entropy

MDPI

Article

An Approach to Canonical Correlation Analysis Based on Rényi's Pseudodistances

María Jaenada [1], Pedro Miranda [1,*], Leandro Pardo [1] and Konstantinos Zografos [2]

[1] Interdisciplinary Mathematics Institute, Complutense University of Madrid, 28040 Madrid, Spain
[2] Mathematics, Probability-Statistics and Operational Research Unit, Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece
* Correspondence: pmiranda@ucm.es

Abstract: Canonical Correlation Analysis (CCA) infers a pairwise linear relationship between two groups of random variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this paper, we present a new procedure based on Rényi's pseudodistances (RP) aiming to detect linear and non-linear relationships between the two groups. RP canonical analysis (RPCCA) finds canonical coefficient vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$, by maximizing an RP-based measure. This new family includes the Information Canonical Correlation Analysis (ICCA) as a particular case and extends the method for distances inherently robust against outliers. We provide estimating techniques for RPCCA and show the consistency of the proposed estimated canonical vectors. Further, a permutation test for determining the number of significant pairs of canonical variables is described. The robustness properties of the RPCCA are examined theoretically and empirically through a simulation study, concluding that the RPCCA presents a competitive alternative to ICCA with an added advantage in terms of robustness against outliers and data contamination.

Keywords: Information Canonical Correlation Analysis; Kullback–Leibler divergence; mutual information; Renyi's pseudodistances; robustness; consistency

Citation: Jaenada, M.; Miranda, P.; Pardo, L.; Zografos, K. An Approach to Canonical Correlation Analysis Based on Rényi's Pseudodistances. *Entropy* **2023**, *25*, 713. https://doi.org/10.3390/e25050713

Academic Editors: Carlos Alberto De Bragança Pereira, Antonio M. Scarfone and Christian H. Weiss

Received: 30 January 2023
Revised: 16 March 2023
Accepted: 21 April 2023
Published: 25 April 2023

1. Introduction

Canonical Correlation Analysis (CCA) is a statistical technique used to identify and measure associations among two sets of variables; in the following, denoted by $\boldsymbol{X}_{q\times 1}$ and $\boldsymbol{Y}_{p\times 1}$ $(q \leq p)$. It is appropriate in situations where multiple regression would be used but where there are multiple intercorrelated outcome variables. Hence, it allows us to summarize relationships into a lesser number of statistics while preserving the main facets of those relationships. CCA was first considered in [1] and has been widely used in the statistical literature; for example, to summarize relationships between sets of variables, to reduce the dimensionality of data or to transform two sets of variables into a new dataset of uncorrelated variables as a preprocessing step for the multiple linear regression model. More insight about CCA can be found, e.g., in [2,3].

CCA looks for two direction vectors $\boldsymbol{a}, \boldsymbol{b}$ (canonical vectors) such that the linear combinations $U = \boldsymbol{a}^T\boldsymbol{X}$ and $V = \boldsymbol{b}^T\boldsymbol{Y}$, so called canonical variables, are (linearly) correlated as much as possible. However, if a linear relationship does not exist between the pairs $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$, CCA could fail in detecting these pairs of canonical vectors. In other words, CCA can only detect *linear* relations between the canonical variables, but other functional relationships may exist.

The linear restriction is a significant drawback of CCA when analyzing some real data with highly non-linear relationships. For example, Oulai et al. [4] presented a real situation with non-linear relationships between variables regarding the representation of a hydrological process in the delineation of homogeneous regions. In their context, the two groups of variables under consideration were hydrological variables and meteorological

and/or graphical characteristics of watersheds, and their non-linear relationship depended essentially on the physiographic characteristics of the watersheds. Additionally, Ref. [5] presented a nice application of non-linear CCA to seasonal climate forecasting. In [6], some real life data with complex non-linear relationships that cannot be properly captivated by classical CCA are also presented. There is an extensive bibliography addressing non-linear CCA. Without wishing to cite all the existing literature on the topic, we would like to mention some interesting works on the subject: [7] (Chapter 6), [8–11] and references therein.

To shed light on this problem, let us consider the following situation described in [12]: let $\boldsymbol{X} = (X_1, X_2)^T$ and $\boldsymbol{Y} = (Y_1, Y_2)^T$ be a pair of random vectors such that

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad Y_1 = X_1^2 + Z,\ Y_2 = Z,$$

with $Z \sim \chi_1^2$ and independent from $\boldsymbol{X}$. In this case, $Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}_{2\times 2}$, and so the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are uncorrelated (so they are linearly independent). Consequently, classical CCA cannot detect that, indeed, Y_1 is related (although not linearly) to X_1, even if the variables are not *fully* independent. On the other hand, as the pair $(\boldsymbol{X}, \boldsymbol{Y})$ does not follow a normal distribution (and therefore uncorrelation does not imply independence), a hidden relationship may exist (and indeed it does exist!) that has not been detected by CCA. Of course, under normality, rejecting any linear correlation using CCA implies independence between both variables.

It is not surprising that CCA fails in the previous example, as CCA focuses on "linear trends", but the true relation underlying it is quadratic. To overcome this drawback, in a pioneer paper, Yin [12] proposed the use of the Kullback–Leibler divergence and developed a new procedure called *Informational Canonical Correlation Analysis* (ICCA), aiming to also detect non-linear relationships for linear combinations of the components.

Let $U = \boldsymbol{a}^T\boldsymbol{X}$ and $V = \boldsymbol{b}^T\boldsymbol{Y}$ be linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ defining a pairwise of canonical variables, with $\boldsymbol{a} \in \mathbb{R}^q$ and $\boldsymbol{b} \in \mathbb{R}^p$. We denote by $f_{UV}(u, v)$ the joint probability density function (PDF) of (U, V), and, by $f_U(u)$ (resp. $f_V(v)$), the marginal unidimensional PDF of U (resp. V). From a statistical point of view, both canonical variables U and V would be independent if their joint distribution coincides with the product of the marginal PDF's, $f_{UV}(u, v) = f_U(u) \times f_V(v)$, and, conversely, a strong dependence between U and V would result in a large statistical distance between the joint PDF and the product of the marginals. A suitable divergence should then be adopted to measure such statistical closeness of the two PDFs. The Kullback–Leibler divergence is the most commonly used measure for distinguishing two distributions, and it has a great statistical importance in the field of information theory.

The Kullback–Leibler divergence between $f_{UV}(u, v)$ and $f_U(u) \times f_V(v)$ is given as

$$D_{KL}(\boldsymbol{a}, \boldsymbol{b}) := D_{KL}(f_{UV}, f_U \times f_V) = \int_{\mathbb{R}^2} f_{UV}(u, v) \ln \frac{f_{UV}(u, v)}{f_U(u) f_V(v)} du dv. \tag{1}$$

The above divergence is not symmetric, so it quantifies the expected inaccuracy excess from using $f_U \times f_V$ as a model when the actual PDF is f_{UV}. That is, the inaccuracy caused by assuming independence between the pair of canonical variables. Consequently, truly independent canonical variables should minimize the Kullback–Leibler divergence in Equation (1); conversely, functionally dependent canonical variables should maximize the divergence. For more details about the Kullback–Leibler divergence, see [13].

In this vein, ICCA aims to identify q pairwise canonical variables $\boldsymbol{a}_i \in \mathbb{R}^q$ and $\boldsymbol{b}_i \in \mathbb{R}^p$, $i \leq q \leq p$ such that $D_{KL}(\boldsymbol{a}_i, \boldsymbol{b}_i) = \max_{\boldsymbol{a}, \boldsymbol{b}} D_{KL}(\boldsymbol{a}, \boldsymbol{b})$. However, the Kullback–Leibler divergence is invariant under linear transformations, and so there are infinitely many ways to define canonical vectors yielding the same objective function. Then, for identification, we constrain the canonical variables to have unit variance. Moreover, once a relationship is identified by a pair of canonical variables, we expect to exclude its effect from the consecutive canonical variables. For such a purpose, we also require that pairs of canonical variables are uncorrelatated with any other pair. That is, ICCA finds q linearly

independent pairs of canonical variables with unit variance maximizing (in decreasing order) the corresponding Kullback–Leibler divergence. Mathematically, to compute each pair of variables, we need to solve the optimization problem $D_{KL}(\boldsymbol{a}_i, \boldsymbol{b}_i) = \max_{\boldsymbol{a},\boldsymbol{b}} D_{KL}(\boldsymbol{a}, \boldsymbol{b})$ subject to $\boldsymbol{a}_i^T \boldsymbol{\Sigma}_X \boldsymbol{a}_i = \boldsymbol{b}_i^T \boldsymbol{\Sigma}_Y \boldsymbol{b}_i = 1$ and $\boldsymbol{a}_j^T \boldsymbol{\Sigma}_X \boldsymbol{a}_i = \boldsymbol{b}_j^T \boldsymbol{\Sigma}_Y \boldsymbol{b}_i = 0$ for $j = 1, \ldots, i-1$, where $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ denote the variance–covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We apply the same for RP.

From Yin's (2004) reinterpretation of the canonical analysis, several procedures based on divergence and entropy measures have been proposed to reduce the limitations of CCA. For example, Mandal et al. [14] considered (α, β) divergence measures defined in [15], and Iaci and Sriram [6] used the density power divergence measures defined in [16] as a measure of statistical closeness. In [17], canonical dependence based on the squared-loss mutual information was studied. Other interesting results regarding ICCA can be seen in [18–23].

Despite its popularity, the Kullback–Leibler divergence association measure is quite sensitive to outlying observations, as pointed out in [24]. For outliers, we mean data that behave very differently to expectations according to the law modeling the relation. The main purpose of this paper is to extend the ICCA procedure to a wider family of robust methods based on RP divergence, which remains competitive to ICCA in terms of efficiency but provides a more stable estimation of the canonical vectors in the presence of contamination in the data.

The RP family, parameterized by a tuning parameter τ controlling the trade-off between robustness and efficiency, was considered for the first time in Jones et al. [25]. Later, Broniatowski et al. [26] demonstrated that RP is a proper divergence, positive for any two densities and for all values of the tuning parameter [26,27], and it is null if (and only if) both densities are the same. The theory in [26] for independent and identically distributed random variables was extended to the case of independent but not identically distributed random variables in [28]. They termed this family of pseudodistances as RP because of their similarities with Renyi's divergence measures Rényi (1961) [29]. Rényi's pseudodistance has shown promising behavior in other statistical problems, providing robust minimum RP estimators with good asymptotic and robustness properties, and it includes the Kullback–Leibler divergence as a particular case at $\tau = 0$. For example, Toma and Leoni-Aubin [30] considered efficient and robust measures for general parametric models based on RP and, Toma et al. [31] later developed a new criterion for model selection based on the RP. In [27], Castilla et al. introduced a family of Wald-type tests for testing the parameters in linear regression models, and these results were later extended for generalized linear regression models in [32,33]. Wald-type tests based on minimum RP estimators in bidimensional normal populations were considered in [34]. Jaenada et al. [35] introduced and studied the minimum RP estimators under restricted parameter spaces, which are of great statistical interest in many practical applications such as hypothesis testing. Under the name of γ-*entropy*, Fujisawa and Eguchi [36] applied RP to introduce robust estimators of general parametric families. Motivated for the great performance of the minimum RP estimator on those different statistical models in terms of robustness, we have adopted the RP divergence to extend the ICCA procedure.

The rest of the paper is organized as follows. The Rényi's Pseudodistance Canonical Correlation Analysis (RPCCA) is introduced in Section 2, and some of its properties are studied. Next, an estimation design for computing the canonical vectors in practice using RPCCA is described in Section 3. In Section 4, the robustness of the RPCCA is theoretically established. Section 5 describes a permutation test to determine the number of significant canonical variables and thereby provide a dimension reduction method. In Section 6, a Monte Carlo simulation study is carried out to empirically evaluate the performance of the RPCCA and compare the proposed method with the ICCA in terms of estimation accuracy and robustness. An example with real data is studied in Section 6.3. Finally, some conclusions are drawn in Section 7.

2. Rényi's Pseudodistance Canonical Correlation Analysis

Given two multidimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, the RPCCA aims to identify two direction vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ (the canonical vectors), such that the corresponding canonical variables $U = \boldsymbol{a}^T\boldsymbol{X}$ and $V = \boldsymbol{b}^T\boldsymbol{Y}$ are as dependent as possible. Such dependency is measured in terms of RP between their joint distribution and the product of their marginal distributions. The **RP of tuning parameter** τ between the joint distribution of the bidimensional random variable (U, V) and the product of their marginals, $f_U(u) \times f_V(v)$, is given for $\tau > 0$ by (cf. [26]).

$$\begin{aligned} d_\tau(\boldsymbol{a},\boldsymbol{b}) &= d_\tau(f_U \times f_V, f_{UV}) \\ &= \frac{1}{\tau+1}\ln\int_{\mathbb{R}^2} f_U^{\tau+1}(u) f_V^{\tau+1}(v)dudv - \frac{1}{\tau}\ln\int_{\mathbb{R}^2} f_U^{\tau}(u) f_V^{\tau}(v) f_{UV}(u,v)dudv \\ &\quad + \frac{1}{\tau(\tau+1)}\ln\int_{\mathbb{R}^2} f_{UV}^{\tau+1}(u,v)dudv. \end{aligned}$$

Hence, the RP measures the statistical discrepancy between the joint PDF of the canonical variables, f_{UV} and the marginal PDF's product $f_U \times 4_V$, or, in other words, the loss in accuracy that comes with assuming independence.

For $\tau = 0$, the RP can be defined as the corresponding limit, $\tau \to 0$, yielding the Kullback–Leibler divergence:

$$d_0(\boldsymbol{a},\boldsymbol{b}) = \lim_{\tau\downarrow 0} d_\tau(\boldsymbol{a},\boldsymbol{b}) = \lim_{\tau\downarrow 0} d_\tau(f_U \times f_V, f_{UV}) = D_{KL}(f_{UV}, f_U \times f_V). \tag{2}$$

As earlier discussed, independent canonical variables lead to $d_\tau(\boldsymbol{a},\boldsymbol{b}) = 0$, and, contrarily, strong dependency should result in large RP distances. Then, the **RPCCA** procedure aims to identify pairwise canonical vectors $\boldsymbol{a}_i \in \mathbb{R}^q$ and $\boldsymbol{b}_i \in \mathbb{R}^p$, $i \leq q \leq p$ such that

$$d_\tau(\boldsymbol{a}_i,\boldsymbol{b}_i) = \max_{\boldsymbol{a},\,\boldsymbol{b}} d_\tau(\boldsymbol{a},\boldsymbol{b}),$$

and, as before for identification, the canonical variables should have unit variance and be uncorrelated with any previous pairwise of canonical variables:

$$\boldsymbol{a}_i^T\boldsymbol{\Sigma}_X\boldsymbol{a}_i = \boldsymbol{b}_i^T\boldsymbol{\Sigma}_Y\boldsymbol{b}_i = 1, \quad \forall i,$$

$$\boldsymbol{a}_j^T\boldsymbol{\Sigma}_X\boldsymbol{a}_i = \boldsymbol{b}_j^T\boldsymbol{\Sigma}_Y\boldsymbol{b}_i = 0, \quad \forall j = 1,\ldots,i-1,$$

where $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are the variance–covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

Note that, by Equation (2), the ICCA procedure presented in [12] is recovered at $\tau = 0$, and so the RPCCA generalizes ICCA.

Remark 1. *Given the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, RPCCA finds the vectors $\boldsymbol{a}_1, \boldsymbol{b}_1$ such that $\boldsymbol{a}_1^T\boldsymbol{X}$ and $\boldsymbol{b}_1^T\boldsymbol{Y}$ are maximally related. This maximal relation is measured via $d_\tau(\boldsymbol{a}_i,\boldsymbol{b}_i)$, as previously defined. Once these vectors $\boldsymbol{a}_1, \boldsymbol{b}_1$ are obtained, the procedure looks for a new pair of vectors $\boldsymbol{a}_2, \boldsymbol{b}_2$ such that $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are incorrelated, and the same applies for $\boldsymbol{b}_1, \boldsymbol{b}_2$, and $\boldsymbol{a}_2^T\boldsymbol{X}$ and $\boldsymbol{b}_2^T\boldsymbol{Y}$ are maximally related. Consequently,*

$$d_\tau(\boldsymbol{a}_1,\boldsymbol{b}_1) \geq d_\tau(\boldsymbol{a}_2,\boldsymbol{b}_2).$$

Next, the procedure looks for $\boldsymbol{a}_3, \boldsymbol{b}_3$ being incorrelated to $\boldsymbol{a}_1, \boldsymbol{a}_2$ and $\boldsymbol{b}_1, \boldsymbol{b}_2$, respectively, and so on. Hence, it follows that

$$d_\tau(\boldsymbol{a}_i,\boldsymbol{b}_i) \geq d_\tau(\boldsymbol{a}_{i+1},\boldsymbol{b}_{i+1}),$$

for any $i = 1,\ldots,q-1$. If $d_\tau(\boldsymbol{a}_i,\boldsymbol{b}_i) = 0$. Then, independence arises and the procedure stops. In practice, we will have an estimation of $d_\tau(\boldsymbol{a}_i,\boldsymbol{b}_i)$, and we will stop the procedure if this value does not exceed a certain threshold. This will be applied in Section 5 in order to determine the number of components.

Let us consider, again, the example described in the introduction, where $\boldsymbol{X} = (X_1, X_2)^T$ and $\boldsymbol{Y} = (Y_1, Y_2)^T$ satisfy

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad Y_1 = X_1^2 + Z,\ Y_2 \sim Z, \quad Z \sim \chi_1^2.$$

The true value of the first pair of canonical vectors are then $\boldsymbol{a}_1 = \boldsymbol{b}_1 = (1,0)^T$. Under the described setup, it follows that

$$f_{UV}(u,v) = f_{X_1,Y_1}(x,y) = \frac{1}{\pi}\frac{1}{2}\exp\left(-\frac{1}{2}y\right)(y-x^2)^{-\frac{1}{2}}, \quad y > 0, x \in \mathbb{R},$$

and

$$f_U(u) = f_{X_1}(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right), \quad f_V(v) = f_{Y_1}(y) = \frac{1}{2}\exp\left(-\frac{1}{2}y\right), \quad y > 0.$$

Clearly,

$$f_{UV}(u,v) \neq f_U(u) \times f_V(v)$$

and because of the properties of the divergence

$$d_\tau(\boldsymbol{a}_1, \boldsymbol{b}_1) > 0.$$

The last inequality holds because the RP divergence, $d_\tau(\cdot,\cdot)$, only reaches the value zero if both arguments coincide, as discussed in Section 1 (see [26] for more details). In this case, RPCCA should identify a pair $\boldsymbol{a}_1, \boldsymbol{b}_1$ with a non-zero informational coefficient of canonical correlation defining the canonical variables $\boldsymbol{a}_1^T\boldsymbol{X}$ and $\boldsymbol{b}_1^T\boldsymbol{Y}$.

For practical use of RPCCA, it is interesting to note that RPCCA is equivariant under invertible linear transformations. This equality does not hold for other extensions of ICCA, but proportionality arises instead.

Proposition 1. *Consider two random variables U and V, and take $R = cU$ and $S = eV$, where c and e are non-zero real numbers (Indeed, the result also holds if we consider two random vectors $\boldsymbol{U}, \boldsymbol{V}$, and consider $\boldsymbol{R} = C\boldsymbol{U}$ and $\boldsymbol{S} = D\boldsymbol{V}$, where C and D are two invertible matrices. In this case, RP is computed considering multidimensional integrals.). Then,*

$$d_\tau(f_U \times f_V, f_{UV}) = d_\tau(f_R \times f_S, f_{RS}).$$

Proof. By definition,

$$\begin{aligned}
d_\tau(f_R \times f_S, f_{RS}) &= \frac{1}{\tau+1}\ln\int f_R^{\tau+1}(r) f_S^{\tau+1}(s)\, drds + \frac{1}{\tau(\tau+1)}\ln\int f_{RS}^{\tau+1}(r,s)\, drds \\
&\quad -\frac{1}{\tau}\ln\int f_R^{\tau}(r) f_S^{\tau}(s) f_{RS}(r,s)\, drd\, s \\
&= \frac{1}{\tau+1}\ln\int f_U^{\tau+1}(u)\left(\frac{1}{c}\right)^{\tau+1} f_V^{\tau+1}(v)\left(\frac{1}{e}\right)^{\tau+1} ce\, dudv \\
&\quad +\frac{1}{\tau(\tau+1)}\ln\int f_{UV}^{\tau+1}(u,v)\left(\frac{1}{ce}\right)^{\tau+1} ce\, dudv \\
&\quad -\frac{1}{\tau}\ln\int f_U^{\tau}(u)\left(\frac{1}{c}\right)^{\tau} f_V^{\tau}(s)\left(\frac{1}{e}\right)^{\tau} f_{UV}(u,v)\left(\frac{1}{ce}\right) ce\, dud\, v \\
&= \frac{1}{\tau+1}\ln\left(\frac{1}{|c||e|}\right)^{\tau}\int f_U^{\tau+1}(u) f_V^{\tau+1}(v)\, dud\, v \\
&\quad +\frac{1}{\tau(\tau+1)}\ln\left(\frac{1}{|c||e|}\right)^{\tau}\int f_{UV}^{\tau+1}(u,v)\, dudv \\
&\quad -\frac{1}{\tau}\ln\left(\frac{1}{|c||e|}\right)^{\tau}\int f_U^{\tau}(u) f_V^{\tau}(v) f_{UV}(u,v)\, dud\, v \\
&= \left(\frac{1}{\tau+1}+\frac{1}{\tau(\tau+1)}-\frac{1}{\tau}\right)\ln\left(\frac{1}{|c||e|}\right)^{\tau} + d_\tau(f_U \times f_V, f_{UV}) \\
&= d_\tau(f_U \times f_V, f_{UV}).
\end{aligned}$$

□

The next result establishes that the RPCCA is reduced to CCA in the case of normal distributions.

Proposition 2. *In the case of normal distributions, RPCCA coincides with CCA.*

Proof. Consider normal populations, i.e., assume that the multidimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly normally distributed,

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \equiv \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_Y \end{pmatrix} \right).$$

Therefore, the bidimensional random variable $(U, V) = \left(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}\right)$ follows a bidimensional normal distribution whose vector mean is

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T = (E\left[\boldsymbol{a}^T\boldsymbol{X}\right], E\left[\boldsymbol{b}^T\boldsymbol{Y}\right])^T = (\boldsymbol{a}^T\boldsymbol{\mu_X}, \boldsymbol{b}^T\boldsymbol{\mu_Y})^T,$$

and the variance–covariance matrix is given by

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

being

$$\sigma_1^2 = Var\left[\boldsymbol{a}^T\boldsymbol{X}\right] = \boldsymbol{a}^T\boldsymbol{\Sigma}_X\boldsymbol{a},\ \sigma_2^2 = Var\left[\boldsymbol{b}^T\boldsymbol{Y}\right] = \boldsymbol{b}^T\boldsymbol{\Sigma}_Y\boldsymbol{b} \text{ and } \rho = \frac{Cov(U,V)}{\sigma_1\sigma_2} = \frac{\boldsymbol{a}^T\boldsymbol{\Sigma}_{XY}\boldsymbol{b}}{\sigma_1\sigma_2}.$$

On the other hand, the marginal densities $f_{\mu_1,\sigma_1}(u)$ and $f_{\mu_2,\sigma_2}(v)$ of $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$, respectively, are normal distributions,

$$f_U(u) \equiv \mathcal{N}\left(\mu_1, \sigma_1^2\right) \text{ and } f_V(v) \equiv \mathcal{N}\left(\mu_2, \sigma_2^2\right).$$

We first compute the RP between $f_{\mu_1,\sigma_1}(u) \times f_{\mu_2,\sigma_2}(v)$ and $f_{\mu_1,\mu_2,\sigma_1,\sigma_2,\rho}(u,v)$. Considering the results obtained in Supplementary Materials (Appendix A) in [6], we have

$$\int_{\mathbb{R}^2} f_U(u)^{\tau+1} f_V(v)^{\tau+1} du dv = k_1^\tau (1+\tau)^{-1},$$

being $k_1 = (2\pi\sigma_1\sigma_2)^{-1}$ and

$$\int_{\mathbb{R}^2} f_{UV}(u,v)^{\tau+1} du dv = k_1^\tau (1+\tau)^{-1}\left(1-\rho^2\right)^{-\frac{\tau}{2}}.$$

On the other hand, it is not difficult to see that

$$\int_{\mathbb{R}^2} f_U^\tau(u) f_V^\tau(v) f_{UV}(u,v) du dv = k_1^\tau [(1+\tau(1+\rho))(1+\tau(1-\rho))]^{-1/2}.$$

Based on the previous quantities, we have

$$\begin{aligned} d_\tau(\boldsymbol{a},\boldsymbol{b}) &= \frac{1}{\tau+1}\ln k_1^\tau (1+\tau)^{-1} - \frac{1}{\tau}\ln k_1^\tau k_1^\tau [(1+\tau(1+\rho))(1+\tau(1-\rho))]^{-1/2} \\ &\quad + \frac{1}{\tau(\tau+1)}\ln k_1^\tau (1+\tau)^{-1}(1-\rho^2)^{-\tau/2} \\ &= \ln \frac{((1+\tau(1+\rho))(1+\tau(1-\rho)))^{1/2\tau}}{(1+\tau)^{1/\tau}(1-\rho^2)^{1/2(\tau+1)}}. \end{aligned}$$

For fixed τ, it can be seen from the previous expression that $d_\tau(\boldsymbol{a},\boldsymbol{b})$ depends on ρ. Moreover, it is not difficult to show that $d_\tau(\boldsymbol{a},\boldsymbol{b})$ is an increasing function on ρ^2 for any τ (see Figure 1 for $\tau = 0.1, 0.3$ and 0.9). To show this, it suffices to see that

$$f_\tau(\rho) = \frac{[(1+\tau(1+\rho))(1+\tau(1-\rho))]^{1/2\tau}}{(1+\tau)^{1/\tau}(1-\rho^2)^{1/2(\tau+1)}}, \rho \in (-1,1)$$

is increasing in ρ^2, and so it will be its logarithm transform. Now, note that

$$(1+\tau(1+\rho))(1+\tau(1-\rho)) = 1+2\tau+\tau^2(1-\rho^2).$$

So, it suffices to show that the function

$$\frac{\left(1+2\tau+\tau^2(1-\rho^2)\right)^{1/2\tau}}{(1-\rho^2)^{1/2(\tau+1)}}$$

is increasing in ρ^2. Taking derivatives with respect to ρ^2, we obtain

$$f'(\rho^2) = \left[\frac{1}{2\tau}\left(1+2\tau+\tau^2(1-\rho^2)\right)^{1/2\tau-1}(-\tau^2)(1-\rho^2)^{1/2(\tau+1)}\right.$$

$$\left. -\frac{1}{2(\tau+1)}(1-\rho^2)^{1/2(\tau+1)-1}\left(1+2\tau+\tau^2(1-\rho^2)\right)^{1/2\tau}\right]\frac{1}{(1-\rho^2)^{\frac{1}{\tau+1}}}.$$

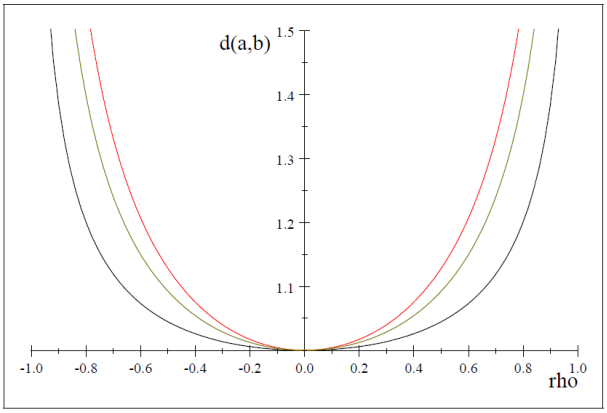

Figure 1. $f_\tau(\rho)$ for different values of τ. $\tau = 0.1$ (red), $\tau = 0.3$ (green) and $\tau = 0.9$ (black).

Thus, it suffices to check the non-negativity of

$$\left[\frac{1}{2\tau}(-\tau^2)(1-\rho^2) + \frac{1}{2(\tau+1)}\left(1+2\tau+\tau^2(1-\rho^2)\right)\right].$$

Finally,

$$\left[\frac{1}{2\tau}(-\tau^2)(1-\rho^2) + \frac{1}{2(\tau+1)}\left(1+2\tau+\tau^2(1-\rho^2)\right)\right] =$$

$$\frac{-\tau^2(1-\rho^2)}{2\tau(\tau+1)} + \frac{1+2\tau}{2(\tau+1)} = \frac{\tau^2(\rho^2+1)+\tau}{2\tau(\tau+1)} > 0.$$

and the result holds. Thus, RPCCA is equivalent to classical CCA in the case of random normal variables. □

It can be seen that $d_\tau(\boldsymbol{a};\boldsymbol{b})$ is an increasing function on ρ^2 for any $\tau > 0$ under normal distributions; hence, RPCCA also extends CCA with a tuning parameter τ determining the sharpness of the distance $d_\tau(\boldsymbol{a},\boldsymbol{b})$ (or the function $f_\tau(\cdot)$ in the proof of Proposition 2).

3. Consistency

We now focus on the practical side of the RPCCA estimation. In practice, the PDFs f_{UV}, f_U and f_V are unknown; thus, they should be empirically estimated. Likewise, the RPCCA should be formulated for an empirical setup.

The RP, $d_\tau(\boldsymbol{a}, \boldsymbol{b})$, can be expressed in terms of expected values as

$$\begin{aligned} d_\tau(\boldsymbol{a}, \boldsymbol{b}) = & \frac{1}{\tau(\tau+1)} \ln E_{f_{UV}}[f_{U,V}(U,V)^\tau] - \frac{1}{\tau} \ln E_{f_{UV}}[f_U^\tau(U) f_V^\tau(V)] \\ & + \frac{1}{\tau+1} \ln E_{f_U}[f_U(U)^\tau] E_{f_V}[f_V(V)]. \end{aligned} \tag{3}$$

This interpretation of $d_\tau(\boldsymbol{a}, \boldsymbol{b})$ makes the definition of its empirical estimator easier. Let $(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \ldots, n$ be a random sample of size n from the multidimensional random variables $(\boldsymbol{X}, \boldsymbol{Y})$. Then, an empirical estimator of $d_\tau(\boldsymbol{a}, \boldsymbol{b})$ is given by

$$\widehat{d}_\tau^n(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{\tau(\tau+1)} \ln \left[\frac{1}{n} \sum_{i=1}^n \widehat{f_{UV}^n}^\tau(u_i, v_i) \right] - \frac{1}{\tau} \ln \left[\frac{1}{n} \sum_{i=1}^n \widehat{f_U^n}^\tau(u_i) \widehat{f_V^n}^\tau(v_i) \right] \tag{4}$$

$$+ \frac{1}{\tau+1} \ln \left[\frac{1}{n} \sum_{i=1}^n \widehat{f_U^n}^\tau(u_i) \frac{1}{n} \sum_{i=1}^n \widehat{f_V^n}^\tau(v_i) \right]. \tag{5}$$

Here, $\widehat{f_U^n}(u)$, $\widehat{f_V^n}(v)$ and $\widehat{f_{UV}^n}(u, v)$ are kernel density estimators of $f_U(u)$, $f_V(v)$ and $f_{UV}(u, v)$, respectively, given by

$$\widehat{f_U^n}(u) = \frac{1}{n a_n^1} \sum_{i=1}^n K\left(\frac{u - u_i}{a_n^1} \right), \ u \in \mathbb{R}, \tag{6}$$

$$\widehat{f_V^n}(v) = \frac{1}{n a_n^2} \sum_{i=1}^n K\left(\frac{v - v_i}{a_n^2} \right), \ v \in \mathbb{R}, \tag{7}$$

and

$$\widehat{f_{UV}^n}(u, v) = \frac{1}{n b_n^1 b_n^2} \sum_{i=1}^n K\left(\frac{u - u_i}{b_n^1} \right) K\left(\frac{v - v_i}{b_n^2} \right). \tag{8}$$

For the PDF's estimators, we will use the univariate Gaussian kernel with $a_n^j = 1.06 n^{-0.2} s_j$ and $b_n^j = n^{-1/6} s_j$ for $j = 1, 2$, and the corresponding sample standard deviations s_1 and s_2. This kernel function was proposed in [37] and adopted in many other extensions of ICCA, but other types of kernels could be considered instead, as long as they satisfy the conditions of Lemma 1 below (When the distribution is known up to a parameter value $f_{\boldsymbol{\theta}}$, this information should be taken into account. Hence, the procedure would be the usual procedure in these situations. First, we estimate the parameter of the distribution $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ and then consider the distribution with the estimated parameters $f_{\hat{\boldsymbol{\theta}}}$. Next, we use $f_{\hat{\boldsymbol{\theta}}}$ instead of $\hat{f}$.). Other interesting results about kernel distributions can be found in [38,39].

Then, the estimated canonical vectors, based on the RP with tuning parameter τ can be computed as

$$\begin{aligned} \left(\widehat{\boldsymbol{a}}_n^\tau, \widehat{\boldsymbol{b}}_n^\tau \right) = & \arg\max_{\boldsymbol{a}, \boldsymbol{b}} \widehat{d}_\tau^n(\boldsymbol{a}, \boldsymbol{b}), \\ & \text{s.t. } (\boldsymbol{a}_n^\tau)^T \widehat{\boldsymbol{\Sigma}}_{11} \boldsymbol{a}_n^\tau = 1 \text{ and } (\boldsymbol{b}_n^\tau)^T \widehat{\boldsymbol{\Sigma}}_{22} \boldsymbol{b}_n^\tau = 1, \end{aligned} \tag{9}$$

where $\widehat{\boldsymbol{\Sigma}}_{11}$ and $\widehat{\boldsymbol{\Sigma}}_{22}$ are the empirical estimators of the variance–covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

We next establish the consistency of the estimated canonical vectors under some regularity conditions. That is, we will prove that the estimated canonical vectors $\left(\widehat{\boldsymbol{a}}_n^\tau, \widehat{\boldsymbol{b}}_n^\tau \right)$ converge for large sample sizes to the true canonical vectors defining the underlying

functional relationship. For such a result, it is necessary to present the following lemma whose proof can be found in [40].

Lemma 1. *Let* $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$ *be i.i.d. replications of the multidimensional random variables* $(\boldsymbol{X}, \boldsymbol{Y})$. *Consider a sequence* $\{a_n\}_{n\in\mathbb{N}}$ *such that* $0 < a_n$ *and* $\lim_{n\to\infty} a_n = 0$. *Assume*

$$\sum_{n=1}^{\infty} e^{-\gamma n a_n^2} < \infty, \sum_{n=1}^{\infty} e^{-\gamma n a_n^4} < \infty, \forall \gamma > 0.$$

Consider a function K of bounded variation (Consider a function $g : \mathbb{R}^k \mapsto \mathbb{R}$, *and let P be the set of finite partitions of* $\mathbb{R}^k$ *in rectangles* $p = \{[x_j, y_j), j = 1, \ldots, u_p\}$. *Then, g is said to be of bounded variation if* $sup_{p\in P}\left\{ \sum_{j=1}^{u_p} \sum_{\epsilon_1,\ldots,\epsilon_k\in\{0,1\}^k} (-1)^{\sum_{i=1}^k \epsilon_i} g(\epsilon_1 x_{j1} + (1-\epsilon_1)y_{j1}, \ldots, \epsilon_k x_{jk} + (1-\epsilon_k)y_{jk}) \right\} < \infty$.*) and suppose* $f_U(\boldsymbol{a}^T\boldsymbol{x})$ *is uniformly continuous in* $\boldsymbol{a}$ *and* $\boldsymbol{x}$, $f_V(\boldsymbol{b}^T\boldsymbol{y})$ *is uniformly continuous in* $\boldsymbol{b}$ *and* $\boldsymbol{y}$, *and* $f_{UV}(\boldsymbol{a}^T\boldsymbol{x}, \boldsymbol{b}^T\boldsymbol{y})$ *is uniformly continuous in* $\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{b}$ *and* $\boldsymbol{y}$. *Then,*

$$\sup_{\boldsymbol{a},\boldsymbol{x}} |\widehat{f_U^n}(\boldsymbol{a}^T\boldsymbol{x}) - f_U(\boldsymbol{a}^T\boldsymbol{x})| \xrightarrow{a.s.} 0.$$

$$\sup_{\boldsymbol{b},\boldsymbol{y}} |\widehat{f_V^n}(\boldsymbol{b}^T\boldsymbol{y}) - f_V(\boldsymbol{b}^T\boldsymbol{y})| \xrightarrow{a.s.} 0.$$

$$\sup_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{x},\boldsymbol{y}} |\widehat{f_{UV}^n}(\boldsymbol{a}^T\boldsymbol{x}, \boldsymbol{b}^T\boldsymbol{y}) - f_{UV}(\boldsymbol{a}^T\boldsymbol{x}, \boldsymbol{b}^T\boldsymbol{y})| \xrightarrow{a.s.} 0.$$

Note that the Gaussian kernel functions defined in Equations (6)–(8) satisfy the conditions of Lemma 1. Of course, any other election of the kernel should also satisfy these regularity conditions. Now, let us define for any real value $b > 0$ the set of indices such that the observations $\boldsymbol{a}^T\boldsymbol{x}_i$ and $\boldsymbol{b}^T\boldsymbol{y}_i, i = 1, \ldots, n$ have positive densities

$$\chi_b = \{i : f_{UV}^\tau(\boldsymbol{a}^T\boldsymbol{x}_i, \boldsymbol{b}^T\boldsymbol{y}_i) \geq b,\ f_U^\tau(\boldsymbol{a}^T\boldsymbol{x}_i) \geq b,\ f_V^\tau(\boldsymbol{b}^T\boldsymbol{y}_i) \geq b\}$$

and denote by n_b the number of data outside this set. The next result establishes the consistency of the RPCCA.

Proposition 3. *Suppose the conditions of Lemma 1 hold. Assume* $b \to 0$ *such that*

$$\frac{n_b}{n} \xrightarrow[n\to\infty]{P} 0,$$

and consider the estimated and true pairs of canonical vectors,

$$\left(\widehat{\boldsymbol{a}}_n, \widehat{\boldsymbol{b}}_n\right) = \arg\max_{\boldsymbol{a},\boldsymbol{b}} \widehat{d}_\tau^n(\boldsymbol{a}, \boldsymbol{b}) \text{ and } \quad (\boldsymbol{a}^*, \boldsymbol{b}^*) = \arg\max_{\boldsymbol{a},\boldsymbol{b}} d_\tau(\boldsymbol{a}, \boldsymbol{b}).$$

Further, assume that the maximum $(\boldsymbol{a}^*, \boldsymbol{b}^*)$ *is unique. Then,*

$$(\widehat{\boldsymbol{a}}_n, \widehat{\boldsymbol{b}}_n) \xrightarrow[n\to\infty]{P} (\boldsymbol{a}^*, \boldsymbol{b}^*).$$

Proof. Take $0 < \epsilon, 0 < b$ such that $\epsilon \xrightarrow[n\to\infty]{} 0, b \xrightarrow[n\to\infty]{} 0$ and $\epsilon b^{-1} \xrightarrow[n\to\infty]{} 0$.

By identification, we can assume $\widehat{\boldsymbol{a}}_n^T\boldsymbol{\Sigma}_{11}\widehat{\boldsymbol{a}}_n = \widehat{\boldsymbol{b}}_n^T\boldsymbol{\Sigma}_{22}\widehat{\boldsymbol{b}}_n = 1$. Let us suppose that

$$(\widehat{\boldsymbol{a}}_n, \widehat{\boldsymbol{b}}_n) \xrightarrow[n\to\infty]{P} (\boldsymbol{a}^*, \boldsymbol{b}^*).$$

Hence, there exists a subsequence of $\{(\widehat{a}_n, \widehat{b}_n)\}$ (that will be denoted by $(\widehat{a}_n, \widehat{b}_n)$ to avoid hard notation) and (a_0, b_0) such that $a_0^T \Sigma_{11} a_0 = b_0^T \Sigma_{22} b_0 = 1$, $(a_0, b_0) \neq (a^*, b^*)$ and

$$(\widehat{a}_n, \widehat{b}_n) \longrightarrow (a_0, b_0).$$

Now, applying Lemma 1, we know that

$$\sup |\widehat{f_U^n}(\widehat{a}_n^T x_i) - f_U(\widehat{a}_n^T x_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

$$\sup |\widehat{f_V^n}(\widehat{b}_n^T y_i) - f_V(\widehat{b}_n^T y_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

$$\sup |\widehat{f_{UV}^n}(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i) - f_{UV}(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Thus, for $\tau > 0$,

$$\sup |\widehat{f_U^n}^\tau(\widehat{a}_n^T x_i) - f_U^\tau(\widehat{a}_n^T x_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

$$\sup |\widehat{f_V^n}^\tau(\widehat{b}_n^T y_i) - f_V^\tau(\widehat{b}_n^T y_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

$$\sup |\widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i) - f_{UV}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Hence, for an n large enough,

$$\widehat{f_U^n}^\tau(\widehat{a}_n^T x_i) = f_U^\tau(\widehat{a}_n^T x_i) + \Delta_{1i} = f_U^\tau(a_0^T x_i) + \delta_{1i},$$

$$\widehat{f_V^n}^\tau(\widehat{b}_n^T y_i) = f_V^\tau(\widehat{b}_n^T y_i) + \Delta_{2i} = f_V^\tau(b_0^T y_i) + \delta_{2i},$$

$$\widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i) = f_{UV}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i) + \Delta_{3i} = f_{UV}^\tau(a_0^T x_i, b_0^T y_i) + \delta_{3i}. \tag{10}$$

Here, $|\delta_{1i}|, |\delta_{2i}|, |\delta_{3i}| < \epsilon$. Remark that $\ln\left(\frac{1}{n}\sum_{i=1}^n \widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i)\right)$ can be written as

$$\ln\left(\frac{1}{n}\sum_{i=1}^n \widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i) \frac{\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)}{\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)}\right)$$

$$= \ln\left(\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)\right) + \ln\left(\frac{\frac{1}{n}\sum_{i=1}^n \widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i)}{\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)}\right).$$

Now, applying Equation (10), we obtain

$$\frac{\frac{1}{n}\sum_{i=1}^n \widehat{f_{UV}^n}^\tau(\widehat{a}_n^T x_i, \widehat{b}_n^T y_i)}{\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)} = 1 + \frac{\frac{1}{n}\sum_{i=1}^n \delta_{3i}}{\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i)}.$$

The same can be performed for the two other cases. As $|\delta_{3i}| < \epsilon$, it follows that

$$\frac{1}{n}\sum_{i=1}^n \delta_{3i} \leq \epsilon.$$

On the other hand,

$$\frac{1}{n}\sum_{i=1}^n f_{UV}^\tau(a_0^T x_i, b_0^T y_i) \geq \frac{1}{n}\sum_{i=1}^n I(i \in \chi_b) f_{UV}^\tau(a_0^T x_i, b_0^T y_i) \geq \frac{n - n_b}{n} b.$$

As $\epsilon b^{-1} \to 0$ and $\frac{n_b}{n} \to 0$, we conclude that

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\delta_{3i}}{\frac{1}{n}\sum_{i=1}^{n}f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i,\boldsymbol{b}_0^T\boldsymbol{y}_i)}\to 0.$$

The same can be performed for the two other cases. Hence,

$$\begin{aligned}\widehat{d}_{\tau}^{n}\left(\widehat{\boldsymbol{a}}_n,\widehat{\boldsymbol{b}}_n\right) &= \frac{1}{\tau+1}\ln\Big(\frac{1}{n^2}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)\sum_{i=1}^{n}f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)+\frac{1}{\tau+1}o(1)\\ &\quad+\frac{1}{\tau(\tau+1)}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i,\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)+\frac{1}{\tau(\tau+1)}o(1)\\ &\quad-\frac{1}{\tau}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)-\frac{1}{\tau}o(1)\\ &= \overline{d}_{\tau}^{n}(\boldsymbol{a}_0,\boldsymbol{b}_0)+o(1),\end{aligned}$$

with

$$\begin{aligned}\overline{d}_{\tau}^{n}(\boldsymbol{a}_0,\boldsymbol{b}_0) &= \frac{1}{\tau+1}\ln\Big(\frac{1}{n^2}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)\sum_{i=1}^{n}f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)\\ &\quad+\frac{1}{\tau(\tau+1)}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i,\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)\\ &\quad-\frac{1}{\tau}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big).\end{aligned}$$

Note that $\overline{d}_{\tau}^{n}(\boldsymbol{a}_0,\boldsymbol{b}_0)-d_{\tau}(\boldsymbol{a}_0,\boldsymbol{b}_0)$ is given by

$$\begin{aligned}&\frac{1}{\tau(\tau+1)}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i,\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)-\frac{1}{\tau(\tau+1)}\ln\big(E_{f_{UV}}(f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{X},\boldsymbol{b}_0^T\boldsymbol{Y}))\big)\\ &+\frac{1}{(\tau+1)}\ln\Big(\frac{1}{n^2}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)\sum_{i=1}^{n}I(i\in\chi_b)f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)-\frac{1}{(\tau+1)}\ln\big(E_{f_U}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{X})E_{f_V}(f_V^{\alpha}(\boldsymbol{b}_0^T\boldsymbol{Y}))\big)\\ &-\frac{1}{\tau}\ln\Big(\frac{1}{n}\sum_{i=1}^{n}f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i)f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{y}_i)\Big)+\frac{1}{\tau}\ln\big(E_{f_{UV}}(f_U^{\tau}(\boldsymbol{a}_0^T\boldsymbol{X})f_V^{\tau}(\boldsymbol{b}_0^T\boldsymbol{Y}))\big).\end{aligned}$$

As ln is continuous and applying the Strong Law of Large Numbers, it follows

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n}f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{x}_i,\boldsymbol{b}_0^T\boldsymbol{y}_i)\right)\xrightarrow[n\to\infty]{a.s.}\ln\Big(E_{f_{UV}}(f_{UV}^{\tau}(\boldsymbol{a}_0^T\boldsymbol{X},\boldsymbol{b}_0^T\boldsymbol{Y}))\Big).$$

We can perform this similarly for the two other lines. We conclude that $\overline{d}_{\tau}^{n}(\boldsymbol{a}_0,\boldsymbol{b}_0)\xrightarrow[n\to\infty]{P}$ $d_{\tau}(\boldsymbol{a}_0,\boldsymbol{b}_0)$, and hence $\widehat{d}_{\tau}^{n}(\boldsymbol{a}_0,\boldsymbol{b}_0)\xrightarrow[n\to\infty]{P}d_{\tau}(\boldsymbol{a}_0,\boldsymbol{b}_0)$. On the other hand, $\widehat{d}_{\tau}^{n}(\widehat{\boldsymbol{a}}_n,\widehat{\boldsymbol{b}}_n)\geq\widehat{d}_{\tau}^{n}(\boldsymbol{a}^*,\boldsymbol{b}^*)$ because $(\widehat{\boldsymbol{a}}_n,\widehat{\boldsymbol{b}}_n)$ is the optimum by definition.

Taking limits,

$$d_{\tau}(\boldsymbol{a}_0,\boldsymbol{b}_0)=\lim_{n\to\infty}\widehat{d}_{\tau}^{n}(\widehat{\boldsymbol{a}}_n,\widehat{\boldsymbol{b}}_n)\geq\lim_{n\to\infty}\widehat{d}_{\tau}^{n}(\boldsymbol{a}^*,\boldsymbol{b}^*)=d_{\tau}(\boldsymbol{a}^*,\boldsymbol{b}^*).$$

However, $d_{\tau}(\boldsymbol{a}^*,\boldsymbol{b}^*)\geq d_{\tau}(\boldsymbol{a}_0,\boldsymbol{b}_0)$ because $(\boldsymbol{a}^*,\boldsymbol{b}^*)$ is the optimum. Hence, as $(\boldsymbol{a}^*,\boldsymbol{b}^*)$ is the only maximum, we conclude that

$$(\boldsymbol{a}^*,\boldsymbol{b}^*)=(\boldsymbol{a}_0,\boldsymbol{b}_0),$$

a contradiction. □

4. Robustness

To motivate the inherent robustness property of the RPCCA procedure, we examine the behavior of the estimated divergence in Equation (4) for small values of the tuning parameter. The presented heuristic argument was first discussed in [6] for the density power divergence generalization of ICCA. Consider the estimated RP

$$\begin{aligned}\widehat{d}^n_\tau(\boldsymbol{a},\boldsymbol{b}) \;:=\; & \frac{1}{\tau+1}\ln\left[\left(\frac{1}{n}\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\right)\left(\frac{1}{n}\sum_{i=1}^n \widehat{f^n_V}^\tau(v_i)\right)\right] - \frac{1}{\tau}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\right] \\ & + \frac{1}{\tau(\tau+1)}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\right].\end{aligned}$$

and let the tuning parameter be $\tau \downarrow 0$. Taking limits in the estimated divergence defined in Equation (4), the first term vanishes, and therefore

$$\widehat{d}^n_\tau(\boldsymbol{a},\boldsymbol{b}) \approx -\frac{1}{\tau}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\right] + \frac{1}{\tau(\tau+1)}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\right].$$

We first study the limiting behavior of the first term,

$$l_\tau := -\frac{1}{\tau}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\right].$$

For $\tau \downarrow 0$, this term is a limit of indeterminate form $(0/0)$. Applying L'Hôpital's rule, we obtain

$$l_\tau \approx \frac{\frac{1}{n}\left[\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\ln\left(\widehat{f^n_U}(u_i)\right) + \widehat{f^n_V}^\tau(v_i)\widehat{f^n_U}^\tau(u_i)\ln\left(\widehat{f^n_V}(v_i)\right)\right]}{\frac{1}{n}\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)}.$$

Now, the denominator tends to 1 when $\tau \downarrow 0$ so that

$$l_\tau \approx \frac{1}{n}\left[\sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\ln\left(\widehat{f^n_U}(u_i)\right) + \widehat{f^n_V}^\tau(v_i)\widehat{f^n_U}^\tau(u_i)\ln\left(\widehat{f^n_V}(v_i)\right)\right].$$

Similarly, consider

$$m_\tau := \frac{1}{\tau(\tau+1)}\ln\left[\frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\right]$$

and consider its L'Hôpital approximation given by

$$m_\tau \approx \frac{1}{2\tau+1}\frac{\frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\ln\widehat{f^n_{UV}}(u_i,v_i)}{\frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)} \approx \frac{1}{n}\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\ln\widehat{f^n_{UV}}(u_i,v_i).$$

Consequently,

$$\widehat{d}^n_\tau(\boldsymbol{a},\boldsymbol{b}) \approx \frac{1}{n}\left[\sum_{i=1}^n \widehat{f^n_{UV}}^\tau(u_i,v_i)\ln\widehat{f^n_{UV}}(u_i,v_i) - \sum_{i=1}^n \widehat{f^n_U}^\tau(u_i)\widehat{f^n_V}^\tau(v_i)\ln\left(\widehat{f^n_U}(u_i)\widehat{f^n_V}(v_i)\right)\right].$$

Note that this approximation is valid for τ closed to 0, but, for $\tau = 0$,

$$d_0(\boldsymbol{a},\boldsymbol{b}) = \lim_{\tau\downarrow 0}\tilde{d}_\tau(\boldsymbol{a},\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^n \ln\left(\frac{\widehat{f^n_{UV}}(u_i,v_i)}{\widehat{f^n_U}(u_i)\widehat{f^n_V}(v_i)}\right) = \frac{1}{n}\left(\sum_{i=1}^n \ln\widehat{f^n_{UV}}(u_i,v_i) - \sum_{i=1}^n \ln\widehat{f^n_U}(u_i)\widehat{f^n_V}(v_i)\right).$$

This implies that $\widehat{d}^n_\tau(f_U \times f_V, f_{UV})$ can be seen as a weighted value of the empirical Kullback–Leibler divergence and the weights depend on $\widehat{f^n_{UV}}^\tau(u_i, v_i)$, $\widehat{f^n_U}^\tau(u_i)$ and $\widehat{f^n_V}^\tau(v_i)$. Therefore, if the observations $\boldsymbol{x}_i, \boldsymbol{y}_i$ or both are outliers, the corresponding density estimations would decrease, so the corresponding weights would not be considered as important as other data on the estimated distance, thus making Renyi's pseudodistance more robust to outliers than the Kullback–Leibler.

5. Testing to Determine the Number of Pairs

In this section, a dimension reduction algorithm is described for determining the number of significant pairs of canonical vectors: the non-parametric sequential test [41,42].

In the classical approach of CCA, the maximum number of pairs $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ is determined by the greatest index j such that $(\boldsymbol{a}_j, \boldsymbol{b}_j)$ is the first pair satisfying $\rho(\boldsymbol{a}_j^T\boldsymbol{X}, \boldsymbol{b}_j^T\boldsymbol{Y}) = 0$. That is, the CCA should run until the best estimated pair leads to linear independence. A natural extension for the RPCCA formulation is then replicated in the CCA dimension reduction algorithm, but using the RP divergence as a measure of dependence.

Let us denote by d^i_τ the maximum value achieved at the i-th iteration,

$$d^i_\tau = \max_{\boldsymbol{a}_i, \boldsymbol{b}_i} d_\tau(\boldsymbol{a}_i, \boldsymbol{b}_i),\ i = 1, \ldots, l = \min(q, p),$$

such that $\boldsymbol{a}_i^T\boldsymbol{\Sigma}_{11}\boldsymbol{a}_i = \boldsymbol{b}_i^T\boldsymbol{\Sigma}_{22}\boldsymbol{b}_i = 1$ and $\boldsymbol{a}_j^T\boldsymbol{\Sigma}_{11}\boldsymbol{a}_i = \boldsymbol{b}_j^T\boldsymbol{\Sigma}_{22}\boldsymbol{b}_i = 0$. The sequence of maximums is decreasing and lower-bounded by 0, indicating independence between the estimated canonical variables, $d^1_\tau \geq d^2_\tau \geq \ldots \geq d^l_\tau \geq 0$. Then, a stopping criterion for the maximum number of canonical correlations is naturally determined by the testing problem

$$H_0 : d^i_\tau = 0 \text{ vs } H_1 : d^i_\tau > 0,\ i = 1, \ldots, l.$$

If H_0 is not rejected, then all subsequent canonical variables from the i-th onward are not significantly related. Otherwise, the relation is significant, and the maximum number of significant canonical correlations is at least i. It is difficult to obtain the exact sample distribution of d^i_τ, but a non-parametric permutation test can be applied, as proposed in [24], for estimating the p-value of the test. Let us explain this procedure with some detail. Suppose there is a relationship between $\boldsymbol{a}_i^T\boldsymbol{X}$ and $\boldsymbol{b}_i^T\boldsymbol{Y}$ for some vectors $\boldsymbol{a}_i, \boldsymbol{b}_i$, i.e., H_0 does not hold. This means that there exists a function f such that

$$f(\boldsymbol{a}_i^T\boldsymbol{X}) \approx \boldsymbol{b}_i^T\boldsymbol{Y}.$$

Our procedure will estimate vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ and will consider some (near!) vectors $\widehat{\boldsymbol{a}}_i$ and $\widehat{\boldsymbol{b}}_i$, respectively. Consequently, we expect that for the sample $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$, we will obtain

$$f(\widehat{\boldsymbol{a}}_i^T\boldsymbol{X}_j) \approx \widehat{\boldsymbol{b}}_i^T\boldsymbol{Y}_j, \quad j = 1, \ldots, n.$$

This will translate in a large value of $d^n_\tau(\widehat{\boldsymbol{a}_i}^T\boldsymbol{X}, \widehat{\boldsymbol{b}_i}^T\boldsymbol{Y})$ and, consequently, the corresponding estimation $\widehat{d}^n_\tau(\widehat{\boldsymbol{a}_i}^T\boldsymbol{X}, \widehat{\boldsymbol{b}_i}^T\boldsymbol{Y})$. Now, if we consider a permutation of the data corresponding to $\boldsymbol{X}$ but maintaining the order for the data corresponding to $\boldsymbol{Y}$, any possible relationship is destroyed because the data corresponding to $\boldsymbol{X}_i$ do not correspond to individual i in the sample, so that they have nothing to do with $\boldsymbol{Y}_i$. In other words, if we denote the reordered sample for $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ by $(\boldsymbol{X}^*_1, \ldots, \boldsymbol{X}^*_n)$, it follows that for any $\boldsymbol{c}, \boldsymbol{d}$, then $\widehat{d}^n_\tau(\boldsymbol{c}^T\boldsymbol{X}^*, \boldsymbol{d}^T\boldsymbol{Y}) \approx 0$ showing independence. Consequently, when the procedure looks for some vectors $\widehat{\boldsymbol{a}^*_i}, \widehat{\boldsymbol{b}^*_i}$ s.t.

$$\widehat{d}^n_\tau(\widehat{\boldsymbol{a}^*_i}^T\boldsymbol{X}^*, \widehat{\boldsymbol{b}^*_i}^T\boldsymbol{Y}) = \max \widehat{d}^n_\tau\left(\boldsymbol{a}^T\boldsymbol{X}^*, \boldsymbol{b}^T\boldsymbol{Y}\right),$$

these values are not expected to model a strong relation (because it does not exist), so that we expect $\widehat{d}^n_\tau(\widehat{\boldsymbol{a}_i^*}^T\boldsymbol{X}^*, \widehat{\boldsymbol{b}_i^*}^T\boldsymbol{Y}) \approx 0$. Hence, if H_0 does not hold and a relationship between the canonical variables exists, we expect that for (almost) any permutation

$$\widehat{d}^n_\tau(\widehat{\boldsymbol{a}}_i{}^T\boldsymbol{X}, \widehat{\boldsymbol{b}}_i{}^T\boldsymbol{Y}) > \widehat{d}^n_\tau(\widehat{\boldsymbol{a}_i^*}^T\boldsymbol{X}^*, \widehat{\boldsymbol{b}_i^*}^T\boldsymbol{Y}).$$

On the other hand, if H_0 holds, then there is independence between $\boldsymbol{c}^T\boldsymbol{X}$ and $\boldsymbol{d}^T\boldsymbol{Y}$ for any $\boldsymbol{c}, \boldsymbol{d}$. Consequently, for the best possible estimated vectors $\widehat{\boldsymbol{a}}_i, \widehat{\boldsymbol{b}}_i$, we will obtain

$$\widehat{d}^n_\tau(\widehat{\boldsymbol{a}}_i{}^T\boldsymbol{X}, \widehat{\boldsymbol{b}}_i{}^T\boldsymbol{Y}) \approx 0.$$

When considering a permutation of the values corresponding to $\boldsymbol{X}$, independence will arise again, and hence, in this case, we expect

$$\widehat{d}^n_\tau(\widehat{\boldsymbol{a}}_i{}^T\boldsymbol{X}, \widehat{\boldsymbol{b}}_i{}^T\boldsymbol{Y}) \approx \widehat{d}^n_\tau(\widehat{\boldsymbol{a}_i^*}^T\boldsymbol{X}^*, \widehat{\boldsymbol{b}_i^*}^T\boldsymbol{Y}).$$

Of course, the number of possible permutations is $n!$, and this is not affordable for large values of n. Hence, we are going to consider just a subset of randomly chosen permutations. Then, if $\widehat{d}^{i,w}_\tau$ denotes the value of the index corresponding to the w-th randomly permuted sample, the estimated p-value of the test is given by

$$\frac{1}{R}\sum_{w=1}^{R} I_{\left[\widehat{d}^{i,w}_\tau > \widehat{d}^i_\tau\right]},$$

where R denotes the number of permutations considered. Yin [12] used $R = 1000$ for a permutation test for ICCA. If the p-value is smaller than a certain significance level, then the null hypothesis $d^i_\tau = 0$ should be rejected implying a significant relationship for the i-th canonical variables, and the process should be repeated for $i+1$. Conversely, if the null hypothesis is not rejected, then we should assume that the canonical variables are independent and conclude that there are only i estimated canonical variables exhibiting significant relationships. More details about this dimension reduction method can be seen in [24].

6. Simulation Study

6.1. Computational Methods

Consider $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\mathbb{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ as $p \times n$ and $q \times n$ matrices with n observations of the random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The estimation of the i-th pair of canonical vectors $\widehat{\boldsymbol{a}}^\tau_i, \widehat{\boldsymbol{b}}^\tau_i$ based on the RP with tuning parameter τ is computed through the constrained maximization problem

$$\begin{aligned}\left(\widehat{\boldsymbol{a}}^\tau_n, \widehat{\boldsymbol{b}}^\tau_n\right) &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \widehat{d}^i_\tau(\boldsymbol{a}, \boldsymbol{b}),\\ \text{s.t. } (\boldsymbol{a}^\tau_i)^T\widehat{\boldsymbol{\Sigma}}_X\boldsymbol{a}^\tau_i &= 1 \text{ and } (\boldsymbol{b}^\tau_i)^T\widehat{\boldsymbol{\Sigma}}_Y\boldsymbol{b}^\tau_n = 1,\\ (\boldsymbol{a}^\tau_i)^T\widehat{\boldsymbol{\Sigma}}_X\boldsymbol{b}^\tau_j &= 0 \text{ and } (\boldsymbol{b}^\tau_i)^T\widehat{\boldsymbol{\Sigma}}_Y\boldsymbol{b}^\tau_j = 0,\ j = 1, \ldots, i-1\end{aligned} \tag{11}$$

where $\widehat{\boldsymbol{\Sigma}}_X$ and $\widehat{\boldsymbol{\Sigma}}_Y$ are the empirical estimators of the variance–covariance matrices of the multidimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

The optimization problem constraints can be simplified by scaling the sample matrices to have zero mean and unit variance as follows:

$$\begin{aligned}\widetilde{\mathbb{X}} &= \widehat{\boldsymbol{\Sigma}}_X^{-1/2}(\mathbb{X} - \overline{\mathbb{X}})\\ \widetilde{\mathbb{Y}} &= \widehat{\boldsymbol{\Sigma}}_{22}^{-1/2}(\mathbb{Y} - \overline{\mathbb{Y}}),\end{aligned} \tag{12}$$

where $\overline{\mathbb{X}}$ and $\overline{\mathbb{Y}}$ denote the corresponding sample mean vectors. From Proposition 1, the RPCCA is invariant under such linear transformations, and, consequently, the problem constraints are transformed into

$$\begin{aligned}(\boldsymbol{a}_i^\tau)^T\boldsymbol{a}_i^\tau = 1 \text{ and } (\boldsymbol{b}_i^\tau)^T\boldsymbol{b}_n^\tau = 1,\\ (\boldsymbol{a}_i^\tau)^T\boldsymbol{b}_j^\tau = 0 \text{ and } (\boldsymbol{b}_i^\tau)^T\boldsymbol{b}_j^\tau = 0,\ j = 1,\ldots,i-1.\end{aligned} \tag{13}$$

For empirical covariance matrices, it may appear as disease degeneration resulting in uninvertible matrices. In those cases, we can skip the scaling transform and apply the estimation algorithm under the original restrictions.

From the transformed canonical vectors, $\widetilde{\boldsymbol{a}}_i$ and $\widetilde{\boldsymbol{b}}_i$, the estimated canonical vectors in the original space can be easily recovered as $\widehat{\boldsymbol{a}}_i = \widehat{\boldsymbol{\Sigma}}_X^{-1/2}\widetilde{\boldsymbol{a}}_i$ and $\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{\Sigma}}_Y^{-1/2}\widetilde{\boldsymbol{b}}_i$.

Here, the constrained optimization is carried out iteratively using the non-linear constrained optimizer *optimize* from the *scipy* package in *Python*, which implements a Sequential Quadratic Programming (SQP) method. The source code for the implementation is publicly available on https://github.com/MariaJaenada/Robust-Canonical-Correlations (Github) (accessed on 29 January 2023).

6.2. Monte Carlo Simulation

We empirically examine the robustness of the RPCCA method through a Monte Carlo simulation. We consider a pair of random vectors, $\boldsymbol{X} = (X_1,\ldots,X_8)$ and $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$, whose components satisfy a linear and a non-linear relationship of the form:

$$Y_1 = (2X_1 + X_2 + X_3)^2 \quad \text{and} \quad Y_2 = X_2 - X_3. \tag{14}$$

The rest of the variables are independent and they are defined as follows: X_1, X_2, X_6, X_7 and X_8. They are standard normal variables. X_3 comes from a chi-square distribution with 7 degrees of freedom, X_4 follows a t-Student distribution with 5 degrees of freedom and X_5 comes from a Fisher–Snedecor distribution with 3 and 12 degrees of freedom, respectively. Finally, Y_3 comes from a t-Student with 9 degrees of freedom.

The true underlying canonical vectors are then $\boldsymbol{a}_1 = (0,\ 1, -1,\ 0,\ 0,\ 0,\ 0,\ 0)$, $\boldsymbol{b}_1 = (0,\ 1,\ 0)$ and $\boldsymbol{a}_2 = (2,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0)$, $\boldsymbol{b}_2 = (1,\ 0,\ 0)$. Note that they are orthogonal, and so are the related variables Y_1 and Y_2. Although in the procedure we compute unit-norm vectors, we have considered the description vectors with natural coefficients as they look easier to understand. We named the first canonical vector $\boldsymbol{a}_1$ because we empirically detected that the linear relationship is first captured.

We generate a random sample of the pairs $\boldsymbol{X}$ and $\boldsymbol{Y}$ of size $n = 100$, and we estimate the pairs of canonical vectors $\widehat{\boldsymbol{a}}_i$ and $\widehat{\boldsymbol{b}}_i$, $i = 1, 2$, such that the random variables $U_i = \boldsymbol{a}_i^T\boldsymbol{X}$ and $V_i = \boldsymbol{b}_i^T\boldsymbol{Y}$ are functionally interrelated. To examine the performance of the RPCCA method under contamination, we randomly switch the functional relationships in Equation (14) for an ε% of the observations, with $\varepsilon = 5,\ 10,\ 15$ and 20 denoting the contamination proportion. That is, for a random ε% of the observations, the values of Y_1 and Y_2 are exchanged, generating orthogonal outliers; the functions defining the Y_2 and Y_1 are orthogonal to each other. Therefore, this contamination will worsen both relationships at the same time in orthogonal directions. We repeat the simulations over $R = 500$ replications and compute averages of the following performance measures: We quantify the accuracy of the estimates with the absolute correlations between the estimated and true canonical variables, $|\rho(\boldsymbol{a}_i, \widehat{\boldsymbol{a}}_i)| = |\rho(\boldsymbol{a}_i^T\boldsymbol{X}, \widehat{\boldsymbol{a}}_i^T\boldsymbol{X})|$ and $|\rho(\boldsymbol{b}_i, \widehat{\boldsymbol{b}}_i)| = |\rho(\boldsymbol{b}_i^T\boldsymbol{Y}, \widehat{\boldsymbol{b}}_i^T\boldsymbol{Y})|$. Additionally, to evaluate the robustness of the method, we compute the L_2-norm between the canonical vectors fitted under uncontaminated and contaminated data, $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{a}}^c$,

$$L_2(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{a}}^c) = ||\widehat{\boldsymbol{a}} - \widehat{\boldsymbol{a}}^c||_2$$

as well as the projection of $\widehat{\boldsymbol{a}}^c$ into the orthogonal subspace spanned by the uncontaminated estimate, $\widehat{\boldsymbol{a}}$,

$$P_2(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{a}}^c) = ||(\boldsymbol{I} - \widehat{\boldsymbol{a}}\widehat{\boldsymbol{a}}^T)\widehat{\boldsymbol{a}}_c^T||_2.$$

The distance measures $L_2(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{a}}^c)$ and $P_2(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{a}}^c)$ are smaller the more stable the estimate is, implying that the estimates are not largely affected by the contamination; hence the corresponding method is more robust. Summarizing, the correlations between true and estimated canonical variables $\rho(\cdot, \cdot)$ aim to represent the accuracy of the method, whereas the distance measures between estimated canonical vectors for pure data and for contaminated data, L_2 and P_2, aim to represent the robustness of the method.

Tables 1 and 2 present all performance measures for the RPCCA method over a grid of tuning parameters ranging from 0 (corresponding to ICCA) to 0.8. All methods perform suitably well in terms of accuracy, achieving high absolute correlations between true and estimated canonical variables, even under contaminated scenarios. However, the linear relationship in the first component is captured worse by the ICCA in the presence of contamination, as shown by the lower absolute correlations between the canonical variables, $\rho(\widehat{\boldsymbol{a}}_1^c, \widehat{\boldsymbol{b}}_1^c)$. Moreover, the RPCCA method with positive values of the tuning parameter produces more stable estimations of the canonical vectors, having smaller P_2 and L_2 distances between the uncontaminated and contaminated estimated canonical vectors in both components, $(\widehat{\boldsymbol{b}_1}, \widehat{\boldsymbol{b}}_1^c)$ and $(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$, thus demonstrating the advantage in terms of robustness. Although the differences in performance are not impressive, the gain in robustness with very little loss of accuracy with respect to the ICCA makes the RPCCA very attractive.

On the other hand, if the underlying relationship is easily identified, the proposed robust RPCCA performs as good as the ICCA under pure data and outperforms the ICCA in the presence of contamination (Table 1). However, for $\tau > 0$, the loss in accuracy in the relationship identification under pure data would be unavoidable (although not very significant); hence, the tuning parameter should be chosen sufficiently close to zero (from the literature, less than 1) to provide an adequate compromise between efficiency loss and robustness gain. Moderate values of the tuning parameter, around 0.3, offer the best compromise producing canonical estimators that are robust against data contamination with a small loss of efficiency with respect to the ICCA in the absence of contamination.

6.3. Real Data Application

We finally illustrate the applicability of our method with real-life data on the heredity of head shape in men. For such a purpose, we use a well-known dataset from Frets [43] that collects the head length and head breadth for the first and second sons for $n = 25$ families. Then, the first and second set of variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, have the dimension 2 and represent the head length and head breadth of the corresponding son. From the dataset, we want to analyze whether there is a relationship between the head shape among male offspring. The data have been widely used in the literature, and Mardia et al. [2] and Yin [12] analyzed the canonical correlations between the first and second sons' head shapes using CCA and ICCA, respectively. In their analyses, they found one significant pair of canonical variables with a strong linear relationship. Figure 2 shows the plots of the first (left) and second (right) pair of canonical variables for the head data estimated by RPCCA with $\tau = 0$ (top) and $\tau = 0.5$ (bottom). As shown, both methods coincide on the first pair of canonical variables (estimate the same observations for the first pair of canonical variables), x_1 and y_1, having linear correlation coefficients of $\rho = 78.67\%$ ($\tau = 0$) and $\rho = 76.86\%$ ($\tau = 0.5$) as illustrated on the corresponding plots. For the second pair of canonical variables, none of the methods find any clear functional relationship between linear combinations of the variables, and the two procedures considered estimate very different canonical variables without a clear functional relationship between them (as shown in Figure 2). Thus, we also conclude that there is only one pair of canonical variables.

Table 1. RPCCA error measures for the first canonical vector under different values of the tuning parameter τ.

τ	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
Pure data									
$\rho(\widehat{a}_1, \widehat{b}_1)$	0.92878	0.96111	0.95505	0.97044	0.97099	0.98082	0.97090	0.98256	0.98606
$\rho(a_1, \widehat{a}_1)$	0.99908	0.99907	0.99893	0.99877	0.99856	0.99831	0.99799	0.99767	0.99728
$\rho(b_1, \widehat{b}_1)$	0.99946	0.99951	0.99933	0.99938	0.99936	0.99931	0.99906	0.99916	0.99899
5% contamination									
$\rho(\widehat{a}_1^c, \widehat{b}_1^c)$	0.65796	0.68689	0.71486	0.75556	0.76384	0.80646	0.80876	0.81990	0.81201
$\rho(a_1, \widehat{a}_1^c)$	0.99801	0.99815	0.99805	0.99786	0.99749	0.99645	0.99353	0.98368	0.96927
$\rho(b_1, \widehat{b}_1^c)$	0.99548	0.99571	0.99531	0.99538	0.99461	0.99345	0.99084	0.97936	0.96222
$P_2(\widehat{a}_1, \widehat{a}_1^c)$	0.41125	0.38246	0.36111	0.31247	0.30856	0.25109	0.25272	0.22822	0.23050
$P_2(\widehat{b}_1, \widehat{b}_1^c)$	0.35685	0.32477	0.31442	0.27277	0.26949	0.22345	0.22962	0.20914	0.21654
$L_2(\widehat{a}_1, \widehat{a}_1^c)$	0.56677	0.52712	0.49669	0.42880	0.42197	0.33991	0.34109	0.30249	0.30241
$L_2(\widehat{b}_1, \widehat{b}_1^c)$	0.41496	0.37628	0.36696	0.31681	0.31561	0.26241	0.27298	0.24999	0.26391
10% contamination									
$\rho(\widehat{a}_1^c, \widehat{b}_1^c)$	0.41963	0.43878	0.46604	0.49193	0.53443	0.56155	0.57944	0.59613	0.60565
$\rho(a_1, \widehat{a}_1^c)$	0.99698	0.99714	0.99712	0.99686	0.99563	0.99225	0.98450	0.96631	0.95789
$\rho(b_1, \widehat{b}_1^c)$	0.99054	0.99018	0.98974	0.98968	0.98719	0.98401	0.97274	0.95167	0.93961
$P_2(\widehat{a}_1, \widehat{a}_1^c)$	0.68137	0.65989	0.63978	0.60585	0.57014	0.53623	0.51301	0.48214	0.47437
$P_2(\widehat{b}_1, \widehat{b}_1^c)$	0.56499	0.54370	0.53006	0.51765	0.48283	0.46007	0.44608	0.43238	0.43151
$L_2(\widehat{a}_1, \widehat{a}_1^c)$	0.94237	0.91248	0.88343	0.83560	0.78292	0.73427	0.69877	0.64896	0.63289
$L_2(\widehat{b}_1, \widehat{b}_1^c)$	0.63783	0.61685	0.60406	0.59298	0.55556	0.53199	0.52065	0.51211	0.51843
15% contamination									
$\rho(\widehat{a}_1^c, \widehat{b}_1^c)$	0.26726	0.29776	0.32650	0.34987	0.40472	0.42547	0.43099	0.43524	0.45246
$\rho(\widehat{a}_1, \widehat{a}_1^c)$	0.99662	0.99682	0.99670	0.99631	0.99405	0.98763	0.97539	0.95782	0.93472
$\rho(\widehat{b}_1, \widehat{b}_1^c)$	0.98740	0.98702	0.98627	0.98541	0.98364	0.97541	0.95686	0.93536	0.91078
$P_2(\widehat{a}_1, \widehat{a}_1^c)$	0.83673	0.81093	0.78650	0.75948	0.70320	0.68752	0.68190	0.67188	0.64529
$P_2(\widehat{b}_1, \widehat{b}_1^c)$	0.67519	0.65622	0.64107	0.62355	0.58194	0.57677	0.58054	0.58166	0.57108
$L_2(\widehat{a}_1, \widehat{a}_1^c)$	1.16074	1.12367	1.08848	1.04925	0.96766	0.94157	0.92937	0.90983	0.86243
$L_2(\widehat{b}_1, \widehat{b}_1^c)$	0.75424	0.73561	0.72109	0.70581	0.66064	0.66160	0.67484	0.68431	0.68025
20% contamination									
$\rho(\widehat{a}_1^c, \widehat{b}_1^c)$	0.19852	0.21736	0.22524	0.23800	0.27663	0.28001	0.30514	0.32119	0.31858
$\rho(\widehat{a}_1, \widehat{a}_1^c)$	0.99661	0.99683	0.99652	0.99593	0.99429	0.99003	0.96500	0.94362	0.90237
$\rho(\widehat{b}_1, \widehat{b}_1^c)$	0.98527	0.98503	0.98336	0.98251	0.97950	0.97282	0.94210	0.91641	0.86912
$P_2(\widehat{a}_1, \widehat{a}_1^c)$	0.90198	0.90524	0.89112	0.89756	0.85384	0.85092	0.82023	0.80367	0.79283
$P_2(\widehat{b}_1, \widehat{b}_1^c)$	0.72434	0.71723	0.71807	0.72166	0.69817	0.70049	0.69510	0.69435	0.70935
$L_2(\widehat{a}_1, \widehat{a}_1^c)$	0.45211	0.45326	0.44686	0.45094	0.42851	0.42746	0.41497	0.41101	0.41076
$L_2(\widehat{b}_1, \widehat{b}_1^c)$	0.61339	0.61723	0.60407	0.60759	0.59650	0.59591	0.58441	0.59536	0.59885

Additionally, to illustrate the advantage of our method in terms of robustness (with a small loss of deficiency), we contaminate a single observation (obs. 24) in both vector variables, generating an outlying observation. Then, we apply RPCCA at $\tau = 0$ (corresponding to ICCA) and $\tau = 0.5$ with the uncontaminated and the contaminated data. Table 3 presents P_2 and N_2 distances between the first pair of canonical vectors (identifying the linear relationship) estimated under uncontaminated and contaminated data, with only one outlying observation. Because the sample size is small, an outlying observation heavily influences the ICCA estimation, whereas the RPCCA method with $\tau = 0.5$ shows a great stability in the canonical vector estimation. These results illustrate the advantage of the RPCCA in

real-life applications, producing robust estimates of the canonical variables with a small loss of efficiency with respect to the ICCA estimation in the absence of data contamination.

Table 2. RPCCA error measures for the second canonical vector under different values of the tuning parameter τ.

τ	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
Pure data									
$\rho(\widehat{\boldsymbol{a}}_2, \widehat{\boldsymbol{b}}_2)$	0.22829	0.19656	0.20033	0.18556	0.18386	0.17334	0.17650	0.16407	0.15426
$\rho(\boldsymbol{a}_2, \widehat{\boldsymbol{a}}_2)$	0.99687	0.99704	0.99514	0.99387	0.98976	0.96917	0.94815	0.93493	0.88471
$\rho(\boldsymbol{b}_2, \widehat{\boldsymbol{b}}_2)$	0.99464	0.99490	0.99122	0.98980	0.98375	0.96336	0.93706	0.91885	0.85931
5% contamination									
$\rho(\widehat{\boldsymbol{a}}_2^c, \widehat{\boldsymbol{b}}_2^c)$	0.30303	0.29521	0.27510	0.24815	0.24522	0.21903	0.22096	0.19770	0.20928
$\rho(\boldsymbol{a}_2, \widehat{\boldsymbol{a}}_2^c)$	0.91167	0.92137	0.91977	0.91112	0.91556	0.90693	0.88127	0.84600	0.82650
$\rho(\boldsymbol{b}_2, \widehat{\boldsymbol{b}}_2^c)$	0.88049	0.89166	0.88662	0.88904	0.89241	0.88801	0.87192	0.83197	0.80588
$P_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.41815	0.38347	0.36114	0.32231	0.31958	0.29261	0.33041	0.32611	0.36974
$P_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.44509	0.41272	0.40340	0.36543	0.36662	0.33627	0.36107	0.34812	0.36182
$L_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.56221	0.51686	0.48609	0.43221	0.42597	0.38524	0.42792	0.41586	0.46555
$L_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.54952	0.50631	0.48727	0.43321	0.43065	0.38858	0.41934	0.39996	0.41901
10% contamination									
$\rho(\widehat{\boldsymbol{a}}_2^c, \widehat{\boldsymbol{b}}_2^c)$	0.31085	0.29651	0.28563	0.28324	0.25836	0.24660	0.25507	0.23689	0.22360
$\rho(\boldsymbol{a}_2, \widehat{\boldsymbol{a}}_2^c)$	0.76633	0.76784	0.77030	0.77828	0.75962	0.77061	0.78303	0.75727	0.71201
$\rho(\boldsymbol{b}_2, \widehat{\boldsymbol{a}}_2^c)$	0.77035	0.75135	0.75242	0.75593	0.74394	0.75508	0.74693	0.74538	0.70453
$P_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.69844	0.67264	0.64760	0.62480	0.58957	0.56764	0.57146	0.56234	0.60946
$P_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.70333	0.68232	0.65851	0.63732	0.60797	0.57351	0.55445	0.54736	0.53981
$L_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.94522	0.91180	0.87884	0.84254	0.79453	0.76052	0.75866	0.73627	0.79154
$L_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.87344	0.84137	0.80061	0.76733	0.72179	0.67148	0.64907	0.64158	0.63347
15% contamination									
$\rho(\widehat{\boldsymbol{a}}_2^c, \widehat{\boldsymbol{b}}_2^c)$	0.29041	0.26604	0.25341	0.24867	0.22370	0.22678	0.22668	0.21698	0.20791
$\rho(\widehat{\boldsymbol{a}}_2, \widehat{\boldsymbol{b}}_2^c)$	0.62670	0.62511	0.62775	0.63798	0.63931	0.67347	0.66306	0.63881	0.61437
$\rho(\widehat{\boldsymbol{b}}_2, \widehat{\boldsymbol{b}}_2^c)$	0.67113	0.65281	0.63298	0.64345	0.64183	0.66804	0.66294	0.66655	0.64393
$P_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.86840	0.83480	0.80722	0.78144	0.73384	0.71580	0.71970	0.73208	0.73667
$P_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.83973	0.81076	0.78283	0.74695	0.69787	0.68081	0.65889	0.63941	0.61576
$L_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	1.17901	1.13476	1.09703	1.06197	0.99294	0.96654	0.96359	0.97065	0.96509
$L_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	1.04721	1.00181	0.95442	0.89936	0.83020	0.80219	0.77413	0.74676	0.72249
20% contamination									
$\rho(\widehat{\boldsymbol{a}}_2^c, \widehat{\boldsymbol{b}}_2^c)$	0.23981	0.23457	0.22347	0.22122	0.21361	0.20732	0.20959	0.19139	0.20603
$\rho(\widehat{\boldsymbol{a}}_2, \widehat{\boldsymbol{b}}_2^c)$	0.51483	0.51794	0.52276	0.53520	0.54956	0.53813	0.56693	0.54726	0.56153
$\rho(\widehat{\boldsymbol{b}}_2, \widehat{\boldsymbol{b}}_2^c)$	0.62369	0.59115	0.56066	0.56134	0.58636	0.58110	0.60681	0.61363	0.63613
$P_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.94170	0.92377	0.91740	0.90923	0.86834	0.86758	0.83873	0.82330	0.83708
$P_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.89672	0.87587	0.85688	0.82781	0.78143	0.75837	0.72490	0.70158	0.69711
$L_2(\widehat{\boldsymbol{a}_2}, \widehat{\boldsymbol{a}}_2^c)$	0.47541	0.46413	0.46168	0.45820	0.43798	0.43988	0.43203	0.42330	0.43539
$L_2(\widehat{\boldsymbol{b}_2}, \widehat{\boldsymbol{b}}_2^c)$	0.76644	0.73624	0.72641	0.69977	0.67411	0.63579	0.61011	0.61020	0.59094

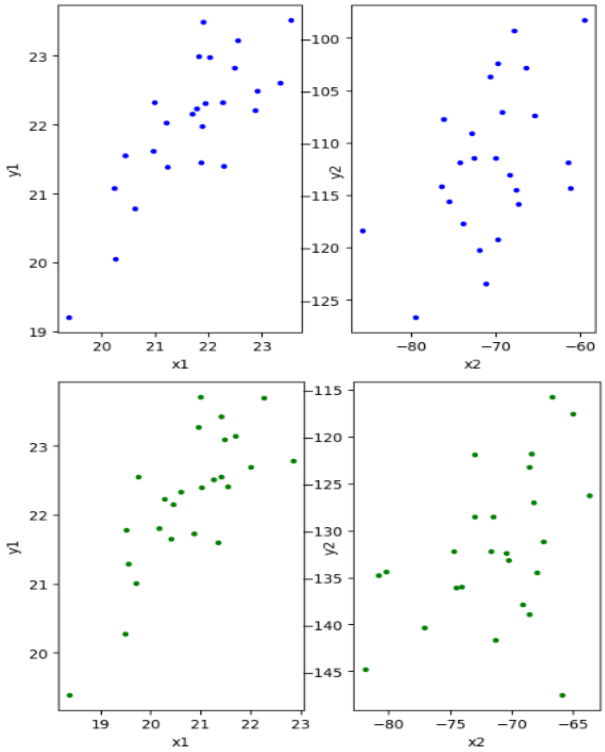

Figure 2. Pairs of canonical variables obtained from RPCCA with $\tau = 0$ (**top**) and $\tau = 0.5$ (**bottom**) for the head dataset.

Table 3. P_2 and N_2 distances between the estimated canonical vectors under uncontaminated and contaminated data.

τ	**0**	**0.5**
$P_2(\widehat{a_1}, \widehat{a}_1^c)$	0.222	0.064
$P_2(\widehat{b_1}, \widehat{b}_1^c)$	0.131	0.059
$L_2(\widehat{a_1}, \widehat{a}_1^c)$	0.224	0.064
$L_2(\widehat{b_1}, \widehat{b}_1^c)$	1.995	0.059

7. Conclusions

We have presented a robust generalization of the ICCA based on RP for identifying linear and non-linear relationships between two sets of variables. We have derived sample versions for estimating the canonical vectors in practice, and we have demonstrated the consistency of such estimators. Further, the robustness advantage of the RPCCA has been examined theoretically and empirically, concluding that the proposed RPCCA offers an appealing alternative to ICCA, competitive in terms of estimation accuracy and more robust against data contamination. The method manages to detect hidden functional relationships between linear combinations of the variables and suitably approximates the true underlying relationships, even under contaminated scenarios. Moreover, a permutation test for determining the number of significant pairs of canonical vectors is presented. Since the RPCCA is a parametric family, a data-driven algorithm for determining optimal values of the tuning parameter is a worthwhile pursuit for future research. Also, the methodology presented here can be extended in future works for identifying relationships between more than two sets of variables. The idea is to consider not only two random vectors but k random vectors as considered in [24] and to look for the linear combinations in all of them so that the RP between the marginal distributions and the whole distribution is as large as possible.

Author Contributions: Conceptualization, M.J., P.M., L.P. and K.Z.; methodology, M.J., P.M., L.P. and K.Z.; software, M.J., P.M., L.P. and K.Z.; validation, M.J., P.M., L.P. and K.Z.; formal analysis, M.J., P.M., L.P. and K.Z.; investigation, M.J., P.M., L.P. and K.Z.; resources, M.J., P.M., L.P. and K.Z.; data curation, M.J., P.M., L.P. and K.Z.; writing—original draft preparation, M.J., P.M., L.P. and K.Z.; writing—review and editing, M.J., P.M., L.P. and K.Z.; visualization, M.J., P.M., L.P. and K.Z.; supervision, M.J., P.M., L.P. and K.Z.; project administration, M.J., P.M., L.P. and K.Z.; funding acquisition, M.J., P.M., L.P. and K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Spanish Grants PID2021-124933NB-I00 and FPU/018240.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are very grateful to the referees and associate editor for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The founders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CCA	Canonical Analysis
ICCA	Informational Canonical Analysis
RP	Rényi Pseudodistance
RPCCA	Rényi Pseudodistance Canonical Analysis

References

1. Hotelling, H. Relations between two sets of variables. *Biometrika* **1936**, *28*, 321–377. [CrossRef]
2. Mardia, K.; Kent, J.; Bibby, J. *Multivariate Analysis*; Academic Press: New York, NY, USA, 1979.
3. Rencher, A.C.; Christensen, W.F. *Methods of Multivariate Analysis*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012.
4. Ouali, D.; Chebana, F.; Ouarda, T.B.M.J. Non-linear canonical correlation analysis in regional frequency analysis. *Stoch. Environ. Res. Risk Assess* **2016**, *30*, 449–462. [CrossRef]
5. Cannon, A.J.; Hsieh, W.W. Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting. *Nonlinear Process. Geophys.* **2008**, *15*, 221–232. [CrossRef]
6. Iaci, R.; Sriram, T.N. Robust multivariate association and dimension reduction using density divergences. *J. Multivar. Anal.* **2013**, *117*, 281–295. [CrossRef]
7. Gifi, A. *Nonlinear Multivariate Analysis*; Wiley-Blackwell: Hoboken, NJ, USA, 1990.
8. Breiman, L.; Friedman, J.H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
9. Lai, P.L.; Fyfe, C. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.* **2000**, *10*, 365–377. [CrossRef] [PubMed]
10. Painsky, A.; Feder, M.; Tishby, N. Nonlinear canonical correlation analysis: A compressed representation approach. *Entropy* **2020**, *22*, 208. [CrossRef] [PubMed]
11. Van Der Burg, E.; de Leeuw, J. Non-linear canonical correlation. *Br. J. Math. Stat. Psychol.* **1983**, *36*, 54–80. [CrossRef]
12. Yin, X. Canonical correlation analysis based on information theory. *J. Multivar. Anal.* **2004**, *91*, 161–176. [CrossRef]
13. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall: Boca Raton, FL, USA, 2006.
14. Mandal, A.; Cichocki, A. Non-Linear Canonical Correlation Analysis Using Alpha-Beta Divergence. *Entropy* **2013**, *15*, 2788–2804. [CrossRef]
15. Cichocki, A.; Cruces, S.; Amari, S.I. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170. [CrossRef]
16. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559. [CrossRef]
17. Karasuyama, M.; Sugiyama, M. Canonical dependence analysis based on squared-loss mutual information. *Neural Netw.* **2012**, *34*, 46–55. [CrossRef] [PubMed]
18. Nielsen, A.; Vestergaard, J.S. Canonical analysis based on mutual information. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26-31 July 2015; pp. 1068–1071.
19. Romanazzi, M. Influence in canonical correlaion analysis. *Psychometrika* **1992**, *57*, 237–259. [CrossRef]

20. Sakar, C.O.; Kursun, O. An hybrid method for feature selection based on mutual information and canonical correlation analysis. In Proceedings of the 20th International Conference on Pattern Recognition, Istambul, Turkey, 23–26 August 2010.
21. Sakar, C.O.; Kursun, O. A method for combining mutual information and canonical correlation analysis: Predictive mutual information and its use in feature selection. *Expert Syst. Appl.* **2012**, *39*, 3333–3344. [CrossRef]
22. Wang, Y.; Cang, S.; Yu, H. Mutual information inspired on feature selection using kernel canonical correlation analysis. *Expert Syst.* **2019**, *4*, 100014. [CrossRef]
23. Bell, C.B. Mutual information and maximal correlation as measures of dependence. *Ann. Math. Stat.* **1962**, *33*, 587–595. [CrossRef]
24. Iaci, R.; Yin, X.; Sriram, T.N.; Klingerberg, C.P. An informational measure of association and dimension reduction for multiple sets and groups with applications in morphometric analysis. *J. Am. Stat. Assoc.* **2008**, *103*, 1166–1176. [CrossRef]
25. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* **2001**, *88*, 865–873. [CrossRef]
26. Broniatowski, M.; Toma, A.; Vajda, I. Decomposable pseudodistance and applications in statistical estimation. *J. Stat. Plan. Inference* **2012**, *142*, 2574–2585. [CrossRef]
27. Castilla, E.; Martín, N.; Muñoz, S.M.; Pardo, L. Robust Wald-type tests based on minimum Rényi pseudodistances estimators for the multiple regresion model. *J. Stat. Comput. Simul.* **2020**, *90*, 2655–2680. [CrossRef]
28. Castilla, E.; Jaenada, M.; Pardo, L. Estimation and testing on independent not identically distributed observations based on Rényi's pseudodistances. *IEEE Trans. Inf. Theory* **2022**, *68*, 4588-4609. [CrossRef]
29. Rényi, A. On measures of entropy and information. In *Proceeding of the 4th Symposium on Probability and Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
30. Toma, A.; Leoni-Aubin, S. Optimal robust M-estimators using Rényi pseudodistances. *J. Multivar. Anal.* **2013**, *115*, 259–273. [CrossRef]
31. Toma, A.; Karagrigoriou, A.; Trentou, P. Robust model selection criteria based on pseudodistances. *Entropy* **2020**, *22*, 304. [CrossRef] [PubMed]
32. Jaenada, M.; Pardo, L. The minimum Renyi's Pseudodistances estimators for Generalized Linear Models. In *Data Analysis and Related Applications: Theory and Practice*; Proceeding of the ASMDA; Wiley: Athens, Greece, 2021.
33. Jaenada, M.; Pardo, L. Robust statistical inference in generalized linear models based on minimum Renyi pseudistance estimators. *Entropy* **2022**, *24*, 123. [CrossRef] [PubMed]
34. Castilla, E.; Jaenada, M.; Martín, N.; Pardo, L. Robust approach for comparing two dependent normal populations through Wald-type tests based on Rényi's pseudodistance estimators. *Stat. Comput.* **2023**, *32*, 100. [CrossRef]
35. Jaenada, M.; Miranda, P.; Pardo, L. Robust Test Statistics Based on Restricted Minimum Rényi's Pseudodistance Estimators. *Entropy* **2022**, *24*, 616. [CrossRef]
36. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
37. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, UK, 1986.
38. Kim, J.S.; Scott, C. Robust kernel density estimation. *J. Mach. Learn. Res.* **2012**, *13*, 2529–2565.
39. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; Wiley: New York, NY, USA, 1992; Volume 1.
40. Rüschendorf, L. Consistency of estimators for multivariate density functions and for the mode. *Sankhya Ser. A* **1977**, *39*, 243–250.
41. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997; Volume 1.
42. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1993; Volume 57.
43. Frets, G.P. Heredity of head form in man. *Genetica* **1921**, *3*, 193–384. [CrossRef]