sparsity is imposed directly on each α_k and β_k. The nonzero elements in α and β correspond to the active variables in X and Y, respectively. In our simulations, we have the symmetry in X and Y, and thereby set $\alpha_k = \beta_k$, which implies that $s_x^\star = s_y^\star$.

We consider three scenarios of the form (12). The first scenario is a model satisfying the null hypothesis (Model N), where $\Sigma_{XY} = 0$, $K = 0$, $s_x^\star = s_y^\star = 0$. The next two scenarios are alternative hypothesis models (Models A1 and A2), with the true numbers of active variables $s_x^\star = s_y^\star = 3$. For methods that require prescribed sparsity levels, we set $s_x = s_y = s \in \{1, 2, 3, 4\}$ under Model N, and $s_x = s_y = 3$ under Models A1 and A2. Without loss of generality, the active variables are taken as the first three components of X and Y. Model A1 is the single pair CCA model with $K = 1$, so $\tau = \rho$. The sparse CCA direction $\alpha_1 = \beta_1$ is set as $v_1/\surd v_1^{\mathrm{T}}\Sigma_X v_1$ to satisfy the length constraint, where $v_1 = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$. For Model A2, a general sparse CCA model, we take the number of components $K = 3$ and set $(\rho_1, \rho_2, \rho_3) = (\tau, 2\tau, 3\tau)/\surd 14$. The sparse CCA directions $\alpha_k = \beta_k$, $k = 1, 2, 3$, are set to have 1 in the kth component and 0s elsewhere. Under Models A1 and A2, we vary $\tau \in \{0.1, 0.2, 0.3, 0.4\}$ to study the effect of changes in the strength of the correlation.

4.2. *Simulation results for hypothesis testing*

We compared various methods, whenever they are applicable, for the 5%-level test of $\tau_{\max} = 0$ versus $\tau_{\max} > 0$: the proposed testing procedure using the stabilized one-step estimator; the classical F-test for the Pillai trace without variable selection, as implemented in the `manova` R package (R Development Core Team, 2022); the F-test on selected variables with Bonferroni correction; and the higher criticism method (Donoho & Jin, 2004, 2015) based on p-values computed from the F-test for all $\binom{p}{s_x}\binom{q}{s_y}$ combinations of variables. The higher criticism statistic was calculated following the procedure described in Donoho & Jin (2015, § 1.1 and § 2.1) with the critical value calculated using the Gumbel distribution. For methods that require variable selection, the Bonferroni corrected F-test and the stabilized one-step estimator, the variables were selected using Algorithm 1. All of the F-tests considered, as well as the higher criticism procedure, are based on p-values for the multivariate analysis-of-variance F-test that targets the Pillai trace, whereas our approach targets the root-Pillai trace. For the Bonferroni corrected F-test, although we only used the F-statistic based on the $s_x + s_y$ variables selected from Algorithm 1, the Bonferroni correction covers all $\binom{p}{s_x}\binom{q}{s_y}$ combinations of variables potentially involved in the F-test.

We also considered the naive application of the F-test on selected variables, which comes without any adjustment for variable selection. When $p \geqslant 30$, this approach always rejected the null in our simulations even under the null hypothesis. This is not surprising, as such an F-test fails to adjust for spurious correlations. Another simple approach is to use multiple testing of $\mathrm{corr}(X_k, Y_j) = 0$ for all pq pairs of variables, in conjunction with the normal approximation test described by DiCiccio & Romano (2017). Rejection of one or more of these pq null hypotheses gives a rejection of $\tau_{\max} = 0$. We implemented this test by controlling the false discovery rate at the 5% level using the Benjamini & Hochberg (1995) procedure, as well as controlling the Type-I error at the 5% level using a Bonferroni correction. Either correction gives similar results to our proposed method for moderate p and q, but for large p or q, we find that the Benjamini–Hochberg-adjusted normal approximation test is computationally intractable. See the Supplementary Material for computation time comparisons. Both these methods are expected to have much less power than our method, as they can only detect correlations between pairs of variables, i.e., only examine $s_x = s_y = 1$.

In Table 1, we report the proportion of rejections under each simulation setting, based on 500 simulation replications for each case. For the higher criticism procedure, the total number of test statistics in one replication is $\binom{p}{s_x}\binom{q}{s_y}$. Therefore, it was only included for $p = 30, s < 3$ scenarios, and was shown to have unsatisfactory Type-I error control. The multivariate analysis-of-variance F-test worked adequately for $p = 30$, but is not applicable for large p. The feasible methods for high-dimensional settings are seen to be the proposed test based on the stabilized one-step

Table 1. *Simulation under the null Model N and the two alternative Models A1 and A2. The reported numbers are the rejected proportion based on* 500 *replicated datasets for each of the simulation settings. A dash indicates that the test is computationally intractable.*

		Model N				Model A1				Model A2			
p		$s = 1$	2	3	4	$\tau = 0.1$	0.2	0.3	0.4	$\tau = 0.1$	0.2	0.3	0.4
30	OS	0.058	0.054	0.074	0.056	0.068	0.312	0.830	0.996	0.064	0.234	0.720	0.980
	HC	0.134	0.628	–	–	–	–	–	–	–	–	–	–
	MF	0.034	0.034	0.034	0.034	0.040	0.072	0.216	0.470	0.042	0.074	0.216	0.496
	BF	0.048	0.008	0.002	0	0.002	0.074	0.662	0.996	0.004	0.062	0.530	0.966
	BH-DR	0.038	0.038	0.038	0.038	0.070	0.424	0.950	1	0.048	0.282	0.872	0.998
100	OS	0.054	0.072	0.070	0.080	0.074	0.190	0.660	0.982	0.058	0.136	0.588	0.946
	BF	0.046	0.010	0.002	0	0.002	0.018	0.366	0.962	0	0.024	0.304	0.906
	BH-DR	0.040	0.040	0.040	0.040	0.040	0.192	0.782	0.998	0.040	0.144	0.67	0.980
1000	OS	0.066	0.056	0.050	0.064	0.066	0.076	0.274	0.866	0.072	0.074	0.334	0.838
	BF	0.046	0.010	0.002	0	0.002	0.002	0.072	0.664	0.002	0.002	0.068	0.646
	B-DR	0.030	0.030	0.030	0.030	0.030	0.048	0.224	0.932	0.030	0.040	0.234	0.78
5000	OS	0.082	0.068	0.072	0.066	0.072	0.076	0.154	0.670	0.066	0.072	0.174	0.732
	BF	0.052	0.002	0	0	0	0	0.016	0.330	0	0	0.016	0.428
	B-DR	0.030	0.030	0.030	0.030	0.003	0.030	0.102	0.408	0.030	0.032	0.096	0.498

OS, our proposed testing procedure using the stabilized one-step estimator; HC, the higher criticism method; MF, the classical F-test for the Pillai trace without variable selection; BF, the F-test on selected variables with Bonferroni correction; BH-DR, multiple testing of $\mathrm{corr}(X_k, Y_j) = 0$ for all pq pairs of variables, in conjunction with the normal approximation test described in DiCiccio & Romano (2017) controlling the false discovery rate at the 5% level.

estimator, the Bonferroni corrected F-test and the Bonferroni corrected normal approximation test. Clearly, for large p, the proposed method has much better Type-I error control, under Model N, than the F-test with Bonferroni correction, and much larger power, under Models A1 and A2, than Bonferroni corrected tests when $p = 5000$. Overall, the proposed stabilized one-step testing procedure adequately controlled the Type-I error around the nominal level $\alpha = 0.05$. Specifically, the Type-I error is always between 0.05 and 0.1 for all different $p \in \{30, 100, 1000, 5000\}$ and $s \in \{1, 2, 3, 4\}$ combinations. Although our test procedure is asymptotically valid, its slight anti-conservative behaviour appears to be caused by the stabilized one-step estimator using subsamples of size as small as 250 in this case. In contrast, higher criticism fails to control the Type-I error; the normal approximation test, either using Bonferroni correction or the Benjamini–Hochberg procedure, is conservative, and the Bonferroni corrected F-test is extremely conservative if we use $s > 1$. The proposed test is more powerful than the multivariate analysis-of-variance F-test, even in low dimensions ($p = 30$), and is able to detect weak signals, when the canonical correlations are no larger than 0.4 in all models, in very high dimensions ($p = 5000$).

4.3. *Simulation results for parameter estimation*

Although our theory and implementation are equally applicable to stabilized one-step estimators of $\tau_{\max}$ and $\tau^2_{\max}$, the empirical results for $\tau_{\max}$ are generally better than those of $\tau^2_{\max}$. The stabilized one-step estimator of $\tau^2_{\max}$ is not simply $\hat{\tau}^2_{\max}$.

Histograms of the estimated $\tau_{\max}$ and $\tau^2_{\max}$ from 500 independent samples, of size $n = 500$, under the null Model N are presented in Fig. 2, where we vary $s_x = s_y = s \in \{1, 2, 3, 4\}$. The stabilized one-step estimates for $\tau_{\max}$ and $\tau^2_{\max}$ are both seen to be approximately normal. For $\tau_{\max}$, the estimates are all centred around the truth, $\tau_{\max} = 0$, regardless of the choice of s. However,

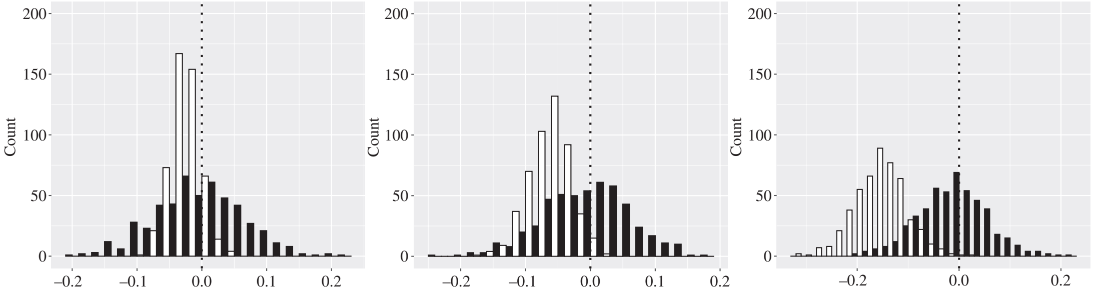

Fig. 2. Histograms of $\hat{\tau}_{\max}$ (black) under the null Model N based on 500 independent simulated datasets. The plots from left to right correspond to the sparsity levels $s_x = s_y = s = 1, 2$ and 4, respectively. The stabilized one-step estimator of $\tau^2_{\max}$ (white) shows a negative bias that becomes increasingly pronounced as s increases.

for $\tau^2_{\max}$, there is a severe underestimation phenomenon, which becomes more pronounced as s increases. This is because the number of parameters in $C_{Y_{\mathcal{J}}X_{\mathcal{K}}}$ is s^2, and requires a larger sample size n for the asymptotic properties to come into effect as s increases.

Under the alternative model (Model A1) with correlation strength $\tau_{\max}$ varying from 0.2 to 0.8, the histograms of $\hat{\tau}_{\max}$, again based on 500 independent samples of size $n = 500$, are given in the top panel of Fig. 3. Recall that the true sparsity levels are $s^\star_x = s^\star_y = 3$ in this model. We also set $s_x = s_y = s = 3$. Although there is an issue of underestimation for both $\tau_{\max}$ and $\tau^2_{\max}$ when the signal is weak ($\tau_{\max} = 0.2$ and 0.4), there is a substantial improvement when the correlation is strong enough. The improvement is more pronounced in the histogram of $\hat{\tau}_{\max}$ compared with that of the stabilized one-step estimator of $\tau^2_{\max}$ (bottom panel). It is worth noting that $\tau_{\max} = 0.8$ is still a relatively weak correlation, the estimated $\tau_{\max}$ exceeds 1.5 in the real data example of § 5, but both estimators worked well at $\tau_{\max} = 0.8$. An explanation for the underestimation is that the stabilizing procedure tends to attenuate the estimates to some extent, at least in the neighbourhood of $\tau_{\max} = 0$. However, as seen in Fig. 2, the behaviour of $\hat{\tau}_{\max}$ under the null model is unaffected by such attenuation, being approximately zero-mean normal.

5. Analysis of glioblastoma multiforme data

Glioblastoma, also called glioblastoma multiforme, is a type of fast-growing brain tumour and the most common primary form of brain tumour in adults. Data were collected by The Cancer Genome Atlas project (Weinstein et al., 2013) on 490 patients with glioblastoma, including data on $q = 534$ microRNA expression and 17 472 gene expression measurements for each patient. It is of interest to find associations between microRNA and gene expression. Following previous studies (Wang, 2015; Molstad, 2021), we analyse the $p = 1000$ genes with the largest median absolute deviations in gene expression, and pre-process the data by removing 93 subjects whose gene expression is substantially different from the majority. The resulting sample size in our data analysis is then $n = 397$.

We applied our Algorithm 1 to this dataset and obtained estimates of the maximal root-Pillai trace $\tau_{\max}$ over a range of values of $s = s_x = s_y$. The results are displayed in Fig. 4. Without adjusting for the postselection, the sample estimate $\hat{\tau}_{\text{samp}}$ of $\tau_{\max}$ increases almost linearly due to spurious correlations. On the other hand, the stabilized one-step estimator $\hat{\tau}_{\max}$ gives reasonable estimates that settle down beyond $s = 15$. The confidence intervals suggest that there is a highly significant association between microRNA and gene expression, with p-value less than 10^{-10}, which is consistent with previous studies. The results for the stabilized one-step estimator are based on 10 random reorderings of the data, because we do not know if the samples are

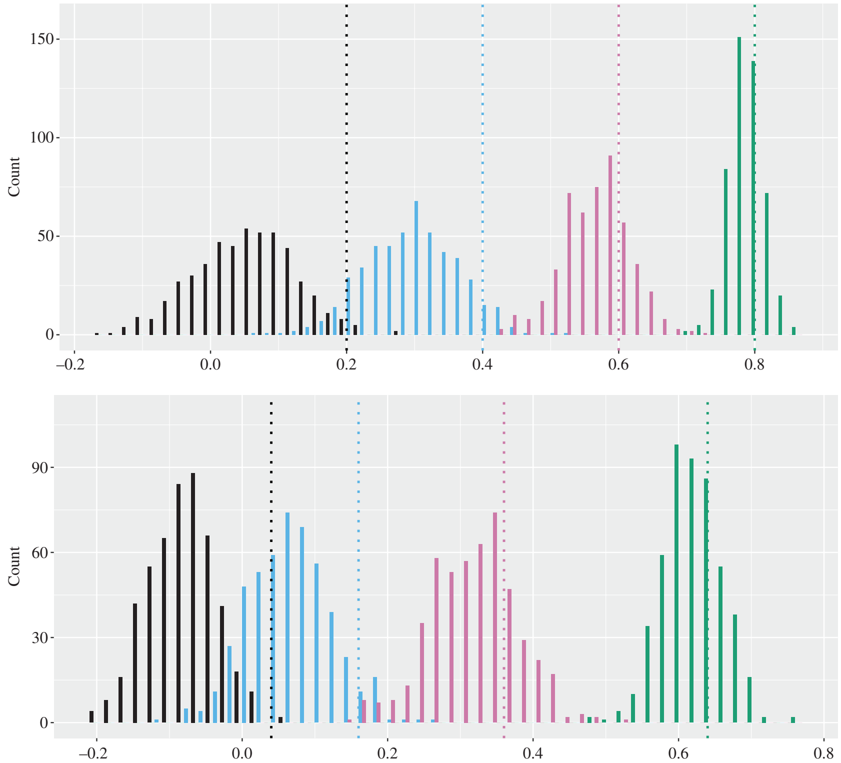

Fig. 3. Histograms of the stabilized one-step estimators for $\tau_{\max}$ (top panel) and $\tau^2_{\max}$ (bottom panel) under Model A1. In each plot, the four coloured histograms from left to right correspond to $\tau_{\max} = 0.2$, 0.4, 0.6 and 0.8, respectively. The vertical dashed lines are the true values of the targeted parameters $\tau_{\max}$ (top) and $\tau^2_{\max}$ (bottom).

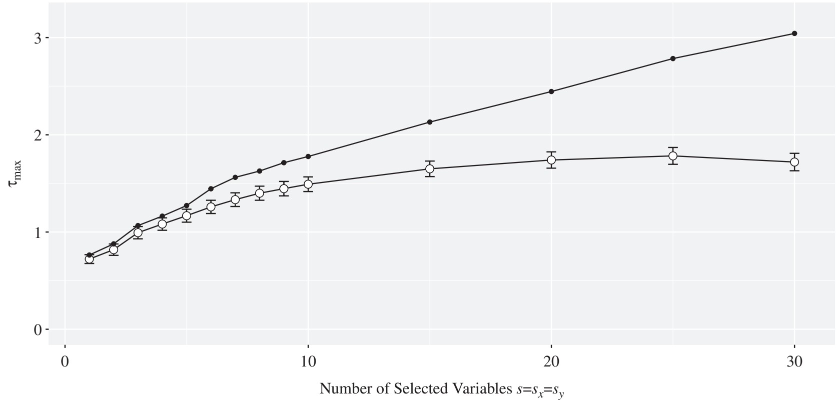

Fig. 4. Glioblastoma data analysis. Estimates of the maximal root-Pillai trace $\tau_{\max}$ are plotted against $s = s_x = s_y$ varying from 1–30, with the selected variables at each step found using Algorithm 1. The filled circles are $\hat{\tau}_{\text{samp}}$, without adjustment for postselection, giving inflated estimates of $\tau_{\max}$. The open circles and associated 95% confidence intervals are based on the stabilized one-step estimator $\hat{\tau}_{\max}$, with 10 random reorderings and averaged point estimates and averaged lower/upper confidence interval endpoints over these reorderings.

independent and identically distributed. The results without random reordering are very similar. The random ordering of the samples has little affect on the results; see further figures in the Supplementary Material for more details.

Under sparsity level $s = 3$, Table S.1 in the Supplementary Material lists the most correlated variables and their marginal correlations, while the stabilized one-step estimator $\hat{\tau}_{\max} = 0.931$ with standard error 0.085. Interestingly, the first two microRNA measurements, *hsa.miR.219* and *hsa.miR.222*, also appear in a reported dependency network of important microRNAs obtained by precision matrix estimation (Wang, 2015, Fig. 1). The top 25 microRNA and top 25 gene expressions in our analysis are provided in the Supplementary Material.

Acknowledgement

The authors thank Dr. Aaron Molstad from the University of Florida for sharing the pre-processed glioblastoma multiforme dataset. Research for this paper was partly supported by the U.S. National Science Foundation and National Institutes of Health.

Supplementary material

The Supplementary Material includes proofs and additional numerical results.

References

Andrews, D. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68**, 399–405.

Bao, Z., Hu, J., Pan, G. & Zhou, W. (2019). Canonical correlation coefficients of high-dimensional Gaussian vectors: finite rank case. *Ann. Statist.* **47**, 612–40.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300.

Bodnar, T., Dette, H. & Parolya, N. (2019). Testing for independence of large dimensional vectors. *Ann. Statist.* **47**, 2977–3008.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–54.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

Davies, R. B. (2002). Hypothesis testing when a nuisance parameter is present only under the alternative: linear model case. *Biometrika* **89**, 484–9.

Devlin, S. J., Gnanadesikan, R. & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–45.

DiCiccio, C. J. & Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *J. Am. Statist. Assoc.* **112**, 1211–20.

Donoho, D. & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–94.

Donoho, D. & Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30**, 1–25.

Gao, C., Ma, Z. & Zhou, H. H. (2017). Sparse CCA: adaptive estimation and computational barriers. *Ann. Statist.* **45**, 2074–101.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**, 413–30.

Hardoon, D. R. & Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* **83**, 331–53.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–77.

Khim, J. T., Jog, V. & Loh, P.-L. (2016). Computing and maximizing influence in linear threshold and triggering models. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett, eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 4545–53.

Krause, A. & Golovin, D. (2014). Submodular function maximization. In L. Bordeaux, Y. Hamadi, & P. Kohli, eds. *Tractability: Practical Approaches to Hard Problems*, pp. 71–104. Cambridge: Cambridge University Press.

Leeb, H. & Pötscher, B. M. (2017). Testing in the presence of nuisance parameters: some comments on tests post-model-selection and random critical values. In S. Ejaz Ahmed, ed. *Big and Complex Data Analysis: Methodologies and Applications*, pp. 69–82. Cham: Springer.

Luedtke, A. R. & van der Laan, M. (2018). Parametric-rate inference for one-sided differentiable parameters. *J. Am. Statist. Assoc.* **113**, 780–8.

MAI, Q. & ZHANG, X. (2019). An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics* **75**, 734–44.

MCKEAGUE, I. W. & QIAN, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *J. Am. Statist. Assoc.* **110**, 1422–33.

MOLSTAD, A. J. (2021). New insights for the multivariate square-root lasso. *arXiv:* 1909.05041v4.

NAYLOR, M. G., LIN, X., WEISS, S. T., RABY, B. A. & LANGE, C. (2010). Using canonical correlation analysis to discover genetic regulatory variants. *PLoS One* **5**, 1–6.

NEMHAUSER, G. L. & WOLSEY, L. A. (1978). Best algorithms for approximating the maximum of a submodular set function. *Math. Oper. Res.* **3**, 177–88.

PARKHOMENKO, E., TRITCHLER, D. & BEYENE, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statist. Applic. Genet. Molec. Biol.* **8** Doi: 10.2202/1544-6115.1406.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*, Lecture Notes in Statistics, vol. 13, Ed. D. Brillinger, S. Fienberg, J. Gani, J. Hartigan & K. Krickeberg, eds. New York: Springer.

PFANZAGL, J. (1990). *Estimation in Semiparametric Models*. New York: Springer.

PILLAI, K. C. S. (1955). Some new test criteria in multivariate analysis. *Ann. Math. Statist.* **26**, 117–21.

QADAR, M. & SEGHOUANE, A.-K. (2019). A projection CCA method for effective fMRI data analysis. *IEEE Trans. Biomed. Eng.* **66**, 3247–56.

R DEVELOPMENT CORE TEAM (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

SEGHOUANE, A.-K. & SHOKOUHI, N. (2019). Estimating the number of significant canonical coordinates. *IEEE Access* **7**, 108806–17.

SHI, H., DRTON, M. & HAN, F. (2021). Distribution-free consistent independence tests via center-outward ranks and signs. *J. Am. Statist. Assoc.*, DOI: 10.1080/01621459.2020.1782223.

SHU, H., WANG, X. & ZHU, H. (2020). D-CCA: a decomposition-based canonical correlation analysis for high-dimensional datasets. *J. Am. Statist. Assoc.* **115**, 292–306.

SONG, Y., SCHREIER, P. J., RAMIREZ, D. & HASIJA, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. *Sig. Proces.* **128**, 449–58.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.

VAN DER LAAN, M. J. & LENDLE, S. D. (2014). Online targeted learning. Technical Report 330, Division of Biostatistics, University of California, Berkeley. Available at `http://www.bepress.com/ucbbiostat/`.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

WAAIJENBORG, S. & ZWINDERMAN, A. H. (2007). Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers. *BMC Proc.* **1**, S122.

WAAIJENBORG, S., VERSELEWEL DE WITT HAMER, P. C. & ZWINDERMAN, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statist. Applic. Genet. Molec. Biol.* **7**, 1–29.

WANG, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statist. Sinica* **25**, 831–51.

WANG, Y. R., JIANG, K., FELDMAN, L. J., BICKEL, P. J. & HUANG, H. (2015). Inferring gene–gene interactions and functional modules using sparse canonical correlation analysis. *Ann. Appl. Statist.* **9**, 300–23.

WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C., STUART, J. M. & NETWORK, C. G. A. R. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genet.* **45**, 1113.

WIESEL, A., KLIGER, M. & HERO, A. O. (2008). A greedy approach to sparse canonical correlation analysis. *arXiv:* 0801.2748.

WITTEN, D. M., TIBSHIRANI, R. & HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–34.

YANG, Y. & PAN, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43**, 467–500.

ZHENG, S., CHENG, G., GUO, J. & ZHU, H. (2019). Test for high-dimensional correlation matrices. *Ann. Statist.* **47**, 2887–921.

ZHU, L., XU, K., LI, R. & ZHONG, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829–43.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–20.

[*Received on* 2 *March* 2021. *Editorial decision on* 19 *October* 2021]