Statistica Sinica **18**(2008), 335-353

COMMON CANONICAL VARIATES FOR INDEPENDENT GROUPS USING INFORMATION THEORY

Xiangrong Yin and T. N. Sriram

University of Georgia

Abstract: Suppose that data on $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is a $q \times 1$ vector and $\mathbf{Y}$ is $p \times 1$ vector, are collected from C independent but closely related populations, and that one is interested in measuring the amount of relationship between sets of variables $\mathbf{Y}$ and $\mathbf{X}$ within each population. Goria and Flury (1996) argued that in these situations it is more meaningful to construct common canonical variates that are identical across populations, while the canonical correlations themselves may vary. Here we construct common information canonical variates based on Kullback-Leibler information. The proposed method does not require specific distributional assumptions and is useful in measuring true relations, whether linear or nonlinear. Simulations and dataset examples are presented. We also contrast our findings, in some instances, with those of Goria and Flury (1996).

Key words and phrases: Common information canonical variates, kernel density estimators, sequential permutation test.

1. Introduction

Canonical correlation analysis (CCA), a theory pioneered by Hotelling (1935, 1936), is a useful multivariate statistical technique for measuring the amount of linear relationship between sets of multiple variables $\mathbf{Y}$ and $\mathbf{X}$. The purpose of canonical correlation is to identify the optimum structure or dimensionality of each variable set that maximizes the linear relationship between them. In the last three decades or so there has been a proliferation of literature on generalizations and modifications of the classical two-set theory of CCA. Kettenring (1971, 1985) investigated simultaneous consideration of more than two sets of random variables. Van der Burg and De Leeuw (1983) presented an alternating least squares algorithm, termed as nonlinear canonical correlation analysis, to find an optimal scale for each variable in multivariate settings. Van der Burg, De Leeuw and Verdegaal (1988) extended the latter work to several sets of variables. Shi and Taam (1992) used a conditional mean and a nonparametric estimation method to find nonlinear structures between two sets of the variables. Leurgans, Moyeed and Silverman (1993) extended the CCA to situation where data are curves, while Luijtens, Symons and Vuylsteke-wauters (1994) developed linear and nonlinear CCA for group-structured data. Recently, Yin (2004) proposed a new canonical analysis based on Kullback-Leibler (KL) information, which is useful in measuring true relations, whether linear or nonlinear.

As Kettenring (1971) remarks, canonical correlation analysis results are often difficult to interpret but that this can be overcome by imposing restrictions on the coefficients of the canonical variates. Neuenschwander and Flury (1995) developed this idea of common canonical variates further, under the restriction that the canonical variates have the same coefficients in all the sets of the variables. Das and Sen (1994) have studied restricted canonical correlations obtained under nonnegativity constraints on the coefficients. In their pioneering article, Goria and Flury (1996) proposed a new concept called common canonical analysis in which the coefficients of the canonical variates are identical across the C populations, while the canonical correlations themselves may vary across populations. Such common canonical models, when appropriate for given data, are more parsimonious and hence are preferred over one with many parameters. In this paper, we revisit the problem studied in Goria and Flury (1996) and propose a new common canonical analysis for C independent populations using a recently developed KL information approach of Yin (2004).

We formally describe our method in Section 2 and study its basic properties in Section 2.1. A computational algorithm is described in Section 2.2, and a permutation test for determining the number of canonical variates is developed in Section 2.3. In Section 2.4, under certain regularity conditions, we obtain strong consistency of the proposed estimators. In Section 3 we present simulations and revisit two examples, of which one with machine data brings in some interesting differences between our approach and that of Goria and Flury (1996). Concluding remarks are given in Section 4.

2. The Common CCA Method

Suppose that $\{(\mathbf{Y}_i, \mathbf{X}_i, W_i), i = 1, \cdots, n\}$ is a random sample distributed as $(\mathbf{Y}, \mathbf{X}, W)$, where $W = 1, \ldots, C$. As in Goria and Flury (1996), we want to construct canonical variates $\eta = \mathbf{a}^T\mathbf{X}$ and $\psi = \mathbf{b}^T\mathbf{Y}$ such that η and ψ have the largest possible information within each of the C groups, and the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ are identical across all the independent groups. These, so called common canonical variates, provide a parsimonious summary of data while describing the relationships between two sets of variables across independent populations. More specifically, we define KL information by

$$\mathcal{I}(\mathbf{a},\mathbf{b}) = \mathrm{E}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{Y},W)}{p(\mathbf{a}^T\mathbf{X}|W)}\Big) \tag{1}$$

$$= \mathrm{E}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X},\mathbf{b}^T\mathbf{Y}|W)}{p(\mathbf{b}^T\mathbf{Y}|W)p(\mathbf{a}^T\mathbf{X}|W)}\Big) \tag{2}$$

$$= \mathrm{E}\Big(\log\frac{p(\mathbf{b}^T\mathbf{Y}|\mathbf{a}^T\mathbf{X},W)}{p(\mathbf{b}^T\mathbf{Y}|W)}\Big), \tag{3}$$

where $p(\cdot|\cdot)$ is the conditional density. For each $i \leq k = \min(q,p)$, we find the coefficient vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ such that

$$\mathcal{I}_i = \mathcal{I}(\mathbf{a}_i,\mathbf{b}_i) = \max_{\mathbf{a},\mathbf{b}}\mathcal{I}(\mathbf{a},\mathbf{b}) \tag{4}$$

subject to the constraints $\mathbf{a}_i^T\mathbf{\Sigma}_1\mathbf{a}_i = \mathbf{b}_i^T\mathbf{\Sigma}_2\mathbf{b}_i = 1$ and $\mathbf{a}_j^T\mathbf{\Sigma}_1\mathbf{a}_i = \mathbf{b}_j^T\mathbf{\Sigma}_2\mathbf{b}_i = 0$, for $j = 1,\ldots,k$, $i \neq j$, and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_\mathbf{X}^w$, $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_\mathbf{Y}^w$ for $w = 1,\ldots,C$. Here, $\mathbf{\Sigma}_\mathbf{X}^w$ and $\mathbf{\Sigma}_\mathbf{Y}^w$ are conditional covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, given $W = w$. Note that the constraints in (4) are not essential. However, we impose these in order to be consistent with the classical CCA method. Also, in practice, by the nature of the common canonical covariate problem, we would expect the covariance matrices across the groups to be the same. Even if this is not the case, one can standardize the variables within each group and have an identity covariance matrix across the groups.

Here, we refer to the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ as the ith common information canonical coefficient vectors for $\mathbf{X}$ and $\mathbf{Y}$, respectively. The random variables $\eta_i = \mathbf{a}_i^T\mathbf{X}$ and $\psi_i = \mathbf{b}_i^T\mathbf{Y}$ are called the ith common information canonical variates, and $\mathcal{I}_i$ is called the ith mutual common canonical information. Note that $\mathcal{I}_i$, $1 \leq i \leq k = \min(q,p)$, plays a somewhat similar role as the ith canonical correlation in the classical CCA as shown in part (2) of Proposition 2 below.

In general, the information numbers $\mathcal{I}_i$ are much harder to interpret than the canonical correlations because the former measure dependence through the likelihood, including dependence via mean functions or variance functions; see Yin (2004, Sec. 2.3 and Sec. 7.1) for more details. However, when $(\mathbf{X},\mathbf{Y})|W$ is jointly normally distributed, it can be shown that $\mathcal{I}(\mathbf{a},\mathbf{b}) = (-1/2)\sum_{w=1}^C p_w \log[1 - \rho_w^2(\mathbf{a},\mathbf{b})]$, where $p_w = P(W = w)$ and $\rho_w(\mathbf{a},\mathbf{b})$ is the correlation coefficient between $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ within group w. If $C = 1$, then the usual CCA is equivalent to the information CCA (Yin (2004)). However, if $C > 1$ and $\rho_w^2(\mathbf{a},\mathbf{b})$ varies with w for the same pair of variates, but with the same order within w for different pairs, then our common method agrees with the usual CCA for individual groups. If $C > 1$ and $\rho_w^2(\mathbf{a},\mathbf{b})$ varies with w for the same pair of variates, but with different order within w for different pairs, then our common method does not agree with usual CCA for individual groups. In such a case, since the significance of our information value only points to some significant relationship but not the actual structure of it (unlike the common CCA method where a significant correlation coefficient speaks to a linear relationship), we can use plots to reveal the actual structure of relationship. These plots may help us to detect whether a common model holds. For instance, when $C = 2$, suppose that there is one significant pair (η_1, ψ_1) for group 1 ($w = 1$), and there is one significant pair (η_2, ψ_2) for group 2 ($w = 2$), but $\mathrm{Cov}\,(\eta_1, \eta_2) = 0$ and $\mathrm{Cov}\,(\psi_1, \psi_2) = 0$. Then a common model may conclude that there are two significant pairs, while the plots may reveal otherwise. In this instance, our information CCA will detect that a common model may not be appropriate, see the machine data example in Section 3.2.

Throughout, we assume that the maximization in (4) has a unique solution. Also, assuming that all the required densities exist, we maximize the sample version of $\mathcal{I}(\mathbf{a}, \mathbf{b})$ with respect to $\mathbf{a}$ and $\mathbf{b}$:

$$\begin{aligned}\mathcal{I}_n(\mathbf{a}, \mathbf{b}) &= \frac{1}{n} \sum_{w=1}^{C} \sum_{j=1}^{n_w} \log \Big(\frac{p(\mathbf{a}^T\mathbf{X}_j | \mathbf{b}^T\mathbf{Y}_j, w)}{p(\mathbf{a}^T\mathbf{X}_j | w)} \Big) \\ &= \frac{1}{n} \sum_{w=1}^{C} \sum_{j=1}^{n_w} \log \Big(\frac{p(\mathbf{b}^T\mathbf{Y}_j | \mathbf{a}^T\mathbf{X}_j, w)}{p(\mathbf{b}^T\mathbf{Y}_j | w)} \Big),\end{aligned}$$

where n_w is the sample size for group w and $n = \sum_{w=1}^{C} n_w$. The successive common coefficient vectors are obtained by maximizing $\mathcal{I}_n(\mathbf{a}, \mathbf{b})$ subject the constraints in (4). If $\mathbf{a}$ is known, then the first equality above is equivalent to maximizing the conditional log likelihood over $\mathbf{b}$, and vice versa. Thus, our method maximizes a conditional log likelihood function in order to find the common canonical coefficient vectors. Our approach is more general than the one proposed in Goria and Flury (1996) because the latter method can only detect linear relationships. Note that the density functions in the above equalities are unknown and hence need to be estimated using appropriate kernel density estimators. Details are in Section 2.2.

2.1. Properties of the method

For the information measure $\mathcal{I}(\mathbf{a}, \mathbf{b})$ defined in (1) we establish some basic properties, the first of which gives the invariance of the information measure under affine transformations.

Proposition 1. *Let* $\mathbf{U} = \mathbf{A}^{-1}\mathbf{X} + \mathbf{a}_0$ *and* $\mathbf{V} = \mathbf{B}^{-1}\mathbf{Y} + \mathbf{b}_0$ *where* $\mathbf{A}$ *and* $\mathbf{B}$ *are two nonsingular matrices with appropriate dimensions, and* $\mathbf{a}_0$ *and* $\mathbf{b}_0$ *are fixed* $q \times 1$ *and* $p \times 1$ *vectors. Then*

1. $\mathcal{I}_{\mathbf{XY}}(\mathbf{a},\mathbf{b}) = \mathcal{I}_{\mathbf{UV}}(\mathbf{A}^T\mathbf{a},\mathbf{B}^T\mathbf{b})$.
2. *The common canonical information between* $\mathbf{U}$ *and* $\mathbf{V}$ *is the same as that of* $\mathbf{X}$ *and* $\mathbf{Y}$.
3. *The common information canonical vectors for* $\mathbf{U}$ *and* $\mathbf{V}$ *are given by* $\mathbf{A}^T\mathbf{a}_i$ *and* $\mathbf{B}^T\mathbf{b}_i$, $i = 1,\dots,k$, *where* $\mathbf{a}_i$ *and* $\mathbf{b}_i$ *are the common information canonical coefficient vectors for* $\mathbf{X}$ *and* $\mathbf{Y}$.

The first part of the next proposition gives a necessary and sufficient condition for the random variables $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ to be conditionally independent, given W. Incidentally, if $\mathcal{I}(\mathbf{a},\mathbf{b}) = 0$ for every $\mathbf{a}$ and $\mathbf{b}$, then the first part of Proposition 2 also implies that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent, given W. Note that in the Common CCA, $\rho_w(\mathbf{a},\mathbf{b}) = 0$ for every $\mathbf{a},\mathbf{b}$ and fixed w does not necessarily imply independence of $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ given w, unless $(\mathbf{X},\mathbf{Y})|W$ is normal. The second part of Proposition 2 shows that $\mathcal{I}_i$ also satisfies similar properties as the classical canonical correlations in that the common canonical information decreases, with $\mathcal{I}_1$ providing the most information.

Proposition 2. *The following hold for the information measure defined above.*

1. $\mathcal{I}(\mathbf{a},\mathbf{b}) \geq 0$, *for any* $\mathbf{a}$ *and* $\mathbf{b}$. *Moreover,* $\mathcal{I}(\mathbf{a},\mathbf{b}) = 0$ *iff* $\mathbf{b}^T\mathbf{Y}$ *is conditionally independent of* $\mathbf{a}^T\mathbf{X}$, *given* W.
2. $\mathcal{I}_1 > \cdots > \mathcal{I}_i \cdots > \mathcal{I}_m > 0 = \mathcal{I}_{m+1} = \cdots = \mathcal{I}_k$, *where* $m \leq k = min(q,p)$.

2.2. Computational algorithm

As in Yin (2004), we use the sample version of (2) to obtain $(\mathbf{a},\mathbf{b})$. For our computations, as suggested in Scott (1992) and Silverman (1986), we use a two-dimensional Gaussian kernel density estimate (with product kernels) to estimate $p(\mathbf{a}^T\mathbf{X},\mathbf{b}^T\mathbf{Y}|W)$, and a one-dimensional Gaussian kernel density estimate to estimate $p(\mathbf{a}^T\mathbf{X}|W)$ and $p(\mathbf{b}^T\mathbf{Y}|W)$, respectively. For all these density estimates we use the optimal bandwidth suggested in Scott (1992, p. 150). Step 1 below gives definitions of the density estimates. Step 2 gives the sample index that needs to be maximized.

Step 1. Construct the following one- and two- dimensional density estimates using a univariate Gaussian kernel K:

$$p_1(u|w) = \frac{1}{n_w h}\sum_{i=1}^{n_w} K\Big(\frac{u-u_i}{h}\Big) \text{ for }, u \in R,$$

$$p_2(u_1,u_2|w) = \frac{1}{n_w h_1 h_2}\sum_{i=1}^{n_w} K\Big(\frac{u_1-u_{i1}}{h_1}\Big) K\Big(\frac{u_2-u_{i2}}{h_2}\Big) \text{ for }, (u_1,u_2) \in R^2,$$

where $h = 1.06 s n_w^{-.2}$ and $h_j = s_j n_w^{-1/6}$ for $j = 1,2$, with the corresponding sample standard deviations s, s_1, and s_2 of u, u_1, and u_2, respectively.

Step 2. Obtain

$$(\hat{\mathbf{a}}_{i,n}, \hat{\mathbf{b}}_{i,n}) = \arg\max_{\mathbf{a},\mathbf{b}} \hat{\mathcal{I}}_n(\mathbf{a}, \mathbf{b}) = \arg\max_{\mathbf{a},\mathbf{b}} \frac{1}{n} \sum_{w=1}^{C} \sum_{j=1}^{n_w} \log \frac{p_2(\mathbf{a}^T\mathbf{x}_j, \mathbf{b}^T\mathbf{y}_j|w)}{p_1(\mathbf{a}^T\mathbf{x}_j|w)p_1(\mathbf{b}^T\mathbf{y}_j|w)}.$$

Any iteratively re-weighted least squares algorithm can be used to solve this equation. However, a preferred method that incorporates the constraint in (4) (see Hernández and Velilla (2005)) is known as Sequential Quadratic Programming (SQP) (Gill, Murray and Wright (1981, Chap. 6)). This method imposes nonlinear constraint during the search. In our computations, we use the corresponding sample versions of pooled covariance matrices for the data on $\mathbf{X}$ and $\mathbf{Y}$, respectively (see (4)), and the function 'fmincon' in *Matlab* that uses SQP procedure. With this function, given initial values, we use the default convergence criterion in *Matlab*. This processes is repeated until $i = \min(q, p)$.

2.3. Sequential permutation test and graphical plot

One of the goals of CCA is to determine the number of canonical variates. In our present context, this essentially means that we need to determine the number of common canonical information values I_i (starting from $i = 1$) that are significantly different from zero. This is possible here, because part 1 of Proposition 2 guarantees that $\mathcal{I}(\mathbf{a}, \mathbf{b}) = 0$ iff $\mathbf{b}^T\mathbf{Y}$ is conditionally independent of $\mathbf{a}^T\mathbf{X}$, given W. In view of this, we develop the following sequential permutation test to estimate the number of pairs of the canonical variates.

For notational convenience, let $\mathbf{A}_0$ and $\mathbf{B}_0$ be null vectors. Let $\mathbf{A}_i = (\mathbf{a}_1, \cdots, \mathbf{a}_i)$ and $\mathbf{B}_i = (\mathbf{b}_1, \cdots, \mathbf{b}_i)$ be the first i common information canonical coefficient vectors for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $\mathbf{A}_{c,i} = (\mathbf{a}_{i+1}, \cdots, \mathbf{a}_q)$ and $\mathbf{B}_{c,i} = (\mathbf{b}_{i+1}, \cdots, \mathbf{b}_p)$ be matrices so that $\mathbf{A}_q = (\mathbf{A}_i, \mathbf{A}_{c,i})$ and $\mathbf{B}_p = (\mathbf{B}_i, \mathbf{B}_{c,i})$ satisfy $\mathbf{A}_q^T\mathbf{\Sigma}_1\mathbf{A}_q = I_q$ and $\mathbf{B}_p^T\mathbf{\Sigma}_2\mathbf{B}_p = I_p$. For $\mathcal{I}_i$ defined in (4), we want to sequentially test the following hypotheses: $H_0 : \mathcal{I}_i = 0; H_1 : \mathcal{I}_i > 0$, for $i = 1, \cdots, k = \min(q, p)$.

Fix $(\mathbf{A}_{c,i-1}^T\mathbf{x}, \mathbf{B}_{c,i-1}^T\mathbf{y})$, use the algorithm in Section 2.2 to calculate $\hat{\mathcal{I}}_{i,n} = \hat{\mathcal{I}}_n(\hat{\mathbf{a}}_{i,n}, \hat{\mathbf{b}}_{i,n})$. Let $\mathbf{x}_w$ and $\mathbf{y}_w$ be the corresponding $\mathbf{x}$ and $\mathbf{y}$ data for $W = w$, respectively. For fixed w, permute the n_w points of the $\mathbf{B}_{c.i}^T\mathbf{y}_w$ data and repeat this for each $w = 1, \ldots, C$. This forms the new data $\mathbf{B}_{c,i-1}^T\mathbf{y}_{new}$. Combine this with $\mathbf{A}_{c,i-1}^T\mathbf{x}$ to find $\max_{\mathbf{a},\mathbf{b}} \hat{\mathcal{I}}_n(\mathbf{a}, \mathbf{b})$ using the algorithm in Section 2.2 again. Denote this maximum value as $\mathcal{I}_{i,n}^*$. Repeat this process B times, each time calculating $\mathcal{I}_{i,n}^*$. Let $\mathcal{I}_{i,0.05}^*$ be the 95th percentile of $\mathcal{I}_{1,n}^*$ values. We reject H_0 if $\hat{\mathcal{I}}_{i,n} > \mathcal{I}_{i,0.05}^*$ and proceed to the case $i + 1$. If we do not reject H_0, then we stop and conclude that there are $(i - 1)$ significant common information canonical variates. In our numerical studies we use $B = 1{,}000$. For more details, see Davison and Hinkley (1997), and Efron and Tibshirani (1993).

In practice, we use the corresponding sample versions to carry out the permutation test. Our choice of kernel or bandwidth may introduce some bias in the estimated densities used above. However, these will be canceled by our permutation test since we use the same kernel and bandwidths during each permutation. Undoubtedly, the sequential permutation tests given above are computationally intensive.

As an alternative, it is possible to use graphical methods to estimate number of pairs. After finding the pairs of common information variates, we can plot each common pair conditioning on w and use the structure of the plots to visually decide on the number of pairs; see Example 2 in Section 3.1. This is particularly reasonable in a canonical analysis because we focus on relationship within the pair (η_i, ψ_i) rather than between the pairs (η_i, ψ_j) for $i \neq j$. In addition, these plots can help us identify even nonlinear relationships between common information variates across the groups.

Before we proceed further, we illustrate our common information canonical method and sequential permutation test through a simulated example where the true relationships are nonlinear for each group. This example also brings about a sharp contrast between our method and the ones available in the literature.

An Example: Suppose $P[W = w] = 0.5$ for $w = 0, 1$, and X_1, X_2, Y_2, ϵ_1, ϵ_2 and W are independent random variables, where X_1, X_2, Y_2, ϵ_1 and ϵ_2 are standard normal random variables. We let $\mathbf{X} = (X_1, X_2)^T$ and $\mathbf{Y} = (Y_1, Y_2)^T$, where $Y_1 = (X_1 + X_2)^2 + .2\epsilon_1$ if $W = 0$, and $Y_1 = 8 - (X_1 + X_2)^2 + .2\epsilon_2$ if $W = 1$. It is clear that there is only one true common pair for each group given by $(\eta = \mathbf{a}^T\mathbf{X}, \psi = \mathbf{b}^T\mathbf{Y})$ with $\mathbf{a}^T = (1, 1)$ and $\mathbf{b}^T = (1, 0)$.

Denote by $(\hat{\mathbf{a}}^T\mathbf{x}, \hat{\mathbf{b}}^T\mathbf{y})$ the estimate based on a random sample of size $n = 60$. The correlation coefficient between $\mathbf{a}^T\mathbf{x}$ and $\hat{\mathbf{a}}^T\mathbf{x}$ values, and between $\mathbf{b}^T\mathbf{y}$ and $\hat{\mathbf{b}}^T\mathbf{y}$ values were 0.998 and 0.9983, respectively. Moreover, we performed the permutation test of $H_0 : \mathcal{I}_1 = 0$ vs $H_1 : \mathcal{I}_1 > 0$ which yielded a p-value of 0, and a permutation test of $H_0 : \mathcal{I}_2 = 0$ vs $H_1 : \mathcal{I}_2 > 0$ which yielded a p-value of 0.679. From these, we can conclude that there is only one significant common pair and our estimates are rather accurate. Note in this example that the classical CCA for each individual group or the common canonical analysis of Goria and Flury (1996) would fail to detect the true relationship due to nonlinearity.

2.4. Asymptotic theory

For the estimates defined in Section 2.2, it is possible to establish strong consistency by essentially following the steps in Yin (2004), along with conditioning on $W = w$. Here, we prove the strong consistency of slightly modified estimates of common information coefficient vectors. The modified estimates defined below are obtained by replacing the kernel density estimators defined in Section 2.2 with the following *leave-one-out* type kernel density estimators. Incidentally, these *leave-one-out* density estimators have been used by Powell, Stock and Stoker (1989) in their work. We prove the consistency result for the case $i = 1$ only. The proof is given in the Appendix.

For a univariate kernel K, define the *leave-one-out* kernel density estimates as

$$\hat{f}_i(u|w) = \frac{1}{(n_w - 1)h_{11}} \sum_{j=1, j\neq i}^{n_w} K\Big(\frac{u - u_j}{h_{11}}\Big) \quad \text{for} \quad u \in \mathbb{R}^1, \tag{5}$$

$$\hat{f}_i(u_1, u_2|w) = \frac{1}{(n_w - 1)h_{21}h_{22}} \sum_{j=1, j\neq i}^{n_w} K\Big(\frac{u_1 - u_{j1}}{h_{21}}\Big) K\Big(\frac{u_2 - u_{j2}}{h_{22}}\Big) \ \text{for} \ (u_1, u_2) \in \mathbb{R}^2, \tag{6}$$

where $h_{ij} = c_{ij} n_w^{-\delta_{ij}}$ for some $c_{ij}, \delta_{ij} > 0$ to be specified later.

Define

$$(\boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta}_{1,n}) = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{w=1}^{C} \sum_{i=1}^{n_w} \log \frac{\hat{f}_i(\boldsymbol{\beta}^T \mathbf{Y}_i, \boldsymbol{\alpha}^T \mathbf{X}_i|w)}{\hat{f}_i(\boldsymbol{\beta}^T \mathbf{Y}_i|w) \hat{f}_i(\boldsymbol{\alpha}^T \mathbf{X}_i|w)}.$$

Let Ω be the support of $\mathbf{V} = (\mathbf{Y}^T, \mathbf{X}^T)^T$ and let
$\Omega_y = \{\mathbf{y} : \text{there exist } \mathbf{x} \text{ such that } (\mathbf{y}, \mathbf{x}) \in \Omega\}$, $\Omega_x = \{\mathbf{x} : \text{there exist } \mathbf{y} \text{ such that } (y, \mathbf{x}) \in \Omega\}$, $T_\alpha = \{t : \text{there exist } \mathbf{x} \in \Omega_x \text{ and } \boldsymbol{\alpha} \in \Theta \text{ such that } t = \boldsymbol{\alpha}^T \mathbf{x}\}$ and $T_\beta = \{s : \text{there exist } \mathbf{y} \in \Omega_y \text{ and } \boldsymbol{\beta} \in \Theta \text{ such that } s = \boldsymbol{\beta}^T \mathbf{y}\}$, where Θ denotes the parameter space for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

We state the regularity conditions needed to prove the consistency theorem. Assume the following.

(C.0) $(\mathbf{a}, \mathbf{b})$ is unique and identifiable.

(C.1) $(\mathbf{Y}_i, \mathbf{X}_i) \in \mathbb{R}^p \times \mathbb{R}^q$ are i.i.d. random vectors with $\mathrm{E}(||\mathbf{X}_i||) < \infty$ and $\mathrm{E}(||\mathbf{Y}_i||) < \infty$.

(C.2) Θ is a compact subset of $R^p \times R^q$.

(C.3) The set Ω is compact such that $\inf_{(\mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Omega \times \Theta} p(\boldsymbol{\beta}^T \mathbf{y}, \boldsymbol{\alpha}^T \mathbf{x}|w) > 0$, $\inf_{(\mathbf{x}, \boldsymbol{\alpha}) \in \Omega_x \times \Theta} p(\boldsymbol{\alpha}^T \mathbf{x}|w) > 0$, and $\inf_{(\mathbf{y}, \boldsymbol{\beta}) \in \Omega_y \times \Theta} p(\boldsymbol{\beta}^T \mathbf{y}|w) > 0$.

(C.4) $p(s, t|w)$, $p(t|w)$ and $p(s|w)$ satisfy Lipschitz conditions, for $s \in T_\beta$, and $t \in T_\alpha$, uniformly in $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Theta$ for any fixed w.

(C.5) K is a real, symmetric, and differentiable kernel with $|K(u)| < c_0$ and $|(\partial K(u))/\partial u| < c_0$ for some $c_0 > 0$.

In condition (C.0), even if $(\mathbf{a}, \mathbf{b})$ is the unique population direction, it may not be identifiable. Note that for any nonzero constants c_1 and c_2, $(c_1\mathbf{a}, c_2\mathbf{b})$ will give the same solution. Thus for identifiability, we need to impose some constraints. For instance, without loss of generality, we can let the first element of $\mathbf{a}, \mathbf{b}$ to be 1, then the rest of the $q - 1$ and $p - 1$ parameters, respectively, are uniquely determined and identifiable.

Note that condition (C.3) seems at first glance restrictive. It says that the joint and the marginal pdfs are bounded below by 0. However, we may confine it to a relevant compact region, as we do in condition (C.2), where the pdfs are strictly positive. We use condition (C.3) for technical simplicity.

Lemma 1. *Under the assumptions* (C.0)−(C.5)*, with* $\delta_{ij} \in (0, 1/4)$ *in* (6)*, we have*

$$\sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta} |\log \hat{f}_i(s,t|w) - \log p(s,t|w)| \to 0 \ a.s.$$

$$\sup_{\mathbf{x}\in\Omega_x}\ \sup_{\boldsymbol{\alpha}\in\Theta} |\log \hat{f}_i(t|w) - \log p(t|w)| \to 0 \ a.s.$$

$$\sup_{\mathbf{y}\in\Omega_y}\ \sup_{\boldsymbol{\beta}\in\Theta} |\log \hat{f}_i(s|w) - \log p(s|w)| \to 0 \ a.s..$$

Lemma 2. *Under* (C.0)−(C.5)*, with* $\delta_{ij} \in (0, 1/4)$ *in* (6)*, we have*

$$\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta} |\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha},\boldsymbol{\beta}) - \mathcal{I}(\boldsymbol{\alpha},\boldsymbol{\beta})| \to 0 \ a.s..$$

Theorem 1. *Under* (C.0)–(C.5)*, with* $\delta_{ij} \in (0, 1/4)$ *in* (6)*, we have* $(\boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta}_{1,n}) \to (\mathbf{a}, \mathbf{b})$ *a.s.*.

Theorem 1 provides the consistency result for the first pair. The second pair is obtained by maximizing $\hat{\mathcal{I}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ over the set $\{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in R^q \times R^p : \boldsymbol{\alpha} \perp \boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta} \perp \boldsymbol{\beta}_{1,n}\}$ which is a random set depending on $(\boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta}_{1,n})$. If $(\boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta}_{1,n})$ were the population pair, then the proof for the second pair is exactly the same. However, since $(\boldsymbol{\alpha}_{1,n}, \boldsymbol{\beta}_{1,n})$ is a consistent estimator, we can modify the proof in Theorem 1 appropriately to obtain the consistency of the second pair. The same logic can then be applied to further pairs. We omitted these proofs here.

3. Numerical studies

3.1. Simulations

We present two simulation studies to assess the performance of our method. We simulated 100 datasets for each of the two studies where there are two groups, i.e, $C = 2$, but the relationship between the true pairs is linear and/or nonlinear. We use the distance $m = ||(I_t - \hat{\mathbf{B}}\hat{\mathbf{B}}^T)\mathbf{B}||$ if $s \geq k$, or $m = ||(I_t - \mathbf{B}\mathbf{B}^T)\hat{\mathbf{B}}||$ if $s < k$ (Xia, Tong, Li and Zhu (2002)), to measure the goodness of two k-dimensional subspaces $\mathcal{S}(\mathbf{B}_{t\times k})$ and $\mathcal{S}(\hat{\mathbf{B}}_{t\times s})$, where $\mathbf{B}^T\mathbf{B} = I_k$ and $\hat{\mathbf{B}}^T\hat{\mathbf{B}} = I_s$. Also see Ye and Weiss (2003) and Li, Zha and Chiaromonte (2005) for similar measures. If $m = 0$, then $\mathcal{S}(\mathbf{B}) = \mathcal{S}(\hat{\mathbf{B}})$. The smaller is the m value, the better is the estimate.

Example 1. Let $P[W = w] = 0.5$ for $w = 0, 1$. Suppose X_1, X_2, Y_2, ϵ_1, ϵ_2 and W are independent random variables, where X_1, X_2, Y_2, ϵ_1 and ϵ_2 are standard normal random variables. Let $\mathbf{X} = (X_1, X_2)^T$ and $\mathbf{Y} = (Y_1, Y_2)^T$, where $Y_1 = X_1 + X_2 + .2\epsilon_1$ if $W = 0$, and $Y_1 = \sin(X_1 + X_2) + .3\epsilon_2$ if $W = 1$. It is clear that there is only one true common pair for each group given by $(\eta = \mathbf{a}^T\mathbf{X}, \psi = \mathbf{b}^T\mathbf{Y})$ with $\mathbf{a}^T = (1, 1)$ and $\mathbf{b}^T = (1, 0)$. However, note that the relationship between the pair is linear when $W = 0$ and nonlinear when $W = 1$.

Table 1 reports the *mean* ± *stand.error* for 100 values of m^2 with $\hat{\mathbf{B}} = \hat{\mathbf{a}}$ and $\mathbf{B} = \mathbf{a}$, and $\hat{\mathbf{B}} = \hat{\mathbf{b}}$ and $\mathbf{B} = \mathbf{b}$, respectively. We take $\mathbf{a}_0 = (1/\sqrt{5}, -2/\sqrt{5})^T$ and $\mathbf{b}_0 = (1/\sqrt{2}, 1/\sqrt{2})^T$. From Table 1, one can see that the results are more accurate for larger sample sizes as expected. Even when sample size $n = 60$, the results seem reasonable.

Table 1. Accuracy of the estimated dimensions in Example 1.

m^2	$n = 60$	$n = 120$	$n = 240$
a	0.1615 ± 0.3456	0.0316 ± 0.1651	0.0176 ± 0.1071
b	0.1301 ± 0.2995	0.0253 ± 0.1386	0.0120 ± 0.0664

To see the effect of the initial values, we choose $\mathbf{a}_0 = (1/\sqrt{2}, 1/\sqrt{2})^T$ and $\mathbf{b}_0 = (0, 1)^T$. The results are 0.0220 ± 0.0880 and 0.0785 ± 0.2358 for **a** and **b** respectively, when $n = 60$; 0.0139 ± 0.0927 and 0.0126 ± 0.0824 for **a** and **b**, respectively, when $n = 240$. This shows that there is some early effect of initial values, possibly due to the procedure being trapped around a local maximum, or slower convergence to the solution etc., but these problems disappear when the sample size is large. We suggest that when the sample size is small, one might use prior information for choosing starting values, and several of them to check whether a proper global solution is reached. Another way for choosing initial values is to take that $(\mathbf{a}_0, \mathbf{b}_0)$ with the biggest index $\hat{\mathcal{I}}(\mathbf{a}_0, \mathbf{b}_0)$ among a set of random selected initials values.

In addition, based on a random sample of size $n = 60$, we performed the permutation test of $H_0 : \mathcal{I}_1 = 0$ vs $H_1 : \mathcal{I}_1 > 0$, which yielded a p-value of 0, and a permutation test of $H_0 : \mathcal{I}_2 = 0$ vs $H_1 : \mathcal{I}_2 > 0$, which yielded a p-value of 0.858. From this we conclude that there is only one common pair, which indeed is the case.

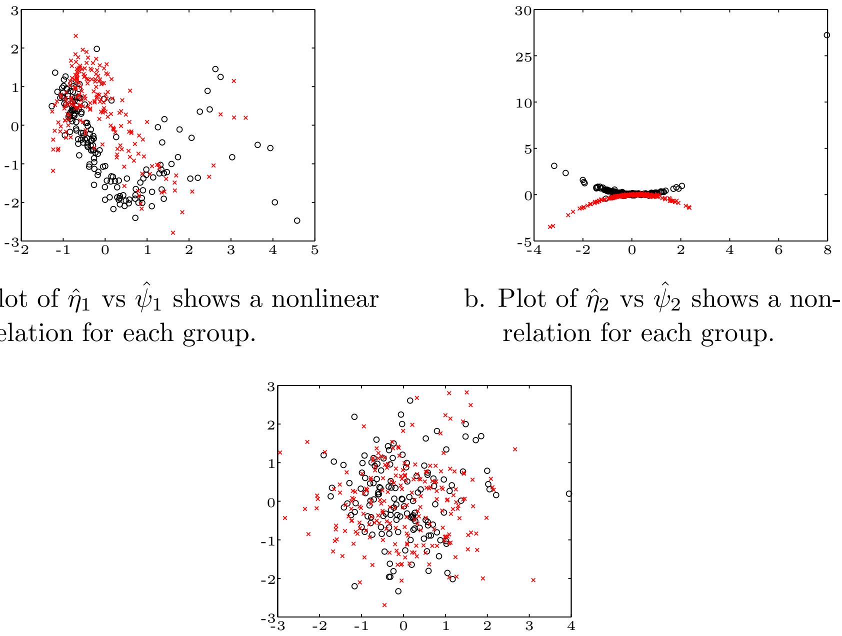
Example 2. Let $P[W = 0] = 0.4 = 1 - P[W = 1]$. Suppose that $X_1 \sim N(0, 1)$, $X_2 \sim t_{13}$, $X_3 \sim F(2, 10)$, $X_4 \sim \chi^2(2)$, $X_5 \sim F(2, 12)$, $X_6 \sim \chi^2(3)$, $Y_2 \sim U(0, 1)$, $Y_3 \sim N(0, 1)$, $Y_4 \sim \chi^2(3)$, $Y_3 \sim t_{15}$, and ϵ_i's for $i = 1, 2, 3, 4$, are iid standard normal. Let $(\eta_1 = \mathbf{a}_1^T\mathbf{X}, \psi_1 = \mathbf{b}_1^T\mathbf{Y})$ with $\mathbf{a}_1^T = (0, 0, 0, 0, 1, .3)$ and $\mathbf{b}_1^T = (1, -2, 0, 0, 0, 0)$, and $(\eta_2 = \mathbf{a}_2^T\mathbf{X}, \psi_2 = \mathbf{b}_2^T\mathbf{Y})$ with $\mathbf{a}_2^T = (1, -1, 1, -1, 0, 0)$ and $\mathbf{b}_2^T = (0, 0, 0, 1, -1, 1)$. Furthermore, we let $\mathbf{b}_1^T\mathbf{Y} = \cos(\mathbf{a}_1^T\mathbf{X}) + 0.2\epsilon_1$ and

$\mathbf{b}_2^T\mathbf{Y} = (\mathbf{a}_2^T\mathbf{X})^2 + 0.2\epsilon_3$ if $W = 0$, and $\mathbf{b}_1^T\mathbf{Y} = \sin(\mathbf{a}_1^T\mathbf{X}) + 0.3\epsilon_2$ and $\mathbf{b}_2^T\mathbf{Y} = -(\mathbf{a}_2^T\mathbf{X})^2 + 0.3\epsilon_3$ if $W = 1$, where $\mathbf{X} = (X_1, \ldots, X_6)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_3)^T$. Thus there are two true pairs.

Table 2 reports the *mean* ± *stand.error* for 100 values of m^2. Initial values are chosen in a way that the convergence is reached quickly. To illustrate our graphical way of determining the number of pairs, a random sample of size $n = 360$ was drawn. Figures 1a and 1b show a strong nonlinear relation between $\hat{\eta}_1$ and $\hat{\psi}_1$, and between $\hat{\eta}_2$ and $\hat{\psi}_2$, while Figure 1c shows no-apparent relation between $\hat{\eta}_3$ and $\hat{\psi}_3$ for third pair. This example serves as an evidence that the graphical methods can be used to determine the number of pairs. Finally, note that the non-normality of some of the variables has little affect on our method.

Table 2. Accuracy of the estimated dimensions in Example 2.

m^2	$n = 240$	$n = 360$
$\mathbf{a}_1$	0.0149 ± 0.0641	0.0104 ± 0.0757
$\mathbf{b}_1$	0.0010 ± 0.0016	0.0006 ± 0.0017
$\mathbf{a}_2$	0.0171 ± 0.0133	0.0162 ± 0.0215
$\mathbf{b}_2$	0.0624 ± 0.0873	0.0556 ± 0.0548



a. Plot of $\hat{\eta}_1$ vs $\hat{\psi}_1$ shows a nonlinear relation for each group.

b. Plot of $\hat{\eta}_2$ vs $\hat{\psi}_2$ shows a non-linear relation for each group.

c. Plot of $\hat{\eta}_3$ vs $\hat{\psi}_3$ shows a noapparent relation for either group.

Figure 1. Graphs of first three common estimated pairs in Example 2.

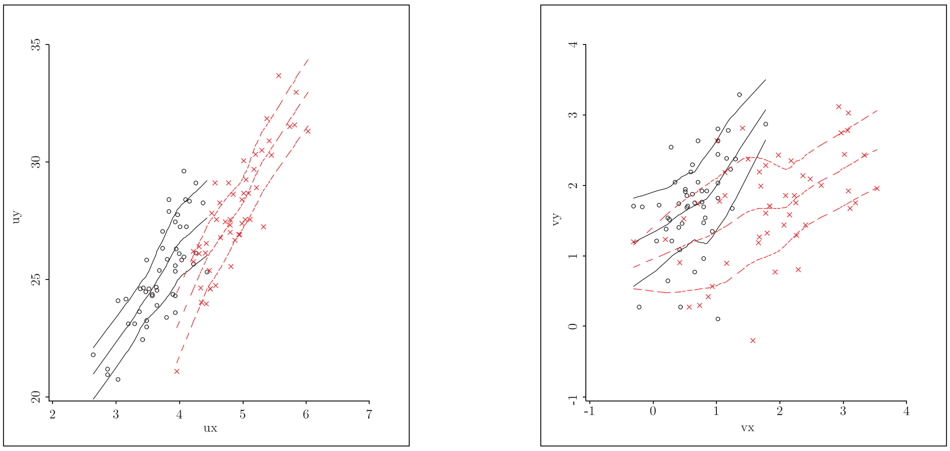
3.2. Iris and machine datasets

Iris data: This well-known dataset, discussed by Anderson (1935), has three species of Iris flowers known as *Iris Versicolor*, *Iris Virginica* and *Iris Seton.* The following four variables were measured from 50 plants of each species : $\mathbf{X}^T = (X_1, X_2)$ where X_1 = sepal length and X_2 = sepal width, and $\mathbf{Y}^T = (Y_1, Y_2)$ where Y_1 = petal length and Y_2 = petal width. Goria and Flury (1996) concluded that a common canonical variate model fits well for data belonging to the two groups *Iris Versicolor* and *Iris Virginica*, with two common pairs denoted as (U_1, V_1) and (U_2, V_2).

We carried out the common information canonical analysis for the Iris data belonging to the two groups *Iris Versicolor* and *Iris Virginica.* The first and second common information canonical pairs $(\hat{\mathbf{a}}_1^T\mathbf{x}, \hat{\mathbf{b}}_1^T\mathbf{y})$ and $(\hat{\mathbf{a}}_2^T\mathbf{x}, \hat{\mathbf{b}}_2^T\mathbf{y})$ are plotted in Figures 2a and 2b, respectively. Comparing with Goria and Flury's analysis, we obtained the following correlation coefficients: $Corr(\hat{\mathbf{a}}_1^T\mathbf{x}, U_1) = 0.9998$, $Corr(\hat{\mathbf{b}}_1^T\mathbf{y}, V_1) = 0.99995$, $Corr(\hat{\mathbf{a}}_2^T\mathbf{x}, U_2) = 0.9879$, and $Corr(\hat{\mathbf{b}}_2^T\mathbf{y}, V_2) = 0.9973$. These show that our method yields very similar result as in Goria and Flury (1996). Furthermore, our sequential permutation test of $H_0 : \mathcal{I}_1 = 0$ vs $H_1 : \mathcal{I}_1 > 0$ yielded a p-value of 0, followed by a permutation test of $H_0 : \mathcal{I}_2 = 0$ vs $H_1 : \mathcal{I}_2 > 0$ which yielded a p-value of 0. Hence, we conclude that two pairs of common information canonical variates are needed. In addition, we also carried out our method for all three groups and found that the results are very similar. These results are not given here.

Figure 2a shows a strong linear relationship between $\hat{\mathbf{a}}_1^T\mathbf{x}$ and $\hat{\mathbf{b}}_1^T\mathbf{y}$ for both the species. Figure 2b shows a mild linear relationship between $\hat{\mathbf{a}}_2^T\mathbf{x}$ and $\hat{\mathbf{b}}_2^T\mathbf{y}$ for both the species. These conclusions are consistent with those in Goria and Flury (1996). The superimposed curves over the scatter plots in Figures 2a and 2b are the mean curves (in the middle) with lower and upper bounds of one-standard deviations obtained using LOWESS method, conditioned on each species. This example clearly illustrates that our results agree with those of Goria and Flury (1996), when the linear relationship between the pairs is the dominant one for both species.

Machine data: The electrode data (Flury and Riedwyl (1988)) are of five measurements on 50 electrodes produced by two different machines: the first three $\mathbf{X}^T = (X_1, X_2, X_3)$ are widths and the other two $\mathbf{Y}^T = (Y_1, Y_2)$ are lengths. Goria and Flury (1996) concluded that a common canonical variate model fits well for *Machine* 1 and *Machine* 2. Their first two common canonical correlations are $\hat{\rho}_{1,1} = 0.748$ and $\hat{\rho}_{2,1} = 0.101$ for *Machine* 1, and $\hat{\rho}_{1,2} = 0.139$ and $\hat{\rho}_{2,2} = -0.322$ for *Machine* 2. The values $\hat{\rho}_{1,1}$ and $\hat{\rho}_{1,2}$ indicate that the linear relationship between the length and width variables is much stronger for *Machine* 1 than for the *Machine* 2. For *Machine* 2, the common correlations $\hat{\rho}_{1,2}$ and $\hat{\rho}_{2,2}$



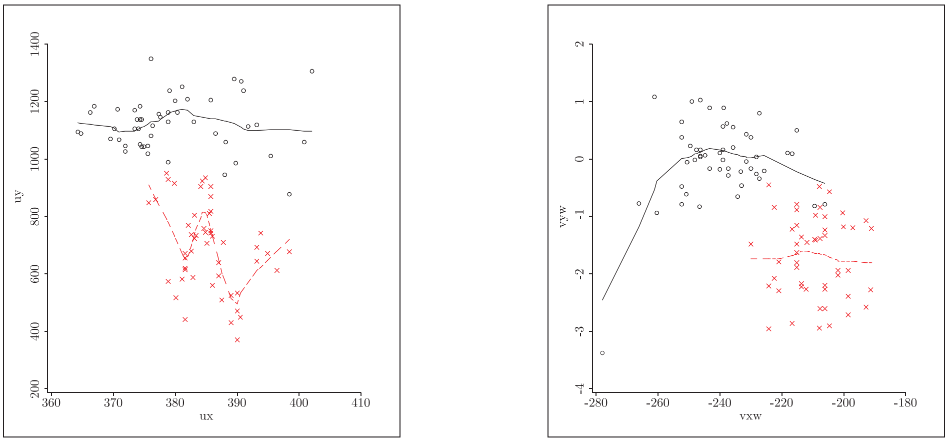
a. Plot of ux=$\hat{\mathbf{a}}_1^T\mathbf{x}$ vs uy=$\hat{\mathbf{b}}_1^T\mathbf{y}$ shows a strong linear relationship for both species.

b. Plot of vx=$\hat{\mathbf{a}}_2^T\mathbf{x}$ vs vy=$\hat{\mathbf{b}}_2^T\mathbf{y}$ shows a mild linear relationship for both species.

Figure 2. Iris flowers data for species *Versicolor* and *Virginica*.

seem to indicate a very weak linear relationship between the length and width variables. Finally, as noted in Goria and Flury (1996), these correlation values suggest a potential conflict in which the first common canonical variate has stronger correlation for *Machine* 1 but a weaker correlation for *Machine* 2. In our view, all these observations indicate that there may not be a significant second common canonical variate and, more importantly, a common canonical variate model may not even be appropriate for the machine data. Next we show that our common information analysis for the machine data supports this view.

Our sequential permutation test showed that the first common information canonical pair is significant with a p-value 0.003, but the test of $H_0 : \mathcal{I}_2 = 0$ vs $H_1 : \mathcal{I}_2 > 0$ produced a p-value of 0.048, which raises strong doubts about the necessity of second common information canonical pair. The plot of our first common information canonical variates in Figure 3a shows a nonlinear relationship for *Machine* 2 but no apparent relationship for *Machine* 1, where $ux \propto \hat{\eta}_1 = 0.723x_1 + 0.333x_2 + 0.605x_3$ and $uy \propto \hat{\psi}_1 = 0.596y_1 + 0.803y_2$. It seems that $\hat{\eta}_1$ is weighted width with approximate weights $(2.2, 1, 1.8)$, while $\hat{\psi}_1$ is weighted length with approximate weights $(3, 4)$. The plot of our second common information canonical variates in Figure 3b shows no relationship for *Machine* 2, but perhaps a mild nonlinear relation for *Machine* 1, where $uxw \propto \hat{\eta}_2 = 0.697x_1 - 0.702x_2 - 0.148x_3$ and $vyw \propto \hat{\psi}_2 = -0.767y_1 + 0.642y_2$. It seems that $\hat{\eta}_2$ is weighted width with approximate weights $(-4.7, 4.7, 1)$ while $\hat{\psi}_2$ is weighted length with approximate weights $(1.2, -1)$. These seem to indicate that *Machine* 1 and *Machine* 2 may have different pairs of significant variates.



a. Plot of the first common canonical pair: ux=$\hat{\mathbf{a}}_1^T\mathbf{x}$ vs uy=$\hat{\mathbf{b}}_1^T\mathbf{y}$ shows strong nonlinear relationship for *Machine* 2 and no relationship for *Machine* 1.

b. Plot of the second common canonical pair: vxw=$\hat{\mathbf{a}}_2^T\mathbf{x}$ vs vyw=$\hat{\mathbf{b}}_2^T\mathbf{y}$ shows mild nonlinear relationship for *Machine* 1 and no relationship for *Machine* 2.

Figure 3. Common Information canonical variates for Machine data.

Therefore, instead of a common model, an individual information canonical analysis may be more appropriate. This is the case that we discussed in the third paragraph in Section 2.

Clearly, our conclusions do not agree at all with those of Goria and Flury (1996). One simple way to resolve this is to carry out a (comparable) analysis individually for each machine. That is, to carry out an individual analysis using the classical CCA and Yin (2004)'s method, and to compare the results to those in Goria and Flury (1996) and ours above. If the conclusion of a common analysis approach is consistent with the comparable individual analysis, then the conclusions based on that approach may be considered valid; otherwise common analysis may not be valid.

Yin (2004) analyzed this data for *Machine* 2 using his information-based canonical analysis and determined that only the first pair of information canonical variate, say $(\hat{\eta}_1^2, \hat{\psi}_1^2)$, is needed, and that the relationship between the elements of the first pair is nonlinear. Yin (2004) also showed that a sine curve fits the scatter plot of the first pair reasonably well, which agrees with our conclusions here. Incidentally, the classical CCA for *Machine* 2 shows, on the contrary, that there is no significant pair. As for *Machine* 1, we carried out an information canonical analysis and the associated sequential permutation tests of Yin (2004). These showed that the first information-based canonical variate pair, say $(\hat{\eta}_1^1, \hat{\psi}_1^1)$, for *Machine* 1 is significant with a p-value of 0, but the second pair has a p-value of 0.428, implying that only the first pair is necessary for the *Machine*

1 data. It is interesting to note that the classical CCA for *Machine* 1 data has essentially the same result, as indicated by the correlations $Corr(\hat{\eta}_1^1, \hat{\eta}_1^{CCA})$ $= 0.9869$ and $Corr(\hat{\psi}_1^1, \hat{\psi}_1^{CCA}) = 0.999986$, where $\hat{\eta}_i^{CCA}$ is the canonical variate based on the classical CCA. We also note that the pairs $(\hat{\eta}_1^2, \hat{\psi}_1^2)$ for *Machine* 2 and $(\hat{\eta}_1^1, \hat{\psi}_1^1)$ for *Machine* 1 are in different directions, but not completely uncorrelated. That is why our common informational method finds the first pair $(\hat{\eta}_1^2, \hat{\psi}_1^2)$, but marginally recovers $(\hat{\eta}_1^1, \hat{\psi}_1^1)$. With the plots, our method essentially leads to using a separate information CCA analysis for each machine.

In view of all these, an individual informational canonical analysis seems more appropriate for the machine data. Also, only the first canonical variate is necessary for each machine. Finally, there seems to be a strong linear relationship between the first canonical variate pair for *Machine* 1, while there seems to be a strong nonlinear relationship between the first canonical variate pair for *Machine* 2.

4. Concluding Remarks

We have described a common canonical analysis based on Kullback-Leibler information which is useful in measuring true relations, whether linear or nonlinear. We have also proposed a sequential permutation test as well as graphical plot to determine the number of pairs of canonical information variates to use in practice. In addition, graphical plots can identify the true relationship between the linear combinations. As shown by our simulations and examples, our method is competitive with the method of Goria and Flury (1996) when the relation is linear. More importantly, our method offers a general dimension reduction technique which can help identify nonlinear relationships.

Inherent in our method are maximizations and permutation tests which require choice of appropriate kernel density estimators, bandwidths and initial values. In our computations we found Gaussian kernels and the bandwidth selections made in Section 2.2 to be reasonable choices. Since our primary focus is on finding common canonical variates that provide maximum information and dimension reduction, optimal choices of kernel density estimators and bandwidth do not seem very crucial as they only play a role in our intermediate steps.

Appendix: Proofs

Proposition 1. *As for the first assertion, use the notations to write*

$$\begin{aligned}\mathcal{I}_{\mathbf{XY}}(\mathbf{a},\mathbf{b}) &= \mathrm{E}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{Y},W)}{p(\mathbf{a}^T\mathbf{X}|W)}\Big) = \mathrm{E}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{BV},W)}{p(\mathbf{a}^T\mathbf{X}|W)}\Big)\\ &= \mathrm{E}\Big(\log\frac{p(\mathbf{b}^T\mathbf{BV}|\mathbf{a}^T\mathbf{X},W)}{p(\mathbf{b}^T\mathbf{BV}|W)}\Big) = \mathrm{E}\Big(\log\frac{p(\mathbf{b}^T\mathbf{BV}|\mathbf{a}^T\mathbf{AU},W)}{p(\mathbf{b}^T\mathbf{BV}|W)}\Big)\\ &= \mathcal{I}_{\mathbf{UV}}(\mathbf{A}^T\mathbf{a},\mathbf{B}^T\mathbf{b}).\end{aligned}$$

This proves Assertion 1.

Assertion 2 *follows from Assertion* 1 *because* $\mathcal{I}_{i,\mathbf{XY}}(\mathbf{a}_i, \mathbf{b}_i) = \mathcal{I}_{i,\mathbf{UV}}(\mathbf{A}^T\mathbf{a}_i, \mathbf{B}^T\mathbf{b}_i)$.

Assertion 3 *follows directly from Assertion* 2 *and uniqueness. Note that the constraints in* (4) *still hold.*

Proposition 2. *Rewriting the information in terms of conditional densities and using the definition of KL information* (Kullback (1959)) *we get*

$$\begin{aligned}\mathcal{I}(\mathbf{a},\mathbf{b}) &= \mathrm{E}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{Y},W)}{p(\mathbf{a}^T\mathbf{X}|W)}\Big)\\ &= \mathrm{E}_{(\mathbf{b}^T\mathbf{Y},W)}\Big[\mathrm{E}_{\mathbf{a}^T\mathbf{X}|(\mathbf{b}^T\mathbf{Y},W)}\Big(\log\frac{p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{Y},W)}{p(\mathbf{a}^T\mathbf{X}|W)}\Big)\Big] \geq 0.\end{aligned}$$

This yields the first part of the first assertion. Also, by the definition of KL information, the equality holds above if and only if $p(\mathbf{a}^T\mathbf{X}|\mathbf{b}^T\mathbf{Y},W) = p(\mathbf{a}^T\mathbf{X}|W)$, *which yields the second part of the first assertion.*

Assertion 2 *follows from the definition and by the assumption of uniqueness of the maximization. Without the uniqueness assumption, the strict inequalities become equalities. It is possible that Assertion* 2 *holds for* $m < k$. *If not,* $m = k$.

Lemma 1. *We first prove that*

$$\sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta} |\hat{f}_i(s,t|w) - p(s,t|w)| \to 0 \text{ a.s.}. \tag{7}$$

Using the result of Klein and Spady (1993), it suffices to show that

$$\sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\left|\mathrm{E}_{|W}\Big[\frac{1}{h_{21}}K\Big(\frac{s-s_j}{h_{21}}\Big)\frac{1}{h_{22}}K\Big(\frac{t-t_j}{h_{22}}\Big)\Big] - p(s,t|w)\right| \to 0.$$

But the left hand side is less than or equal to

$$\begin{aligned}&\sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big|\int_{W=w}\frac{1}{h_{21}}K\Big(\frac{s-u_1}{h_{21}}\Big)\frac{1}{h_{22}}K\Big(\frac{t-u_2}{h_{22}}\Big)(p(u_1,u_2)-p(s,t)du_1du_2\Big|\\ &\quad\leq \sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big|\int_{W=w}K(v_1)K(v_2)[p(s-h_{21}v_1,t-h_{22}v_2)-p(s,t)]dv_1dv_2\Big|\\ &\quad\leq \int_{W=w}|K(v_1)K(v_2)|\sup_{\mathbf{v}\in\Omega}\ \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big[|p(s-h_{21}v_1,t-h_{22}v_2)-p(s,t)|\Big]dv_1dv_2\\ &\quad\leq L\int_{W=w}|K(v_1)K(v_2)(h_{21}|v_1|+h_{22}|v_2|)dv_1dv_2\\ &\quad\leq n^{-\delta}L\int_{W=w}|K(v_1)K(v_2)|v_1|dv_1dv_2 \leq n^{-\delta}L,\end{aligned}$$

where L is some generic positive constant.

Note that

$$\sup_{\mathbf{v}\in\Omega}\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big|\log\hat{f}_i(s,t|w)-\log p(s,t|w)\Big|$$
$$\leq\Big(\inf_{\mathbf{v}\in\Omega}\inf_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}[\hat{f}_i(s,t|w),p(s,t|w)]\Big)^{-1}\times\sup_{\mathbf{v}\in\Omega}\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big|\hat{f}_i(s,t|w)-p(s,t|w)\Big|.$$

The first term is bounded above by (C.3) *and* (7)*. The conclusion follows from* (7) *again. Similarly, the other two results follow.*

Lemma 2. *Since*

$$\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big|\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha},\boldsymbol{\beta})-\mathcal{I}_n(\boldsymbol{\alpha},\boldsymbol{\beta})\Big|$$
$$\leq\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\frac{1}{n}\sum_{w=1}^{C}\sum_{i=1}^{n_w}\Big[\Big|\log\hat{f}_i(s_i,t_i|w)-\log p(s_i,t_i|w)\Big|$$
$$+\Big|\log\hat{f}_i(t_i|w)-\log p(t_i|w)\Big|+\Big|\log\hat{f}_i(s_i|w)-\log p(s_i|w)\Big|\Big]$$
$$\leq\sup_{\mathbf{v}\in\Omega}\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}\Big[\Big|\log\hat{f}_i(s,t|w)-\log p(s,t|w)\Big|$$
$$+\Big|\log\hat{f}_i(t|w)-\log p(t|w)\Big|+\Big|\log\hat{f}_i(s|w)-\log p(s|w)\Big|\Big]$$

following Lemma 1*, we have* $\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}|\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha},\boldsymbol{\beta})-\mathcal{I}_n(\boldsymbol{\alpha},\boldsymbol{\beta})|\to 0$*, a.s..*

Under our assumptions, the assumptions $A1$, $A2$, and $A4$ of Andrews (1987) are satisfied. Therefore, $\sup_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta}|\mathcal{I}_n(\boldsymbol{\alpha},\boldsymbol{\beta})-\mathcal{I}(\boldsymbol{\alpha},\boldsymbol{\beta})|\to 0$ a.s.. Hence the conclusion follows.

Theorem 1. *Under our condition, $\mathcal{I}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is continuous in Θ. Let $\epsilon>0$. Since Θ is compact, there is a number $\delta>0$ such that*

$$\sup\{\mathcal{I}(\boldsymbol{\alpha},\boldsymbol{\beta}):|\mathbf{a}-\boldsymbol{\alpha}|+|\mathbf{b}-\boldsymbol{\beta}|>\epsilon,(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta\}<\mathcal{I}(\mathbf{a},\mathbf{b})-\delta. \tag{8}$$

Based on Lemma 2*, with probability tending to* 1*, we have*

$$|\hat{\mathcal{I}}_n^{(1)}(\mathbf{a},\mathbf{b})-\mathcal{I}(\mathbf{a},\mathbf{b})|<\frac{\delta}{4}\ \text{ and }\ |\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})-\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})|<\frac{\delta}{4}.$$

By construction, $\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})\geq\hat{\mathcal{I}}_n^{(1)}(\mathbf{a},\mathbf{b})$ which, combined with the first of the above inequalities, implies with probability tending to 1 *that, $\hat{\mathcal{I}}_n^{(1)}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})>\mathcal{I}(\mathbf{a},\mathbf{b})-\delta/4$. Using the second inequality above we see that, with probability tending to 1, $\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})>\mathcal{I}(\mathbf{a},\mathbf{b})-\delta/2$. By* (8)*, $P\big(|\boldsymbol{\alpha}_{1,n}-\mathbf{a}|+|\boldsymbol{\beta}_{1,n}-\mathbf{b}|>\epsilon\big)\leq$*

$P\left(\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})<\mathcal{I}(\mathbf{a},\mathbf{b})-\delta\right)$. *Therefore, the limit of the left hand side is no more than*

$$\begin{aligned}
&\limsup_{n\to\infty} P\Big(\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})<\mathcal{I}(\mathbf{a},\mathbf{b})-\delta\Big)\\
&\quad=\limsup_{n\to\infty} P\Big(\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})<\mathcal{I}(\mathbf{a},\mathbf{b})-\delta,\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})>\mathcal{I}(\mathbf{a},\mathbf{b})-\frac{\delta}{2}\Big)\\
&\qquad+\limsup_{n\to\infty} P\Big(\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})<\mathcal{I}(\mathbf{a},\mathbf{b})-\delta,\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})\le\mathcal{I}(\mathbf{a},\mathbf{b})-\frac{\delta}{2}\Big)\\
&\quad\le 0+\limsup_{n\to\infty} P\Big(\mathcal{I}(\boldsymbol{\alpha}_{1,n},\boldsymbol{\beta}_{1,n})\le\mathcal{I}(\mathbf{a},\mathbf{b})-\frac{\delta}{2}\Big)\\
&\quad=0.
\end{aligned}$$

Hence the result.

Acknowledgement

The authors would like to thank the Editor, An associate editor, and two referees whose suggestions led to a greatly improved paper. The work of Sriram was supported in part by by the NSF grant SES-0241651.

References

Anderson, E. (1935). The Irises of the gaspe peninsula. *Bull. Amer. Iris Soc.* **59**, 2-5.

Andrews, D. W. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica*, **55**, 1465-1471.

Das, S. and Sen, P. K. (1994). Restricted canonical correlation. *Linear Algebra Appl.* **210**, 29-47.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach.* Chapman and Hall, London.

Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization.* Academic Press, New York.

Goria, M. N. and Flury, B. D. (1996). Common canonical variates in k independent groups. *J. Amer. Statist. Assoc.* **91**, 1735-1742.

Hernández, A. and Velilla, S. (2005). Dimension reduction in nonparametric kernel discriminant analysis. *J. Comput. Graph. Statist.* **14**, 847-866.

Hotelling, H. (1935). The most predictable criterion, *Journal of Educational Psychology* **26**, 139-142.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28**, 321–377.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58**, 433-51.

Kettenring, J. R. (1985). Canonical correlation analysis. In *Encyclopedia of Statistical Sciences*, **1** (Edited by S. Kotz and N. L. Johnson), 354-65. John Wiley, New York.

Klein, R. W. and Spady, R. H. (1993). An effect semiparametric estimator for binary response models. *Econometrica*, **61**, 387-421.

Kullback, S. (1959). *Information Theory and Statistics.* John Wiley, New York.

Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55**, 725-740.

Li, B., Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Luijtens, K., Symons, F. and Vuylsteke-wauters, M. (1994). Linear and non-linear canonical correlation analysis: an exploratory tool for the analysis of group-structured data. *J. Appl. Stat.* **21**, 43-61.

Neuenschwander, B. E. and Flury, B. D. (1995). Common canonical variates. *Biometrika* **82**, 553-560.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

Shi, S. G. and Taam, W. (1992). Non-linear canonical correlation analysis with a simulated annealing solution. *J. Appl. Stat.* **19**, 155-165.

Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Van der Burg, E. and De Leeuw, J. (1983). Non-linear canonical correlation. *British J. Math. Statist. Psych.* **36**, 54-80.

Van der Burg, E., De Leeuw, J. and Verdegaal, R. (1988). Non-linear canonical correlation with m sets of variables. *Psychometrika* **2**, 171-197.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Yin, X. (2004). Canonical correlation analysis based on information theory. *J. Multivariate Anal.* **91**, 161-176.

Department of Statistics, University of Georgia, 204 Statistics Building, Athens, GA 30602, U.S.A.

E-mail: xryin@stat.uga.edu

Department of Statistics, University of Georgia, 204 Statistics Building, Athens, GA 30602, U.S.A.

E-mail: tn@stat.uga.edu

(Received February 2005; accepted July 2006)