Canonical Correlation Analysis based on Hilbert-Schmidt Independence Criterion and Centered Kernel Target Alignment

Billy Chang BILLY.CHANG@MAIL.UTORONTO.CA
Dalla Lana School of Public Health, University of Toronto, M5T 3M7, Canada

Uwe Kruger UWEKRUGER@SQU.EDU.OM
Dept. Mechanical & Industrial Engineering, Sultan Qaboos University, Al Khoud, Sultanate of Oman

Rafal Kustra R.KUSTRA@UTORONTO.CA
Dalla Lana School of Public Health, University of Toronto, M5T 3M7, Canada

Junping Zhang JPZHANG@FUDAN.EDU.CN
Shanghai Key Lab of Intelligent Information Processing & School of Computer Science, Fudan University, China

Abstract

Canonical correlation analysis (CCA) is a well established technique for identifying linear relationships among two variable sets. Kernel CCA (KCCA) is the most notable nonlinear extension but it lacks interpretability and robustness against irrelevant features. The aim of this article is to introduce two nonlinear CCA extensions that rely on the recently proposed Hilbert-Schmidt independence criterion and the centered kernel target alignment. These extensions determine linear projections that provide maximally dependent projected data pairs. The paper demonstrates that the use of linear projections allows removing irrelevant features, whilst extracting combinations of strongly associated features. This is exemplified through a simulation and the analysis of recorded data that are available in the literature.

1. Introduction

CCA, developed for discovering linear associations between two multivariate data sets, dates back to the early to mid 1930s (Hotelling, 1936). For given high-dimensional random vectors $\mathbf{x} \in \mathbb{R}^P$ and $\mathbf{y} \in \mathbb{R}^Q$, let x_i and y_i, $i = 1, \dots N$, be N independent realizations of $\mathbf{x}$ and $\mathbf{y}$, respectively. As a multivariate data analysis tool, CCA extracts canonical vectors u and v such that $u^T\mathbf{x}$ and $v^T\mathbf{y}$ possess a maximum correlation coefficient. These pairs of vectors reveal different linear associations that are encapsulated within $\mathbf{x}$ and $\mathbf{y}$. During the early development phase, CCA has seen applications predominantly in the field of psychology. To date, CCA has a wide range of applications including, for example, functional data analysis (Leurgans et al., 1993) and bioinformatics (Cao et al., 2009).

More recently, the research community proposed regularized approaches to CCA for simultaneous analysis of multiple high-dimension data sets (Witten & Tibshirani, 2009; Parkhomenko et al., 2009; Hardoon & Shawe-Taylor, 2011). Also the introduction of nonlinear extensions of CCA has received attention. Initially based on neural networks (Hsieh, 2000), using kernel methods (Bach & Jordan, 2002) has become a notable approach for extracting complex non-linear associations between data sets, such as those arising in image analysis and industrial process modelling (Hardoon et al., 2004; Sharma et al., 2006).

Kernel approaches to CCA, however, are often compromised by the following two key issues. Firstly, KCCA attempts to find canonical scalar functions f and g such that the correlation between the transformed variables $f(\mathbf{x})$ and $g(\mathbf{y})$ are maximized. Utilizing the kernel trick, KCCA operates in a high-dimensional reproducing kernel Hilbert space of functions, providing no interpretable results for subsequent exploratory analysis. Secondly, KCCA considers all P and Q elements stored in the x_i and y_i vectors for model estimation with no filtering procedure to remove

Proceedings of the 30^th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

irrelevant features. This can potentially affect the robustness of KCCA against redundant variables. While an appropriate choice of the regularization parameter may overcome the latter issue (Fukumizu et al., 2007), the tuning parameter selection problem for KCCA has not been adequately addressed in the literature.

(Balakrishnan et al., 2012) provides an attempt to address these issues by utilizing non-linear transformations that are modelled by sparse generalized additive models (Ravikumar et al., 2009). Introducing variable weights and a variable elimination step has allowed easier model interpretation and the removal of irrelevant variables. However, the use of generalized additive models inherently ignores interactions between variables, hence limiting the flexibility of the model for non-linear structure discovery. Furthermore, it is not clear how this method can produce multiple orthogonal functional transforms or weight vectors, properties respectively enjoyed by KCCA and CCA.

Another approach that attempts addressing these issues is described in (Sharma et al., 2006), where by modelling f and g using neural networks, the correlation between non-linearly transformed canonical variates, i.e. $f(u^T\mathbf{x})$ and $g(v^T\mathbf{y})$, is maximized. However, highly non-linear signals require complexly structured networks with multiple layers of neurons to capture; the need for an appropriate network design and the computational burden of neural network fitting may hamper its practical usefulness.

This article presents extensions of CCA that (i) can be robust against spurious dimensions and (ii) allow for multiple and interpretable canonical vectors. These extensions are motivated by noticing that the estimated non-linear transforms f and g (obtained using KCCA) can hardly provide any insight regarding the interrelationships between $\mathbf{x}$ and $\mathbf{y}$. This, however, is one of the benefits of linear CCA, as it utilizes u and v to formulate hypotheses or to conduct exploratory data analysis. Hence, a reliable estimation of u and v in a nonlinear context can reveal important dependencies between two variables sets. The determination of u and v is achieved by constructing objective functions for u and v based on two general measures of dependence: (i) the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), and (ii) the centered kernel target alignment (KTA) (Cortes et al., 2012). Since u and v are projection vectors in the original data spaces, they can provide interpretable insights into the relationships between $\mathbf{x}$ and $\mathbf{y}$ (first issue). Furthermore, inspecting the canonical variates u^Tx_i and v^Ty_i as one-dimensional compression guided by HSIC or KTA, only relevant features are utilized for non-linearity search, thereby decreasing the sensitivity of our approaches against irrelevant, noisy features (second issue). Finally, multiple canonical weight vectors can be easily obtained by iteratively performing the proposed algorithm on sequential orthogonal subspaces.

2. Preliminaries

This section provides a brief summary of CCA and KCCA first and introduces HSIC and KTA in Subsection 2.2, which form the basis of the proposed work.

2.1. CCA and KCCA

Defining $X \in \mathbb{R}^{N\times P}$ and $Y \in \mathbb{R}^{N\times Q}$, where the ith rows are x_i and y_i, respectively, which have column means of zero, CCA computes two canonical vectors u and v to maximize the following sample correlation:

$$cor(u^TX, v^TY) = \frac{u^TX^TYv}{\sqrt{(u^TX^TXu)(v^TY^TYv)}} \tag{1}$$

$$\text{subject to } ||u|| = ||v|| = 1$$

The unity constraint is superfluous here, but the algorithms in Section 4 require its incorporation.

KCCA maximizes $cor(f(\mathbf{x}), g(\mathbf{y}))$, i.e. the correlation between the nonlinear transformations of $\mathbf{x}$ and $\mathbf{y}$, by determining f and g, that is $f(\mathbf{x}), g(\mathbf{y})$, within respective reproducing kernel Hilbert spaces $\mathbb{H}_\mathbf{x}$ and $\mathbb{H}_\mathbf{y}$. Defining $k_x(\cdot,\cdot)$ and $k_y(\cdot,\cdot)$ by the associated kernel functions for $\mathbb{H}_\mathbf{x}$ and $\mathbb{H}_\mathbf{y}$ respectively, and let K_x and K_y be the Gram matrices for the N samples of $\mathbf{x}$ and $\mathbf{y}$, i.e. the (i,j)th entry of K_x and K_y are respectively $k_x(x_i,x_j)$ and $k_y(y_i,y_j)$, then the maximum of $cor(f(\mathbf{x}), g(\mathbf{y}))$ can be formulated as:

$$\max_{\alpha,\beta\in\mathbb{R}^N} \frac{\alpha^T\tilde{K}_x\tilde{K}_y\beta}{\sqrt{\alpha^T(\tilde{K}_x-\eta I)^2\alpha\beta^T(\tilde{K}_y-\eta I)^2\beta)}} \tag{2}$$

where $K^2 = K^TK$, η is a regularization constant, and $\tilde{K}$ is the centered Gram matrix of K, i.e. the (i,j)th entry of $\tilde{K}$ is:

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{N}\sum_{i=1}^{N}K_{ij} - \frac{1}{N}\sum_{j=1}^{N}K_{ij} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K_{ij}$$

The solution to the optimization problem (2) is the largest generalized eigenvalue of a system of generalized eigenvalue problem (Bach & Jordan, 2002).

As x_i and y_i are non-linearly transformed prior to sample correlation evaluation, (2) can be seen as a measure of non-linear correlation. However, evaluating (2) is practically inconvenient as it requires a careful choice of η, and solving the generalized eigenvalue problem can be computationally intensive. To overcome these issues, this article utilizes the HSIC and KTA measures which are introduced next.

2.2. HSIC and KTA

These are two criteria for determining non-linear associations that do not require the solution of generalized eigenvalue problems nor rely on regularization parameter by virtue of their constructions. Associated with $\mathbb{H}_{\mathbf{x}}$ and $\mathbb{H}_{\mathbf{y}}$, HSIC is the squared Hilbert-Schmidt norm of the cross-covariance operator between the probability space of $\mathbf{x}$ and $\mathbf{y}$ (Gretton et al., 2005). The empirical HSIC measure is defined as:

$$\rho_h = \frac{1}{(N-1)^2} tr(K_x \tilde{K}_y) \tag{3}$$

The fact that HSIC can be used as a measure of (in)dependence when associated with a universal kernel has been justified in (Gretton et al., 2005). This allows HSIC to be used for constructing kernel variants of independent component analysis (Shen et al., 2009), supervised and unsupervised dimension reduction (Fukumizu et al., 2009; Wang et al., 2010), and feature selection (Song et al., 2012b).

The KTA criterion has the following definition:

$$\rho_a = \frac{tr(K_x \tilde{K}_y)}{\sqrt{tr(K_x \tilde{K}_x) tr(K_y \tilde{K}_y)}} \tag{4}$$

Comparing (3) and (4), the KTA criterion is simply a normalized version of HSIC. Despite this similarity between KTA and HSIC, KTA has mainly been used for kernel selection and kernel learning (Cortes et al., 2012), areas of application rather dissimilar to the applications of HSIC mentioned above.

3. CCA based on HSIC and KTA

This section motivates the rationale behind the two algorithms to overcome the inherent limitations of KCCA, i.e. (i) the lack of interpretability due to the transformation of the data vectors into abstract Hilbert spaces and (ii) the inability of removing irrelevant features from the original variables. The two algorithms rely on the following objective function to estimate canonical vectors with respect to pre-defined measures of empirical non-linear correlation $ncor(.,.)$ between $u^T\mathbf{x}$ and $v^T\mathbf{y}$:

$$\max_{u,v;||u||=||v||=1} ncor(u^T\mathbf{x}, v^T\mathbf{y}) \tag{5}$$

Although KCCA provides a nonlinear correlation measure using $f(\mathbf{x})$ and $g(\mathbf{y})$, the proposed algorithms utilize the conceptually simpler HSIC and KTA criteria. Incorporating the HSIC criterion, the first proposed CCA variant, hsicCCA, computes u and v to maximize:

$$\rho_h(u,v) = \frac{1}{(N-1)^2} tr(K^u \tilde{K}^v) \quad s.t.\ ||u|| = ||v|| = 1 \tag{6}$$

where K^u and K^v are Gram matrices for projected data $u^T x_i$ and $v^T y_i$, i.e. their (i,j)th entry are $k^u(x_i, x_j) = k_x(u^T x_i, u^T x_j)$ and $k^v(y_i, y_j) = k_y(v^T y_i, v^T y_j)$ respectively. To highlight the relationship between CCA and hsicCCA, it is easy to show that using a linear kernel, (6) reduces to:

$$\rho_h(u,v) = (u^T X^T Y v)^2 \quad s.t.\ ||u|| = ||v|| = 1 \tag{7}$$

(7) is equivalent (up to squaring) to the diagonal CCA (DCCA) criterion introduced by (Witten & Tibshirani, 2009), which has been further developed into sparse CCA for high-dimensional genomic data analysis. It should be noted that DCCA assumes that elements within $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated, which makes DCCA a regularized form of CCA, particularly suitable for low-sample and high-dimensional data analysis when the correlation among the variables within each data set cannot be effectively estimated.

Next, incorporating the KTA criterion into (5), the second proposed algorithm of this article, ktaCCA, maximizes:

$$\rho_a(u,v) = \frac{tr(K^u \tilde{K}^v)}{\sqrt{tr(K^u \tilde{K}^u) tr(K^v \tilde{K}^v)}} \tag{8}$$

$$\text{subject to } ||u|| = ||v|| = 1$$

which gives rise to the ktaCCA algorithm, i.e. the second algorithms proposed in this article. Utilizing linear kernel functions, (8) reduces to the squared criterion of classical CCA (1). The objective functions for hsicCCA and ktaCCA are, hence, extensions of DCCA and CCA, respectively, to extract non-linear associations. As both algorithms rely on the original variables instead of their nonlinear transformations, the interpretability concern of KCCA does not arise for the hsicCCA and ktaCCA variates.

Moreover, the fact that (i) DCCA is a regularized version of CCA and (ii) hsicCCA and ktaCCA reduce to DCCA and CCA, respectively, if linear kernels functions are used, also highlights that hsicCCA is a regularized version of ktaCCA. According to (Gretton et al., 2005) and (Cortes et al., 2012), HSIC and KTA have different concentration bounds, depending on the choice of the kernel. The above considerations suggest that hsicCCA and ktaCCA can be viewed as natural nonlinear extensions of CCA. In section 5 we introduce the simulation study and other data experiments to shed light on the similarities and differences between the two methods.

4. Estimation Procedure

This section introduces a gradient-descent algorithm for solving (6) and (8). While other choices of kernel are available, this article employs the Gaussian RBF kernel exclusively. Allowing different bandwidth parameters for $k_x(\cdot)$ and $k_y(\cdot)$, these kernel functions are:

$$k_x(x_i, x_j) = \exp(-\sigma_x ||x_i - x_j||^2)$$

$$k_y(y_i, y_j) = \exp(-\sigma_y ||y_i - y_j||^2)$$

The kernel functions for the projected representations therefore become:

$$k^u(x_i, x_j) = \exp(-\sigma_x ||u^T(x_i - x_j)||^2)$$

$$k^v(y_i, y_j) = \exp(-\sigma_y ||v^T(y_i - y_j)||^2)$$

Ignoring the constant term, the gradient of (6) with respect to u^T and v^T are:

$$\frac{\partial \rho_h(u,v)}{\partial u^T} = -2\sigma_x u^T \sum_{i=1}^{N}\sum_{j=1}^{N} K^u_{ij}\tilde{K}^v_{ij}(x_i - x_j)(x_i - x_j)^T \tag{9}$$

$$\frac{\partial \rho_h(u,v)}{\partial v^T} = -2\sigma_y v^T \sum_{i=1}^{N}\sum_{j=1}^{N} \tilde{K}^u_{ij}K^v_{ij}(y_i - y_j)(y_i - y_j)^T \tag{10}$$

For ktaCCA (8), the gradients of $\log(\rho_a(u,v))$ will instead be considered. The gradients of $\log(\rho_a(u,v))$ with respect to u^T and v^T are:

$$\frac{\partial \log(\rho_a(u,v))}{\partial u^T} = u^T \sum_{i=1}^{N}\sum_{j=1}^{N} W^u_{ij}(x_i - x_j)(x_i - x_j)^T \tag{11}$$

$$\frac{\partial \log(\rho_a(u,v))}{\partial v^T} = v^T \sum_{i=1}^{N}\sum_{j=1}^{N} W^v_{ij}(y_i - y_j)(y_i - y_j)^T \tag{12}$$

where:

$$W^u_{ij} = -2\sigma_x K^u_{ij}\left(\frac{\tilde{K}^v_{ij}}{tr(K^u\tilde{K}^v)} - \frac{\tilde{K}^u_{ij}}{tr(K^u\tilde{K}^u)}\right)$$

$$W^v_{ij} = -2\sigma_y K^v_{ij}\left(\frac{\tilde{K}^u_{ij}}{tr(K^u\tilde{K}^v)} - \frac{\tilde{K}^v_{ij}}{tr(K^v\tilde{K}^v)}\right)$$

Detailed derivations of (9-12) will be presented in the supplementary.

For estimating u and v, this article considers a gradient descent algorithm, with a modified gradient to ensure the unit length constraint is satisfied at each step (Edelman et al., 1998). See Algorithm 1 for the training algorithm of hsicCCA. The algorithm for training ktaCCA is based on replacing the objective function (6) with the log-based objective function of (8), and by replacing the gradients (9,10) with the gradients (11,12). Optimal step-sizes θ_u, θ_v in Algorithm 1 can be found numerically, e.g. using the Nelder-Mead method.

Similar to classical CCA, upon finding an estimate for u and v, one can obtain further canonical weight vectors by solving (6) or (8) with $x_i - uu^T x_i$ and $y_i - vv^T y_i$ to obtain u_2 and v_2, i.e. solving (6) or (8) using the residuals resulting from projecting the data x_i and y_i to the orthogonal subspace of u and v respectively. This procedure can be repeated for finding u_3, v_3, u_4, v_4, etc. The maximum number of projections that can be obtained is therefore $\min(rank(X), rank(Y))$. It should be noted that the resulting sets of vectors $u_1, u_2, \ldots$ are mutually orthogonal, and the same property applies to $v_1, v_2, \ldots$. To implement this procedure, however, one must initialize the new projection vectors such that the initialized vector is orthogonal to all the previously obtained projection vectors in Algorithm 1.

The bandwidth parameter σ_x is chosen using the "median trick" (Song et al., 2012a), i.e. the median Euclidean distance between all pairs of (x_i, x_j), and σ_y is chosen similarly. Note that the median distance can change after the orthogonal residual subspace projection step described in the previous paragraph, and therefore the bandwidth parameters for estimating $u_1, v_1, u_2, v_2, \ldots$ are different.

5. Experiments

We use simulated and real datasets which have been used in the literature to investigate our two proposed methods in contrast with CCA, KCCA and DCCA. To avoid the problem of local minima, multiple random restarts are employed in each experiment.

5.1. Simulation

The simulation example, where source signals are artificially generated, along with some noise variables, tests the performance of KCCA and the various CCA variants to recover the source signals in the presence of noise and irrelevant variables. Here $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)^T$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4)^T$ are de-

Algorithm 1 Gradient Descent for hsicCCA

```
Input: data $x_i, y_i$, number of factors $M$, stopping criterion $e$
for $m = 1$ to $M$ do
  $\sigma_x = median(dist(x_i)), \sigma_y = median(dist(y_i))$
  $t = 1$
  initialize $u_{mt}, v_{mt}$
  repeat
    $obj = \rho_h(u_{mt}, v_{mt})$
    Compute $K^{u_{mt}}, K^{v_{mt}}, \tilde{K}^{u_{mt}}, \tilde{K}^{v_{mt}}$
    $\eta = -\frac{\partial \rho_h(u,v)}{\partial u}\big|_{u=u_{mt}}, \gamma = -\frac{\partial \rho_h(u,v)}{\partial v}\big|_{v=v_{mt}}$
    $h^{u_{mt}} = \eta - (\eta^T u_{mt})u_{mt}, n^{u_{mt}} = h^{u_{mt}}/|h^{u_{mt}}|$
    $h^{v_{mt}} = \gamma - (\gamma^T v_{mt})v_{mt}, n^{v_{mt}} = h^{v_{mt}}/|h^{v_{mt}}|$
    $u(\theta_u) = u_{mt}\cos\theta_u + n^{u_{mt}}\sin\theta_u$
    $v(\theta_v) = v_{mt}\cos\theta_v + n^{v_{mt}}\sin\theta_v$
    $(\theta_u, \theta_v) = argmin_{\theta_u,\theta_v}\{-\rho_h(u(\theta_u), v(\theta_v))\}$
    $t = t + 1$
    update $u_{mt} = u(\theta_u), v = v(\theta_v)$
  until $(\rho_h(u_{mt}, v_{mt}) - obj)/obj < e$
  $u_m = u_{mt}, v_m = v_{mt}$
  $x_i = x_i - u_m u_m^T x_i, y_i = y_i - v_m v_m^T y_i$
end for
Output: $u_1, \ldots, u_M, v_1, \ldots, v_M$
```

fined as follows:

$$z \sim \text{uniform}(-\pi, \pi)$$

$$\mathbf{x}_1 = \sin(z) + \epsilon_{x_1}; \mathbf{y}_1 = \cos(z) + \epsilon_{y_1}$$

$$\mathbf{x}_2 \sim N(0,1), \mathbf{x}_3 \sim N(0,1); \mathbf{y}_2 = \cos(\mathbf{x}_2 + \mathbf{x}_3) + \epsilon_{y_2}$$

$$\mathbf{x}_4 \sim N(0,1); \mathbf{y}_3 = \mathbf{x}_4 + \epsilon_{y_3}$$

$$\mathbf{x}_5 \sim N(0,1)$$

$$\mathbf{y}_4 \sim N(0,1)$$

$$\epsilon_{\mathbf{x}_1}, \epsilon_{\mathbf{y}_1} \sim N(0, \sigma = 0.1); \epsilon_{\mathbf{y}_2}, \epsilon_{\mathbf{y}_3} \sim N(0, \sigma = 0.5)$$

Here, $\mathbf{x}_1$ and $\mathbf{y}_1$ form a circle, while $\mathbf{x}_2 + \mathbf{x}_3$ and $\mathbf{y}_2$ form a cosine curve, and $\mathbf{x}_4$ and $\mathbf{y}_3$ are linearly correlated. N=100 samples are generated from $\mathbf{x}$ and $\mathbf{y}$ described above. The objective here is to examine how well hsicCCA, ktaCCA, CCA, DCCA, and KCCA can extract the three source signals.

The canonical variates resulting from the four CCA variants are shown in Figure 1. Whilst each method can discover the linear relationship between $\mathbf{x}_3$ and $\mathbf{y}_4$, only ktaCCA and hsicCCA can discover the remaining two non-linear patterns. Indeed, the three pairs of projection vectors $u_1, v_1, u_2, v_2, u_3, v_3$ obtained from ktaCCA are (rounded to one significant digit):

$$u_1 = (0, -0.1, -0.2, -\mathbf{1}, 0)^T; v_1 = (0, -0.1, -\mathbf{1}, 0)^T$$

$$u_2 = (0, -\mathbf{0.7}, -\mathbf{0.7}, 0.2, 0)^T; v_2 = (0.1, -\mathbf{1}, 0.1, -0.2)^T$$

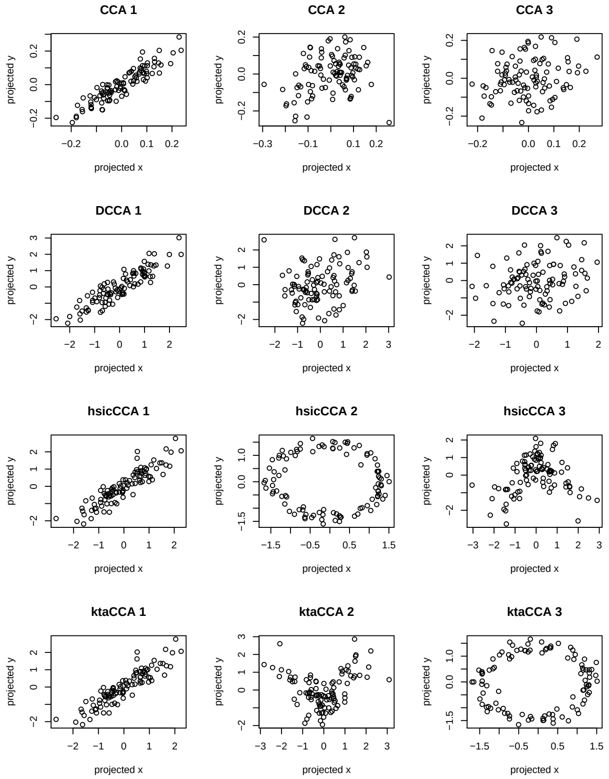

Figure 1. The projected canonical variates for the simulation study. The title of each plot represents the method and the order of the canonical variates.

$$u_3 = (\mathbf{1}, 0, 0, 0, 0)^T; v_3 = (-\mathbf{1}, -0.1, 0, 0)^T$$

The canonical weight vectors obtained by hsicCCA are similar, up to certain sign changes[1].

More precisely, (u_1, v_1) above reveals that $\mathbf{x}_4$ is correlated with $\mathbf{y}_3$ (i.e. the linear relation), (u_2, v_2) suggests that a combination of $\mathbf{x}_2$ and $\mathbf{x}_3$ is strongly related to $\mathbf{y}_2$ (i.e. the cosine relation), and (u_3, v_3) indicates the dependency between $\mathbf{x}_1$ and $\mathbf{y}_1$ (i.e. the circle). It should be noted that the vectors alone are unable to describe the type of relationships (linear or nonlinear), but they allow identifying the linear combinations of elements between the two data sets which are strongly associated.

Note from Figure 1 that the cosine signal and the circular signal are discovered by hsicCCA and ktaCCA in different orders, suggesting that hsicCCA and ktaCCA algorithms may differ in performance, depending on the structure of the interrelationship between the two sets of variables.

For the analysis of KCCA's results, note that although

[1] Sign changes preserve both the HSIC and KTA measures here, due to simple invariant properties of the Gaussian kernel. Details are omitted for brevity.

KCCA can only provide the transformed variables $f_1(x_i), f_2(x_i), \dots$ and $g_1(y_i), g_2(y_i), \dots$, it is still possible to investigate whether KCCA has identified the true source signals by investigating the sample correlation between the true signals and the KCCA transformed variables (Table 1). The circular signal is difficult to interpret, as there are many transformations for z which can generate a circle. A more detailed analysis has shown that KCCA has learned the $\cos(2z)$ function through its 5th pair of canonical functions. Furthermore, the 2nd pair of canonical functions have extracted the linear signal between $\mathbf{x}_3$ and $\mathbf{y}_4$, and the 3rd pair have discovered the cosine signal between $\mathbf{x}_2 + \mathbf{x}_3$ and $\mathbf{y}_2$. The results of Table 1 also suggest that the first pair of canonical functions has partially extracted the cosine signal, as indicated by the moderate sample correlation in the first column of Table 1. However, it is difficult to argue intuitively why there are two pairs of canonical functions (i.e. the first pair and the third pair) extracting the same cosine signal.

These results suggest that although KCCA has identified all of the three signals, they are not being discovered by the first three pairs of canonical functions $(f_1, g_1), (f_2, g_2)$, and (f_3, g_3). As the first three pairs of KCCA canonical functions are supposed to capture the three strongest associations, and that there are only three signals being generated by the simulation model, KCCA has apparently extracted certain redundant signals, demonstrating its lack of robustness against irrelevant variables.

Table 1. Absolute sample correlation between true signal and KCCA transformed signal. High correlations are boldfaced.

	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$
$\cos(2z)$	0.13	0.08	0.12	0.05	**0.92**
$\cos(x_2 + x_3)$	0.51	0.21	**0.83**	0.06	0.06
x_4	0.06	**0.94**	0.18	0.02	0.08
	$g_1(y)$	$g_2(y)$	$g_3(y)$	$g_4(y)$	$g_5(y)$
$\cos(2z)$	0	0.09	0.11	0.05	**0.91**
y_2	0.45	0.28	**0.85**	0.09	0.05
y_3	0.1	**0.95**	0.15	0.02	0.03

5.2. Canadian Weather Data

This data set records $P = Q = 12$ monthly average temperatures and the associated monthly average precipitations from $N = 35$ weather stations located at different parts of Canada (Ramsay & Silverman, 1997). As the true underlying variable interrelationships are unknown in this study, the application of KCCA is not considered here due to its lack of interpretability. The performance of CCA, DCCA, hsicCCA and ktaCCA in extracting the association between temperature and precipitation will be compared based on the cross-validation principle to be statistically sound.

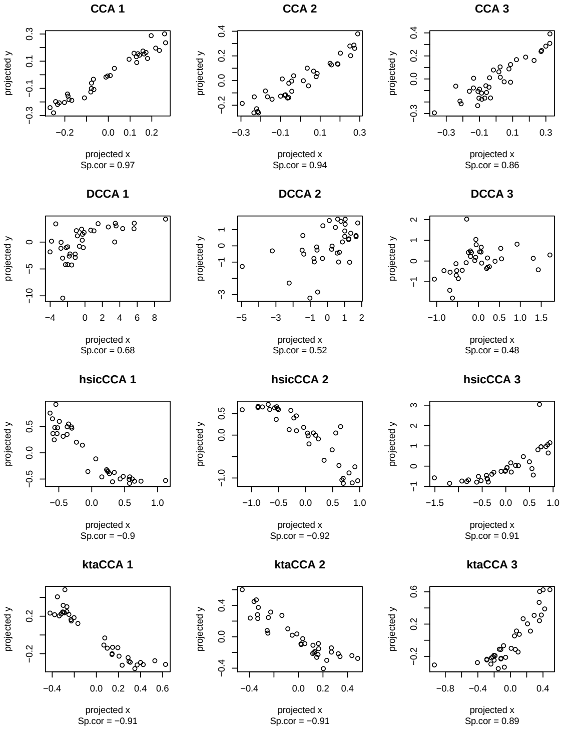

Figure 2. Canonical variates for the Canadian weather data. See caption of Figure 1. The subtitles present the Spearman correlation for the projected canonical variates.

Note that finding a good non-linear correlation measure for quantitative comparison is not trivial; the HSIC and KTA measures are not on the same scale, which circumvents a direct comparison of the computed dependence measures. Further, using HSIC and KTA as performance measures may provide a biased advantage for hsicCCA and ktaCCA, respectively, as these are methods built upon HSIC and KTA. Other measures such as KCCA contain tuning parameters, which are difficult to determine *a priori*. Fortunately, most non-linear associations discovered by the compared methods are monotone, and hence the Spearman correlation (Kruskal, 1958), a principled measure for monotone non-linear correlation, is a suitable measure for quantitative comparisons in this experiment.

Figure 2 presents the results obtained from fitting the four CCA variants on all $N = 35$ samples. The Spearman correlations are recorded in the subtitles of each plot. This figure shows that DCCA can only identify one moderately correlated pattern, while hsicCCA and ktaCCA can discover a series of strong non-linear patterns. Furthermore, the patterns discovered by hsicCCA and ktaCCA, based on the Spearman statistic, are all highly correlated. On the surface, CCA may have identified certain strongly correlated linear signals, but in a low-sample and high-dimensional setting, the above observation may be a result of over-fitting. The cross-validation analysis below will demonstrate that this is indeed the case.

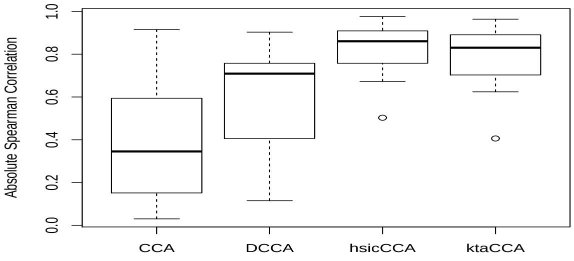

Figure 3. Canadian weather data: cross-validated absolute Spearman correlations for the 1st set of canonical variates.

The application of cross-validation relied on 20 random-splits with the inclusion of 25 samples to identify models using each of the four CCA variants. The Spearman correlation was then determined using the remaining 10 samples for each of the 20 random splits. Figure 3 shows box plots of the absolute Spearman correlation based on these splits for the first set of canonical vectors. The large range of low Spearman correlation values indicates over-fitting for standard CCA, while DCCA, being a regularized variant of CCA, has produced higher Spearman correlations with slightly less variation. As the strongest signal between the temperature and precipitation data is non-linear, however, ktaCCA and hsicCCA have produced significantly larger Spearman values than CCA and DCCA.

5.3. Boston Housing Data

This case study reported here relates to the Boston Housing Data (Harrison & Rubinfeld, 1978), a data set known to contain various non-linear signals. After removing variables containing categorical and discrete values, the variables “nitric oxide concentration” (nox), “proportion of units built prior to 1940” (age) and “percentage of lower status population” (lstat) will form one data set, and the variables “average room number” (rm), “distances to employment centers” (dis), and “median home value” (medv) will form the other data set. The number of samples in this data set is $N = 506$, while $P = Q = 3$.

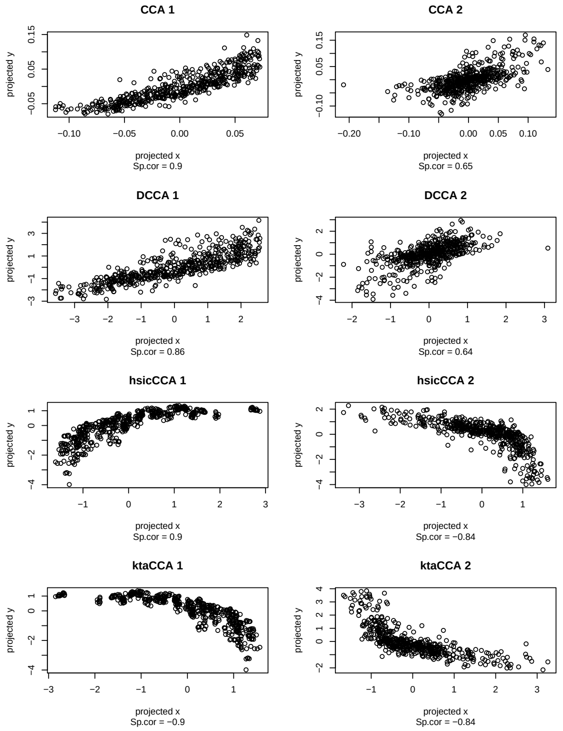

Figure 4. The two canonical variates for the Boston data. See caption of Figures 1 and 2.

Figure 4 shows the results of applying the four CCA variants using all 506 samples. KCCA was not included in this comparison, given that the nonlinear relationships are unknown. Figure 4 highlights that all four methods can identify a correlated signal through their first canonical projections. However, the next strongest signal is a non-linear one between “lstat” and “medv”, extracted only by ktaCCA and hsicCCA. For example, the 2nd pair of canonical weight vectors obtained by hsicCCA are (rounded to one significant digit):

$$u_2 = (0.1, -0.2, \mathbf{1}), v_2 = (-0.3, 0.1, -\mathbf{1})$$

where the last element of u_2 corresponds to the weight of “lstat”, and the last element of v_2 represents the weight of “medv”. The five-fold cross-validated results (Figure 5) further suggests that all CCA variants can stably extract a strongly correlated pattern by their first projections, but only ktaCCA and hsicCCA can extract the non-linear pattern through their second projections.

Further comparisons of hiscCCA and ktaCCA with CCA and DCCA based on explicit quadratic features are available in the supplementary.

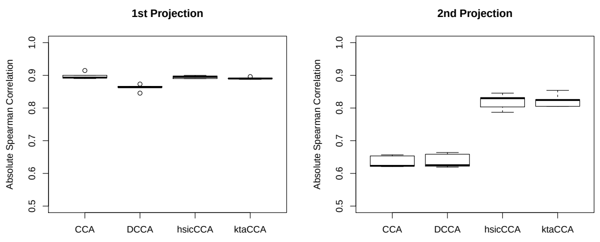

Figure 5. Cross-validated absolute Spearman correlations for the Boston housing data.

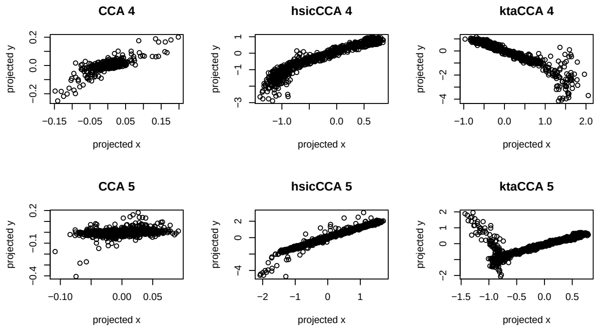

Figure 6. The 4th and 5th canonical variates for the Gaia data. See caption of Figures 1 and 2.

5.4. Gaia Data

To investigate the two proposed algorithms' performance on a larger-scale data, this final analysis considers a data set including 1000 randomly selected samples out of a total of 8286 samples of photon emission measures for 16 bands of wavelength intervals, generated through a computer simulation experiment conducted in (Bailer-Jones, 2010). To explore the interconnection between the bands, the first 8 bands are used to construct the x-variable set, and the other 8 bands form the y-variable set, i.e. $N = 1000$ and $P = Q = 8$ in this experiment. As the focus in this application relates to the large-sample performance, only the hsicCCA and the ktaCCA algorithms are contrasted with the standard linear CCA algorithm. All these methods have discovered three strong linear connections between the two groups of bands through their first three canonical projections (results not shown here). Figure 6, however, shows the 4th and 5th canonical projected variates of the three compared techniques and outlines that CCA is unable to extract further significant associations beyond its 3rd canonical projections, while hsicCCA and ktaCCA can discover further non-linear association patterns between the two groups of variables.

6. Discussion

This paper has proposed two nonlinear CCA extensions which discover multiple pairs of orthogonal canonical vectors that capture non-linear relationship. Our proposal has the advantage of interpretability of results and of dimensionality reduction which can ignore irrelevant dimensions. These are in contrast to other proposed methods, notably KCCA, whose results are not directly interpretable and rely on all input dimensions regardless of their relevance to the underlying structure.

The proposed algorithms carry out linear projections of the original variables first, which are subsequently used to determine objective functions that relate to HSIC and KTA. In contrast to KCCA, relying on orthogonal projections maintains the ability in identifying and interpreting variable interrelationships. Moreover, both algorithms are capable of extracting linear and nonlinear interrelationships and are therefore nonlinear extensions to DCCA and CCA.

Through a total of four experiments, the reported work has demonstrated hsicCCA and ktaCCA's robustness against noise and redundant variables, and their ability to provide interpretable results through the canonical weight vectors. Hence, these algorithms directly address and overcome the limitation of earlier work on nonlinear CCA.

The computational complexity for computing the gradients for ktaCCA and hsicCCA are both $O(N^2(P^2 + Q^2))$ (consult the supplementary for detailed speed trials). Approximation techniques, such as those described in (Jegelka & Gretton, 2007), may be used to reduce the computational burden of both algorithms. For model selection and regularization purposes, methods to choose M, σ_x and σ_y may be developed. Further empirical and theoretical analysis may also be pursued to shed light on the intrinsic similarity and differences between hsicCCA and ktaCCA.

An R package, "hsicCCA", that contains the implementation of the proposed algorithms, is available at the CRAN R-Repository.

References

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

Bailer-Jones, C. A. L. The ilium forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from gaia spectrophotometry. *Monthly Notices of the Royal Astronomical Society*, 403:96–116, 2010.

Balakrishnan, S., Puniyani, K., and Lafferty, J. Sparse additive functional and kernel CCA. In *ICML*, 2012.

Cao, K. L., Gonzalez, I., and Dejean, S. integromics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–2856, 2009.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *JMLR*, 13:795–828, 2012.

Edelman, A., Arias, T. A., and Smith, S. The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, 20(2):303–353, 1998.

Fukumizu, K., Bach, F. R., and Gretton, A. Statistical consistency of kernel canonical correlation analysis. *JMLR*, 8:361–383, 2007.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

Gretton, A., Bousquet, O., Smola, A., and Scholkopf, B. Measuring statistical dependence with hilbert-schmidt norm. In *Algorithmic learning theory*. 2005.

Hardoon, D. R. and Shawe-Taylor, J. Sparse canonical correlation analysis. *Machine Learning*, 83:331–353, 2011.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with applications to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Harrison, D. and Rubinfeld, D. L. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

Hotelling, H. Relations between two sets of variables. *Biometrika*, 28(3/4):321–377, 1936.

Hsieh, W. W. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13(10):1095–1105, 2000.

Jegelka, S. and Gretton, A. *Large-Scale Kernel Machines*, chapter 225, pp. 225–250. The MIT Press, 2007.

Kruskal, W. H. Ordinal measures of association. *JASA*, 53(284):814–861, 1958.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. Canonical correlation anlaysis when the data are curves. *JRSS Series B*, 55:725–740, 1993.

Parkhomenko, E., Tritchler, D., and Beyene, J. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

Ramsay, J. O. and Silverman, B. W. *Functional Data Analysis*. New York: Springer-Verlag, 1997.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *JRSS Series B*, 71(5): 1009–1030, 2009.

Sharma, S. K., Kruger, U., and Irwin, G. W. Deflation based nonlinear canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems*, 83:34–43, 2006.

Shen, H, Jegelka, S, and Gretton, A. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57(9):3498–3511, 2009.

Song, L., Boots, B., Siddiqi, S., Gordon, G., and Smola, A. Hilbert space embeddings of hidden markov models. In *ICML*, 2012a.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012b.

Wang, M., Sha, F., and Jordan, M. I. Unsupervised kernel dimension reduction. In *NIPS*, 2010.

Witten, D. M. and Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27, 2009.