2013

Robust multivariate association and dimension reduction using density divergences

Ross Iaci
William & Mary, riaci@wm.edu

T. N. Sriram

Recommended Citation

Iaci, R., & Sriram, T. N. (2013). Robust multivariate association and dimension reduction using density divergences. Journal of Multivariate Analysis, 117, 281-295.

Journal of Multivariate Analysis 117 (2013) 281–295

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Robust multivariate association and dimension reduction using density divergences

Ross Iaci [a,*], T.N. Sriram [b]

[a] Department of Mathematics, The College of William and Mary, Williamsburg, VA, 23185, United States

[b] Department of Statistics, University of Georgia, Athens, GA, 30602, United States

ARTICLE INFO

Article history:
Received 16 December 2011
Available online 18 March 2013

AMS subject classifications:
62G07
62G20
62G35
62H12
62H20

Keywords:
Multivariate association measures
Density power divergence
Density alpha divergence
Dimension reduction
Permutation test
Robustness

ABSTRACT

In this article, we introduce two new families of multivariate association measures based on power divergence and alpha divergence that recover both linear and nonlinear dependence relationships between multiple sets of random vectors. Importantly, this novel approach not only characterizes independence, but also provides a smooth bridge between well-known distances that are inherently robust against outliers. Algorithmic approaches are developed for dimension reduction and the selection of the optimal robust association index. Extensive simulation studies are performed to assess the robustness of these association measures under different types and proportions of contamination. We illustrate the usefulness of our methods in application by analyzing two socioeconomic datasets that are known to contain outliers or extreme observations. Some theoretical properties, including the consistency of the estimated coefficient vectors, are investigated and computationally efficient algorithms for our nonparametric methods are provided.

1. Introduction

Canonical Correlation Analysis (CCA) pioneered by Hotelling [10] is a classical method for determining pair-wise linear relationships between two sets of random vectors $\mathbf{X}_{p\times 1}$ and $\mathbf{Y}_{q\times 1}$. One main drawback of CCA is that the canonical coefficients computed based on a classical estimator of the covariance matrix are vulnerable to outlying observations, which in turn affects the recovered relationships; see [20]. The presence of outliers or extreme observations severely impacts the performance of multivariate association measures, such as CCA, because of their potential to mask the true relationships or even identify spurious ones. Therefore, robust methods are fundamental to the study of multivariate associations, since manual cleaning is often not feasible and it is a challenge just to detect outliers in multivariate contexts. Even if it is possible to detect outliers, the fact that these observations may genuinely be part of the dataset makes it essential to consider inherently robust approaches that are capable of extracting the underlying true relationships in the presence of such values.

Karnel [14] adopted an obvious robust version of CCA by estimating a covariance matrix using an M-estimator, whereas [7] used a minimum covariance determinant estimator. Importantly, canonical variates obtained using robust estimates of a covariance matrix lose their natural interpretations. Branco et al. [3] proposed two approaches for robust CCA based on projection pursuit and robust alternating regressions. Also, see [16,8,4] for other robust modifications of CCA. While

* Corresponding author.
E-mail addresses: riaci@wm.edu (R. Iaci), tn@stat.uga.edu (T.N. Sriram).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.03.004

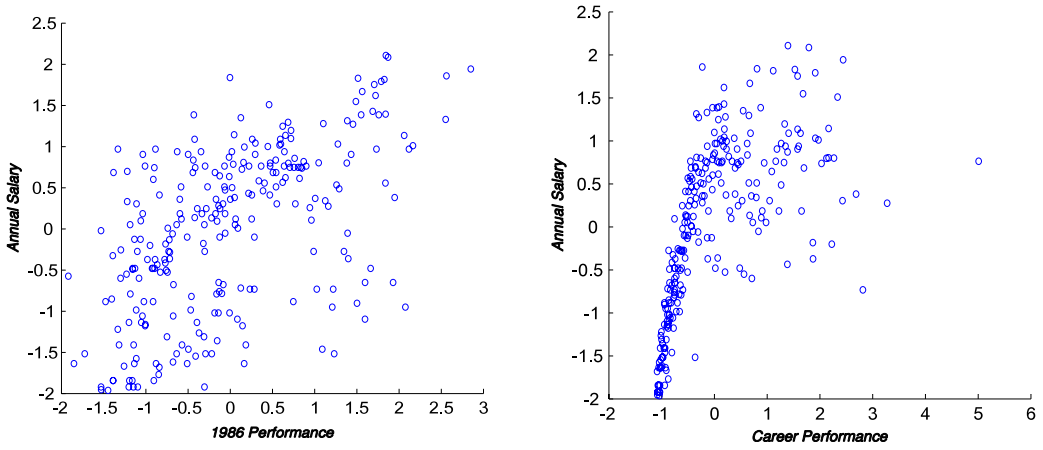

Fig. 1. Variate plots z (*annual salary*) vs: $\widehat{\mathbf{a}}_1^T\mathbf{x}$ (left panel) and $\widehat{\mathbf{b}}_1^T\mathbf{y}$ (right panel) (example 5.2).

these methods address the issue of robustness in association studies, they are limited to detecting only linear relationships between two sets.

In the last decade, there has been a renewed interest in developing methods for multivariate association and dimension reduction involving multiple sets. For instance, Yin and Sriram [28] extended the Kullback–Leibler (KL) divergence method of Yin [26] to recover linear/nonlinear relationships between groups of multiple sets of vectors. Note that the method of Yin [26] was motivated by the work of Yin and Cook [27] for dimension reduction in regression. Iaci et al. [13] then introduced an overall measure of association based on KL divergence for application in morphological integration studies. More recently, Iaci, Sriram and Yin [12] developed Generalized Canonical Analysis (GCA) based on an L_2 measure that generalizes CCA and applied it to extract linear and nonlinear relationships between a set of mortality, air pollution and weather vectors, for an environmental dataset. However, as shown later, these association measures are also sensitive to outliers.

In this article, we construct a continuum of multivariate association measures based on density power divergence (DPD) parameterized by a tuning parameter α, which helps balance infinitesimal robustness and efficiency measured in terms of the recovered amount of dependence between sets of variables. Our method, termed Power Divergence Canonical Analysis (PDCA), is shown to identify both linear and nonlinear relationships between multiple sets, even in the presence of extreme observations. In fact, for $\alpha \in [0, 1]$, the PDCA index based on DPD provides a smooth bridge between the KL divergence method ($\alpha = 0$) and an L_2 distance ($\alpha = 1$) and thus, encompasses the methods of Yin [26] and Iaci et al. [13]. Moreover, for two sets our method is shown to be equivalent to CCA under multivariate normality. A second continuum of multivariate association measures is also proposed based on density alpha divergence (DAD), which includes the well-known Neyman's Chi-square ($\alpha = -1$) and Pearson's Chi-square ($\alpha = 2$) and the Hellinger distance ($\alpha = 1/2$). The latter method, termed Alpha Divergence Canonical Analysis (ADCA), is developed not only to provide another family of robust measures, but also for comparison purposes. A novelty here is, rather than pre-selecting any one value of α and considering the resulting robust association measure, we propose a continuum of robust measures and let the data determine an optimal estimate of α which balances robustness and efficiency.

We motivate the need for our new robust methodologies through two socioeconomic datasets that are known to contain outliers or extreme observations. The first dataset concerns consumer expenditures and was extracted from the Consumer Expenditure (CE) survey. We first analyze the entire dataset and identify a relationship between a *household expenditure* vector and a *socioeconomic* vector using PDCA, GCA and CCA and then study the effect of outliers on the recovered association.

The second is the baseball *Hitters' Salary* dataset collected for a data analysis exposition sponsored by the *American Statistical Association*. This dataset has been well-studied from a regression perspective to answer the question, "are players paid according to their performance?". The presence of extreme observations in this dataset has been handled in a variety of ways in past analyses. For instance, after identifying and removing outlying observations, Xia et al. [25] reanalyzed the dataset to illustrate their effective dimension reduction (EDR) method, termed Minimum Average Variance Estimation (MAVE). We take an entirely different approach and use our robust PDCA method to simultaneously study the joint association between the three sets of variables; *annual salary*, *1986 performance* and *career performance*. The benefit of such an approach is seen in the plots in Fig. 1, which show that our multiple set PDCA analysis simultaneously recovers a relationship between the *annual salary* and *1986 performance*, and the *annual salary* and the *career performance* variables. Such a distinction of the influence that the *performance* variables have on *annual salary* is not apparent in [25]. Importantly, with our robust methods a preliminary analysis to detect and remove the outliers from this dataset is not necessary.

The article is organized as follows. In Section 2 we introduce the association measures based on DPD and DAD used to recover the joint linear/nonlinear relationships between multiple sets. We show analytically the equivalence of our DPD based method to CCA under a multivariate normal assumption in Section 2.2. Computational aspects of our methods are discussed in Section 2.3 and a consistency result for the estimated coefficient vectors is stated in Section 2.4 with

a proof provided in the Appendix. The robustness of PDCA is motivated through an illustrative example in Section 2.5. We use a permutation test in Section 3.1 to determine the number of significant relationships and thereby, provide a method for dimension reduction. In Section 3.2 we develop a data driven algorithm to determine the optimal level of the tuning parameter that parameterizes the most robust index. Simulation studies are performed in Section 4 to compare the performance of PDCA, ADCA and GCA in the presence of different types and proportions of gross-error contamination. The additional simulations given in Web Appendices E–M are briefly described in Section 4.2. The two real datasets are analyzed in Sections 5.1 and 5.2 and concluding remarks are given in Section 6.

2. Methodology

In this section, we propose and study two association measures based on density divergence, which provide two families of robust methods to recover relationships between multiple sets of random vectors.

2.1. Density divergences and families of association measures

For $\alpha > 0$, we state the multivariate extension of the Basu et al. [1] *density power divergence* (DPD) between two density functions g_1 and g_2 as

$$D_\alpha(g_2, g_1) = \int_{\mathbf{u}} \left[g_1^{1+\alpha}(\mathbf{u}) - \left(1 + \frac{1}{\alpha}\right) g_1^{\alpha}(\mathbf{u}) g_2(\mathbf{u}) + \left(\frac{1}{\alpha}\right) g_2^{1+\alpha}(\mathbf{u}) \right] d\mathbf{u}, \tag{1}$$

where $\mathbf{u} = (u_1, \ldots, u_m)^T$ is an m-dimensional vector with $m \geq 1$ and α is a tuning parameter. Since $D_\alpha(g_2, g_1)$ is a divergence, we have that

$$D_\alpha(g_2, g_1) \geq 0 \quad \text{for all } g_1 \quad \text{and} \quad g_2, \text{ and } D_\alpha(g_2, g_1) = 0 \quad \text{if and only if} \quad g_1 = g_2. \tag{2}$$

When $\alpha = 0$, the integrand in (1) is undefined and thus, we define $D_0(g_2, g_1) = \lim_{\alpha \to 0} D_\alpha(g_2, g_1) = \int g_2(\mathbf{u}) \ln \left[\frac{g_2(\mathbf{u})}{g_1(\mathbf{u})} \right] d\mathbf{u}$, which is the KL divergence. When $\alpha = 1$, $D_1(g_2, g_1) = \int |g_2(\mathbf{u}) - g_1(\mathbf{u})|^2 d\mathbf{u}$ is the L_2 distance between g_1 and g_2, and thus, for $0 < \alpha < 1$, the DPD is a smooth bridge between the KL divergence and the L_2 distance between two densities.

Note that, the DPD does not include some well-known distances, such as the Hellinger distance (HD) that was exploited by Beran [2] to construct minimum HD estimators that are as efficient as the MLE and simultaneously robust against error contaminations. This, and the work of Cressie and Read [6], motivate us to consider another family of density power divergences. For $\alpha \in (-\infty, \infty) \setminus \{0, 1\}$, the *density alpha divergence* (DAD) between two multivariate densities g_1 and g_2 is defined to be

$$d_\alpha(g_2, g_1) = \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\mathbf{u}} [g_1(\mathbf{u})/g_2(\mathbf{u})]^\alpha \, g_2(\mathbf{u}) d\mathbf{u} \right]. \tag{3}$$

Note that, d_α is the Hellinger distance when $\alpha = 1/2$, Neyman's Chi-square when $\alpha = -1$, and is equivalent to Pearson's Chi-square when $\alpha = 2$; see [19,5]. When $\alpha = 0$ or 1, we define d_α in (3) to be the KL divergence. Also, for $\alpha \in (-\infty, \infty)$, $d_\alpha(g_2, g_1)$ is a divergence and (2) holds.

Next, consider $m \geq 2$ sets of random vectors $\mathbf{X}^{(k)}$ with dimension p_k, $k = 1, \ldots, m$, with associated coefficient vectors $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)})$, and suppose that $g_2 = f_{(\mathbf{a}^{(1)T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)T}\mathbf{X}^{(m)})}$ and $g_1 = \prod_{k=1}^m f_{\mathbf{a}^{(k)T}\mathbf{X}^{(k)}}$ denote the joint and the product of the marginal densities of $(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})$ and $\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}$, $k = 1, \ldots, m$, respectively. For $\alpha \geq 0$, we define a power divergence multivariate association index by substituting g_1 and g_2 into (1) as

$$\mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) = D_\alpha\left(f_{(\mathbf{a}^{(1)T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)T}\mathbf{X}^{(m)})}, \prod_{k=1}^m f_{\mathbf{a}^{(k)T}\mathbf{X}^{(k)}} \right). \tag{4}$$

Analogously, an alpha divergence index is defined using (3) as

$$\mathcal{A}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) = d_\alpha\left(f_{(\mathbf{a}^{(1)T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)T}\mathbf{X}^{(m)})}, \prod_{k=1}^m f_{\mathbf{a}^{(k)T}\mathbf{X}^{(k)}} \right), \tag{5}$$

where $\alpha \in (-\infty, \infty)$. Note that, for every fixed $\alpha \geq 0$, (2) gives the result that

$$D_\alpha\left(f_{(\mathbf{a}^{(1)T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)T}\mathbf{X}^{(m)})}, \prod_{k=1}^m f_{\mathbf{a}^{(k)T}\mathbf{X}^{(k)}} \right) = 0 \Leftrightarrow f_{(\mathbf{a}^{(1)T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)T}\mathbf{X}^{(m)})} = \prod_{k=1}^m f_{\mathbf{a}^{(k)T}\mathbf{X}^{(k)}};$$

that is, the $\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}$, $k = 1, \ldots, m$, are mutually independent. This also holds for d_α in (3).

To jointly recover linear and nonlinear relationships between the random vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$, for each $\alpha \geq 0$, we search successively for the coefficient vectors $(\mathbf{a}_i^{(1)}, \ldots, \mathbf{a}_i^{(m)})$ such that the projected vectors $(\mathbf{a}_i^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}_i^{(m)^T}\mathbf{X}^{(m)})$ have the most dependence by maximizing $\mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ with respect to $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)})$. The maximization is performed under the constraints that the projected vectors, or variates, have unit variance, $\mathbf{a}_i^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{a}_i^{(k)} = 1$ for all $i = 1, \ldots, \min(p_k)$, and are uncorrelated, $\mathbf{a}_i^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{a}_j^{(k)} = 0$ for all $j = 1, \ldots, i-1$. Note that, we could also formulate our index using coefficient matrices and provide an overall measure of association as in [13], but instead focus on identifying the coefficient vectors that extract the relationships between multiple sets.

Not only is our method able to extract both linear and nonlinear dependence relationships between multiple sets, but under a two-set multivariate normal assumption our method is equivalent to CCA; see Section 2.2. Therefore, we term our method *Power Divergence Canonical Analysis* (PDCA). Similarly, when the alpha divergence index, $\mathcal{A}_\alpha$, is used, we term our method *Alpha Divergence Canonical Analysis* (ADCA). Thus, for $m \geq 2$, $\{\mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right), \alpha \geq 0\}$ and $\{\mathcal{A}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right), \alpha \in (-\infty, \infty)\}$ provide two new families of multivariate association measures that characterize independence.

2.2. Equivalence to CCA under normality

Let $(\mathbf{X} = \mathbf{X}^{(1)}, \mathbf{Y} = \mathbf{X}^{(2)})$ be multivariate normal, then the PDCA index in (4) reduces to

$$\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b}) = \frac{1}{\alpha(\alpha+1)(2\pi)^\alpha \sigma_1^\alpha \sigma_2^\alpha}\left[\alpha - \frac{(\alpha+1)^2}{\sqrt{\left[1+2\alpha+\alpha^2(1-\rho^2)\right]}} + \frac{1}{(1-\rho^2)^{\alpha/2}}\right], \tag{6}$$

where $\sigma_1 = \mathrm{Var}(\mathbf{a}^T\mathbf{X})$, $\sigma_2 = \mathrm{Var}(\mathbf{b}^T\mathbf{Y})$ and $\rho = \rho(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{b}/(\sigma_1\sigma_2)$; see Web Appendix A. For each $\alpha > 0$, it can be shown using calculus and algebra that $\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b})$ is a decreasing function of $(1-\rho^2)$, which implies that maximizing $\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b})$ with respect to $(\mathbf{a}, \mathbf{b})$ is achieved by maximizing $\rho(\mathbf{a}, \mathbf{b})$ and thus, equivalent to CCA. For $\alpha = 0$, PDCA is defined to be the KL divergence method of Yin [26], which was shown to be equivalent to CCA and hence, PDCA is equivalent to CCA for all $\alpha \geq 0$. Moreover, the GCA method of Iaci et al. [12] is equivalent to CCA under normality and thus, equivalent to PDCA.

2.3. Computational methods

The population version $\mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ in (4) can be written using the integral representation in (1) as

$$\begin{aligned}\mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) &= \prod_{k=1}^m \mathrm{E}\left[f^\alpha_{\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}}(\mathbf{a}^{(k)^T}\mathbf{X}^{(k)})\right] - (1+1/\alpha)\mathrm{E}\left[\prod_{k=1}^m f^\alpha_{\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}}(\mathbf{a}^{(k)^T}\mathbf{X}^{(k)})\right] \\ &\quad + (1/\alpha)\mathrm{E}\left[f^\alpha_{(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})}(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})\right],\end{aligned}$$

where the expectation in the last two terms is with respect to the joint density of $(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})$. Therefore, letting $\{\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(m)}, i = 1, \ldots, n\}$ denote a random sample from $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$, we estimate the PDCA index as

$$\begin{aligned}\widehat{\mathcal{R}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) &= \prod_{k=1}^m \left[n^{-1}\sum_{i=1}^n \widehat{f}_n^\alpha(\mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)})\right] - (1+1/\alpha)n^{-1}\sum_{i=1}^n\left[\prod_{k=1}^m \widehat{f}_n^\alpha(\mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)})\right] \\ &\quad + (1/\alpha)\left[n^{-1}\sum_{i=1}^n \widehat{f}_n^\alpha(\mathbf{a}^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}_i^{(m)})\right],\end{aligned} \tag{7}$$

where the functions $\widehat{f}_n(\mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)})$ and $\widehat{f}_n(\mathbf{a}^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}_i^{(m)})$ are kernel density estimates of $f_{\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}}(\mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)})$ and $f_{(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})}(\mathbf{a}^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}_i^{(m)})$, respectively. Similarly, we define a sample version of the ADCA index as $\widehat{\mathcal{A}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) = \frac{1}{\alpha(1-\alpha)}\left(1 - \frac{1}{n}\sum_{i=1}^n\left[\frac{\prod_{j=1}^m \widehat{f}_n(\mathbf{a}^{(j)^T}\mathbf{x}_i^{(j)})}{\widehat{f}_n\left(\mathbf{a}^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}_i^{(m)}\right)}\right]^\alpha\right)$. The steps for obtaining the estimated coefficient vectors and computational notes pertaining to each step are discussed next.

Step 0:

Whiten each data matrix $\mathbf{D}_{\mathbf{X}^{(k)}}$ corresponding to $\mathbf{X}^{(k)}$ using the nonsingular transformation $\mathbf{Z}^{(k)} = \Sigma_{\mathbf{X}^{(k)}}^{-1/2}\left(\mathbf{X}^{(k)} - \mathrm{E}\mathbf{X}^{(k)}\right)$, $k = 1, \ldots, m$. For ease in exposition, the notation $\mathbf{X}^{(k)}$ is maintained throughout this section.

Note that, for any $\alpha > 0$ and $\mathbf{U}^{(k)} = \mathbf{C}^{(k)}\mathbf{X}^{(k)} + b_k$, $k = 1, \ldots, m$, where $\mathbf{C}^{(k)}$ are nonsingular matrices and b_k is a fixed $p_k \times 1$ vector, then the following holds,

$$\mathcal{R}_{\alpha,(\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(m)})}\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) = M * \mathcal{R}_{\alpha,(\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(m)})}\left(\mathbf{C}^{(1)^T}\mathbf{a}^{(1)}, \ldots, \mathbf{C}^{(m)^T}\mathbf{a}^{(m)}\right),$$

where M is a constant. Thus, the PDCA index is equivariant under invertible linear transformation; see Web Appendix D for the proof. The main motivation for this scale transformation is to simplify the constraints in Section 2.1 to the orthonormal constraints $\mathbf{a}_i^{(k)^T}\mathbf{a}_i^{(k)^T} = 1, i = 1, \ldots, \min(p_k)$ and $\mathbf{a}_i^{(k)^T}\mathbf{a}_j^{(k)^T} = 0, j = 1, \ldots, i-1$, and ease computation. This transformations changes the scale, but not the relationships between the random vectors, rescales the variables to have equivalent magnitude, and importantly allows the estimated coefficient vectors to be transformed back to the original scale.

Step 1:

For a given set of coefficient vectors calculate the product kernel density estimate

$$\widehat{f}_n(\mathbf{a}^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}_i^{(m)}) = \frac{1}{nh_1h_2\cdots h_m}\sum_{j=1}^{n}\prod_{k=1}^{m} K\left[(\mathbf{a}^{(k)^T}\mathbf{x}_j^{(k)} - \mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)})/h_k\right]$$

and evaluate the sample PDCA (or ADCA) index $\widehat{\mathcal{R}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ in (7).

Scott [22] and Silverman [23] suggested the use of Gaussian product kernels for density estimation and the simulation studies performed in [26,13] confirm that this choice works well for their KL divergence based methods. The same was also shown to be true for the L_2 type index in [12]. The success of these methods and the fact that the PDCA index is a smooth bridge between the KL and an L_2 distances, for $0 < \alpha < 1$, induced our selection of Gaussian product kernels. However, our methods hold for any kernel of bounded variation. Finally, motivated by the results of Yin [26] and Iaci et al. [13], and the discussion in [12] on bandwidth selection for multivariate association methods whose main goal is the estimation of the directions $\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}$, as is the focus here, we use the bandwidths $h_k = (4/(d+2))^{1/(d+4)}\, s_k n^{-1/(d+4)}$, where d is the dimension of the density being estimated ($d = 1$ or m here) and s_k is the sample standard deviation of $\{\mathbf{a}^{(k)^T}\mathbf{x}_i^{(k)}, i = 1, \ldots, n\}$.

Step 2:

Under the orthonormal constraints imposed in *Step* 0, the estimated coefficient vectors are the solutions

$$(\widehat{\mathbf{a}}^{(1)}, \ldots, \widehat{\mathbf{a}}^{(m)}) = \operatorname{argmax} \widehat{\mathcal{R}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) \ \left(\text{or } \widehat{\mathcal{A}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)\right).$$

Here, the maximization is carried out iteratively using the nonlinear constrained minimizer (or maximizer) *fmincon* available in Matlab, which implements a Sequential Quadratic Programming (SQP) method. The SQP is a nonlinear constrained minimizing algorithm, which closely mimics Newton's method for constrained optimization. This method maximizes the sample PDCA or ADCA index while incorporating the nonlinear constraints $\mathbf{a}^{(k)^T}\mathbf{a}^{(k)} = 1$ simultaneously. Details of the SQP procedure can be found in [17].

Step 3:

After the ith estimated coefficient vectors $\widehat{\mathbf{a}}_i^{(k)}$, $k = 1, \ldots, m$, are found, each data matrix $\mathbf{D}_{\mathbf{X}^{(k)}}$ is projected into a subspace orthogonal to $\mathbf{A}_{p_k\times i}^{(k)} = \left[\widehat{\mathbf{a}}_1^{(k)}, \ldots, \widehat{\mathbf{a}}_i^{(k)}\right]$, $i < \min(p_k)$, and Steps 1–2 are repeated to find $(\widehat{\mathbf{a}}_{i+1}^{(1)}, \ldots, \widehat{\mathbf{a}}_{i+1}^{(m)})$.

A detailed algorithm for implementing this step is given in Web Appendix P.

2.4. Consistency of the estimated coefficient vectors

In this section, we state a consistency result for the 1st estimated coefficient vectors $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)})$ defined in Section 2.3 for the PDCA index. The 2nd estimated coefficient vectors $(\widehat{\mathbf{a}}_2^{(1)}, \ldots, \widehat{\mathbf{a}}_2^{(m)})$ are obtained by maximizing $\widehat{\mathcal{R}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ over the set $\{\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) : \mathbf{a}^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\mathbf{a}^{(k)} = 1 \text{ and } \mathbf{a}^{(k)^T}\Sigma_{\mathbf{X}^{(k)}}\widehat{\mathbf{a}}_1^{(k)} = 0,\ k = 1, \ldots, m\}$, which is a random set depending on $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)})$. Since the 1st estimated coefficient vectors are shown to be consistent, we can modify the proof of Theorem 1 suitably to establish the consistency of the 2nd estimated coefficient vectors. The same argument applies for successive coefficient vectors.

Theorem 1 (*Consistency*). *Assume the conditions of Lemma 1 in Appendix A.1. Let* $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) = \operatorname{argmax} \widehat{\mathcal{R}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ *and* $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)}) = \operatorname{argmax} \mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$, *for each* $\alpha \geq 0$. *Then,* $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) \to (\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$ *almost surely as* $n \to \infty$.

The proof is given in Appendix A.1. Similar results hold for the estimated coefficient vectors based on the sample ADCA index, $\widehat{\mathcal{A}}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$.

Table 1
Average L_2 distances and correlations for uncontaminated and contaminated data.

	(α)	$\underline{\mathbf{a}}$				$\underline{\mathbf{b}}$			
		$\overline{L_2^*}$	$(\overline{\Delta})$	$\overline{\lvert\rho(\widehat{\mathbf{a}}, \mathbf{a})\rvert}$	$\overline{\lvert\rho(\widehat{\mathbf{a}}_c, \mathbf{a})\rvert}$	$\overline{L_2^*}$	$(\overline{\Delta})$	$\overline{\lvert\rho(\widehat{\mathbf{b}}, \mathbf{b})\rvert}$	$\overline{\lvert\rho(\widehat{\mathbf{b}}, \mathbf{b}_c)\rvert}$
PDCA	(0.6)	0.1410		0.9979	0.9841	0.0281		0.9994	0.9986
ADCA	(0.9)	0.3822	(2.71)	0.9980	0.8231	0.2866	(10.20)	0.9994	0.8450
GCA		0.4502	(3.19)	0.9983	0.8389	0.1511	(5.380)	0.9995	0.9372
CCA		0.7285	(5.17)	0.9984	0.6170	0.5299	(18.86)	0.9995	0.7464

2.5. Robustness

To motivate the inherent robustness property of the PDCA index, again consider two sets of vectors $\mathbf{X}$ and $\mathbf{Y}$. Note that, for α close to 0, a heuristic argument shows that the sample PDCA index $\widehat{\mathcal{R}}_\alpha(\mathbf{a}, \mathbf{b})$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left[\ln\left(\widehat{f}_n(\mathbf{a}^T\mathbf{x}_i, \mathbf{b}^T\mathbf{y}_i)\right)\widehat{f}_n^{\alpha}(\mathbf{a}^T\mathbf{x}_i, \mathbf{b}^T\mathbf{y}_i) - \ln\left(\widehat{f}_n(\mathbf{a}^T\mathbf{x}_i)\widehat{f}_n(\mathbf{b}^T\mathbf{y}_i)\right)\widehat{f}_n^{\alpha}(\mathbf{a}^T\mathbf{x}_i)\widehat{f}_n^{\alpha}(\mathbf{b}^T\mathbf{y}_i)\right];$$

see Web Appendix B. However, when $\alpha = 0$, $\widehat{\mathcal{R}}_\alpha(\mathbf{a}, \mathbf{b}) \approx n^{-1}\sum_{i=1}^{n}\ln\left(\frac{\widehat{f}_n(\mathbf{a}^T\mathbf{x}_i, \mathbf{b}^T\mathbf{y}_i)}{\widehat{f}_n(\mathbf{a}^T\mathbf{x}_i)\widehat{f}_n(\mathbf{b}^T\mathbf{y}_i)}\right)$, which is the estimating function for the Kullback–Leibler (KL) index of Yin [26] and Iaci et al. [13]. Therefore, for α near 0, the PDCA index can be viewed as a weighted version of the KL index. Next, suppose that $\mathbf{x}_i$ and/or $\mathbf{y}_i$ are outliers, then the PDCA index naturally down-weights these values through $\widehat{f}_n^{\alpha}(\mathbf{a}^T\mathbf{x}_i, \mathbf{b}^T\mathbf{y}_i)$, $\widehat{f}_n^{\alpha}(\mathbf{a}^T\mathbf{x}_i)$ and $\widehat{f}_n^{\alpha}(\mathbf{b}^T\mathbf{y}_i)$, and hence, the estimated coefficient vectors of $\mathbf{a}$ and $\mathbf{b}$ are less affected. Whereas, the KL index assigns a weight of one to each observation, including the outliers, and hence, the resulting estimates of $\mathbf{a}$ and $\mathbf{b}$ will be more affected. Consequently, we investigate the range of values $\alpha = 0.1, \ldots, 1.0$ to determine the optimal value that parameterizes the most robust PDCA index. This simple heuristic motivation reveals the rationale for the expected robustness using PDCA.

Next, we illustrate the above motivation through an example that shows that for many values of $\alpha > 0$, the PDCA method is considerably more robust against gross-error contamination than is GCA, ADCA and CCA. Since CCA can only detect linear relationships between two sets, we will conduct a comparative robustness study by contaminating a linear relationship between $\mathbf{X}$ and $\mathbf{Y}$. We consider a simple scenario with two random vectors $\mathbf{X} = (X_1, \ldots, X_8)^T$ and $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$, where $X_1, \ldots, X_8, Y_2, Y_3$ are independent $N(0, 1)$ random variables, and $Y_1 = (2X_1 + X_2 + X_3) + \epsilon$. Suppose that a dataset of size $n = 100$ is randomly generated with $\epsilon \sim N(0, \sigma = 0.5)$. We create a 10% contamination of the linear relationship by randomly replacing $n^* = 10$ of the ϵ's with ϵ^*'s drawn from a $U(0, 50)$.

Let $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$ and $(\widehat{\mathbf{a}}_c, \widehat{\mathbf{b}}_c)$ correspond to the estimated coefficient vectors of the true vectors $\mathbf{a} = (2, 1, 1, 0, 0, 0, 0, 0)^T$ and $\mathbf{b} = (1, 0, 0)^T$ calculated from the uncontaminated and contaminated datasets, respectively. In order to measure the effect of contamination on the estimated coefficient vectors (given by PDCA, GCA and CCA), we consider the L_2-norm, $L_2^* = \|(\mathbf{I} - \widehat{\mathbf{a}}\widehat{\mathbf{a}}^T)\widehat{\mathbf{a}}_c\|_2$, which is the projection of $\widehat{\mathbf{a}}_c$ into the orthogonal subspace spanned by $\widehat{\mathbf{a}}$. The L_2^* value is small if $\widehat{\mathbf{a}} \cong \widehat{\mathbf{a}}_c$, implying that the estimates are not largely affected by the contamination, and hence, the corresponding method is more robust; we calculate this measure analogously for $(\widehat{\mathbf{b}}, \widehat{\mathbf{b}}_c)$. To compare the methods, with PDCA as the standard, we consider the relative change in the difference $\Delta = L_2^*/L_{2;\mathcal{R}}^*$. Finally, the accuracy of the estimates can be quantified by the absolute correlations between the estimated and true variates, $|\rho(\widehat{\mathbf{a}}, \mathbf{a})| = |\rho(\widehat{\mathbf{a}}^T\mathbf{x}, \mathbf{a}^T\mathbf{x})|$, $|\rho(\widehat{\mathbf{a}}_c, \mathbf{a})|$, $|\rho(\widehat{\mathbf{b}}, \mathbf{b})|$ and $|\rho(\widehat{\mathbf{b}}_c, \mathbf{b})|$.

Table 1 reports the following average values computed using 500 datasets generated with the above specifications: $\overline{L_2^*}$, $\overline{|\rho(\cdot, \cdot)|}$ and $\overline{\Delta}$. For PDCA and ADCA, respectively, the table reports only the value of α for which the $\overline{L_2^*}$ value is the smallest, while the figures in Web Appendix C give the plots of the $\overline{L_2^*}$ values for $\alpha = 0.1, 0.2, \ldots, 1.0$, and additionally, $\alpha = -1, 2$ for ADCA. It is evident from the reported $\overline{L_2^*}$ and $\overline{\Delta}$ values, and the plots, that PDCA is considerably more robust than ADCA, GCA and CCA, especially when estimating $\mathbf{b}$. In terms of the $\overline{L_2^*}$ distance values in the plots, $\alpha = 0.6$ generates the most robust PDCA index, whereas for the ADCA method, all values $\alpha \geq 0.4$ produce similar results, with $\alpha = 0.9$ narrowly parameterizing the most robust ADCA index.

For this simple example, Fig. 1 in Web Appendix C shows that PDCA is more resistant to contamination when α is near 0 than when it approaches 1. This suggests that we can algorithmically select a value of α that yields maximum robustness. In Section 3.2, a data driven procedure for selecting the value of α that achieves the most robustness is given. It is important to note that the PDCA method achieves robustness against contamination without having to use robust versions of estimated covariance matrices.

In terms of the average of the absolute correlations between the estimated and true vectors, PDCA has near perfect correlation for the datasets with and without contamination, whereas for the other methods the correlations decrease considerably under contamination, with CCA showing the most decline. Note that, the values in Table 1 are reported in the transformed scale, where $\mathbf{X}$ and $\mathbf{Y}$ have identity covariance matrices; see Section 2.3.

Using a robust CCA method would expectedly improve performance, however, if a nonlinear relationship existed between the two sets, then CCA would be less successful in identifying this type of association whether or not a robust version is used.

Moreover, CCA is unable to jointly study the associations between multiple sets. All this motivates our study of density divergences for recovering multivariate associations with a particular focus on the PDCA method.

3. Dimension reduction and alpha selection

In this section, we develop a method for dimension reduction and the selection of an optimal level of the tuning parameter α for the PDCA index. The same methods can be adopted for the ADCA index.

3.1. Dimension reduction

To provide a parsimonious summary of the relationships that exist between the m-sets of random vectors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$, for each fixed α, we search successively for the minimum number $l_\alpha \leq p^* = \min(p_k)$, $k = 1, \ldots, m$, of linear combinations $\{(\mathbf{a}_i^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}_i^{(m)^T}\mathbf{X}^{(m)}); i = 1, \ldots, l_\alpha\}$ having significant relationships that collectively describe the associations between the vectors. To this end, first note that $\mathcal{R}_\alpha = \mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right) = 0$ if and only if $(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})$ are mutually independent, which follows directly from (2) with $g_1 = \prod_{k=1}^m f_{\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}}$ and $g_2 = f_{(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)})}$. Second, for $1 \leq i \leq p^*$, let $\mathcal{R}_{(\alpha,i)} = \mathcal{R}_\alpha\left(\mathbf{a}_i^{(1)}, \ldots, \mathbf{a}_i^{(m)}\right) = \max \mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$ denote the PDCA index corresponding to the ith coefficient vectors, then $\mathcal{R}_{(\alpha,1)} > \cdots > \mathcal{R}_{(\alpha,w)} > 0 = \mathcal{R}_{(\alpha,w+1)} = \cdots = \mathcal{R}_{(\alpha,p^*)}$, where $w \leq p^*$; this follows from the definition of $\mathcal{R}_\alpha$ and the assumption that the maximizers are unique.

Since, $\mathcal{R}_{(\alpha,i)}$ characterizes independence and $\mathcal{R}_{(\alpha,i)} \geq \mathcal{R}_{(\alpha,j)}$ for $i \leq j$, as in [13], we can test $H_0 : \mathcal{R}_{(\alpha,i)} = 0$ vs $H_1 : \mathcal{R}_{(\alpha,i)} > 0$, for $i = 1, \ldots, p^*$, using a permutation test. For a fixed α and i, let $\widehat{\mathcal{R}}_{(\alpha,i)} = \max \widehat{\mathcal{R}}_\alpha(\widehat{\mathbf{a}}_i^{(1)}, \ldots, \widehat{\mathbf{a}}_i^{(m)})$ denote the sample PDCA index corresponding to the ith coefficient vectors. If $(\widehat{\mathbf{a}}_i^{(1)^T}\mathbf{x}^{(1)}, \ldots, \widehat{\mathbf{a}}_i^{(m)^T}\mathbf{x}^{(m)})$ are independent, then H_0 is true and a permutation of the rows of the data matrices corresponding to $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}$ should preserve the independence. However, if H_0 is false, then a permutation of the data matrices will destroy the dependence relationship and result in a smaller index value. Therefore, letting $\widehat{\mathcal{R}}^w_{(\alpha,i)}$ denote the value of the PDCA index corresponding to the wth randomly permuted dataset, the observed level of significance is given by p-value$= \sum_{w=1}^{n_p} \mathrm{I}_{\left[\widehat{\mathcal{R}}^w_{(\alpha,i)} > \widehat{\mathcal{R}}_{(\alpha,i)}\right]}/n_p$, where I is an indicator function and n_p is the number of permutations. If the p-value is small, then reject H_0 and proceed to the case $(i+1)$. If H_0 is not rejected for $i = l_\alpha + 1$, then stop further testing and conclude that there are l_α estimated variates exhibiting significant relationships. Hence, we refer to this as a dimension reduction method. An alternative, but subjective approach, is to plot the estimated variate pairs to visually determine the number of significant relationships.

3.2. Alpha selection

After determining the number of $l_{\alpha_j} = l_j$ significant relationships for each $\alpha_j, j = 1, \ldots, A$, we use the quantities computed in the above permutation tests to select an optimal value of α that parameterizes the most robust index and thus, enables PDCA to recover the maximum dependence. To compare the sample index values for different levels of α_j, we standardize $\widehat{\mathcal{R}}_{(\alpha_j,i)}$ as $\widehat{d}_{(\alpha_j,i)} = (\widehat{\mathcal{R}}_{(\alpha_j,i)} - \overline{\mathcal{R}}^*_{(\alpha_j,i)})/S^*_{\mathcal{R}_{(\alpha_j,i)}}$, $i = 1, \ldots, l_j$, where $S^*_{\mathcal{R}_{(\alpha_j,i)}}$ is the sample standard deviation, and $\overline{\mathcal{R}}^*_{(\alpha_j,i)}$ the mean, of the permuted index values $\widehat{\mathcal{R}}^*_{(\alpha_j,i)}$. Next, the value α_j that yields the largest scaled index $\widehat{d}_{(\alpha_j,i)}$ parameterizes the index that recovers the ith largest amount of dependence between the vectors. Finally, we define an estimator of the optimal tuning parameter to be $\widehat{\alpha} = \mathrm{argmax}(\sum_{i=1}^{l_1} \widehat{d}_{(\alpha_1,i)}, \sum_{i=1}^{l_2} \widehat{d}_{(\alpha_2,i)}, \ldots, \sum_{i=1}^{l_l} \widehat{d}_{(\alpha_A,i)})$. That is, we choose the level α_j that allows PDCA to collectively recover the largest amount of dependence between the sets. However, in practice the individual scaled indices can be used subjectively to determine a range of optimal values. The use of scaled index values in a permutation-based algorithm was also suggested by Witten and Tibsharani [24] to determine the best value of the tuning parameter for their sparse CCA procedure.

In the procedure for determining $\widehat{\alpha}$ we specify fixed values $\alpha = \alpha_j, j = 1, \ldots, A$, and then select an optimal value from among these and thus, cannot investigate the consistency of $\widehat{\alpha}$. This raises the question of the possibility of simultaneous consistent estimation of α and the coefficient vectors. This a challenging problem for many reasons, including the fact that α appears not only in multiple terms of the PDCA index, but also in various forms, as a power of density estimators. Consequently, studying the behavior of $\widehat{\mathcal{R}}_\alpha$ as a function of α on a continuous interval is critical. Nevertheless, this is a worthwhile pursuit for future research.

4. Numerical studies

In the simulations that follow we investigate various scenarios involving different sample sizes, combinations of linear and nonlinear relationships between sets, vectors with variables following a variety of distributions, and finally, different types of outlier contamination.

For ease in describing the ways in which outliers are generated, consider two multivariate random vectors $\mathbf{X} = (X_1, \ldots, X_p)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_q)^T$, $\mathbf{X}, \mathbf{Y} \sim F, G$. Next, suppose that the random variables Y_1 and Y_2 have the functional relationships $Y_1 = h_1(\mathbf{X}) + \epsilon_1$ and $Y_2 = h_2(\mathbf{X}) + \epsilon_2$ with $\mathbf{X}$, for some h_1 and h_2. The three types of outliers investigated in our simulation studies are:

Asymmetric: The errors, $\epsilon_j \sim \pi_1 N(0, \sigma) + (1 - \pi_1)U(0, \theta)$, $j = 1, 2$, and $\pi_1 \in [0, 1]$.

Orthogonal: The functional relationships are switched, $Y_1 = h_2(\mathbf{X}) + \epsilon_2$ and $Y_2 = h_1(\mathbf{X}) + \epsilon_1$, with probability π_2, where $\epsilon_j \sim N(0, \sigma)$, $j = 1, 2$. We term these orthogonal outliers since the functions h_1 and h_2 define orthogonal relationships between Y_1 and Y_2 in our simulations. Thus, we expect a $2(1-\pi_2)$ univariate contamination proportion.

Mixture: Asymmetric outliers are generated with probability $(1 - \pi_1)$ and orthogonal outliers with probability π_2.

In discussion, we will often refer to any of these types of outliers simply as contamination, but the specific kind will be clear from the context. The *orthogonal* outliers are created to contaminate the actual relationships, since we are in a multivariate association setting.

Similar to Section 2.5, we quantify the accuracy of the estimated coefficient vectors in the presence of varying proportions of contamination with the absolute correlations $|\rho_j| = |\rho(\widehat{\mathbf{a}}_j^{(k)^T}\mathbf{x}, \mathbf{a}_j^{(k)^T}\mathbf{x})|$, $j = 1, \ldots, l$, where l is the number of true relationships, and report the mean over the number of simulations, denoted $|\overline{\rho}|$ for brevity. Additionally, we calculate the L_2 normed distance between the subspaces spanned by $\widehat{\mathbf{a}}_j^{(k)}$ and the true coefficient vector $\mathbf{a}^{(k)}$ as $\|(\mathbf{I} - \mathbf{a}_j^{(k)}\mathbf{a}_j^{(k)^T})\widehat{\mathbf{a}}_j^{(k)}\|_2$, and report the mean, denoted $\|\bar{\cdot}\|_2$.

The simulation results are reported in the whitened scale, due to the equivariance property stated in Section 2.3. For clarity, the notations $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{Y}$ and $\mathbf{X}^{(3)} = \mathbf{Z}$ are used. The distance measure is calculated in the transformed scale by transforming the true coefficient vectors as $\boldsymbol{\Sigma}^{1/2}\mathbf{a}^{(k)}$. As discussed in Section 2.5, we set the values of the tuning parameter to be $\alpha = 0.1, \ldots, 1.0$, and include $\alpha = -1$ and 2 for the ADCA index.

4.1. Simulation

In this simulation we test the robustness of the PDCA, ADCA and GCA methods for two sets of vectors, composed of variables with a wide range of distributions and a moderate to small sample size. We consider both linear and nonlinear relationships between the sets with *orthogonal* contamination. We investigate the contamination proportions $\pi_2 = 0.05, 0.10, 0.15, 0.20$ and 0.25. The simulation can be summarized as follows:

Simulation: For a sample size of $n = 100$, we define the multivariate random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_8)^T$, and $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$, where $X_1, X_2, X_3 \sim N(0, 1)$, $X_4 \sim \chi^2_{(7)}$, $X_5 \sim t(5)$, $X_6 \sim F(3, 12)$, $X_7, X_8 \sim N(0, 1)$, $Y_3 \sim t(9)$, and $\epsilon_j \sim N(0, 1)$, $j = 1, 2$. The remaining variables are defined as

$$Y_1 = (2X_1 + X_2 + X_3)^2 + 0.5\epsilon_1 \quad \text{and} \quad Y_2 = X_2 - X_3 + 0.2\epsilon_2.$$

The true coefficient vectors are: $\mathbf{a} = (2, 1, 1, 0, 0, 0, 0, 0)^T$, $\mathbf{b} = (1, 0, 0)^T$ and $\widetilde{\mathbf{a}} = (0, 1, -1, 0.0, 0, 0, 0, 0)^T$, $\widetilde{\mathbf{b}} = (0, 1, 0)^T$.

For a dataset drawn according to the above specifications, we estimate the coefficient vectors using the three methods, PDCA, ADCA, and GCA and repeat the process 500 times. We compute estimates of the 1st and 2nd variates as $\widehat{\mathbf{a}}_j^T\mathbf{x}$ and $\widehat{\mathbf{b}}_j^T\mathbf{y}$, $j = 1, 2$, where $\mathbf{x}$ and $\mathbf{y}$ denote a random sample from $\mathbf{X}$ and $\mathbf{Y}$, respectively. The means and estimated standard errors of the absolute correlations and distances for each of the contamination proportions π_2 are given in Web Table 20 for GCA. For the PDCA and ADCA methods, the entire results are reported in Web Tables 21–24. For ease in comparing the performance of the methods, Table 2 reports the correlation results for subjectively selected values of α. In Section 4.1.1, the dimension reduction and optimal α selection methods of Section 3 are performed on a randomly selected simulated dataset for each contamination proportion.

Referencing Table 2, for PDCA when $\pi_2 = 0.05$, the $|\overline{\rho}|$ values are large, which indicate that this method is very robust at this contamination proportion in recovering both relationships. In terms of this measure, the most robustness is attained for the middle ranged values of α, with $\alpha = 0.5$ arguably the best with $|\overline{\rho}|$ values that range from 0.9817 to 0.9975. These results are repeated for the next two proportions, $\pi_2 = 0.10$ and 0.15, with the most robustness achieved on average when $\alpha = 0.5$ or 0.6. The $|\overline{\rho}|$ values range from 0.9612 to 0.9953 when $\alpha = 0.5$ and $\pi_2 = 0.10$, and from 0.9278 to 0.9871 when $\alpha = 0.6$ and $\pi_2 = 0.15$. These results demonstrate that PDCA is also very robust for moderate to high levels of orthogonal contamination. Moreover, even for the highest levels of contamination, the range of $|\overline{\rho}|$ values is from 0.8552 to 0.9859 for $\alpha = 0.7$ and $\pi_2 = 0.20$, and three are between 0.9132 and 0.9513 when $\alpha = 0.7$ and $\pi_2 = 0.25$.

In Web Table 22, the results indicate that the best performance for ADCA occurs for low to middle ranges of α. The $|\overline{\rho}|$ values are reasonable for low levels of contamination, but quickly worsen, especially in recovering the 2nd relationship, for the moderate to very high proportions of contamination, $\pi_2 = 0.15$, 0.20 and 0.25. For example, in Table 2 when $\pi_2 = 0.20$, the $|\overline{\rho}_2|$ values for the 2nd estimated variates for ADCA are 0.6707 and 0.6981; the maximum correlation is 0.7226 over all levels of α. In comparison, the PDCA values are much higher than those of ADCA, with values between 0.8790 and 0.9609 when $\alpha = 0.7$, indicating that PDCA is far more robust comparatively.

Table 2
Mean absolute correlations (standard errors), $|\overline{\rho}|$ (simulation 4.1).

Simulation 4.1					
π_2	0.05	0.10	0.15	0.20	0.25
$\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b})$	$\alpha = 0.5$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.7$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.9975 (0.0001)	0.9953 (0.0006)	0.9871 (0.0027)	0.9609 (0.0059)	0.9513 (0.0062)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9909 (0.0011)	0.9788 (0.0018)	0.9562 (0.0028)	0.9489 (0.0035)	0.9214 (0.0042)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.9817 (0.0032)	0.9612 (0.0056)	0.9278 (0.0073)	0.8790 (0.0108)	0.8141 (0.0128)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.9832 (0.0035)	0.9690 (0.0048)	0.9565 (0.0051)	0.9394 (0.0068)	0.9132 (0.0078)
$\mathcal{G}(\mathbf{a}, \mathbf{b})$					
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.9567 (0.0050)	0.9237 (0.0070)	0.9173 (0.0066)	0.8851 (0.0084)	0.8931 (0.0080)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9747 (0.0042)	0.9582 (0.0048)	0.9411 (0.0048)	0.9095 (0.0063)	0.9139 (0.0051)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.8852 (0.0106)	0.8259 (0.0121)	0.7337 (0.0141)	0.6692 (0.0141)	0.6368 (0.0140)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.9353 (0.0083)	0.8972 (0.0105)	0.8467 (0.0124)	0.8167 (0.0129)	0.8001 (0.0133)
$\mathcal{A}_\alpha(\mathbf{a}, \mathbf{b})$	$\alpha = 0.4$	$\alpha = 0.4$	$\alpha = 0.4$	$\alpha = 0.4$	$\alpha = 0.4$
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.9947 (0.0019)	0.9955 (0.0008)	0.9953 (0.0007)	0.9956 (0.0007)	0.9919 (0.0018)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9735 (0.0017)	0.9446 (0.0023)	0.9035 (0.0030)	0.8593 (0.0033)	0.8246 (0.0035)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.9338 (0.0083)	0.8793 (0.0108)	0.7622 (0.0141)	0.6707 (0.0147)	0.5890 (0.0148)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.9348 (0.0081)	0.8735 (0.0112)	0.7747 (0.0140)	0.6981 (0.0149)	0.6418 (0.0143)

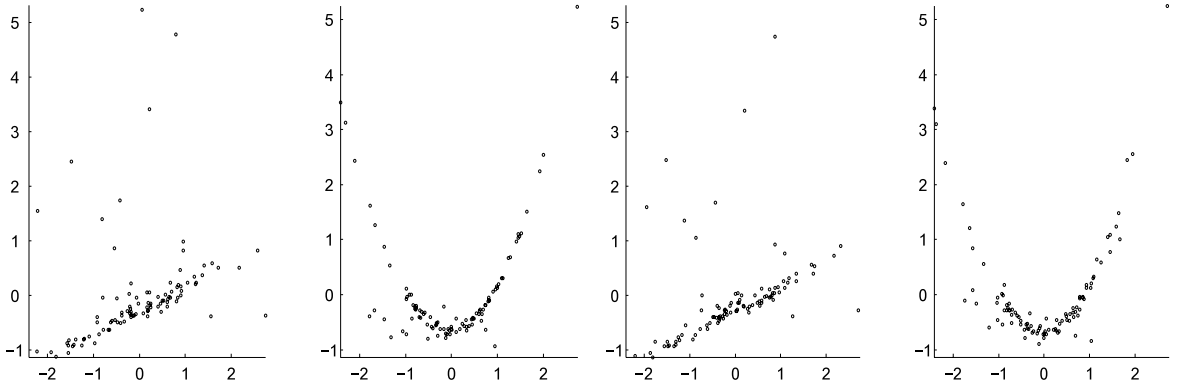

Fig. 2. $\pi_2 = 0.20$ $\mathcal{R}_{0.5}(\mathbf{a}, \mathbf{b})$ Left panel: actual variates. Right panel: estimated variates (Ex. 4.1).

In Table 2, the results show that when the contamination level is $\pi_2 = 0.05$, the $|\overline{\rho}|$ values for the GCA method are comparable to those of ADCA, but far lower than the PDCA values. As π_2 is increased, the GCA performance is far worse in comparison to PDCA and thus, as expected, is far less robust than PDCA in this scenario.

Next, to visually investigate the performance of PDCA, since it outperformed both GCA and ADCA in this simulation, for each contamination proportion π_2 we select a simulated dataset at random and plot the estimated variates $\widehat{\mathbf{a}}_1^T\mathbf{x}$ versus $\widehat{\mathbf{b}}_1^T\mathbf{y}$ and $\widehat{\mathbf{a}}_2^T\mathbf{x}$ versus $\widehat{\mathbf{b}}_2^T\mathbf{y}$ in the far right two panels of Web Figs. 7–11, respectively, for $\mathcal{R}_{0.5}(\mathbf{a}, \mathbf{b})$. In the adjacent left two panels, the true variates $\widetilde{\mathbf{a}}^T\mathbf{x}$ versus $\widetilde{\mathbf{b}}^T\mathbf{y}$ and $\mathbf{a}^T\mathbf{x}$ versus $\mathbf{b}^T\mathbf{y}$ are plotted for comparison. The plots corroborate the tabulated simulation results for these datasets and show that both relationships are recovered using PDCA for each proportion of contamination. The variate plots for the dataset with $\pi_2 = 0.20$ are also given in Fig. 2.

This simulation shows that not only is the PDCA method able to identify complex relationships between these two vectors, but is also resistant to contamination. In addition, the PDCA method firmly outperforms both GCA and ADCA.

4.1.1. Dimension reduction and alpha selection

For the above randomly selected simulated datasets, the permutation test described in Section 3 is performed for each level of α to test the significance of the recovered relationships and to determine an optimal value of α. The permutation p-values, denoted p_1 and p_2, and the individual and summed scaled index statistics, $\widehat{d}_{(\alpha,i)}$ and $\sum \widehat{d}_{(\alpha,i)}$, for the ith significantly identified relationship ($p_i < 0.05$) are reported in Table 3. The results together with the correlations between, $|\rho_1|$ and $|\rho_2|$, are given in Web Table 25.

In Table 3, for the contamination proportions $\pi_2 = 0.05$ and 0.10, the correlations between the estimated and actual variates are very high, with all corresponding p-values $= 0$ except for the second identified relationship when $\alpha = 0.1$ and $\pi_2 = 0.10$. Ignoring this exception, we conclude that all levels of α enable the PDCA index to correctly identify

Table 3
Permutation test results (simulation 4.1).

$\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b})$ Simulation 4.1										
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\pi_2 = 0.05$										
$p_1(p_2)$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
$\widehat{d}_{(\alpha,1)}$	26.7675	26.8614	26.9744	26.3700	26.0635	28.0382	31.4893	33.8380	36.1624	14.2937
$\widehat{d}_{(\alpha,2)}$	16.0925	15.8834	14.6311	14.1575	13.7319	13.1779	12.7619	13.1770	13.6133	27.9038
$\Sigma\, \widehat{d}_{(\alpha,i)}$	42.8600	42.7448	41.6055	40.5275	39.7954	41.2161	44.2512	47.0150	49.7758	42.1975
$\pi_2 = 0.10$										
$p_1(p_2)$	0 (0.4760)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
$\widehat{d}_{(\alpha,1)}$	15.0610	14.7198	15.0836	12.5239	13.3862	14.0317	14.1304	13.0141	15.6475	16.5607
$\widehat{d}_{(\alpha,2)}$	–	12.4395	13.6946	15.8601	15.6507	14.9215	13.6298	15.4721	13.7638	12.9208
$\Sigma\, \widehat{d}_{(\alpha,i)}$	15.0610	27.1593	28.7781	28.3840	29.0369	28.9532	27.7602	28.4861	29.4113	29.4815
$\pi_2 = 0.15$										
$p_1(p_2)$	0 (0.4640)	0 (0.3010)	0 (0.2750)	0 (0.2180)	0 (0.1360)	0 (0.1810)	0 (0)	0 (0)	0.0100 (0)	0.042 (0.692)
$\widehat{d}_{(\alpha,1)}$	12.8256	12.2542	12.0025	11.9642	11.6688	13.3987	8.5174	7.5098	1.6972	0.9140
$\widehat{d}_{(\alpha,2)}$	–	–	–	–	–	–	8.4085	9.5830	9.9049	–
$\Sigma\, \widehat{d}_{(\alpha,i)}$	12.8256	12.2542	12.0025	11.9642	11.6688	13.3987	16.9259	17.0928	11.6021	0.9410
$\pi_2 = 0.20$										
$p_1(p_2)$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.6740 (0)
$\widehat{d}_{(\alpha,1)}$	12.9434	9.5426	11.7453	1.0628	1.3667	1.9691	13.0730	12.0698	16.5580	–
$\widehat{d}_{(\alpha,2)}$	4.7587	12.0954	6.3138	11.2472	1.4653	1.4147	1.4822	9.5543	1.2992	12.0621
$\Sigma\, \widehat{d}_{(\alpha,i)}$	17.7021	21.6379	18.0591	21.3100	20.8319	21.3838	23.5552	21.6240	26.8571	12.0621
$\pi_2 = 0.25$										
$p_1(p_2)$	0 (0.0160)	0 (0.0580)	0 (0.0350)	0 (0.1540)	0 (0.1970)	0 (0.1220)	0 (0.0170)	0 (0.0140)	0.5930 (0)	0.5480 (0)
$\widehat{d}_{(\alpha,1)}$	10.8078	10.2371	9.6778	8.4413	8.5870	13.4795	17.2207	14.9173	–	–
$\widehat{d}_{(\alpha,2)}$	2.4069	–	2.0379	–	–	–	2.5233	2.4959	8.5017	8.2665
$\Sigma\, \widehat{d}_{(\alpha,i)}$	13.2148	10.2371	11.7158	8.4413	8.5870	13.4795	19.7440	17.4133	8.5017	8.2665

the two defined relationships. The optimal level of the tuning parameter is $\alpha = 0.9$ when $\pi_2 = 0.05$, and $\alpha = 0.9$ or 1.0 when $\pi_2 = 0.10$, since these values generate the largest $\sum_{i=1}^{2} \widehat{d}_{(\alpha,i)}$ statistics. When the contamination level is increased to $\pi_2 = 0.15$, the values of α that correctly determine the associations and parameterize the most robust PDCA index are $\alpha = 0.7, 0.8$ and 0.9. However, $\sum_{i=1}^{2} \widehat{d}_{(0.7,i)} = 16.9259$ and $\sum_{i=1}^{2} \widehat{d}_{(0.8,i)} = 17.0928$ are much larger than $\sum_{i=1}^{2} \widehat{d}_{(0.9,i)} = 11.6021$, indicating that $\alpha = 0.7$ or 0.8 are more optimal. This conclusion is supported by noting that both correlations, $|\rho_1|$ and $|\rho_2|$, are near 0.99 when $\alpha = 0.7$ or 0.8. Next, for $\pi_2 = 0.20$, the PDCA index identifies the relationships with p-values $= 0$ for nearly every α. Again, as was the case when $\pi_2 = 0.05$, $\alpha = 0.9$ produces the largest $\sum_{i=1}^{2} \widehat{d}_{(\alpha,i)}$ statistic, with all other levels generating comparable numbers except when $\alpha = 0.1$ or 1.0. Finally, when the mean contamination is raised to $\pi_2 = 0.25$, the values $\alpha = 0.7$ and 0.8 recover the relationships with the lowest p-values and have the largest $\sum_{i=1}^{2} \widehat{d}_{(\alpha,i)}$ statistics. Note that, for each proportion of contamination, $\alpha = 0.7$ or 0.8 were nearly always the most optimal values of α and thus, we infer that the middle to upper range of α values for these selected datasets are likely to provide the most robust PDCA estimates of the defined relationships.

4.1.2. Extension of simulation 4.1

Under the same vector parameterizations as above, we create a *mixture* contamination by independently adding *asymmetric* outliers from a $U(0, 30)$ distribution with probability $1 - \pi_1$. We investigate the pairs, $(\pi_1, \pi_2) = (0.95, 0.05)$, $(0.975, 0.05)$ and $(0.975, 0)$, which give a 15%, 10% and 5% univariate contamination rate and up to a 15%, 10% and 5% multivariate contamination rate. The results are reported in Table 4 for subjectively selected values of α; the entire results are given in Web Appendix K. Note that, the case $(0.975, 0)$ contains only asymmetric outliers. Simulations with only asymmetric outliers are performed in the Web Appendix.

The first two columns of Table 4 give the $|\overline{\rho}|$ and $\|\dot{\cdot}\|_2$ values for the $(\pi_1, \pi_2) = (0.95, 0.05)$ case. For $\alpha = 0.6$, PDCA has $|\overline{\rho}|$ values above 0.99 and 0.95 for the 1st and 2nd estimated variates, respectively. For GCA, the correlations are markedly lower than those of PDCA for both variates. Next, the $|\overline{\rho}_1|$ values for the 1st variates using ADCA when $\alpha = 0.3$ are 0.9914 and 0.9579, respectively, which are lower than PDCA values, but higher than GCA values, as might be expected. However, for the 2nd variates, the $|\overline{\rho}_2|$ values are notably lower with values 0.7641 and 0.7949. These results are repeated for all values of α for both PDCA and ADCA in Web Tables 49 and 50, with the best performance for both methods in general attained for the

Table 4
Mean absolute correlations, $|\overline{\rho}|$, and distances, $\|\cdot\|_2$, (standard errors) (simulation 4.1.2).

Simulation 4.1.2						
(π_1, π_2)	(0.95, 0.05)		(0.975, 0.05)		(0.975, 0)	
	$\|\overline{\rho}\|$	$\|\cdot\|_2$	$\|\overline{\rho}\|$	$\|\cdot\|_2$	$\|\overline{\rho}\|$	$\|\cdot\|_2$
$\mathcal{R}_\alpha(\mathbf{a}, \mathbf{b})$	$\alpha = 0.6$		$\alpha = 0.6$		$\alpha = 0.6$	
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.9908 (0.0010)	0.1130 (0.0031)	0.9939 (0.0004)	0.0970 (0.0023)	0.9964 (0.0002)	0.0782 (0.0015)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9927 (0.0011)	0.0711 (0.0042)	0.9931 (0.0009)	0.0717 (0.0040)	0.9993 (0.0000)	0.0314 (0.0009)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.9529 (0.0062)	0.2047 (0.0078)	0.9590 (0.0062)	0.1808 (0.0075)	0.9825 (0.0037)	0.1281 (0.0048)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.9737 (0.0048)	0.1329 (0.0067)	0.9755 (0.0048)	0.1269 (0.0064)	0.9879 (0.0030)	0.0933 (0.0046)
$\mathcal{G}(\mathbf{a}, \mathbf{b})$						
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.8953 (0.0082)	0.3271 (0.0108)	0.9297 (0.0062)	0.2743 (0.0092)	0.9407 (0.0070)	0.2183 (0.0094)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9633 (0.0053)	0.1719 (0.0078)	0.9769 (0.0038)	0.1382 (0.0063)	0.9740 (0.0055)	0.1112 (0.0071)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.8048 (0.0124)	0.4372 (0.0130)	0.8400 (0.0117)	0.3828 (0.0127)	0.9049 (0.0096)	0.2748 (0.0110)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.9002 (0.0101)	0.2613 (0.0119)	0.9173 (0.0094)	0.2303 (0.0111)	0.9468 (0.0075)	0.1863 (0.0090)
$\mathcal{A}_\alpha(\mathbf{a}, \mathbf{b})$	$\alpha = 0.3$		$\alpha = 0.3$		$\alpha = 0.3$	
$\widehat{\mathbf{a}}_1^T\mathbf{x}$	0.9914 (0.0011)	0.1057 (0.0033)	0.9924 (0.0017)	0.0904 (0.0033)	0.9934 (0.0015)	0.0786 (0.0034)
$\widehat{\mathbf{b}}_1^T\mathbf{y}$	0.9579 (0.0031)	0.2192 (0.0077)	0.9599 (0.0028)	0.2134 (0.0076)	0.9811 (0.0023)	0.1148 (0.0066)
$\widehat{\mathbf{a}}_2^T\mathbf{x}$	0.7641 (0.0142)	0.4500 (0.0150)	0.8605 (0.0116)	0.3292 (0.0130)	0.9374 (0.0080)	0.2031 (0.0098)
$\widehat{\mathbf{b}}_2^T\mathbf{y}$	0.7949 (0.0129)	0.4115 (0.0152)	0.8659 (0.0110)	0.3208 (0.0132)	0.9383 (0.0074)	0.2005 (0.0102)

middle ranged values $\alpha = 0.3$ to 0.8. Thus, once again, PDCA is clearly more robust in recovering the defined relationships at this level of mixture contamination.

In the middle column the asymmetric contamination rate is reduced to $1 - \pi_1 = 0.025$ and the same pattern of results for the previously discussed cases with $1 - \pi_1 = 0.5$ is observed. Note that, in the far right column with the asymmetric contamination removed, all methods improve, with PDCA remaining the most robust and GCA the least. The results based on the $\|\cdot\|_2$ values are consistent with those based on the $|\overline{\rho}|$ values.

4.2. Additional simulations

Extensive simulation studies were conducted to investigate the performance of our proposed methods. In this section, we briefly summarize the additional simulations found in Web Appendices E–M.

The simulations of Web Appendices E and F, repeat the two-set scenario and the multiple set simulation with heavy-tailed distributed variables in [12], respectively. Web Appendices G and K give the entire set of tables and graphs for Simulations 4.1 and 4.1.2, respectively. The remaining Web appendices, H, I, J, and L, give simulations that involve a linear and nonlinear (quartic) relationships between sets with varying types and proportions of contamination. Finally, a three set simulation with complicated relationships between the vectors that are composed of 13 non-normal variables among the 15 considered using PDCA is performed in Web Appendix M in the presence of *asymmetric* contamination. Note that, in all simulations PDCA substantially outperforms ADCA and GCA and thus, only the results are presented and comparative discussions omitted.

5. Data analysis

The following real data analyses are performed in the whitened scale. However, the notations $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{Y}$ and $\mathbf{X}^{(3)} = \mathbf{Z}$ are maintained for clarity. Given the existence of extreme observations in the two datasets, and the improved performance of PDCA over ADCA in simulation, we use PDCA in Section 5.1 for brevity. Due to the superior performance over all methods, we focus on PDCA exclusively in Section 5.2.

5.1. Consumer expenditure (CE) survey

This dataset has been derived from the Quarterly Interview Survey of the Consumer Expenditure (CE) survey undertaken by the US Department of Labor, Bureau of Labor Statistics. This nationally representative survey is a sample of nearly 5000 households, with each household being interviewed five times, the first time to gather basic data about the household, and four other times at quarterly intervals to gather data on expenditures. The households in the present dataset entered the quarterly survey at the beginning of 1995 and the data give the total expenditure by category over the 1995 calendar year. We use the same eight variables and $n_1 = 866$ observations used by Hubert et al. [11] to illustrate their robust Principal Component Analysis (PCA) method for datasets with outliers.

For our analysis we select four variables to make up the $\mathbf{X}$ vector; X_1 = total household expenditure (EXP), X_2 food consumed at home (FDHO), X_3 = housing and household equipment (SHEL), X_4 = telephone services (TELE), and the

remaining the **Y** vector; Y_1 = food consumed away from home (FDAW), Y_2 = clothing (CLOT), Y_3 = health care (HEAL), and Y_4 = entertainment (ENT). The vector **X** is termed the *household expenditure* vector. The vector **Y** is viewed as measure of socioeconomic status and thus, termed the *socioeconomic* vector.

An interesting question is: "What type of multivariate association, if any, exists between these vectors, and what effect do outliers have on recovering the relationships?" To this end, we start by using the kurtosis-based method developed by Peña and Prieto [18] to identify multivariate outliers. The motivation for their method stems from noting that symmetric contamination of univariate data increases the kurtosis coefficient, while asymmetric contamination increases the coefficient when the proportion of outliers is small, but may decrease it if the proportion is large. Therefore, to find multivariate outliers they suggest projecting the dataset into one dimensional orthogonal subspaces that both maximize and minimize their kurtosis index. Next, a scaled median type distance is calculated for each projected observation and labeled an outlier if the value exceeds a critical value determined through simulation. Combining the random vectors **X** and **Y** into one vector, $\mathbf{U} = (\mathbf{X}, \mathbf{Y})^T$, and implementing their algorithm identifies 255 suspected outliers. To determine the most extreme of these outliers, we use the Mahalanobis distance calculated in their procedure and label a sample point to be extreme if $(\mathbf{u}_j - \bar{\mathbf{u}})^T \tilde{\mathbf{S}}^{-1} (\mathbf{u}_j - \bar{\mathbf{u}}) > \chi^2_{p+q}(0.95)$. There were 15 such observations in the dataset and are referred to as the *outliers* hereafter.

To study the effect extreme observations have in recovering the relationships between **X** and **Y**, we begin by considering the full dataset, denoted $\mathbf{D}_{n_1}$, and remove the 15 *outliers* to produce a second dataset, denoted $\mathbf{D}^*_{n_1}$, and then study the change in the estimated coefficient vectors computed from each dataset. Next, a second dataset $\mathbf{D}_{n_2}$ of size $n_2 = 600$ is created by combining the 15 *outliers* with a dataset $\mathbf{D}^*_{n_2}$ containing 585 randomly selected sample points. This is repeated to produce two more datasets, $\mathbf{D}_{n_3}$ and $\mathbf{D}_{n_4}$, of size $n_3 = 400$ and $n_4 = 100$, that have at least a 3.75% and 15% contamination rate, respectively.

Data analysis:

For the full dataset $\mathbf{D}_{n_1}$, the 1st estimated variates using PDCA, GCA, and CCA identify a linear relationship between the *household expenditure* and *socioeconomic* vectors. This can be seen in the plot of the 1st PDCA variates for $\mathbf{D}_{n_1}$ in the left panel of Web Fig. 28; the right panel is the analogous plot for $\mathbf{D}^*_{n_1}$. The loadings for the coefficient vectors are in the left quadrant and leftmost column for each method in Table 5. For PDCA when $\alpha = 0.1$, the coefficient vector for **X** weighs heaviest on X_1, $\widehat{\mathbf{a}}_1 = (1.0, 0.0025, 0.0057, 0.0007)^T$, and for **Y** are a weighted average, to some degree, of all the variables, $\widehat{\mathbf{b}}_1 = (0.5357, 0.4571, 0.2430, 0.6671)^T$. A similar interpretation holds for the CCA and GCA methods. This value of α was selected based on the robustness analysis performed below. It seems reasonable to expect that the total household expenditure variable X_1 would be strongly associated with at least one, or all, of the four *socioeconomic* variables. For example, a family with a low household expenditure naturally spends less on clothing, health care, entertainment, and consumes less food away from home. Plots of the 2nd estimated variates indicated that at most a weak relationship is recovered and thus, focus only on the 1st estimated directions hereafter.

Robustness analysis:

As in Section 2.5 the distance $L^*_{(2,n_i)} = \|(\mathbf{I} - \widehat{\mathbf{a}}_1^{(k)} \widehat{\mathbf{a}}_1^{(k)^T}) \widehat{\mathbf{a}}_1^{*(k)}\|_2$, $k = 1, 2$, where $\widehat{\mathbf{a}}_1^{(k)}$ and $\widehat{\mathbf{a}}_1^{*(k)}$ are the estimated 1st coefficient vectors from $\mathbf{D}_{n_i}$ and $\mathbf{D}^*_{n_i}$, respectively, is used to quantify the effect the outliers have in recovering the relationship. For each of the methods PDCA, CCA and GCA, the top left quadrant of Table 5 gives the loadings for $\widehat{\mathbf{a}}_1$ and $\widehat{\mathbf{a}}^*_1$ for $\mathbf{D}_{n_1}$ and $\mathbf{D}^*_{n_1}$ in the left most column and the adjacent right column, respectively. For PDCA when $\alpha = 0.1$, the L_2 norm distance between the spanned subspaces of $\widehat{\mathbf{a}}_1$ and $\widehat{\mathbf{a}}^*_1$ is $L^*_{(2,n_1)} = 0.0149$, and for CCA and GCA it is 0.0655 and 0.0363, respectively. In the next row the relative change in the distance is reported with PDCA as the standard, $\Delta = L^*_{(2,\cdot)}/L^*_{(2,\cdot)\mathcal{R}}$. The relative change using CCA is $\Delta = 0.0655/0.0149 = 4.396$ times larger than PDCA, and analogously, 2.436 times more using GCA. In the bottom left quadrant the results for the relative change in distance between $\widehat{\mathbf{b}}_1$ and $\widehat{\mathbf{b}}^*_1$ are 1.907 and 1.466 times larger using CCA and GCA, respectively. Similar results are obtained when this analysis is repeated on the datasets with increased contamination, $\mathbf{D}_{n_2}$ and $\mathbf{D}_{n_3}$. The right quadrants of Table 5 give the results for the dataset with the highest level of contamination and notably, the relative change between $\widehat{\mathbf{a}}_1$ and $\widehat{\mathbf{a}}^*_1$ is almost 6 times more using CCA and 10 times more with GCA.

In conclusion, the results quantifying the change in the estimated loadings as the contamination proportion is increased show, as did the simulations, that PDCA is more robust than GCA and the moment based method CCA.

5.2. Baseball salaries

The dataset analyzed in this section has been well-studied in the literature from a regression perspective and was initially presented as a data analysis exposition sponsored by the section on statistics and graphics of the American Statistical Association in 1988. Fifteen groups performed statistical analyses to answer the question "are players paid according to their performance?" Hoaglin and Velleman [9] wrote a review of the successes and failures of these analyses, and commented on the authors' considerations of the effect outliers and extreme values had on their results. More recently, Xia et al. [25] reanalyzed this dataset to illustrate their effective dimension reduction (EDR) method. Important in their analysis was that they first identified and removed outliers, not only to improve their results, but also to argue that the deleted observations were indeed influential points.

Table 5
1st estimated coefficient vector comparison (example 5.1).

	$n_1 = 866$ $\mathbf{D}_{n_1}\|\mathbf{D}^*_{n_1}$						$n_4 = 100$ $\mathbf{D}_{n_4}\|\mathbf{D}^*_{n_4}$					
	$\mathcal{R}_{0.1}(\mathbf{a},\mathbf{b})$		CCA		GCA		$\mathcal{R}_{0.5}(\mathbf{a},\mathbf{b})$		CCA		GCA	
$\widehat{\mathbf{a}}_{11}$	1.000	0.9999	0.9981	0.9955	0.9176	0.9062	0.9924	0.9934	−0.9519	0.9882	0.9598	0.8951
$\widehat{\mathbf{a}}_{12}$	0.0025	−0.0011	0.0466	0.0887	0.2461	0.2793	−0.0311	−0.0560	0.0240	0.0558	0.2023	0.3040
$\widehat{\mathbf{a}}_{13}$	0.0057	−0.0058	0.0312	−0.0237	0.2504	0.2591	−0.0914	−0.0634	0.2714	−0.0678	−0.1768	0.1387
$\widehat{\mathbf{a}}_{14}$	0.0007	−0.0082	−0.0238	0.0223	0.1866	0.1837	−0.0764	−0.0770	0.1399	−0.1255	0.0812	0.2951
$L^*_{(2,\cdot)}$	0.0149		0.0655		0.0363		0.0376		0.2208		0.3917	
Δ	1		4.396		2.436		1		5.872		10.420	
$\widehat{\mathbf{b}}_{11}$	0.5357	0.5430	0.5535	0.5004	0.5543	0.5138	0.4830	0.5286	−0.5149	0.3463	0.3881	0.2991
$\widehat{\mathbf{b}}_{12}$	0.4571	0.4641	0.5230	0.4837	0.5410	0.5236	0.5720	0.4998	−0.4121	0.5643	0.4645	0.6958
$\widehat{\mathbf{b}}_{13}$	0.2430	0.2779	0.3075	0.2981	0.2488	0.2678	0.3445	0.4322	−0.3576	0.5255	0.2941	0.4665
$\widehat{\mathbf{b}}_{14}$	0.6671	0.6423	0.5705	0.6533	0.5815	0.6247	0.5664	0.5329	−0.6611	0.5342	0.7397	0.4569
$L^*_{(2,\cdot)}$	0.0440		0.0839		0.0645		0.1266		0.3060		0.4047	
Δ	1		1.907		1.466		1		2.417		3.197	

We consider two random vectors composed of the predictor variables used by Xia et al. [25] and the response variable, *annual salary*. First, we define the random vector $\mathbf{X} = (X_1, \ldots, X_9)^T$ to consist of the variables number of: times at bat X_1, hits X_2, home runs X_3, runs X_4, runs batted in X_5, walks X_6, errors X_7, putouts X_8, and assists X_9, in 1986. We term $\mathbf{X}$ the *1986 performance* vector. The second vector $\mathbf{Y} = (Y_1, \ldots, Y_7)^T$ is composed of the variables number of: times at bat Y_1, hits Y_2, home runs Y_3, runs Y_4, runs batted in Y_5, walks Y_6, and years in the major leagues Y_7 (in 1986), during their entire career up to 1986. We label $\mathbf{Y}$ the *career performance* vector. Finally, we take the variable Z to be the log of the salary in 1986 and refer to this variable simply as the *annual salary*.

Unlike previous analyses, we take an entirely different approach in analyzing this dataset using our PDCA multiple set index in (4), which studies the joint association between the *annual salary*, *1986 performance* and *career performance* variables without deleting any extreme observations. Although not typical in multivariate association studies, here Z is univariate, so studying its association with the vectors $\mathbf{X}$ and $\mathbf{Y}$ naturally classifies it as the response. Therefore, consistent with the original goals of the data exposition, in the analysis below we treat Z as the response. However, different from regression-type analyses, our method studies the association between two separate sets of predictors simultaneously with the response.

Using the methods in Section 3, the permutation p-values $= 0$ for all values of α, indicating that significant relationships exist between Z, $\mathbf{X}$ and $\mathbf{Y}$, and that $\alpha = 0.4$ parameterizes the most robust index to recover them, since $\widehat{d}_{(0.4,1)} = 29.78$ is the largest; see Web Table 64. Hereafter, only results for $\alpha = 0.4$ are presented. The variate plot of Z versus $\widehat{\mathbf{a}}_1^T\mathbf{X}$ in the left panel of Fig. 1 indicates that a weak linear relationship exists between *annual salary* and the *1986 performance* vector when considered simultaneously with the *career performance* variables. However, when considered simultaneously with the *1986 performance* variables, the plot of Z versus $\widehat{\mathbf{b}}_1^T\mathbf{Y}$ in the right panel of Fig. 1 shows that a very strong nonlinear relationship exists between *annual salary* and the *career performance* vector.

Consider the loadings for the *1986 performance* vector: $\widehat{\mathbf{a}}_1 = (0.5812, 0.2963, 0.3230, 0.1747, 0.5088, 0.3859, 0.0312, 0.1094, -0.1372)^T$. Strong weights are given to the batting variables times at bat X_1, hits X_2, and home runs X_3, with coefficients 0.5812, 0.2963 and 0.3230, respectively. Next, runs batted in X_5, and walks X_6, are given the similarly proportioned weights 0.5088 and 0.3859, respectively. For the fielding variables, errors X_7 is near 0 and the coefficients 0.1094 and −0.1372 for putouts X_8, and assists X_9, effectively negate their influence; see Web Appendix O for a detailed explanation. Therefore, we conclude that the coefficients for $\mathbf{X}$ provide an unequally weighted average of the batting variables, where times at bat X_1, and runs batted in X_5, have the most influence on the linear relationship with *annual salary*.

Next, consider the coefficients for the *career performance* vector: $\widehat{\mathbf{b}}_1 = (0.9312, 0.2514, 0.0594, 0.1154, 0.1264, 0.1872, 0.0199)^T$. The career total times at bat Y_1 has the largest mass 0.9312, followed by 0.2514 for career hits Y_2, and then runs Y_4, runs batted in Y_5, and walks Y_6, are given similar low weights. Of note, home runs Y_3, and the variable number of years in the league Y_7, are weighted near 0. Therefore, we assert that the loadings for the *career* vector are disproportionate and the weight placed on career times at bat Y_1 drives the nonlinear relationship with *annual salary*.

Returning to the variate plots in Fig. 1, the left panel shows that an increase in the *1986 performance* variate is most associated with a linear trend in *annual salary*. With our previous interpretation of the coefficients for the *1986 performance* vector, this means in general that as the batting variables times at X_1, hits X_2, runs batted in X_5, and walks X_6 increase, *annual salary* improves in a weakly linear manner. Next, the right panel plot shows that *annual salary* is smallest when the *career performance* variate is near −1 and increases in a sharp linear fashion with little variation as the variate approaches 0, which likely corresponds to players in the early to middle part of their careers. Finally, as the variate gets close to and exceeds 1, there is a plateau and increased variation with a slight quadratic downturn in *annual salary*, which is likely attributable to players who have been in the league a number of years. These players have a had a large number of times at bat Y_1, and hits Y_2, but there is substantial variation in how much they make.

Note that, the plots in Fig. 1 are similar in nature to those in Fig. 6 of Xia et al. [25] using MAVE. However, there are many striking differences between our results and theirs. First, their plots are based on revised estimates obtained after removing seven outliers, whereas our plots are the result of using PDCA on the entire dataset. Second, their dimension reduction

regression analysis studies the influence of $\mathbf{U} = (\mathbf{X}, \mathbf{Y})^T$ on Z through the two directions, $\widehat{\beta}_1^T\mathbf{U}$ and $\widehat{\beta}_2^T\mathbf{U}$, whereas ours is a simultaneous association study of Z, $\mathbf{X}$, and $\mathbf{Y}$ through $\widehat{\mathbf{a}}_1^T\mathbf{X}$ and $\widehat{\mathbf{b}}_1^T\mathbf{Y}$. Consequently, our analysis reveals the joint relationships between the *annual salary*, *1986 performance* and *career performance* variables. Furthermore, our variate plots separately identify the influence of 1986 and career performance on *annual salary*. However, such a clear division is neither apparent in Fig.6 of Xia et al. [25] nor recoverable through the coefficient vectors $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Thus, our analysis more definitively answers the main interest stated in the data exposition sponsored by the ASA, which shows the versatility and potential of the PDCA method.

6. Discussion

We have introduced two new families of multivariate association measures, PDCA and ADCA, based on density divergences. The families are indexed by a tuning parameter α, which plays a critical role in obtaining an optimally robust association index that is least sensitive to outliers and recovers the most amount of dependence between multiple sets. We have also shown the consistency of the PDCA and ADCA estimates. Furthermore, through extensive simulations, it is shown that PDCA is superior to ADCA, GCA and CCA in recovering the true associations despite the presence of gross-error contamination. PDCA maintains its superior performance in the analysis of two socioeconomic datasets that are known to contain outliers. Moreover, the PDCA method is appealing because it does not require preliminary outlier detection to extract associations between multiple sets. This raises the possibility of theoretically studying the robustness of our method through influence functions and such studies will be taken up elsewhere.

Acknowledgments

We would like to thank the two referees for a careful reading of the article and insightful comments that greatly improved many parts of the paper. T.N. Sriram was supported in part by a National Security Agency grant H98230-11-1-0188.

Appendix A

A.1. Consistency

Let $\mathbf{V}_i$ be a sequence of d-dimensional random vectors with distribution function F and Lebesgue measurable density f. Define the kernel density estimate of f as:

$$\widehat{f}_n(\mathbf{u}) = \frac{1}{na_n^d}\sum_{i=1}^{n} \mathrm{K}\left(\frac{\mathbf{v}-\mathbf{V}_i}{a_n^d}\right), \quad \text{for } \mathbf{v} \in \mathbb{R}^d,$$

where $\mathrm{K} : \mathbb{R}^d \to \mathbb{R}^+$ is a probability density on $\mathbb{R}^d$, uniformly for $\|\mathbf{v}\| \to \infty$ and where $a_n > 0$ and $\lim_{n\to\infty} a_n = 0$. The following lemma follows as direct application of Theorem 1-m of Kiefer [15] and Theorem 1 of Ruschendorf [21]; hence, we omit the proof. Here, for notational convenience, $f(\mathbf{a}^{(k)^T}\mathbf{x}^{(k)})$ denotes the density of $\mathbf{a}^{(k)^T}\mathbf{X}^{(k)}$ for $k = 1, \ldots, m$ and $f\left(\mathbf{a}^{(1)^T}\mathbf{x}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}^{(m)}\right)$ denotes the density of $\left(\mathbf{a}^{(1)^T}\mathbf{X}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{X}^{(m)}\right)$.

Lemma 1. *Let* $\{(\mathbf{X}_i^{(1)}, \ldots, \mathbf{X}_i^{(m)})\}$, $i = 1, \ldots, n$, *be i.i.d, and*

$$\sum_{n=1}^{\infty} e^{-\gamma n a_n^{2d}} < \infty, \quad \textit{for all } \gamma > 0.$$

Let K *be of bounded variation and, for each* $k = 1, \ldots, m$, *let*

$f(\mathbf{a}^{(k)^T}\mathbf{x}^{(k)})$ *be uniformly continuous in* $\mathbf{a}^{(k)}$ *and* $\mathbf{x}^{(k)}$,

$f(\mathbf{a}^{(1)^T}\mathbf{x}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}^{(m)})$ *be uniformly continuous in* $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}$ *and* $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$.

Under these conditions we have for each $k = 1, \ldots, m$, *as* $n \to \infty$

$$\sup_{\mathbf{a}^{(k)},\, \mathbf{x}^{(k)} \in \mathbb{R}^{p_k}} \left|\widehat{f}_n(\mathbf{a}^{(k)^T}\mathbf{x}^{(k)}) - f(\mathbf{a}^{(k)^T}\mathbf{x}^{(k)})\right| \to 0 \quad \textit{almost surely (a.s.)}$$

$$\sup_{\mathbf{a}^{(k)},\, \mathbf{x}^{(k)} \in \mathbb{R}^{p_k},\, \forall k} \left|\widehat{f}_n\left(\mathbf{a}^{(1)^T}\mathbf{x}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}^{(m)}\right) - f\left(\mathbf{a}^{(1)^T}\mathbf{x}^{(1)}, \ldots, \mathbf{a}^{(m)^T}\mathbf{x}^{(m)}\right)\right| \to 0 \quad \textit{a.s.}$$

Proof of Theorem 1. Assume the conditions of Lemma 1. Suppose $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)})$ does not converge to $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$ *a.s.* Then, there exists a subsequence (still denoted by n) and vectors $(\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)})$ satisfying the constraints in Section 2.1 such that $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) \to (\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)})$ *a.s.* and $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)}) \neq (\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)})$. Therefore, for any $\epsilon > 0$ and n large enough, we have from Lemma 1 that for each $k = 1, \ldots, m$

$$\begin{aligned}
\widehat{f}_n\left(\widehat{\mathbf{a}}_1^{(k)^T}\mathbf{x}_i^{(k)}\right) &= f\left(\widehat{\mathbf{a}}_1^{(k)^T}\mathbf{x}_i^{(k)}\right) + \Delta_{1,i} = f\left(\mathbf{a}_0^{(k)^T}\mathbf{x}_i^{(k)}\right) + \delta_{k,i},\\
\widehat{f}_n\left(\widehat{\mathbf{a}}_1^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)^T}\mathbf{x}_i^{(m)}\right) &= f\left(\widehat{\mathbf{a}}_1^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)^T}\mathbf{x}_i^{(m)}\right) + \Delta_{(m+1),i}\\
&= f\left(\mathbf{a}_0^{(1)^T}\mathbf{x}_i^{(1)}, \ldots, \mathbf{a}_0^{(m)^T}\mathbf{x}_i^{(m)}\right) + \delta_{(m+1),i},
\end{aligned}$$

such that $|\delta_{j,i}| < \epsilon$ for all i and $j = 1, \ldots, m+1$. Here, the first set of equalities follow from the conclusion of Lemma 1 and the second set of equalities follow from the assumed uniformly continuity in Lemma 1. Using these and algebraic manipulations, it can be shown that $\widehat{\mathcal{R}}_\alpha\left(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}\right) = \mathcal{R}_\alpha\left(\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)}\right) + o(1)$. Note that, by assumption, $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) = \operatorname{argmax} \widehat{\mathcal{R}}_\alpha(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)})$ and $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)}) = \operatorname{argmax} \mathcal{R}_\alpha\left(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}\right)$. Therefore, $\widehat{\mathcal{R}}_\alpha(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) \geq \widehat{\mathcal{R}}_\alpha(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$, which implies

$$\mathcal{R}_\alpha(\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)}) = \lim_{n\to\infty} \widehat{\mathcal{R}}_\alpha(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) \geq \lim_{n\to\infty} \widehat{\mathcal{R}}_\alpha(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)}) = \mathcal{R}_\alpha(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$$

and also, by assumption, $\mathcal{R}_\alpha(\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)}) \leq \mathcal{R}_\alpha(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$. Thus, $\mathcal{R}_\alpha(\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(m)}) = \mathcal{R}_\alpha(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$, which contradicts the uniqueness of $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$. Therefore, as $n \to \infty$ $(\widehat{\mathbf{a}}_1^{(1)}, \ldots, \widehat{\mathbf{a}}_1^{(m)}) \to (\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_1^{(m)})$ almost surely. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2013.03.004.

References

[1] A. Basu, I. Harris, N. Hjort, M. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (1998) 549–599.
[2] R. Beran, Minimum Hellinger distance estimates for parametric models, The Annals of Statistics 5 (1977) 445–463.
[3] J.A. Branco, C. Croux, P. Filzmoser, M.R. Oliveira, Robust canonical correlations: a comparative study, Computational Statistics 20 (2005) 203–229.
[4] A.J. Cannon, W.W. Hsieh, Robust nonlinear canonical correlation analysis: application to seasonal climate forecasting, Nonlinear Processes in Geophysics 15 (2008) 221–232.
[5] H.S. Chen, K. Lai, Z. Ying, Goodness-of-fit tests and minimum power divergence estimators for survival data, Statistica Sinica 14 (2004) 231–248.
[6] N. Cressie, C. Read, Multinomial goodness-of-fit tests, Journal of the Royal Statistical Society: Series B 46 (1984) 440–464.
[7] C. Croux, C. Dehon, Robust canonical correlations using high breakdown scatter matrices, Preprint, Universitie Libre de Bruxelles, 1999.
[8] C. Dehon, P. Filzmoser, C. Croux, Robust methods for canonical correlation analysis, Available online, 2008.
[9] D. Hoaglin, P. Velleman, A critical look as some analyses of major league baseball salaries, The American Statistician 49 (1995) 277–285.
[10] H. Hotelling, Relations between two sets of variables, Biometrika 58 (1936) 433–451.
[11] M. Hubert, P. Rouseeuw, T. Verdonck, Robust PCA for skewed data and its outlier map, Journal of Computational Statistics and Data Analysis 53 (2009) 2264–2274.
[12] R. Iaci, T.N. Sriram, X. Yin, Multivariate association and dimension reduction: a generalization of canonical correlation analysis, Biometrics 66 (2010) 1107–1118.
[13] R. Iaci, X. Yin, T.N. Sriram, C.P. Klingenberg, An informational measure of association and dimension reduction for multiple sets and groups with applications in morphometric analysis, Journal of the American Statistical Association 103 (2008) 1166–1176.
[14] G. Karnel, Robust canonical correlation and correspondence analysis, in: Conference Proceedings on the Frontiers of Statistical Scientific Theory & Industrial Applications, vol. 2, 1991.
[15] J. Kiefer, On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, Pacific Journal of Mathematics 11 (1961) 649–659.
[16] D. Meyer, Diagnostics for canonical correlation, in: Research Letters in the Information and Mathematical Sciences, Vol. 4, 2003, pp. 79–89.
[17] J. Nocedal, S. Wright, Numerical Optimization, in: Springer Series in Operations Research, Springer, New York, 1999.
[18] D. Peña, F.J. Prieto, Multivariate outlier detection and robust covariance matrix estimation, Technometrics 43 (3) (2001) 286–300.
[19] T.R.C. Read, N. Cressie, Goodness-of-Fit Statistics for Discrete Multivariate Data, Vol. 57, Springer, New York, 1988, pp. 237–259.
[20] M. Romanazzi, Influence in canonical correlation analysis, Psychometrika 57 (1992) 237–259.
[21] L. Ruschendorf, Consistency of estimators for multivariate density functions and for the mode, Sankhyā, Series A 39 (1977) 243–250.
[22] D.W. Scott, Multivariate Density Estimation: Theory, Practice and Visualization, Wiley, New York, 1992.
[23] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, New York, 1986.
[24] D. Witten, R. Tibsharani, Extensions of sparse canonical correlation analysis, with applications to genomic data, Statistical Applications in Genetics and Molecular Biology 8 (2009) Article 28.
[25] Y. Xia, H. Tong, W.K. Li, L.X. Zhu, An adaptive estimation of dimension reduction, Journal of the Royal Statistical Society: Series B 64 (2002) 363–410.
[26] X. Yin, Canonical correlation analysis based on information theory, Journal of Multivariate Analysis 91 (2004) 161–176.
[27] X. Yin, R.D. Cook, Direction estimation in single-index regressions, Biometrika 92 (2) (2005) 371–384.
[28] X. Yin, T.N. Sriram, Common canonical variates for independent groups using information theory, Statistica Sinica 18 (2008) 335–353.