A Tutorial on Canonical Correlation Methods

VIIVI UURTIO, Aalto University
JOÃO M. MONTEIRO, University College London
JAZ KANDOLA, Imperial College London
JOHN SHAWE-TAYLOR, University College London
DELMIRO FERNANDEZ-REYES, University College London
JUHO ROUSU, Aalto University

Canonical correlation analysis is a family of multivariate statistical methods for the analysis of paired sets of variables. Since its proposition, canonical correlation analysis has for instance been extended to extract relations between two sets of variables when the sample size is insufficient in relation to the data dimensionality, when the relations have been considered to be non-linear, and when the dimensionality is too large for human interpretation. This tutorial explains the theory of canonical correlation analysis including its regularised, kernel, and sparse variants. Additionally, the deep and Bayesian CCA extensions are briefly reviewed. Together with the numerical examples, this overview provides a coherent compendium on the applicability of the variants of canonical correlation analysis. By bringing together techniques for solving the optimisation problems, evaluating the statistical significance and generalisability of the canonical correlation model, and interpreting the relations, we hope that this article can serve as a hands-on tool for applying canonical correlation methods in data analysis.

CCS Concepts: •**Computing methodologies** → *Dimensionality reduction and manifold learning;*

General Terms: Multivariate Statistical Analysis, Machine Learning, Statistical Learning Theory

Additional Key Words and Phrases: Canonical correlation, regularisation, kernel methods, sparsity

ACM Reference Format:
Viivi Uurtio, João M. Monteiro, Jaz Kandola, John Shawe-Taylor, Delmiro Fernandez-Reyes, and Juho Rousu, 2017. A Tutorial on Canonical Correlation Methods. *ACM Comput. Surv.* 50, 6, Article 95 (October 2017), 33 pages.
DOI: 10.1145/3136624

1. INTRODUCTION

When a process can be described by two sets of variables corresponding to two different aspects, or views, analysing the relations between these two views may improve the understanding of the underlying system. In this context, a relation is a mapping of the observations corresponding to a variable of one view to the observations corresponding to a variable of the other view. For example in the field of medicine, one view could comprise variables corresponding to the symptoms of the disease and the other to the risk factors that can have an effect on the disease incidence. Identifying the relations between the symptoms and the risk factors can improve the understanding

Author's addresses: V. Uurtio (viivi.uurtio@aalto.fi) and J. Rousu (juho.rousu@aalto.fi), Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland; J. M. Monteiro (joao.monteiro@ucl.ac.uk), Department of Computer Science, University College London, and Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, Gower Street, London WC1E 6BT, UK; J. Shawe-Taylor (j.shawe-taylor@ucl.ac.uk) and D. Fernandez-Reyes (delmiro.fernandez-reyes@ucl.ac.uk), Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK; J. Kandola (j.kandola@imperial.ac.uk), Division of Brain Sciences, Imperial College London, DuCane Road, London WC12 0NN.

DOI: 10.1145/3136624

of the disease exposure and give indications for prevention and treatment. Examples of these kind of two-view settings, where the analysis of the relations could provide new information about the functioning of the system, occur in several other fields of science. These relations can be determined by means of canonical correlation methods that have been developed specifically for this purpose.

Since the proposition of canonical correlation analysis (CCA) by H. Hotelling [Hotelling 1935; Hotelling 1936], relations between variables have been explored in various fields of science. CCA was first applied to examine the relation of wheat characteristics to flour characteristics in an economics study by F. Waugh in 1942 [Waugh 1942]. Since then, studies in the fields of psychology [Hopkins 1969; Dunham and Kravetz 1975], geography [Monmonier and Finn 1973], medicine [Lindsey et al. 1985], physics [Wong et al. 1980], chemistry [Tu et al. 1989], biology [Sullivan 1982], time-series modeling [Heij and Roorda 1991], and signal processing [Schell and Gardner 1995] constitute examples of the early application fields of CCA.

In the beginning of the 21^{st} century, the applicability of CCA has been demonstrated in modern fields of science such as neuroscience, machine learning and bioinformatics. Relations have been explored for developing brain-computer interfaces [Cao et al. 2015; Nakanishi et al. 2015] and in the field imaging genetics [Fang et al. 2016]. CCA has also been applied for feature selection [Ogura et al. 2013], feature extraction and fusion [Shen et al. 2013], and dimension reduction [Wang et al. 2013]. Examples of application studies conducted in the fields of bioinformatics and computational biology include [Rousu et al. 2013; Seoane et al. 2014; Baur and Bozdag 2015; Sarkar and Chakraborty 2015; Cichonska et al. 2016]. The vast range of application domains emphasises the utility of CCA in extracting relations between variables.

Originally, CCA was developed to extract linear relations in overdetermined settings, that is when the number of observations exceeds the number of variables in either view. To extend CCA to underdetermined settings that often occur in modern data analysis, methods of regularisation have been proposed. When the sample size is small, Bayesian CCA also provides an alternative to perform CCA. The applicability of CCA to underdetermined settings has been further improved through sparsity-inducing norms that facilitate the interpretation of the final result. Kernel methods and neural networks have been introduced for uncovering non-linear relations. At present, canonical correlation methods can be used to extract linear and non-linear relations in both over- and underdetermined settings.

In addition to the already described variants of CCA, alternative extensions have been proposed, such as the semi-paired and multi-view CCA. In general, CCA algorithms assume one-to-one correspondence between the observations in the views, in other words, the data is assumed to be paired. However, in real datasets some of the observations may be missing in either view, which means that the observations are semi-paired. Examples of semi-paired CCA algorithms comprise [Blaschko et al. 2008], [Kimura et al. 2013], [Chen et al. 2012], and [Zhang et al. 2014]. CCA has also been extended to more than two views by [Horst 1961], [Carroll 1968], [Kettenring 1971], and [Van de Geer 1984]. In multi-view CCA the relations are sought among more than two views. Some of the modern extensions of multi-view CCA comprise its regularised [Tenenhaus and Tenenhaus 2011], kernelised [Tenenhaus et al. 2015], and sparse [Tenenhaus et al. 2014] variants. Application studies of multi-view CCA and its modern variants can be found in neuroscience [Kang et al. 2013], [Chen et al. 2014], feature fusion [Yuan et al. 2011] and dimensionality reduction [Yuan et al. 2014]. However, both the semi-paired and multi-view CCA are beyond the scope of this tutorial.

This tutorial begins with an introduction to the original formulation of CCA. The basic framework and statistical assumptions are presented. The techniques for solving the CCA optimisation problem are discussed. After solving the CCA problem, the approaches to interpret and evaluate the result are explained. The variants of CCA are illustrated using worked examples. Of the extended versions of CCA, the tutorial concentrates on the topics of regularised, kernel, and sparse CCA. Additionally, the deep and Bayesian CCA variants are briefly reviewed. This tutorial acquaints the reader with canonical correlation methods, discusses where they are applicable and what kind of information can be extracted.

2. CANONICAL CORRELATION ANALYSIS

2.1. The Basic Principles of CCA

CCA is a two-view multivariate statistical method. In multivariate statistical analysis, the data comprises multiple variables measured on a set of observations or individuals. In the case of CCA, the variables of an observation can be partitioned into two sets that can be seen as the two views of the data. This can be illustrated using the following notations. Let the views a and b be denoted by the matrices X_a and X_b, of sizes $n \times p$ and $n \times q$ respectively. The row vectors $\mathbf{x}_a^k \in \mathbb{R}^p$ and $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$ denote the sets of empirical multivariate observations in X_a and X_b respectively. The observations are assumed to be jointly sampled from a normal multivariate distribution. A reason for this is that the normal multivariate model approximates well the distribution of continuous measurements in several sampled distributions [Anderson 2003]. The column vectors $\mathbf{a}_i \in \mathbb{R}^n$ for $i = 1, 2, \ldots, p$ and $\mathbf{b}_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, q$ denote the variable vectors of the n observations respectively. The inner product between two vectors is either denoted by $\langle \mathbf{a}, \mathbf{b} \rangle$ or $\mathbf{a}^T\mathbf{b}$. Throughout this tutorial, we assume that the variables are standardised to zero mean and unit variance. In CCA, the aim is to extract the linear relations between the variables of X_a and X_b.

CCA is based on linear transformations. We consider the following transformations

$$X_a\mathbf{w}_a = \mathbf{z}_a \quad \text{and} \quad X_b\mathbf{w}_b = \mathbf{z}_b$$

where $X_a \in \mathbb{R}^{n\times p}$, $\mathbf{w}_a \in \mathbb{R}^p$, $\mathbf{z}_a \in \mathbb{R}^n$, $X_b \in \mathbb{R}^{n\times q}$, $\mathbf{w}_b \in \mathbb{R}^q$, and $\mathbf{z}_b \in \mathbb{R}^n$. The data matrices X_a and X_b represent linear transformations of the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ onto the images $\mathbf{z}_a$ and $\mathbf{z}_b$ in the space $\mathbb{R}^n$. The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ are often referred to as canonical weight vectors and the images $\mathbf{z}_a$ and $\mathbf{z}_b$ are also termed as canonical variates or scores. The constraints of CCA on the mappings are that the position vectors of the images $\mathbf{z}_a$ and $\mathbf{z}_b$ are unit norm vectors and that the enclosing angle, $\theta \in [0, \frac{\pi}{2}]$ [Golub and Zha 1995; Dauxois and Nkiet 1997], between $\mathbf{z}_a$ and $\mathbf{z}_b$ is minimised. The cosine of the angle, also referred to as the canonical correlation, between the images $\mathbf{z}_a$ and $\mathbf{z}_b$ is given by the formula $\cos(\mathbf{z}_a, \mathbf{z}_b) = \langle \mathbf{z}_a, \mathbf{z}_b \rangle / ||\mathbf{z}_a|| ||\mathbf{z}_b||$ and due to the unit norm constraint $\cos(\mathbf{z}_a, \mathbf{z}_b) = \langle \mathbf{z}_a, \mathbf{z}_b \rangle$. Hence the basic principle of CCA is to find two positions $\mathbf{w}_a \in \mathbb{R}^p$ and $\mathbf{w}_b \in \mathbb{R}^q$ that after the linear transformations $X_a \in \mathbb{R}^{n\times p}$ and $X_b \in \mathbb{R}^{n\times q}$ are mapped onto an n-dimensional unit ball and located in such a way that the cosine of the angle between the position vectors of their images $\mathbf{z}_a \in \mathbb{R}^n$ and $\mathbf{z}_b \in \mathbb{R}^n$ is maximised.

The images $\mathbf{z}_a$ and $\mathbf{z}_b$ of the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ that result in the smallest angle, θ_1, determine the first canonical correlation which equals $\cos\theta_1$ [Björck and Golub 1973]. The smallest angle is given by

$$\begin{aligned} \cos\theta_1 = \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle, \\ ||\mathbf{z}_a||_2 = 1 \quad ||\mathbf{z}_b||_2 = 1 \end{aligned} \tag{1}$$

Let the maximum be obtained by $\mathbf{z}_a^1$ and $\mathbf{z}_b^1$. The pair of images $\mathbf{z}_a^2$ and $\mathbf{z}_b^2$, that has the second smallest enclosing angle θ_2, is found in the orthogonal complements of $\mathbf{z}_a^1$ and $\mathbf{z}_b^1$. The procedure is continued until no more pairs are found. Hence the r angles $\theta_r \in [0, \frac{\pi}{2}]$ for $r = 1, 2, \cdots, q$ when $p > q$ that can be found are recursively defined by

$$\begin{gathered}\cos\theta_r = \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a^r, \mathbf{z}_b^r \rangle, \\ ||\mathbf{z}_a^r||_2 = 1 \quad ||\mathbf{z}_b^r||_2 = 1 \\ \langle \mathbf{z}_a^r, \mathbf{z}_a^j \rangle = 0 \quad \langle \mathbf{z}_b^r, \mathbf{z}_b^j \rangle = 0, \\ \forall j \neq r: \quad j, r = 1, 2, \ldots, \min(p, q).\end{gathered}$$

The number of canonical correlations, r, corresponds to the dimensionality of CCA. Qualitatively, the dimensionality of CCA can be also seen as the number of patterns that can be extracted from the data.

When the dimensionality of CCA is large, it may not be relevant to solve all the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and images $\mathbf{z}_a$ and $\mathbf{z}_b$. In general, the value of the canonical correlation and the statistical significance are considered to convey the importance of the pattern. The first estimation strategy for finding the number of statistically significant canonical correlation coefficients was proposed in [Bartlett 1941]. The techniques have been further developed in [Fujikoshi and Veitch 1979; Tu 1991; Gunderson and Muirhead 1997; Yamada and Sugiyama 2006; Lee 2007; Sakurai 2009].

In summary, the principle behind CCA is to find two positions in the two data spaces respectively that have images on a unit ball such that the angle between them is minimised and consequently the canonical correlation is maximised. The linear transformations of the positions are given by the data matrices. The number of relevant positions can be determined by analysing the values of the canonical correlations or by applying statistical significance tests.

2.2. Finding the positions and the images in CCA

The position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ having images $\mathbf{z}_a$ and $\mathbf{z}_b$ in the new coordinate system of a unit ball that have a maximum cosine of the angle in between can be obtained using techniques of functional analysis. The eigenvalue-based methods comprise solving a standard eigenvalue problem, as originally proposed by Hotelling in [Hotelling 1936], or a generalised eigenvalue problem [Bach and Jordan 2002; Hardoon et al. 2004]. Alternatively, the positions and the images can be found using the singular value decomposition (SVD), as introduced in [Healy 1957]. The techniques can be considered as standard ways of solving the CCA problem.

Solving CCA Through the Standard Eigenvalue Problem. In the technique of Hotelling, both the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and the images $\mathbf{z}_a$ and $\mathbf{z}_b$ are obtained by solving a standard eigenvalue problem. The Lagrange multiplier technique [Hotelling 1936; Hooper 1959] is employed to obtain the characteristic equation. Let X_a and X_b denote the data matrices of sizes $n \times p$ and $n \times q$ respectively. The sample covariance matrix C_{ab} between the variable column vectors in X_a and X_b is $C_{ab} = \frac{1}{n-1} X_a^T X_b$. The empirical variance matrices between the variables in X_a and X_b are given by $C_{aa} = \frac{1}{n-1} X_a^T X_a$ and $C_{bb} = \frac{1}{n-1} X_b^T X_b$ respectively. The joint covariance matrix is then

$$\begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix}. \tag{2}$$

The first and greatest canonical correlation that corresponds to the smallest angle is between the first pair of images $\mathbf{z}_a = X_a\mathbf{w}_a$ and $\mathbf{z}_b = X_b\mathbf{w}_b$. Since the correlation between $\mathbf{z}_a$ and $\mathbf{z}_b$ does not change with the scaling of $\mathbf{z}_a$ and $\mathbf{z}_b$, we can constrain $\mathbf{w}_a$ and $\mathbf{w}_b$ to be such that $\mathbf{z}_a$ and $\mathbf{z}_b$ have unit variance. This is given by

$$\mathbf{z}_a^T\mathbf{z}_a = \mathbf{w}_a^T X_a^T X_a \mathbf{w}_a = \mathbf{w}_a^T C_{aa}\mathbf{w}_a = 1, \tag{3}$$

$$\mathbf{z}_b^T\mathbf{z}_b = \mathbf{w}_b^T X_b^T X_b \mathbf{w}_b = \mathbf{w}_b^T C_{bb}\mathbf{w}_b = 1. \tag{4}$$

Due to the normality assumption and comparability, the variables of X_a and X_b should be centered such that they have zero means. In this case, the covariance between $\mathbf{z}_a$ and $\mathbf{z}_b$ is given by

$$\mathbf{z}_a^T\mathbf{z}_b = \mathbf{w}_a^T X_a^T X_b \mathbf{w}_b = \mathbf{w}_a^T C_{ab}\mathbf{w}_b. \tag{5}$$

Substituting (5), (3) and (4) into the algebraic problem in Equation (1), we obtain:

$$\cos\theta = \max_{\mathbf{z}_a,\mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle = \max_{\mathbf{w}_a \in \mathbb{R}^p, \mathbf{w}_b \in \mathbb{R}^q} \mathbf{w}_a^T C_{ab}\mathbf{w}_b,$$

$$||\mathbf{z}_a||_2 = \sqrt{\mathbf{w}_a^T C_{aa}\mathbf{w}_a} = 1 \quad ||\mathbf{z}_b||_2 = \sqrt{\mathbf{w}_b^T C_{bb}\mathbf{w}_b} = 1.$$

In general, the constraints (3) and (4) are expressed in squared form, $\mathbf{w}_a^T C_{aa}\mathbf{w}_a = 1$ and $\mathbf{w}_b^T C_{bb}\mathbf{w}_b = 1$. The problem can be solved using the Lagrange multiplier technique. Let

$$L = \mathbf{w}_a^T C_{ab}\mathbf{w}_b - \frac{\rho_1}{2}(\mathbf{w}_a^T C_{aa}\mathbf{w}_a - 1) - \frac{\rho_2}{2}(\mathbf{w}_b^T C_{bb}\mathbf{w}_b - 1) \tag{6}$$

where ρ_1 and ρ_2 denote the Lagrange multipliers. Differentiating L with respect to $\mathbf{w}_a$ and $\mathbf{w}_b$ gives

$$\frac{\delta L}{\delta \mathbf{w}_a} = C_{ab}\mathbf{w}_b - \rho_1 C_{aa}\mathbf{w}_a = \mathbf{0} \tag{7}$$

$$\frac{\delta L}{\delta \mathbf{w}_b} = C_{ba}\mathbf{w}_a - \rho_2 C_{bb}\mathbf{w}_b = \mathbf{0} \tag{8}$$

Multiplying (7) from the left by $\mathbf{w}_a^T$ and (8) from the left by $\mathbf{w}_b^T$ gives

$$\mathbf{w}_a^T C_{ab}\mathbf{w}_b - \rho_1 \mathbf{w}_a^T C_{aa}\mathbf{w}_a = 0$$

$$\mathbf{w}_b^T C_{ba}\mathbf{w}_a - \rho_2 \mathbf{w}_b^T C_{bb}\mathbf{w}_b = 0.$$

Since $\mathbf{w}_a^T C_{aa}\mathbf{w}_a = 1$ and $\mathbf{w}_b^T C_{bb}\mathbf{w}_b = 1$, we obtain that

$$\rho_1 = \rho_2 = \rho. \tag{9}$$

Substituting (9) into Equation (7) we obtain

$$\mathbf{w}_a = \frac{C_{aa}^{-1}C_{ab}\mathbf{w}_b}{\rho}. \tag{10}$$

Substituting (10) into (8) we obtain

$$\frac{1}{\rho}C_{ba}C_{aa}^{-1}C_{ab}\mathbf{w}_b - \rho C_{bb}\mathbf{w}_b = 0$$

which is equivalent to the generalised eigenvalue problem of the form

$$C_{ba}C_{aa}^{-1}C_{ab}\mathbf{w}_b = \rho^2 C_{bb}\mathbf{w}_b.$$

If C_{bb} is invertible, the problem reduces to a standard eigenvalue problem of the form

$$C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab}\mathbf{w}_b = \rho^2\mathbf{w}_b.$$

The eigenvalues of the matrix $C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab}$ are found by solving the characteristic equation

$$|C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab} - \rho^2 I| = 0.$$

The square roots of the eigenvalues correspond to the canonical correlations. The technique of solving the standard eigenvalue problem is shown in Example 2.1.

Example 2.1. We generate two data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 60$, $p = 4$ and $q = 3$, respectively as follows. The variables of X_a are generated from a random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4 \sim N(0,1)$. We generate the following linear relations

$$\begin{aligned}\mathbf{b}_1 &= \mathbf{a}_3 + \boldsymbol{\xi}_1\\ \mathbf{b}_2 &= \mathbf{a}_1 + \boldsymbol{\xi}_2\\ \mathbf{b}_3 &= -\mathbf{a}_4 + \boldsymbol{\xi}_3\end{aligned}$$

where $\boldsymbol{\xi}_1 \sim N(0, 0.2), \boldsymbol{\xi}_2 \sim N(0, 0.4)$, and $\boldsymbol{\xi}_3 \sim N(0, 0.3)$ denote vectors of normal noise. The data is standardised such that every variable has zero mean and unit variance. The joint covariance matrix C in (2) of the generated data is given by

$$C = \left(\begin{array}{cccc|ccc} 1.00 & 0.34 & -0.11 & 0.21 & -0.10 & 0.92 & -0.21\\ 0.34 & 1.00 & -0.08 & 0.03 & -0.10 & 0.34 & 0.06\\ -0.11 & -0.08 & 1.00 & -0.30 & 0.98 & -0.03 & 0.30\\ 0.21 & 0.03 & -0.30 & 1.00 & -0.25 & 0.12 & -0.94\\ \hline -0.10 & -0.10 & 0.98 & -0.25 & 1.00 & -0.03 & 0.25\\ 0.92 & 0.34 & -0.03 & 0.12 & -0.03 & 1.00 & -0.13\\ -0.21 & 0.06 & 0.30 & -0.94 & 0.25 & -0.13 & 1.00 \end{array}\right) = \left(\begin{array}{c|c} C_{aa} & C_{ab}\\ \hline C_{ba} & C_{bb}\end{array}\right).$$

Now we compute the eigenvalues of the characteristic equation

$$|C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab} - \rho^2 I| = 0.$$

The square roots of the eigenvalues of $C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab}$ are $\rho_1 = 0.99$, $\rho_2 = 0.94$, and $\rho_3 = 0.92$. The eigenvectors $\mathbf{w}_b$ satisfy the equation

$$(C_{bb}^{-1}C_{ba}C_{aa}^{-1}C_{ab} - \rho^2 I)\mathbf{w}_b = 0.$$

Hence we obtain

$$\mathbf{w}_b^1 = \begin{pmatrix}-0.97\\-0.04\\-0.22\end{pmatrix} \mathbf{w}_b^2 = \begin{pmatrix}-0.39\\-0.37\\0.85\end{pmatrix} \mathbf{w}_b^3 = \begin{pmatrix}0.19\\-0.86\\-0.46\end{pmatrix}$$

and $\mathbf{w}_a$ vectors satisfy

$$\mathbf{w}_a^1 = \frac{C_{aa}^{-1}C_{ab}\mathbf{w}_b^1}{\rho_1} = \begin{pmatrix}-0.04\\-0.00\\-0.99\\0.18\end{pmatrix} \mathbf{w}_a^2 \quad = \frac{C_{aa}^{-1}C_{ab}\mathbf{w}_b^2}{\rho_2} = \begin{pmatrix}-0.41\\0.09\\-0.41\\-0.83\end{pmatrix} \mathbf{w}_a^3 = \frac{C_{aa}^{-1}C_{ab}\mathbf{w}_b^3}{\rho_3} = \begin{pmatrix}-0.84\\-0.10\\0.14\\0.52\end{pmatrix}.$$

The vectors $\mathbf{w}_b^1, \mathbf{w}_b^2$, and $\mathbf{w}_b^3$ and $\mathbf{w}_a^1, \mathbf{w}_a^2$, and $\mathbf{w}_a^3$ correspond to the pairs of positions $(\mathbf{w}_a^1, \mathbf{w}_b^1), (\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ that have the images $(\mathbf{z}_a^1, \mathbf{z}_b^1), (\mathbf{z}_a^2, \mathbf{z}_b^2)$ and $(\mathbf{z}_a^3, \mathbf{z}_b^3)$. In linear CCA, the canonical correlations equal to the square roots of the eigenvalues, that is $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.99$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.94$, and $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.92$. □

Solving CCA Through the Generalised Eigenvalue Problem. The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and their images $\mathbf{z}_a$ and $\mathbf{z}_b$ can also be solved through a generalised eigenvalue problem [Bach and Jordan 2002; Hardoon et al. 2004]. The equations in (7) and (8) can be represented as simultaneous equations

$$\begin{aligned} C_{ab}\mathbf{w}_b &= \rho C_{aa}\mathbf{w}_a \\ C_{ba}\mathbf{w}_a &= \rho C_{bb}\mathbf{w}_b \end{aligned}$$

that are equivalent to

$$\begin{pmatrix} \mathbf{0} & C_{ab} \\ C_{ba} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \rho \begin{pmatrix} C_{aa} & \mathbf{0} \\ \mathbf{0} & C_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}. \tag{11}$$

The equation (11) represents a generalised eigenvalue problem of the form $\beta A\mathbf{x} = \alpha B\mathbf{x}$ where the pair $(\beta, \alpha) = (1, \alpha)$ is an eigenvalue of the pair (A, B) [Saad 2011; Golub and Van Loan 2012]. The pair of matrices $A \in \mathbb{R}^{(p+q)\times(p+q)}$ and $B \in \mathbb{R}^{(p+q)\times(p+q)}$ is also referred to as matrix pencil. In particular, A is symmetric and B is symmetric positive-definite. The pair (A, B) is then called the symmetric pair. As shown in [Watkins 2004], a symmetric pair has real eigenvalues and $(p+q)$ linearly independent eigenvectors. To express the generalised eigenvalue problem in the form $A\mathbf{x} = \rho B\mathbf{x}$, the generalised eigenvalue is given by $\rho = \frac{\alpha}{\beta}$. Since the generalised eigenvalues come in pairs $\{\rho_1, -\rho_1, \rho_2, -\rho_2, \ldots, \rho_p, -\rho_p, 0\}$ where $p < q$, the positive generalised eigenvalues correspond to the canonical correlations.

Example 2.2. Using the data in Example 2.1, we apply the formulation of the generalised eigenvalue problem to obtain the positions $\mathbf{w}_a$ and $\mathbf{w}_b$. The resulting generalised eigenvalues are

$$\{0.99, 0.94, 0.92, 0.00, -0.92, -0.94, -0.99\}.$$

The generalised eigenvectors that correspond to the positive generalised eigenvalues in descending order are

$$\mathbf{w}_a^1 = \begin{pmatrix} -0.04 \\ -0.00 \\ -1.00 \\ 0.18 \end{pmatrix} \mathbf{w}_a^2 = \begin{pmatrix} 0.48 \\ -0.11 \\ 0.48 \\ 0.98 \end{pmatrix} \mathbf{w}_a^3 = \begin{pmatrix} -0.97 \\ -0.11 \\ 0.16 \\ 0.60 \end{pmatrix}$$

$$\mathbf{w}_b^1 = \begin{pmatrix} -0.98 \\ -0.04 \\ -0.23 \end{pmatrix} \mathbf{w}_b^2 = \begin{pmatrix} 0.46 \\ 0.43 \\ -1.00 \end{pmatrix} \mathbf{w}_b^3 = \begin{pmatrix} 0.22 \\ -1.00 \\ -0.54 \end{pmatrix}$$

The vectors $\mathbf{w}_a^1, \mathbf{w}_a^2$, and $\mathbf{w}_a^3$ and $\mathbf{w}_b^1, \mathbf{w}_b^2$, and $\mathbf{w}_b^3$ correspond to the pairs of positions $(\mathbf{w}_a^1, \mathbf{w}_b^1)$, $(\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$. The canonical correlations are $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.99$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.94$, and $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.92$.

The entries of the position pairs differ to some extent from the solutions to the standard eigenvalue problem in the Example 2.1. This is due to the numerical algorithms that are applied to solve the eigenvalues and eigenvectors. Additionally, the signs may also be opposite. This can be seen when comparing the second pairs of positions with the Example 2.1. This results from the symmetric nature of CCA. □

Solving CCA Using the SVD. The technique of applying the SVD to solve the CCA problem was first introduced by [Healy 1957] and described by [Ewerbring and Luk 1989] as follows. First, the variance matrices C_{aa} and C_{bb} are transformed into identity forms. Due to the symmetric positive definite property, the square root factors of the matrices can be found using a Cholesky or eigenvalue decomposition:

$$C_{aa} = C_{aa}^{1/2} C_{aa}^{1/2} \quad \text{and} \quad C_{bb} = C_{bb}^{1/2} C_{bb}^{1/2}.$$

Applying the inverses of the square root factors symmetrically on the joint covariance matrix in (2) we obtain

$$\begin{pmatrix} C_{aa}^{-1/2} & \mathbf{0} \\ \mathbf{0} & C_{bb}^{-1/2} \end{pmatrix} \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix} \begin{pmatrix} C_{aa}^{-1/2} & \mathbf{0} \\ \mathbf{0} & C_{bb}^{-1/2} \end{pmatrix} = \begin{pmatrix} I_q & C_{aa}^{-1/2} C_{ab} C_{bb}^{-1/2} \\ C_{bb}^{-1/2} C_{ba} C_{aa}^{-1/2} & I_p \end{pmatrix}.$$

The position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ can hence be obtained by solving the following SVD

$$C_{aa}^{-1/2} C_{ab} C_{bb}^{-1/2} = U^T S V \tag{12}$$

where the columns of the matrices U and V correspond to the sets of orthonormal left and right singular vectors respectively. The singular values of matrix S correspond to the canonical correlations. The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ are obtained from

$$\mathbf{w}_a = C_{aa}^{-1/2} U \quad \mathbf{w}_b = C_{bb}^{-1/2} V$$

The method is shown in Example 2.3.

Example 2.3. The method of solving CCA using the SVD is demonstrated using the data of Example 2.1. We compute the matrix

$$C_{aa}^{-1/2} C_{ab} C_{bb}^{-1/2} = \begin{pmatrix} -0.02 & 0.90 & -0.06 \\ -0.07 & 0.20 & 0.11 \\ 0.98 & 0.04 & 0.04 \\ 0.01 & -0.02 & -0.93 \end{pmatrix}.$$

The SVD gives

$$C_{aa}^{-1/2} C_{ab} C_{bb}^{-1/2} =$$

$$\underbrace{\begin{pmatrix} -0.03 & -0.03 & 0.95 & -0.30 \\ -0.47 & 0.03 & -0.28 & 0.84 \\ -0.86 & -0.26 & 0.11 & 0.44 \end{pmatrix}}_{U^T} \underbrace{\begin{pmatrix} 0.99 & 0.00 & 0.00 \\ 0.00 & 0.94 & 0.00 \\ 0.00 & 0.00 & 0.92 \\ 0.00 & 0.00 & 0.00 \end{pmatrix}}_{S} \underbrace{\begin{pmatrix} 0.95 & -0.29 & 0.15 \\ 0.01 & -0.44 & -0.90 \\ 0.33 & 0.85 & -0.41 \end{pmatrix}}_{V}.$$

The singular values of the matrix S correspond to the canonical correlations. The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ are given by

$$\mathbf{w}_a^1 = C_{aa}^{-1/2} \mathbf{u}^1 = \begin{pmatrix} 0.04 \\ 0.00 \\ 0.94 \\ -0.17 \end{pmatrix} \mathbf{w}_a^2 = C_{aa}^{-1/2} \mathbf{u}^2 = \begin{pmatrix} -0.43 \\ 0.10 \\ -0.43 \\ -0.87 \end{pmatrix} \mathbf{w}_a^3 = C_{aa}^{-1/2} \mathbf{u}^3 = \begin{pmatrix} -0.91 \\ -0.10 \\ 0.14 \\ 0.56 \end{pmatrix}$$

$$\mathbf{w}_b^1 = C_{bb}^{-1/2} \mathbf{v}^1 = \begin{pmatrix} 0.93 \\ 0.04 \\ 0.21 \end{pmatrix} \mathbf{w}_b^2 = C_{bb}^{-1/2} \mathbf{v}^2 = \begin{pmatrix} -0.40 \\ -0.38 \\ 0.89 \end{pmatrix} \mathbf{w}_b^3 = C_{bb}^{-1/2} \mathbf{v}^3 = \begin{pmatrix} 0.21 \\ -0.93 \\ -0.50 \end{pmatrix}$$

where $\mathbf{u}^i$ and $\mathbf{v}^i$ for $i = 1, 2, 3$ correspond to the left and right singular vectors. The vectors $\mathbf{w}_a^1, \mathbf{w}_a^2$, and $\mathbf{w}_a^3$ and $\mathbf{w}_b^1, \mathbf{w}_b^2$, and $\mathbf{w}_b^3$ correspond to the pairs of positions $(\mathbf{w}_a^1, \mathbf{w}_b^1), (\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$. The canonical correlations are $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.99$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.94$, and $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.92$. □

The main motivation for improving the eigenvalue-based technique was the computational complexity. The standard and generalised eigenvalue methods scale with the cube of the input matrix dimension, in other words, the time complexity is $\mathcal{O}(n^3)$, for a matrix of size $n \times n$. The input matrix $C_{aa}^{-1/2} C_{ab} C_{bb}^{-1/2}$ in the SVD-based technique is rectangular. This gives a time complexity of $\mathcal{O}(mn^2)$, for a matrix of size $m \times n$. Hence the SVD-based technique is computationally more tractable for very large datasets.

To recapitulate, the images $\mathbf{z}_a$ and $\mathbf{z}_b$ of the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ that successively maximise the canonical correlation can be obtained by solving a standard [Hotelling 1936] or a generalised eigenvalue problem [Bach and Jordan 2002; Hardoon et al. 2004] or by applying the SVD [Healy 1957; Ewerbring and Luk 1989]. The CCA problem can also be solved using alternative techniques. The only requirements are that the successive images on the unit ball are orthogonal and that the angle is minimised.

2.3. Evaluating the Canonical Correlation Model

The pair of position vectors that have images on the unit ball with a minimum enclosing angle correspond to the canonical correlation model obtained from the training data. The entries of these position vectors convey the relations between the variables obtained from the sampling distribution. In general, a statistical model is validated in terms of statistical significance and generalisability. To assess the statistical significance of the relations obtained from the training data, Bartlett's sequential test procedure [Bartlett 1941] can be applied. Although the technique was presented in 1941, it is still applied in timely CCA application studies such as [Marttinen et al. 2013; Kabir et al. 2014; Song et al. 2016]. The generalisability of the canonical correlation model determines whether the relations obtained from the training data can be considered to represent general patterns occurring in the sampling distribution. The methods of testing the statistical significance and generalisability of the extracted relations represent standard ways to evaluate the canonical correlation model.

The entries of the position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ can be used as a means to analyse the linear relations between the variables. The linear relation corresponding to the value of the canonical correlation is found between the entries that are of the greatest value. The values of the entries of the position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are visualised in Figure 1. The linear relation that corresponds to the canonical correlation of $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.99$ is found between the variables $\mathbf{a}_3$ and $\mathbf{b}_1$. Since the signs of both entries are negative, the relation is positive. The second pair of positions $(\mathbf{w}_a^2, \mathbf{w}_b^2)$ conveys the negative relation between $\mathbf{a}_4$ and $\mathbf{b}_3$. The positive relation between $\mathbf{a}_1$ and $\mathbf{b}_2$ can be identified from the entries of the third pair of positions $(\mathbf{w}_a^3, \mathbf{w}_b^3)$.

In [Meredith 1964], structure correlations were introduced as a means to analyse the relations between the variables. Structure correlations are the correlations of the original variables, $\mathbf{a}_i \in \mathbb{R}^n$ for $i = 1, 2, \ldots, p$ and $\mathbf{b}_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, q$, with the images, $\mathbf{z}_a \in \mathbb{R}^n$ or $\mathbf{z}_b \in \mathbb{R}^n$. In general, the structure correlations convey how the images $\mathbf{z}_a$ and $\mathbf{z}_b$ are aligned in the space $\mathbb{R}^n$ in relation to the variable axes.

In [Ter Braak 1990], the structure correlations were visualised on a biplot to facilitate the interpretation of the relations. To plot the variables on the biplot, the correlations of the original variables of both sets with two successive images, for example the images $(\mathbf{z}_a^1, \mathbf{z}_a^2)$, of one of the sets are computed. The plot is interpreted by the cosine of the angles between the variable vectors which is given by $\cos(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathbf{b} \rangle / ||\mathbf{a}|| ||\mathbf{b}||$. Hence a positive linear relation is shown by an acute angle while an obtuse angle depicts a negative linear relation. A right angle corresponds to a zero correlation. Three biplots of the data and results of Example 2.1 are shown in Figure 2. In each of the biplots, the same relations that were identified in Figure 1 can be found by analysing the

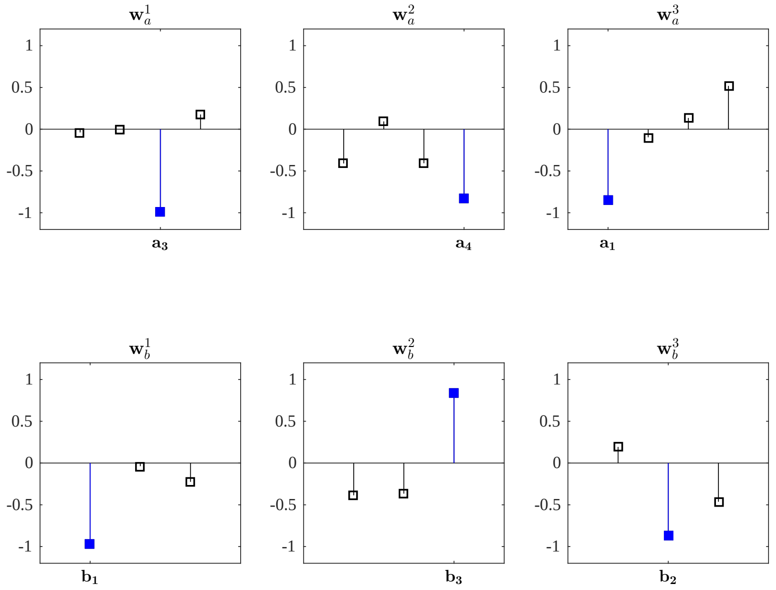

Fig. 1. The entries of the pairs of positions $(\mathbf{w}_a^1, \mathbf{w}_b^1)$, $(\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ are shown. The entry of maximum absolute value is coloured blue.

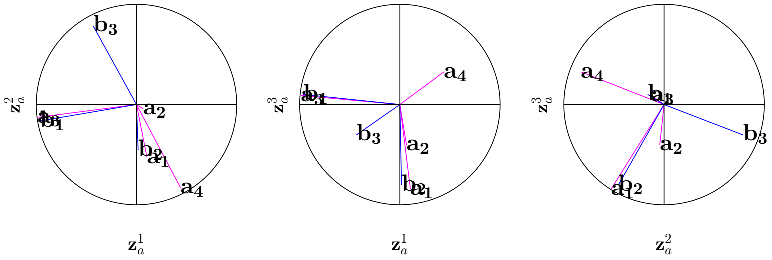

Fig. 2. The biplots are generated using the results of Example 2.1. The biplot on the left shows the relations between the variables when viewed with respect to the images $\mathbf{z}_a^1$ and $\mathbf{z}_a^2$. The biplot in the middle shows the relations between the variables when viewed with respect to the images $\mathbf{z}_a^1$ and $\mathbf{z}_a^3$. The biplot on the right shows the relations between the variables when viewed with respect to the images $\mathbf{z}_a^2$ and $\mathbf{z}_a^3$.

angles between the variable vectors. The extraction of the relations can be enhanced by changing the pairs of images with which the correlations are computed.

The statistical significance tests of the canonical correlations evaluate whether the obtained pattern can be considered to occur non-randomly. The sequential test procedure of Bartlett [Bartlett 1938] determines the number of statistically significant canonical correlations in the data. The procedure to evaluate the statistical significance of the canonical correlations is described in [Fujikoshi and Veitch 1979]. We test the hypothesis

$$H_0 : \min(p, q) = k \text{ against } H_1 : \min(p, q) > k \tag{13}$$

where $k = 0, 1, \ldots, p$ when $p < q$. If the hypothesis $H_0 : \min(p, q) = j$ is rejected for $j = 0, 1, \ldots, k - 1$ but accepted for $H_1 : \min(p, q) > k - 1$ the number of statistically significant canonical correlations can be estimated as k. For the test, the Bartlett-

Lawley statistic, L_k is applied

$$L_k = -\big(n - k - \frac{1}{2}(p+q+1) + \sum_{j=1}^{k} r_j^{-2}\big) \ln\big(\prod_{j=k+1}^{\min(p,q)} (1 - r_j^2)\big). \tag{14}$$

where r_j denotes the j^{th} canonical correlation. The asymptotic null distribution of L_k is the chi-squared with $(p-k)(q-k)$ degrees of freedom. Hence we first test that no canonical relation exists between the two views. If we reject the hypothesis H_0 we continue to test that one canonical relation exists. If all the canonical patterns are statistically significant even the hypothesis $H_0 : \min(p, q) = k - 1$ is rejected.

Example 2.4. We demonstrate the sequential test procedure of Bartlett using the simulated setting of Examples 2.1, 2.2 and 2.3. In the setting, $n = 60$, $p = 4$ and $p = 3$. Hence $\min(p, q) = 3$. First, we test that there are no canonical correlations

$$H_0 : \min(p, q) = 0 \text{ against } H_1 : \min(p, q) > 0 \tag{15}$$

The Bartlett-Lawley statistic is $L_0 = 296.82$. Since $L_0 \sim \chi^2(12)$ the critical value at the significance level $\alpha = 0.01$ is $P(\chi^2 \geq 26.2) = 0.01$. Since $L_0 = 296.82 > 26.2$ the hypothesis H_0 is rejected. Next we test that there is one canonical correlation.

$$H_0 : \min(p, q) = 1 \text{ against } H_1 : \min(p, q) > 1 \tag{16}$$

The Bartlett-Lawley statistic is $L_1 = 154.56$ and $L_1 \sim \chi^2(6)$. The critical value at the significance level $\alpha = 0.01$ is $P(\chi^2 \geq 16.8) = 0.01$. Since $L_1 = 154.56 > 16.8$ the hypothesis H_0 is rejected. We continue to test that there are two canonical correlations

$$H_0 : \min(p, q) = 2 \text{ against } H_1 : \min(p, q) > 2 \tag{17}$$

The Bartlett-Lawley statistic is $L_2 = 70.95$ and $L_2 \sim \chi^2(2)$. The critical value at the significance level $\alpha = 0.01$ is $P(\chi^2 \geq 9.21) = 0.01$. Since $L_1 = 70.95 > 9.21$ the hypothesis H_0 is rejected. Hence the hypothesis $H_1 : \min(p, q) > 2$ is accepted and all three canonical patterns are statistically significant. □

To determine whether the extracted relations can be considered generalisable, or in other words general patterns in the sampling distribution, the linear transformations of the position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ need to be performed using test data. Unlike training data, test data originates from the sampling distribution but were not used in the model computation. Let the matrices $X_a^{test} \in \mathbb{R}^{m \times p}$ and $X_b^{test} \in \mathbb{R}^{m \times q}$ denote the test data of m observations. The linear transformations of the position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are then

$$X_a^{test}\mathbf{w}_a = \mathbf{z}_a^{test} \quad \text{and} \quad X_b^{test}\mathbf{w}_b = \mathbf{z}_b^{test}$$

where the images $\mathbf{z}_a^{test}$ and $\mathbf{z}_b^{test}$ are in the space $\mathbb{R}^m$. The cosine of the angle between the test images $\cos(\mathbf{z}_a^{test}, \mathbf{z}_b^{test}) = \langle \mathbf{z}_a^{test}, \mathbf{z}_b^{test} \rangle$ implies the generalisability. If the canonical correlations computed from test data also result in high correlation values we can deduce that the relations can generally be found from the particular sampling distribution.

Example 2.5. We evaluate the generalisability of the canonical correlation model obtained in Example 2.1. The test data matrices X_a^{test} and X_b^{test} of sizes $m \times p$ and $m \times q$ where $m = 40, p = 4$, and $q = 3$ are from the same distributions as described in Example 2.1. The 40 observations were not included in the computation of the model. The test canonical correlations corresponding to the positions $(\mathbf{w}_a^1, \mathbf{w}_b^1), (\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ are $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.98$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.98$, $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.98$. The high values indicate that the extracted relations can be considered generalisable. □

The canonical correlation model can be evaluated by assessing the statistical significance and testing the generalisability of the relations. The statistical significance of the model can be determined by testing whether the extracted canonical correlations are not non-zero by chance. The generalisability of the relations can be assessed using new observations from the sampling distribution. These evaluation methods can generally be applied to test the validity of the extracted relations obtained using any variant of CCA.

3. EXTENSIONS OF CANONICAL CORRELATION ANALYSIS

3.1. Regularisation Techniques in Underdetermined Systems

CCA finds linear relations in the data when the number of observations exceeds the number of variables in either view. This possibly guarantees the non-singularity of the variance matrices C_{aa} and C_{bb} when solving the CCA problem. In the case of the standard eigenvalue problem, the matrices C_{aa} and C_{bb} should be non-singular so that they can be inverted. In the case of the SVD method, singular C_{aa} and C_{bb} may not have the square root factors. If the number of observations is less than the number of variables it is likely that some of the variables are collinear. Hence a sufficient sample size reduces the collinearity of the variables and guarantees the non-singularity of the variance matrices. The first proposition to solve the problem of insufficient sample size was presented in [Vinod 1976]. A more recent technique to regularise CCA has been proposed in [Cruz-Cano and Lee 2014]. In the following, we present the original method of regularisation [Vinod 1976] due to its popularity in CCA applications [González et al. 2009], [Yamamoto et al. 2008], and [Soneson et al. 2010].

In the work of [Vinod 1976], the singularity problem was proposed to be solved by regularisation. In general, the idea is to improve the invertibility of the variance matrices C_{aa} and C_{bb} by adding arbitrary constants $c_1 > 0$ and $c_2 > 0$ to the diagonal $C_{aa} + c_1 I$ and $C_{bb} + c_2 I$. The constraints of CCA become

$$\mathbf{w}_a^T \big(C_{aa} + c_1 I\big)\mathbf{w}_a = 1$$
$$\mathbf{w}_b^T \big(C_{bb} + c_2 I\big)\mathbf{w}_b = 1$$

and hence the magnitudes of the position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are smaller when regularisation, $c_1 > 0$ and $c_2 > 0$, is applied. The regularised CCA optimisation problem is given by

$$\cos\theta = \max_{\mathbf{w}_a \in \mathbb{R}^p, \mathbf{w}_b \in \mathbb{R}^q} \mathbf{w}_a^T C_{ab} \mathbf{w}_b,$$
$$\mathbf{w}_a^T \big(C_{aa} + c_1 I\big)\mathbf{w}_a = 1 \quad \mathbf{w}_b^T \big(C_{bb} + c_2 I\big)\mathbf{w}_b = 1.$$

The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ can be found by solving the standard eigenvalue problem

$$\big(C_{bb} + c_2 I\big)^{-1} C_{ba} \big(C_{aa} + c_1 I\big)^{-1} C_{ab} \mathbf{w}_b = \rho^2 \mathbf{w}_b.$$

or the generalised eigenvalue problem

$$\begin{pmatrix} \mathbf{0} & C_{ab} \\ C_{ba} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \rho \begin{pmatrix} C_{aa} + c_1 I & \mathbf{0} \\ \mathbf{0} & C_{bb} + c_2 I \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}.$$

As in the case of linear CCA, the canonical correlations correspond to the inner products between the consecutive image pairs $\langle \mathbf{z}_a^i, \mathbf{z}_b^i \rangle$ where $i = 1, 2, \ldots, \min(p, q)$.

The regularisation proposed by [Vinod 1976] makes the CCA problem solvable but introduces new parameters $c_1 > 0$ and $c_2 > 0$ that have to be chosen. The first proposition of applying a leave-one-out cross-validation procedure to automatically select the regularisation parameters was presented in [Leurgans et al. 1993]. Cross-validation is a well-established nonparametric model selection procedure to evaluate the validity of statistical predictions. One of its earliest applications have been presented in [Larson 1931]. A cross-validation procedure entails the partitioning of the observations into subsamples, selecting and estimating a statistic which is first measured on one subsample, and then validated on the other hold-out subsample. The method of cross-validation is discussed in detail for example in [Stone 1974], [Efron 1979], [Browne 2000], and more recently in [Arlot et al. 2010]. The cross-validation approach specifically developed for CCA has been further extended in [Waaijenborg et al. 2008; Yamamoto et al. 2008; González et al. 2009; Soneson et al. 2010].

In cross-validation, the size of the hold-out subsample varies depending on the size of the dataset. A leave-one-out cross-validation procedure is an option when the sample size is small and partitioning of the data into several folds, as is done in k-fold cross-validation, is not feasible. 5-fold cross-validation saves computation time in relation to leave-one-out cross-validation if the sample size is large enough to partition the observations into five folds where each fold is used as a test set in turn.

In general, as demonstrated for example in [Krstajic et al. 2014], a k-fold cross-validation procedure should be repeated when an optimal set of parameters are searched for. Repetitions decrease the variance of the average values measured across the test folds. Algorithm 1 outlines an approach to determine the optimal regularisation parameters in CCA.

ALGORITHM 1: Repeated k-fold cross-validation for regularised CCA

```
Input: Data matrices $X_a$ and $X_b$, number of repetitions $R$, number of folds $F$
Output: Regularisation parameter values $c_1$ and $c_2$ maximising the correlation on test data
Pre-defined ranges for values of $c_1$; $c_2$;
Initialise $r = 1$;
repeat
    Randomly partition the observations into $F$ folds ;
    for all values of $c_1$ do
        for all values of $c_2$ do
            for $i = 1, 2, \dots, F$ do
                Training set: $F - i$ folds, test set: $i$ fold;
                Standardise the variables in the training and test sets;
                For the training data, solve $|C_{bb}^{-1} C_{ba} \big(C_{aa} + c_1 I\big)^{-1} C_{ab} - \rho^2 I| = 0$;
                Find $\mathbf{w}_b$ corresponding to the greatest eigenvalue satisfying
                  $(C_{bb}^{-1} C_{ba} \big(C_{aa} + c_1 I\big)^{-1} C_{ab} - \rho^2 I)\mathbf{w}_b = 0$ ;
                Find $\mathbf{w}_a$ using $\mathbf{w}_a^1 = \frac{\big(C_{aa}+c_1 I\big)^{-1} C_{ab} \mathbf{w}_b^1}{\rho_1}$;
                Transform the training positions $\mathbf{w}_a$ and $\mathbf{w}_b$ using the test observations
                  $X_{a,test}\mathbf{w}_a = \mathbf{z}_a$ and $X_{b,test}\mathbf{w}_b = \mathbf{z}_b$ ;
                Compute $\cos(\mathbf{z}_a, \mathbf{z}_b) = \frac{\langle \mathbf{z}_a, \mathbf{z}_b \rangle}{||\mathbf{z}_a||||\mathbf{z}_b||}$;
            end
            Store the mean of the $F$ values for $\cos(\mathbf{z}_a, \mathbf{z}_b)$ obtained at $c_1$ and $c_2$;
        end
    end
    $r = r + 1$ ;
until $r = R$;
Compute the mean of the $R$ values for $\cos(\mathbf{z}_a, \mathbf{z}_b)$ obtained at $c_1$ and $c_2$ ;
Return the combination $c_1$ and $c_2$ that maximises $\cos(\mathbf{z}_a, \mathbf{z}_b)$
```

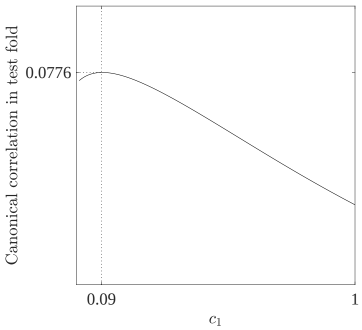

Fig. 3. The maximum test canonical correlation, computed over 50 times repeated 5-fold cross-validation, is obtained at $c_1 = 0.09$.

Example 3.1. To demonstrate the procedure of regularisation in underdetermined settings, we use the same simulated data as in the previous examples but we include additional normally distributed variables. The data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 60$, $p = 70$ and $q = 10$, respectively as follows. The variables of X_a are generated from a random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{70} \sim N(0, 1)$. We generate the following linear relations

$$\begin{aligned} \mathbf{b}_1 &= \mathbf{a}_3 + \boldsymbol{\xi}_1 \\ \mathbf{b}_2 &= \mathbf{a}_1 + \boldsymbol{\xi}_2 \\ \mathbf{b}_3 &= -\mathbf{a}_4 + \boldsymbol{\xi}_3 \end{aligned}$$

where $\boldsymbol{\xi}_1 \sim N(0, 0.01), \boldsymbol{\xi}_2 \sim N(0, 0.03), \boldsymbol{\xi}_3 \sim N(0, 0.02)$ denote vectors of normal noise. The remaining variables of X_b are generated from random univariate normal distribution, $\mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_{10} \sim N(0, 1)$. The data is standardised such that every variable has zero mean and unit variance.

To construct the matrix $C_{bb}^{-1} C_{ba} C_{aa}^{-1} C_{ab}$, the variance matrices C_{aa} and C_{bb} need to be non-singular. Since C_{aa} is obtained from a rectangular matrix, collinearity makes it close to singular. We therefore add a positive constant to the diagonal $C_{aa} + c_1 I$ to make it invertible. C_{bb} is invertible since the data matrix X_b has more rows than columns. The optimal value for the regularisation parameter c_1 can be determined for instance through repeated k-fold cross-validation. As shown in Figure 3, the optimal value $c_1 = 0.09$ was obtained through 50 times repeated 5-fold cross-validation using the procedure presented in the Algorithm 1.

The positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and their respective images $\mathbf{z}_a$ and $\mathbf{z}_b$ on a unit ball are found by solving the eigenvalues of the characteristic equation

$$|C_{bb}^{-1} C_{ba} \big(C_{aa} + c_1 I\big)^{-1} C_{ab} - \rho^2 I| = 0. \tag{18}$$

The number of relations equals $min(p, q) = 10$. The square roots of the first three eigenvalues are $\rho_1 = 0.98$, $\rho_2 = 0.97$ and $\rho_3 = 0.96$. The respective three eigenvectors that correspond to the positions $\mathbf{w}_b$ satisfy the equation

$$(C_{bb}^{-1} C_{ba} \big(C_{aa} + c_1 I\big)^{-1} C_{ab} - \rho^2 I)\mathbf{w}_b = 0. \tag{19}$$

The positions $\mathbf{w}_a$ are found using the formula

$$\mathbf{w}_a^i = \frac{\big(C_{aa} + c_1 I\big)^{-1} C_{ab} \mathbf{w}_b^i}{\rho_i} \tag{20}$$

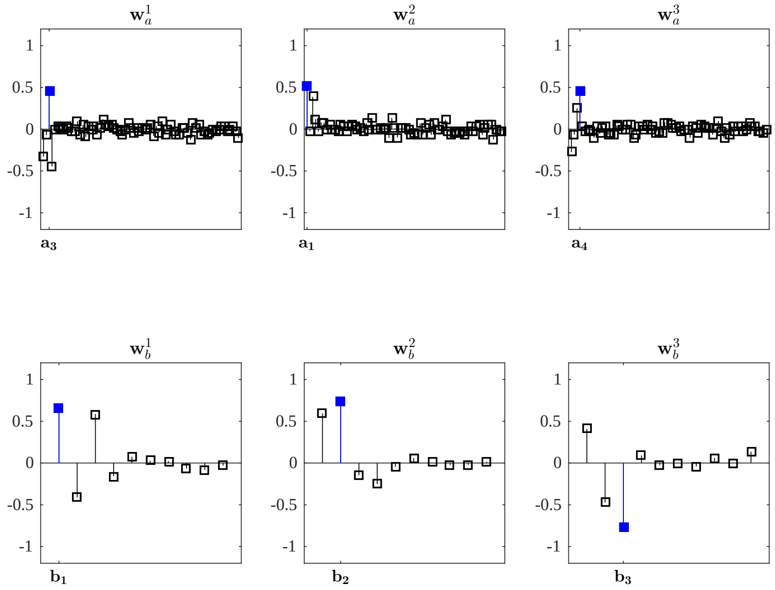

Fig. 4. The entries of the pairs of positions $(\mathbf{w}_a^1, \mathbf{w}_b^1)$, $(\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ are shown. The entry of maximum absolute value is coloured blue. The positive linear relation between $\mathbf{a}_3$ and $\mathbf{b}_1$, the positive linear relation between $\mathbf{a}_1$ and $\mathbf{b}_2$ and the negative linear relation between $\mathbf{a}_4$ and $\mathbf{b}_3$ are extracted by the pairs $(\mathbf{w}_a^1, \mathbf{w}_b^1)$, $(\mathbf{w}_a^2, \mathbf{w}_b^2)$, and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ respectively.

where $i = 1, 2, 3$ corresponds to the sorted eigenvalues and eigenvectors. By rounding correct to three decimal places, the first three canonical correlations are $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.999$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.998$, $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.996$. The extracted linear relations are visualised in Figure 4. □

When either or both of the data views consists of more variables than observations, regularisation can be applied to make the variance matrices non-singular. This involves finding optimal non-negative scalar parameters that, when added to the diagonal entries, improve the invertibility of the variance matrices. After improving the invertibility of the variance matrices, the regularised CCA problem can be solved using the standard techniques.

3.2. Bayesian Approaches for Robustness

Probabilistic approaches have been proposed to improve the robustness of CCA when the sample size is small and to be able to make more flexible distributional assumptions. A robust method generates a valid model regardless of outlying observations. In the following, a brief introduction to Bayesian CCA is provided. A detailed review on Bayesian CCA and its recent extensions can be found in [Klami et al. 2013].

An extension of CCA to probabilistic models was first proposed in [Bach and Jordan 2005]. The probabilistic model contains the latent variables $\mathbf{y}^k \in \mathbb{R}^o$, where $o = \min(p, q)$, that generate the observations $\mathbf{x}_a^k \in \mathbb{R}^p$ and $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$. The latent variable model is defined by

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(0, I_d), \quad && o \geq d \geq 1 \\ \mathbf{x}_a | \mathbf{y} &\sim \mathcal{N}(S_a \mathbf{y} + \boldsymbol{\mu}_a, \Psi_a), \quad && S_a \in \mathbb{R}^{p \times d}, \Psi_a \succeq 0 \\ \mathbf{x}_b | \mathbf{y} &\sim \mathcal{N}(S_b \mathbf{y} + \boldsymbol{\mu}_b, \Psi_b), \quad && S_b \in \mathbb{R}^{q \times d}, \Psi_b \succeq 0 \end{aligned}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes the normal multivariate distribution with mean $\boldsymbol{\mu}$ and covariance Σ. The S_a and S_b correspond to the transformations of the latent variables $\mathbf{y}^k \in \mathbb{R}^o$. The Ψ_a and Ψ_b denote the noise covariance matrices. The maximum likelihood estimates of the parameters $S_a, S_b, \Psi_a, \Psi_b, \boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ are given by

$$\begin{aligned} \hat{S}_a &= C_{aa} W_{ad} M_a & \hat{S}_b &= C_{bb} W_{bd} M_b \\ \hat{\Psi}_a &= C_{aa} - \hat{S}_a \hat{S}_a^{\,T} & \hat{\Psi}_b &= C_{bb} - \hat{S}_b \hat{S}_b^{\,T} \\ \hat{\boldsymbol{\mu}}_a &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_a^k & \hat{\boldsymbol{\mu}}_b &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_b^k \end{aligned}$$

where $M_a, M_b \in \mathbb{R}^{d \times d}$ are arbitrary matrices such that $M_a M_b^T = P_d$ and the spectral norms of M_a and M_b are smaller than one. P_d is the diagonal matrix of the first d canonical correlations. The d columns of W_{ad} and W_{bd} correspond to the positions $\mathbf{w}_a^i$ and $\mathbf{w}_b^i$ for $i = 1, 2, \ldots, d$ obtained using any of the standard techniques described in section 2.1.

The posterior expectations of $\mathbf{y}$ given $\mathbf{x}_a$ and $\mathbf{x}_b$ are $E(\mathbf{y}|\mathbf{x}_a) = M_a^T W_{ad}^T(\mathbf{x}_a - \hat{\mu_a})$ and $E(\mathbf{y}|\mathbf{x}_b) = M_b^T W_{bd}^T(\mathbf{x}_b - \hat{\mu_b})$. As stated in [Bach and Jordan 2005], regardless of what M_a and M_b are, $E(\mathbf{y}|\mathbf{x}_a)$ and $E(\mathbf{y}|\mathbf{x}_b)$ lie in the d-dimensional subspaces of $\mathbb{R}^p$ and $\mathbb{R}^q$ which are identical to those obtained by linear CCA. The generative model of [Bach and Jordan 2005] was further developed in [Archambeau et al. 2006] by replacing the normal noise with the multivariate Student's t distribution. This improves the robustness against outlying observations that are then better modeled by the noise term [Klami et al. 2013].

A Bayesian extension of CCA was proposed by [Klami and Kaski 2007] and [Wang 2007]. To perform Bayesian analysis, the probabilistic model has to be supplemented with prior distributions of the model parameters. In [Klami and Kaski 2007] and [Wang 2007], the prior distribution of the covariance matrices Ψ_a and Ψ_b was chosen to be the inverse-Wishart distribution. The automatic relevance determination [Neal 2012] prior was selected for the linear transformations S_a and S_b. The inference on the posterior distribution was made by applying a variational mean-field algorithm [Wang 2007] and Gibbs sampling [Klami and Kaski 2007].

As in the case of the linear CCA, the variance matrices obtained from high-dimensional data make the inference of the probabilistic and Bayesian CCA models difficult [Klami et al. 2013]. This is because the variance matrices need to be inverted in the inference algorithms. To perform Bayesian CCA on high-dimensional data, dimensionality reduction techniques should be applied as a preprocessing step, as has been done for example in [Huopaniemi et al. 2010].

An advantage of Bayesian CCA, in relation to linear CCA, is the application of the prior distributions that enable to take the possible underlying structure in the data into account. Examples of studies where sparse models were obtained by means of the prior distribution include [Archambeau and Bach 2009] and [Rai and Daume 2009]. In addition to modeling the structure of the data, in [Klami et al. 2012] the Bayesian CCA was extended such that any exponential family distribution could model the noise, not only the normal.

In summary, probabilistic and Bayesian CCA provide alternative ways to interpret the CCA by means of latent variables. Bayesian CCA may be more feasible in settings where knowledge regarding the data can be incorporated through the prior distributions. Additionally, noise can be modelled by other exponential family distribution functions than the normal distribution.

3.3. Uncovering Linear and Non-Linear Relations

CCA [Hotelling 1936] finds linear relations between variables belonging to two views that both are overdetermined. The first proposition to extend CCA to uncover non-linear relations using an optimal scaling method was presented in [Burg and Leeuw 1983]. At the turn of the 21^{st} century, artificial neural networks were incorporated in the CCA framework for finding non-linear relations [Lai and Fyfe 1999; Fyfe and Lai 2000; Hsieh 2000]. Deep CCA [Andrew et al. 2013] is an example of a recent non-linear CCA variant employing artificial neural networks. Shortly after the introduction of the neural networks, propositions of applying kernel methods in CCA were presented in [Lai and Fyfe 2000; Akaho 2001; Van Gestel et al. 2001; Melzer et al. 2001; Bach and Jordan 2002]. Since then, the kernelised version of CCA has received considerable attention in terms of theoretical foundations [Hardoon et al. 2004; Fukumizu et al. 2007; Alam et al. 2008; Blaschko et al. 2008; Hardoon and Shawe-Taylor 2009; Cai 2013] and applications [Melzer et al. 2003; Wang et al. 2005; Hardoon et al. 2007; Larson et al. 2014]. In the following, we present how kernel CCA can be applied to uncover nonlinear relations between the variables. We then provide a brief overview on deep CCA.

To extract linear relations, CCA is performed in the data spaces of $X_a \in \mathbb{R}^{n \times p}$ and $X_b \in \mathbb{R}^{n \times q}$ where the n rows correspond to the observations and the p and q columns correspond to the variables. The relations between the variables are found analysing the positions $\mathbf{w}_a \in \mathbb{R}^p$ and $\mathbf{w}_b \in \mathbb{R}^q$ that have such images $\mathbf{z}_a = X_a\mathbf{w}_a$ and $\mathbf{z}_b = X_b\mathbf{w}_b$ on a unit ball in $\mathbb{R}^n$ that have a minimum enclosing angle. The extracted relations are linear since the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ and their images $\mathbf{z}_a$ and $\mathbf{z}_b$ were obtained in the Euclidean space.

To extract non-linear relations, the positions $\mathbf{w}_a$ and $\mathbf{w}_b$ should be found in a space where the distances, or measures of similarity, between objects are non-linear. This can be achieved using kernel methods, that is by transforming the original observations $\mathbf{x}_a^i \in \mathbb{R}^p$ and $\mathbf{x}_b^i \in \mathbb{R}^q$, where $i = 1, 2, \ldots, n$, to Hilbert spaces $\mathcal{H}_a$ and $\mathcal{H}_b$ through feature maps $\phi_a : \mathbb{R}^p \mapsto \mathcal{H}_a$ and $\phi_b : \mathbb{R}^q \mapsto \mathcal{H}_b$. The similarity of the objects is captured by a symmetric positive semi-definite kernel, corresponding to the inner products in the Hilbert spaces $K_a(\mathbf{x}_a^i, \mathbf{x}_a^j) = \langle \boldsymbol{\phi}_a(\mathbf{x}_a^i), \boldsymbol{\phi}_a(\mathbf{x}_a^j) \rangle_{\mathcal{H}_a}$ and $K_b(\mathbf{x}_b^i, \mathbf{x}_b^j) = \langle \boldsymbol{\phi}_b(\mathbf{x}_b^i), \boldsymbol{\phi}_b(\mathbf{x}_b^j) \rangle_{\mathcal{H}_b}$. The feature maps are typically non-linear and result in high-dimensional intrinsic spaces $\boldsymbol{\phi}_a(\mathbf{x}_a^i) \in \mathcal{H}_a$ and $\boldsymbol{\phi}_b(\mathbf{x}_b^i) \in \mathcal{H}_b$ for $i = 1, 2, \ldots, n$. Through kernels, CCA can be used to extract non-linear correlations, relying on the fact that the CCA solution can always be found within the span of the data [Bach and Jordan 2002; Schölkopf et al. 1998].

The basic principles behind kernel CCA are similar to CCA. First, the observations are transformed to Hilbert spaces $\mathcal{H}_a$ and $\mathcal{H}_b$ using symmetric positive semi-definite kernels

$$K_a(\mathbf{x}_a^i, \mathbf{x}_a^j) = \langle \boldsymbol{\phi}_a(\mathbf{x}_a^i), \boldsymbol{\phi}_a(\mathbf{x}_a^j) \rangle_{\mathcal{H}_a} \text{ and } K_b(\mathbf{x}_b^i, \mathbf{x}_b^j) = \langle \boldsymbol{\phi}_b(\mathbf{x}_b^i), \boldsymbol{\phi}_b(\mathbf{x}_b^j) \rangle_{\mathcal{H}_b}$$

where $i, j = 1, 2, \ldots, n$. As derived in [Bach and Jordan 2002], the original data matrices $X_a \in \mathbb{R}^{n \times p}$ and $X_b \in \mathbb{R}^{n \times q}$ can be substituted by the Gram matrices $K_a \in \mathbb{R}^{n \times n}$ and $K_b \in \mathbb{R}^{n \times n}$. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the positions in the kernel space $\mathbb{R}^n$ that have the images $\mathbf{z}_a = K_a\boldsymbol{\alpha}$ and $\mathbf{z}_b = K_b\boldsymbol{\beta}$ on the unit ball in $\mathbb{R}^n$ with a minimum enclosing angle in between. The kernel CCA problem is hence

$$\cos(\mathbf{z}_a, \mathbf{z}_b) = \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle = \boldsymbol{\alpha}^T K_a^T K_b \boldsymbol{\beta}, \tag{21}$$

$$||\mathbf{z}_a||_2 = \sqrt{\boldsymbol{\alpha}^T K_a^2 \boldsymbol{\alpha}} = 1 \quad ||\mathbf{z}_b||_2 = \sqrt{\boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta}} = 1 \tag{22}$$

As in CCA, the optimisation problem can be solved using the Lagrange multiplier technique.

$$L = \boldsymbol{\alpha}^T K_a^T K_b \boldsymbol{\beta} - \frac{\rho_1}{2}(\boldsymbol{\alpha}^T K_a^2 \boldsymbol{\alpha} - 1) - \frac{\rho_2}{2}(\boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta} - 1) \tag{23}$$

where ρ_1 and ρ_2 denote the Lagrange multipliers. Differentiating L with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ gives

$$\frac{\delta L}{\delta \boldsymbol{\alpha}} = K_a K_b \boldsymbol{\beta} - \rho_1 K_a^2 \boldsymbol{\alpha} = \mathbf{0} \tag{24}$$

$$\frac{\delta L}{\delta \boldsymbol{\beta}} = K_b K_a \boldsymbol{\alpha} - \rho_2 K_b^2 \boldsymbol{\beta} = \mathbf{0} \tag{25}$$

Multiplying (7) from the left by $\boldsymbol{\alpha}^T$ and (8) from the left by $\boldsymbol{\beta}^T$ gives

$$\boldsymbol{\alpha}^T K_a K_b \boldsymbol{\beta} - \rho_1 \boldsymbol{\alpha}^T K_a^2 \boldsymbol{\alpha} = 0 \tag{26}$$

$$\boldsymbol{\beta}^T K_b K_a \boldsymbol{\alpha} - \rho_2 \boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta} = 0. \tag{27}$$

Since $\boldsymbol{\alpha}^T K_a^2 \boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta} = 1$, we obtain that

$$\rho_1 = \rho_2 = \rho. \tag{28}$$

Substituting (28) into Equation (24) we obtain

$$\boldsymbol{\alpha} = \frac{K_a^{-1} K_a^{-1} K_a K_b \boldsymbol{\beta}}{\rho} = \frac{K_a^{-1} K_b \boldsymbol{\beta}}{\rho}. \tag{29}$$

Substituting (29) into (25) we obtain

$$\frac{1}{\rho} K_b K_a K_a^{-1} K_b \boldsymbol{\beta} - \rho K_b^2 \boldsymbol{\beta} = 0 \tag{30}$$

which is equivalent to the generalised eigenvalue problem of the form

$$K_b^2 \boldsymbol{\beta} = \rho^2 K_b^2 \boldsymbol{\beta}. \tag{31}$$

If K_b^2 is invertible, the problem reduces to a standard eigenvalue problem of the form

$$I \boldsymbol{\beta} = \rho^2 \boldsymbol{\beta}. \tag{32}$$

Clearly, in the kernel space, if the Gram matrices are invertible the resulting canonical correlations are all equal to one. Regularisation is therefore needed to solve the kernel CCA problem.

Kernel CCA can be regularised in a similar manner as presented in Section 3.1 [Bach and Jordan 2002; Hardoon et al. 2004]. We constrain the norms of the position vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by adding constants c_1 and c_2 to the diagonals of the Gram matrices K_a and K_b

$$\boldsymbol{\alpha}^T \left(K_a + c_1 I\right)^2 \boldsymbol{\alpha} = 1 \tag{33}$$

$$\boldsymbol{\beta}^T \left(K_b + c_2 I\right)^2 \boldsymbol{\beta} = 1. \tag{34}$$

The solution can then be found by solving the standard eigenvalue problem

$$\left(K_b + c_1 I\right)^{-2} K_b K_a \left(K_a + c_2 I\right)^{-2} K_a K_b \boldsymbol{\alpha} = \rho^2 \boldsymbol{\alpha}.$$

As in the case of CCA, kernel CCA can also be solved through the generalised eigenvalue problem [Bach and Jordan 2002]. Since the data matrices X_a and X_b can be substituted by the corresponding Gram matrices K_a and K_b, the formulation becomes

$$\begin{pmatrix} \mathbf{0} & K_aK_b \\ K_bK_a & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \rho \begin{pmatrix} \left(K_a + c_1 I\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(K_b + c_2 I\right)^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \tag{35}$$

where the constants c_1 and c_2 denote the regularisation parameters. In Example 3.2, kernel CCA, solved through the generalised eigenvalue problem, is performed on simulated data.

Example 3.2. We generate a simulated dataset as follows. The data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 150$, $p = 7$ and $q = 8$, respectively as follows. The seven variables of X_a are generated from a random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_7 \sim N(0, 1)$. We generate the following relations

$$\begin{aligned} \mathbf{b}_1 &= \exp(\mathbf{a}_3) + \boldsymbol{\xi}_1 \\ \mathbf{b}_2 &= \mathbf{a}_1^3 + \boldsymbol{\xi}_2 \\ \mathbf{b}_3 &= -\mathbf{a}_4 + \boldsymbol{\xi}_3 \end{aligned}$$

where $\boldsymbol{\xi}_1 \sim N(0, 0.4)$, $\boldsymbol{\xi}_2 \sim N(0, 0.2)$ and $\boldsymbol{\xi}_3 \sim N(0, 0.3)$ denote vectors of normal noise. The five other variables of X_b are generated from a random univariate normal distribution, $\mathbf{b}_4, \mathbf{b}_5, \ldots, \mathbf{b}_8 \sim N(0, 1)$. The data is standardised such that every variable has zero mean and unit variance.

In kernel CCA, the choice of the kernel function affects what kind of relations can be extracted. In general, a Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = exp(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2})$ is used when the data is assumed to contain non-linear relations. The width parameter σ determines the non-linearity in the distances between the data points computed in the form of inner products. Increasing the value of σ makes the space closer to Euclidean while decreasing makes the distances more non-linear. The optimal value for σ is best determined using a re-sampling method such as a cross-validation scheme, for example procedure similar to the one presented in Algorithm 1. In this example, we applied the ”median trick”, presented in [Song et al. 2010], according to which the σ corresponds to the median of Euclidean distances computed between all pairs of observations. The median distances for the data in this example were $\sigma_a = 3.53$ and $\sigma_b = 3.62$ for the views X_a and X_b respectively. The kernels were centred by $\tilde{K} = K - \frac{1}{n}\mathbf{j}\mathbf{j}^T K - \frac{1}{n}K\mathbf{j}\mathbf{j}^T + \frac{1}{n^2}(\mathbf{j}^T K \mathbf{j})\mathbf{j}\mathbf{j}^T$ where $\mathbf{j}$ contains only entries of value one [Shawe-Taylor and Cristianini 2004].

In addition to the kernel parameters, also the regularisation parameters c_1 and c_2 need to be optimised to extract the correct relations. As in the case of regularised CCA, a repeated cross-validation procedure can be applied to identify the optimal pair of parameters. For the data in this example, the optimal regularisation parameters were $c_1 = 1.50$ and $c_2 = 0.60$ when a 20 times repeated 5-fold cross-validation was applied. The first three canonical correlations at the optimal parameter values were $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.95$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.89$, and $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.87$.

The interpretation of the relations cannot be performed from the positions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ since they are obtained in the kernel spaces. In the case of simulated data, we know what kind of relations are contained in the data. We can compute the linear correlation coefficient between the simulated relations and the transformed pairs of positions $\mathbf{z}_a$ and $\mathbf{z}_b$ [Chang et al. 2013]. The correlation coefficients are shown in Table I. The exponential relation was extracted in the second pair $(\mathbf{z}_a^2, \mathbf{z}_b^2)$, the 3^{rd} order polynomial relation was extracted in the third pair $(\mathbf{z}_a^3, \mathbf{z}_b^3)$ and the linear relation in the first pair $(\mathbf{z}_a^1, \mathbf{z}_b^1)$. □

In [Hardoon et al. 2004], an alternative formulation of the standard eigenvalue problem was presented when the data contains a large number of observations.

Table I. Extracted relations by kernel CCA

	$\mathbf{z}_a^1$	$\mathbf{z}_a^2$	$\mathbf{z}_a^3$
$exp(\mathbf{a}_3)$	0.00	**0.81**	0.09
$\mathbf{a}_1^3$	0.05	0.14	**0.74**
$-\mathbf{a}_4$	**0.99**	0.07	0.04
	$\mathbf{z}_b^1$	$\mathbf{z}_b^2$	$\mathbf{z}_b^3$
$\mathbf{b}_1$	0.02	**0.93**	0.12
$\mathbf{b}_2$	0.08	0.15	**0.87**
$\mathbf{b}_3$	**0.98**	0.01	0.03

If the sample size is large, the dimensionality of the Gram matrices K_a and K_b can cause computational problems. Partial Gram-Schmidt orthogonalization (PGSO) [Cristianini et al. 2002] was proposed as a matrix decomposition method. PGSO results in

$$K_a \simeq R_a R_a^T$$
$$K_b \simeq R_b R_b^T.$$

Substituting these into the Equations (24) and (25) and multiplying by R_a^T and R_b^T respectively we obtain

$$R_a^T R_a R_a^T R_b R_b^T \boldsymbol{\beta} - \rho R_a^T R_a^T R_a R_a^T R_a \boldsymbol{\alpha} = \mathbf{0} \tag{36}$$
$$R_b^T R_b R_b^T R_a R_a^T \boldsymbol{\alpha} - \mu R_b^T R_b^T R_b R_b^T R_b \boldsymbol{\beta} = \mathbf{0}. \tag{37}$$

Let $D_{aa} = R_a^T R_a$, $D_{ab} = R_a^T R_b$, $D_{ba} = R_b^T R_a$, and $D_{bb} = R_b^T R_b$ denote the blocks of the new sample covariance matrix. Let $\tilde{\boldsymbol{\alpha}} = R_a^T \boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\beta}} = R_b^T \boldsymbol{\beta}$ denote the positions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the reduced space. Using these substitutions in (36) and (37) we obtain

$$D_{aa} D_{ab} \tilde{\boldsymbol{\beta}} - \rho D_{aa}^2 \tilde{\boldsymbol{\alpha}} = \mathbf{0} \tag{38}$$
$$D_{bb} D_{ba} \tilde{\boldsymbol{\alpha}} - \rho D_{bb}^2 \tilde{\boldsymbol{\beta}} = \mathbf{0}. \tag{39}$$

If D_{aa} and D_{bb} are invertible we can multiply (38) by D_{aa}^{-1} and (39) by D_{bb}^{-1} which gives

$$D_{ab} \tilde{\boldsymbol{\beta}} - \rho D_{aa} \tilde{\boldsymbol{\alpha}} = \mathbf{0} \tag{40}$$
$$D_{ba} \tilde{\boldsymbol{\alpha}} - \rho D_{bb} \tilde{\boldsymbol{\beta}} = \mathbf{0}. \tag{41}$$

and hence

$$\tilde{\boldsymbol{\beta}} = \frac{D_{bb}^{-1} D_{ba} \tilde{\boldsymbol{\alpha}}}{\rho} \tag{42}$$

which, after a substitution into (38), results in a generalised eigenvalue problem

$$D_{ab} D_{bb}^{-1} D_{ba} \tilde{\boldsymbol{\alpha}} = \rho^2 D_{aa} \tilde{\boldsymbol{\alpha}}. \tag{43}$$

To formulate the problem as a standard eigenvalue problem, let $D_{aa} = SS^T$ denote the complete Cholesky decomposition where S is a lower triangular matrix and let $\hat{\boldsymbol{\alpha}} = S^T \tilde{\boldsymbol{\alpha}}$. Substituting these into (43) we obtain

$$S^{-1} D_{ab} D_{bb}^{-1} D_{ba} S'^{-1} \hat{\boldsymbol{\alpha}} = \rho^2 \hat{\boldsymbol{\alpha}}.$$

If regularisation using the parameter κ is combined with dimensionality reduction the problem becomes

$$S^{-1} D_{ab} \big(D_{bb} + \kappa I\big)^{-1} D_{ba} S'^{-1} \hat{\boldsymbol{\alpha}} = \rho^2 \hat{\boldsymbol{\alpha}}. \tag{44}$$

Table II. Extracted relations by kernel CCA

	$\mathbf{z}_a^1$	$\mathbf{z}_a^2$	$\mathbf{z}_a^3$
$exp(\mathbf{a}_3)$	**0.91**	0.01	0.05
$\mathbf{a}_1^3$	0.01	**0.92**	0.04
$-\mathbf{a}_4$	0.06	0.03	**0.99**
	$\mathbf{z}_b^1$	$\mathbf{z}_b^2$	$\mathbf{z}_b^3$
$\mathbf{b}_1$	**0.91**	0.01	0.04
$\mathbf{b}_2$	0.01	**0.94**	0.05
$\mathbf{b}_3$	0.07	0.04	**0.99**

A numerical example of the method presented by [Hardoon et al. 2004] is given in Example 3.3.

Example 3.3. We generate a simulated dataset as follows. The data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 10000$, $p = 7$ and $q = 8$, respectively as follows. The seven variables of X_a are generated from a random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_7 \sim N(0, 1)$. We generate the following relations

$$\mathbf{b}_1 = \exp(\mathbf{a}_3) + \boldsymbol{\xi}_1$$
$$\mathbf{b}_2 = \mathbf{a}_1^3 + \boldsymbol{\xi}_2$$
$$\mathbf{b}_3 = -\mathbf{a}_4 + \boldsymbol{\xi}_3$$

where $\boldsymbol{\xi}_1 \sim N(0, 0.4)$, $\boldsymbol{\xi}_2 \sim N(0, 0.2)$ and $\boldsymbol{\xi}_3 \sim N(0, 0.3)$ denote vectors of normal noise. The five other variables of X_b are generated from a random univariate normal distribution, $\mathbf{b}_4, \mathbf{b}_5, \ldots, \mathbf{b}_8 \sim N(0, 1)$. The data is standardised such that every variable has zero mean and unit variance.

A Gaussian kernel is used for both views. The width parameter is set using the median trick to $\sigma_a = 3.56$ and $\sigma_b = 3.60$. The kernels were centred. The positions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are found solving the standard eigenvalue problem in (44) and applying the Equation (42). We set the regularisation parameter $\kappa = 0.5$.

The first three canonical correlations at the optimal parameter values were $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.97$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.97$, and $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.96$. The correlation coefficients between the simulated relations and the transformed variables are shown in Table II. The exponential relation was extracted in the first pair $(\mathbf{z}_a^1, \mathbf{z}_b^1)$, the 3^{rd} order polynomial relation was extracted in the second pair $(\mathbf{z}_a^2, \mathbf{z}_b^2)$ and the linear relation in the third pair $(\mathbf{z}_a^3, \mathbf{z}_b^3)$. □

Non-linear relations are also taken into account through neural networks which are employed in deep CCA [Andrew et al. 2013]. In deep CCA, every observation $\mathbf{x}_a^k \in \mathbb{R}^p$ and $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$ is non-linearly transformed multiple times in an iterative manner through a layered network. The number of units in a layer determines the dimension of the output vector which is put in the next layer. As is explained in [Andrew et al. 2013], let the first layer have c_1 units and the final layer o units. The output vector of the first layer for the observation $\mathbf{x}_a^1 \in \mathbb{R}^p$, is $\mathbf{h}_1 = s(S_1^1 \mathbf{x}_a^1 + b_1^1) \in \mathbb{R}^{c_1}$, where $S_1^1 \in \mathbb{R}^{c_1 \times p}$ is a matrix of weights, $b_1^1 \in \mathbb{R}^{c_1}$ is a vector of bias, and $s : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear function applied to each element. The logistic and tanh functions are examples of popular non-linear functions. The output vector $\mathbf{h}_1$ is then used to compute the output of the following layer in similar manner. The final transformed vector $f_1(\mathbf{x}_a^1) = s(S_d^1 h_{d-1} + b_d^1)$ is in the space of $\mathbb{R}^o$, for a network with d layers. The same procedure is applied to the observations $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$.

In deep CCA, the aim is to learn the optimal parameters S_d and b_d for both views such that the correlation between the transformed observations is maximised. Let $H_a \in \mathbb{R}^{o \times n}$ and $H_b \in \mathbb{R}^{o \times n}$ denote the matrices that have the final transformed output vectors in their columns. Let $\tilde{H}_a = H_a - \frac{1}{n} H_a \mathbf{1}$ denote the centered data matrix and let $\hat{C}_{ab} = \frac{1}{m-1} \tilde{H}_a \tilde{H}_b^T$ and $\hat{C}_{aa} = \frac{1}{m-1} \tilde{H}_a \tilde{H}_a^T + r_a I$, where r_a is a regularisation constant, denote the covariance and variance matrices. Same formulae are used to compute the covariance and variance matrices for view b. As in section 2.1, the total correlation of the top k components of H_a and H_b is the sum of the top k singular values of the matrix $T = \hat{C}_{aa}^{-1/2} \hat{C}_{ab} \hat{C}_{bb}^{-1/2}$. If $k = o$, the correlation is given by the trace norm of T, that is

$$corr(H_a, H_b) = tr(T^T T)^{1/2}.$$

The optimal parameters S_d and b_d maximise the trace norm using gradient-based optimisation. The details of the algorithm can be found in [Andrew et al. 2013].

In summary, kernel and deep CCA provide alternatives to the linear CCA when the relations in the data can be considered to be non-linear and the sample size is small in relation to the data dimensionality. When applying kernel CCA on a real dataset, prior knowledge of the relations of interest can help in the analysis of the results. If the data is assumed to contain both linear and non-linear relations a Gaussian kernel could be a first option. The choice of the kernel function depends on what kind of relations the data can be considered to contain. The possible relations can be extracted by testing how the image pairs correlate with the functions of variables. Deep CCA provides an alternative to compute maximal correlation between the views although the neural network makes the identification of the type of relations difficult.

3.4. Improving the Interpretability by Enforcing Sparsity

The extraction of the linear relations between the variables in CCA and regularised CCA relies on the values of the entries of the position vectors that have images on the unit ball with a minimum enclosing angle. The relations can be inferred when the number of variables is not too large for a human to interpret. However, in modern data analysis, it is common that the number of variables is of the order of tens of thousands. In this case, the values of the entries of the position vectors should be constrained such that only a subset of the variables would have a non-zero value. This would facilitate the interpretation since only a fraction of the total number of variables need to be considered when inferring the relations.

To constrain some of the values of the entries of the position vectors to zero, which is also referred to as to enforce sparsity, tools of convex analysis can be applied. In literature, sparsity has been enforced on the position vectors using soft-thresholding operators [Parkhomenko et al. 2007], elastic net regularisation [Waaijenborg et al. 2008], penalised matrix decomposition combined with soft-thresholding [Witten et al. 2009], and convex least squares optimisation [Hardoon and Shawe-Taylor 2011]. The sparse CCA formulations presented in [Parkhomenko et al. 2007; Waaijenborg et al. 2008; Witten et al. 2009] find sparse position vectors that can be applied to infer linear relations between the variables with non-zero entries. The formulation in [Hardoon and Shawe-Taylor 2011] differs from the preceding propositions in terms of the optimisation criterion. The canonical correlation is found between the image obtained from the linear transformation defined by the data space of one view and the image obtained from the linear transformation defined by the kernel of the other view. The selection of which sparse CCA should be applied for a specific task depends on the research question and prior knowledge regarding the variables.

The sparse CCA algorithm of [Parkhomenko et al. 2007] can be applied when the aim is to find sparse position vectors and no prior knowledge regarding the variables is available. The positions and images are solved using the SVD, as presented in Section 2.2. Sparsity is enforced on the entries of the positions by iteratively applying the soft-thresholding operator [Donoho and Johnstone 1995] on the pair of left and right orthonormal singular vectors until convergence. The soft-thresholding operator is a proximal mapping of the L_1 norm [Bach et al. 2011]. The consecutive pairs of sparse left and right singular vectors are obtained by deflating the found pattern from the matrix on which the SVD is computed. The sparse CCA hence results in a sparse set of linearly related variables.

The elastic net CCA [Waaijenborg et al. 2008] finds sparse position vectors but also considers possible groupings in the variables. The elastic net [Zou and Hastie 2005] combines the LASSO [Tibshirani 1996] and the ridge [Hoerl and Kennard 1970] penalties. The elastic net penalty incorporates a grouping effect in the variable selection. The term variable selection refers to that a selected variable has a non-zero entry in the position vector. In the soft-thresholding CCA of [Parkhomenko et al. 2007], the assignment of a non-zero entry is independent of the other entries within the vector. In the elastic net CCA, the ridge penalty groups the variables by the values of the entries and the LASSO penalty either eliminates a group by shrinking the entries of the variables within the group to zero or leaves them as non-zero. The algorithm is based on an iterative scheme of multiple regression. As in [Parkhomenko et al. 2007], the computations are performed in the data space and therefore the extracted relations are also linear.

The penalised matrix decomposition (PMD) formulation of sparse CCA [Witten et al. 2009] is based on finding low-rank approximations of the covariance matrix C_{ab}. An $n \times p$ sized matrix X with rank $K < min(p,q)$ can be approximated using the SVD [Eckart and Young 1936] by

$$\sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T = \underset{\tilde{X} \in M(r)}{\text{argmin}} ||X - \tilde{X}||_F^2$$

where $\mathbf{u}_k$ denotes the column of the matrix U, $\mathbf{v}_k$ denotes the column of the matrix U, σ_k denotes the k^{th} singular value on the diagonal of S, $M(r)$ is the set of rank r $n \times p$ matrices and $r << K$. In the case of CCA, the matrix to be approximated is the covariance matrix $X = C_{ab}$. The optimisation problem in the PMD context is given by

$$\begin{gathered} \min_{\mathbf{w}_a \in \mathbb{R}^p, \mathbf{w}_b \in \mathbb{R}^q} \frac{1}{2} ||C_{ab} - \sigma \mathbf{w}_a \mathbf{w}_b^T||_F^2, \\ ||\mathbf{w}_a||_2 = 1 \quad ||\mathbf{w}_b||_2 = 1, \\ ||\mathbf{w}_a||_1 \leq c_1 \quad ||\mathbf{w}_b||_1 \leq c_2, \quad \sigma \geq 0 \end{gathered}$$

which is equivalent to

$$\begin{gathered} \cos\theta = \max_{\mathbf{w}_a \in \mathbb{R}^p, \mathbf{w}_b \in \mathbb{R}^q} \mathbf{w}_a^T C_{ab} \mathbf{w}_b, \\ ||\mathbf{w}_a||_2 \leq 1 \quad ||\mathbf{w}_b||_2 \leq 1, \\ ||\mathbf{w}_a||_1 \leq c_1 \quad ||\mathbf{w}_b||_1 \leq c_2. \end{gathered}$$

The aim is to find r pairs of sparse position vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ such that their outer products represent low-rank approximations of the original C_{ab} and hence extracts the r linear relations from the data.

The exact derivation of the algorithm to solve the PMD optimisation problem is given in [Witten et al. 2009]. In general, the position vectors, that generate 1-rank approximations of the covariance matrix, are found in an iterative manner. To find one 1-rank approximation, the soft-thresholding operator is applied as follows. Let the soft-thresholding operator be given by

$$S(a, c) = sign(a)(|a| - c)_+$$

where $c > 0$ is a constant. The following formula is applied in the derivation of the algorithm

$$\max_{\mathbf{u}} \langle \mathbf{u}, \mathbf{a} \rangle,$$
$$s.t. \quad ||\mathbf{u}||_2^2 \leq 1, ||\mathbf{u}||_1 < c.$$

The solution is given by $\mathbf{u} = \frac{S(\mathbf{a},\delta)}{||S(\mathbf{a},\delta)||_2}$ with $\delta = 0$ if $||\mathbf{u}_1|| \leq c$. Otherwise, δ is selected such that $||\mathbf{u}_1|| = c$. Sparse position vectors are the obtained by Algorithm 2. At every iteration, the δ_1 and δ_2 are selected by binary search. To obtain several 1-rank approximations, a deflation step is included such that when the converged vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are found, the extracted relations are subtracted from the covariance matrix $C_{ab}^{k+1} \leftarrow C^k - \sigma_k \mathbf{w}_a^k \mathbf{w}_b^{kT}$. In this way, the successive solutions remain orthogonal which is a contstraint of CCA.

ALGORITHM 2: Computation of a 1-rank approximation of the covariance matrix

```
Initialise $||\mathbf{w_b}||_2 = 1$ ;
repeat
    $\mathbf{w_a} \leftarrow \frac{S(C_{ab}\mathbf{w_b},\delta_1)}{||S(C_{ab}\mathbf{w_b},\delta_1)||_2}$ where $\delta_1 = 0$ if $||\mathbf{w}_a||_1 \leq c_1$, otherwise $\delta_1$ is chosen such that
      $||\mathbf{w}_a||_1 = c_1$ and $c_1 > 0$ ;
    $\mathbf{w_b} \leftarrow \frac{S(C_{ab}^T\mathbf{w_a},\delta_2)}{||S(C_{ab}^T\mathbf{w_a},\delta_2)||_2}$ where $\delta_2 = 0$ if $||\mathbf{w}_b||_1 \leq c_2$, otherwise $\delta_2$ is chosen such that
      $||\mathbf{w}_b||_1 = c_2$ and $c_2 > 0$ ;
until convergence;
$\sigma \leftarrow \mathbf{w}_a^T C_{ab} \mathbf{w}_b$
```

Example 3.4. To demonstrate the PMD formulation of sparse CCA, we generate the following data. The data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 50$, $p = 100$ and $q = 150$, respectively as follows. The variables of X_a are generated from a random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{100} \sim N(0, 1)$. We generate the following linear relations

$$\mathbf{b}_1 = \mathbf{a}_3 + \boldsymbol{\xi}_1 \tag{45}$$
$$\mathbf{b}_2 = \mathbf{a}_1 + \boldsymbol{\xi}_2 \tag{46}$$
$$\mathbf{b}_3 = -\mathbf{a}_4 + \boldsymbol{\xi}_3 \tag{47}$$

where $\boldsymbol{\xi}_1 \sim N(0, 0.08), \boldsymbol{\xi}_2 \sim N(0, 0.07)$, and $\boldsymbol{\xi}_3 \sim N(0, 0.05)$ denote vectors of normal noise. The other variables of X_b are generated from a random univariate normal distribution, $\mathbf{b}_4, \mathbf{b}_5, \cdots, \mathbf{b}_{150} \sim N(0, 1)$. The data is standardised such that every variable has zero mean and unit variance.

We apply the R implementation of [Witten et al. 2009] which is available in the PMA package. We extract three rank-1 approximations. The values of the entries of the pairs of position vectors $(\mathbf{w}_a^1, \mathbf{w}_b^1), (\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ corresponding to canonical correlations $\langle \mathbf{z}_a^1, \mathbf{z}_b^1 \rangle = 0.95$, $\langle \mathbf{z}_a^2, \mathbf{z}_b^2 \rangle = 0.92$, $\langle \mathbf{z}_a^3, \mathbf{z}_b^3 \rangle = 0.91$ are shown in Figure 5. The first 1-rank approximation extracted (47), the second (46), and the third (47). □

The sparse CCA of [Hardoon and Shawe-Taylor 2011] is a sparse convex least squares formulation that differs from the preceding versions. The canonical correlation is found between the linear transformations between a data space view and a kernel space view. The aim is to find a sparse set of variables in the data space view that relate to a sparse set of observations, represented in terms of relative similarities, in the kernel space view. An example of a setting, where relations of this type can provide useful

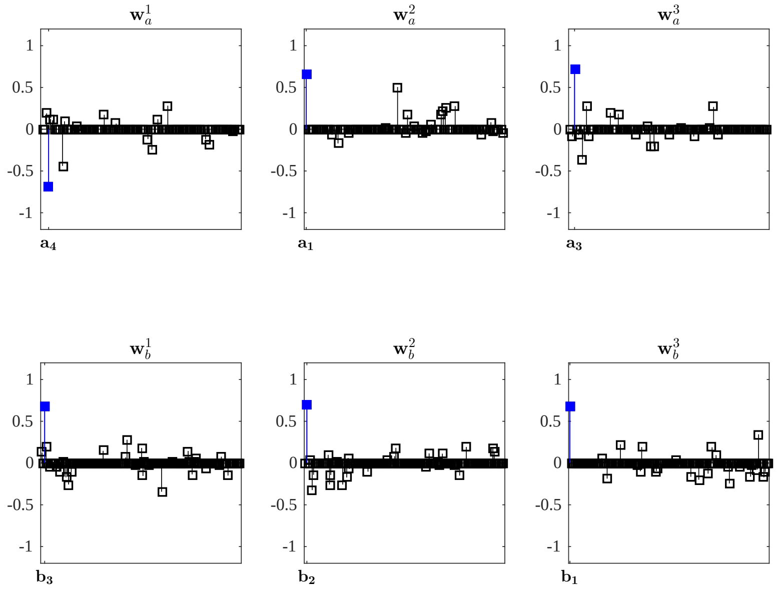

Fig. 5. The values of the entries of the position vector pairs $(\mathbf{w}_a^1, \mathbf{w}_b^1)$, $(\mathbf{w}_a^2, \mathbf{w}_b^2)$ and $(\mathbf{w}_a^3, \mathbf{w}_b^3)$ obtained using the PMD method for sparse CCA are shown. The entry of maximum absolute value is coloured blue. The negative linear relation between $\mathbf{a}_4$ and $\mathbf{b}_3$ is extracted in the first 1-rank approximation. The positive linear relations between $\mathbf{a}_1$ and $\mathbf{b}_2$ and $\mathbf{a}_3$ and $\mathbf{b}_1$ are extracted in the second and third 1-rank approximations.

information, is in bilingual analysis as described in [Hardoon and Shawe-Taylor 2011]. When finding relations between words of two languages, it may be useful to know in what kind of contexts can a word be used in the translated language. The optimisation problem is given by

$$\cos(\mathbf{z}_a, \mathbf{z}_b) = \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle = \mathbf{w}_a^T X_a^T K_b \boldsymbol{\beta},$$

$$||\mathbf{z}_a||_2 = \sqrt{\mathbf{w}_a^T X_a^T X_a \mathbf{w}_a} = 1 \quad ||\mathbf{z}_b||_2 = \sqrt{\boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta}} = 1$$

which is equivalent to the convex sparse least squares problem

$$\min_{\mathbf{w}_a, \boldsymbol{\beta}} ||X_a \mathbf{w}_a - K_b \boldsymbol{\beta}||^2 + \mu ||\mathbf{w}_a||_1 + \gamma ||\tilde{\boldsymbol{\beta}}||_1 \tag{48}$$

$$s.t \quad ||\boldsymbol{\beta}||_\infty = 1 \tag{49}$$

where μ and γ are fixed parameters that control the trade-off between function objective and the level of sparsity of the entries of the position vectors $\mathbf{w}_a$ and $\boldsymbol{\beta}$. The constraint $||\boldsymbol{\beta}||_\infty = 1$ is set to avoid the trivial solution ($\mathbf{w}_a = \mathbf{0}, \boldsymbol{\beta} = \mathbf{0}$). The k^{th} entry of $\boldsymbol{\beta}$ is set to $\beta_k = 1$ and the remaining entries in $\tilde{\boldsymbol{\beta}}$ are constrained by 1-norm. The idea is to fix one sample as a basis for comparison and rank the other similar samples based on the fixed sample. The optimisation problem is solved by iteratively minimising the gap between the primal and dual Lagrangian solutions. The procedure is outlined in Algorithm 3. The exact computational steps can be found in [Hardoon and Shawe-Taylor 2011]. The Algorithm 3 is used to extract one relation or pattern from the data. To extract the successive patterns, deflation is applied to obtain the residual matrices from which the already found pattern is removed. In Example 3.5, the extraction of the first pattern is shown.

ALGORITHM 3: Pseudo-code to solve the convex sparse least squares problem

```
repeat
    1. Use the dual Lagrangian variables to solve the primal variables ;
    2. Check whether all constraints on the primal variables hold ;
    3. Use the primal variables to solve the dual Lagrangian variables ;
    4. Check whether all dual Lagrangian variable constraints hold ;
    5. Check whether 2 holds, if not go to 1 ;
until convergence;
```

Example 3.5. In the sparse CCA of [Hardoon and Shawe-Taylor 2011], the idea is to determine the relations of the variables in the data space view X_a to the observations in the kernel space view K_b where the observations comprise the variables of the view b. This setting differs from all of the previous examples where the idea was to find relations between the variables. Since one of the views is kernelised, the relations cannot be explicitly simulated. We therefore demonstrate the procedure on data generated from random univariate normal distribution as follows. The data matrices X_a and X_b of sizes $n \times p$ and $n \times q$, where $n = 50$, $p = 100$ and $q = 150$, respectively are generated as follows. The variables of X_a and X_b are generated from random univariate normal distribution, $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{100} \sim N(0, 1)$ and $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{150} \sim N(0, 1)$ respectively. The data is standardised such that every variable has zero mean and unit variance.

The Gaussian kernel function $K(\mathbf{x}, \mathbf{y}) = exp(-||\mathbf{x} - \mathbf{y}||^2/2\sigma^2)$ is used to compute the similarities for the view b. The choice of the kernel is justified since the underlying distribution is normal. The width parameter is set to $\sigma = 17.25$ using the median trick. The kernel matrix is centred by $\tilde{K} = K - \frac{1}{n}\mathbf{jj}^T K - \frac{1}{n}K\mathbf{jj}^T + \frac{1}{n^2}(\mathbf{j}^T K\mathbf{j})\mathbf{jj}^T$ where $\mathbf{j}$ contains only entries of value one [Shawe-Taylor and Cristianini 2004].

To find the positions $\mathbf{w}_a$ and $\boldsymbol{\beta}$, we solve

$$f = \min_{\mathbf{w}_a, \boldsymbol{\beta}} ||X_a \mathbf{w}_a - K_b \boldsymbol{\beta}||^2 + \mu||\mathbf{w}_a||_1 + \gamma||\tilde{\boldsymbol{\beta}}||_1$$

$$s.t \quad ||\boldsymbol{\beta}||_\infty = 1$$

using the implementation proposed in [Uurtio et al. 2015]. As stated in [Hardoon and Shawe-Taylor 2011], to determine which variable in the data space view X_a is most related to the observation in K_b, the algorithm needs to be run for all possible values of k. This means that every observation is in turn set as a basis for comparison and a sparse set of the remaining observations $\tilde{\boldsymbol{\beta}}$ is computed. The optimal value of k gives the minimum objective value f.

We run the algorithm by initially setting the value of the entry $\beta_k = 1$ for $k = 1, 2, \ldots, n$. The minimum objective value $f = 0.03$ was obtained at $k = 29$. This corresponds to a canonical correlation of $\langle \mathbf{z}_a, \mathbf{z}_b \rangle = 0.88$. The values of the entries of $\mathbf{w}_a$ and $\boldsymbol{\beta}$ are shown in Figure 6. The observation corresponding to $k = 29$ in the kernelised view K_b is most related to the variables $\mathbf{a}_{15}, \mathbf{a}_{16}, \mathbf{a}_{18}, \mathbf{a}_{20}$, and $\mathbf{a}_{24}$. □

The sparse versions of CCA can be applied to settings when the large number of variables hinders the inference of the relations. When the interest is to extract sparse linear relations between the variables, the proposed algorithms of [Parkhomenko et al. 2007; Waaijenborg et al. 2008; Witten et al. 2009] provide a solution. The algorithm of [Hardoon and Shawe-Taylor 2011] can be applied if the focus is to find how the variables of one view relate to the observations that correspond to the combined sets of the variables in the other view. In other words, the approach is useful if the focus is not to uncover the explicit relations between the variables but to gain insight how a variable relates to a complete set of variables of an observation.

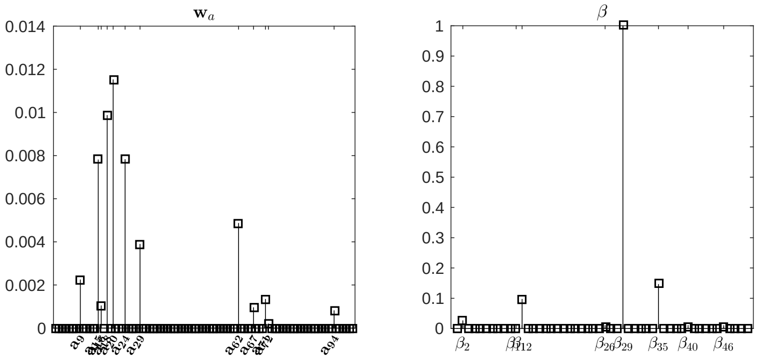

Fig. 6. The values of the entries of the positions $\mathbf{w}_a$ and $\boldsymbol{\beta}$ at the optimal value of k are shown.

4. DISCUSSION

This tutorial presented an overview on the methodological evolution of canonical correlation methods focusing on the original linear, regularised, kernel, and sparse CCA. Succinct reviews were also conducted on the Bayesian and neural network-based deep CCA variants. The aim was to explain the theoretical foundations of the variants using the linear algebraic interpretation of CCA. The methods to solve the CCA problems were described using numerical examples. Additionally, techniques to assess the statistical significance of the extracted relations and the generalisability of the patterns were explained. The aim was to delineate the applicabilities of the different CCA variants in relation to the properties of the data.

In CCA, the aim is to determine linear relations between variables belonging to two sets. From a linear algebraic point of view, the relations can be found by analysing the linear transformations defined by the two views of the data. The most distinct relations are obtained by analysing the entries of the first pair of position vectors in the two data spaces that are mapped onto a unit ball such that their images have a minimum enclosing angle. The less distinct relations can be identified from the successive pairs of position vectors that correspond to the images with a minimum enclosing angle obtained from the orthogonal complements of the preceding pairs of images. This tutorial presented three standard ways of solving the CCA problem, that is by solving either a standard [Hotelling 1935; Hotelling 1936] or a generalised eigenvalue problem [Bach and Jordan 2002; Hardoon et al. 2004], or by applying the SVD [Healy 1957; Ewerbring and Luk 1989].

The position vectors of the two data spaces, that convey the related pairs of variables, can be obtained using alternative techniques than the ones selected for this tutorial. The three methods were chosen because they have been much applied in CCA literature and they are relatively straightforward to explain and implement. Additionally, to understand the further extensions of CCA, it is important to know how it originally has been solved. The extensions are often further developed versions of the standard techniques.

For didactic purposes, the synthetic datasets used for the worked examples were designed to represent optimal data settings for the particular CCA variants to uncover the relations. The relations were generated to be one-to-one, in other words one variable in one view was related with only one variable in the other view. In real datasets, which are often much larger than the synthetic ones in this paper, the relations may not be one-to-one but rather many-to-many (one-to-two, two-to-three, etc.). As in the worked examples, these relations can also be inferred by examining the entries of the position vectors of the two data spaces. However, the understanding of how the one-to-one relations are extracted provides means to uncover the more complex relations.

To apply the linear CCA, the sample size needs to exceed the number of variables of both views which means that the system is required to be overdetermined. This is to guarantee the non-singularity of the variance matrices. If the sample size is not sufficient, regularisation [Vinod 1976] or Bayesian CCA [Klami et al. 2013] can be applied. The feasibility of regularisation has not been studied in relation to the number of variables exceeding the number of observations. Improving the invertibility by introducing additional bias has been shown to work in various settings but the limit when the system is too underdetermined that regularisation cannot assist in recovering the underlying relations has not been resolved. Bayesian CCA is more robust against outlying observations, when compared with linear CCA, due to its generative model structure.

In addition to linear relations, non-linear relations are taken into account in kernelised and neural network-based CCA. Kernel methods enable the extraction of non-linear relations through the mapping to a Hilbert space [Bach and Jordan 2002; Hardoon et al. 2004]. When applying kernel methods in CCA, the disparity between the number of observations and variables can be huge due to very dimensional kernel induced feature spaces, a challenge that is tackled by regularisation. The types of relations that can be extracted, is determined by the kernel function that was selected for the mapping. Linear relations are extracted by a linear kernel and non-linear relations by non-linear kernel functions such as the Gaussian kernel. Although kernelisation extends the range of extractable relations, it also complicates the identification of the type of relation. A method to determine the type of relation involves testing how the image vectors correlate with a certain type of function. However, this may be difficult if no prior knowledge of the relations is available. Further research on how to select the optimal kernel functions to determine the most distinct relations underlying in the data could facilitate the final inference making. Neural network-based deep CCA is an alternative to kernelised CCA, when the aim is to find a high correlation between the final output vectors obtained through multiple non-linear transformations. However, due to the network structure, it is not straightforward to identify the relations between the variables.

As a final branch of the CCA evolution, this tutorial covered sparse versions of CCA. Sparse CCA variants have been developed to facilitate the extraction of the relations when the data dimensionality is too high for human interpretation. This has been addressed by enforcing sparsity on the entries of the position vectors [Parkhomenko et al. 2007; Waaijenborg et al. 2008; Witten et al. 2009]. As an alternative to operating in the data spaces, [Hardoon and Shawe-Taylor 2011] proposed a primal-dual sparse CCA in which the relations are obtained between the variables of one view and observations of the other. The sparse variants of CCA in this tutorial were selected based on how much they have been applied in literature. As a limitation of the selected variants, sparsity is enforced on the entries of the position vectors without regarding the possible underlying dependencies between the variables which has been addressed in the literature of structured sparsity [Chen et al. 2012].

In addition to studying the techniques of solving the optimisation problems of CCA variants, this tutorial gave a brief introduction to evaluating the canonical correlation model. Bartlett's sequential test procedure [Bartlett 1938; Bartlett 1941] was given as an example of a standard method to assess the statistical significance of the canonical correlations. The techniques of identifying the related variables through visual inspection of biplots [Meredith 1964; Ter Braak 1990] were presented. To assess whether the extracted relations can be considered to occur in any data with the same underlying sampling distribution, the method of applying both training and test data was explained. As an alternative method, the statistical significance of the canonical correlation model could be assessed using permutation tests [Rousu et al. 2013]. The visualisation of the results using the biplots is mainly applicable in the case of linear relations. Alternative approaches could be considered to visualise the non-linear relations extracted by kernel CCA.

To conclude, this tutorial compiled the original, regularised, kernel, and sparse CCA into a unified framework to emphasise the applicabilities of the four variants in different data settings. The work highlights which CCA variant is most applicable depending on the sample size, data dimensionality and the type of relations of interest. Techniques for extracting the relations are also presented. Additionally, the importance of assessing the statistical significance and generalisability of the relations is emphasised. The tutorial hopefully advances both the practice of CCA variants in data analysis and further development of novel extensions.

The software used to produce the examples in this paper are available for download at https://github.com/aalto-ics-kepaco/cca-tutorial.

ACKNOWLEDGMENTS

The work by Viivi Uurtio and Juho Rousu has been supported in part by Academy of Finland (grant 295496/D4Health). João M. Monteiro was supported by a PhD studentship awarded by Fundação para a Ciência e a Tecnologia (SFRH/BD/88345/2012). John Shawe-Taylor acknowledges the support of the EPSRC through the C-PLACID project Reference: EP/M006093/1.

REFERENCES

S Akaho. 2001. A Kernel Method For Canonical Correlation Analysis. In *In Proceedings of the International Meeting of the Psychometric Society (IMPS2001*.

Md A Alam, M Nasser, and K Fukumizu. 2008. Sensitivity analysis in robust and kernel canonical correlation analysis. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*. IEEE, 399–404.

TW Anderson. 2003. An introduction to statistical multivariate analysis. (2003).

G Andrew, R Arora, J Bilmes, and K Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.

C Archambeau and FR Bach. 2009. Sparse probabilistic projections. In *Advances in neural information processing systems*. 73–80.

C Archambeau, N Delannay, and M Verleysen. 2006. Robust probabilistic projections. In *Proceedings of the 23rd International conference on machine learning*. ACM, 33–40.

S Arlot, A Celisse, and others. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys* 4 (2010), 40–79.

F Bach, R Jenatton, J Mairal, G Obozinski, and others. 2011. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning* 5 (2011).

FR Bach and MI Jordan. 2002. Kernel independent component analysis. *Journal of machine learning research* 3, Jul (2002), 1–48.

FR Bach and MI Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. (2005).

MS Bartlett. 1938. Further aspects of the theory of multiple regression. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 34. Cambridge Univ Press, 33–40.

MS Bartlett. 1941. The statistical significance of canonical correlations. *Biometrika* 32, 1 (1941), 29–37.

B Baur and S Bozdag. 2015. A canonical correlation analysis-based dynamic bayesian network prior to infer gene regulatory networks from multiple types of biological data. *Journal of Computational Biology* 22, 4 (2015), 289–299.

Å Björck and GH Golub. 1973. Numerical methods for computing angles between linear subspaces. *Mathematics of computation* 27, 123 (1973), 579–594.

MB Blaschko, CH Lampert, and A Gretton. 2008. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 133–145.

MW Browne. 2000. Cross-validation methods. *Journal of mathematical psychology* 44, 1 (2000), 108–132.

E Burg and J Leeuw. 1983. Non-linear canonical correlation. *British journal of mathematical and statistical psychology* 36, 1 (1983), 54–80.

J Cai. 2013. The distance between feature subspaces of kernel canonical correlation analysis. *Mathematical and Computer Modelling* 57, 3 (2013), 970–975.

L Cao, Z Ju, J Li, R Jian, and C Jiang. 2015. Sequence detection analysis based on canonical correlation for steady-state visual evoked potential brain computer interfaces. *Journal of neuroscience methods* 253 (2015), 10–17.

JD Carroll. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th annual convention of the American Psychological Association*, Vol. 3. 227–228.

B Chang, U Krüger, R Kustra, and J Zhang. 2013. Canonical Correlation Analysis based on Hilbert-Schmidt Independence Criterion and Centered Kernel Target Alignment.. In *ICML (2)*. 316–324.

X Chen, S Chen, H Xue, and X Zhou. 2012. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. *Pattern Recognition* 45, 5 (2012), 2005–2018.

X Chen, C He, and H Peng. 2014. Removal of muscle artifacts from single-channel EEG based on ensemble empirical mode decomposition and multiset canonical correlation analysis. *Journal of Applied Mathematics* 2014 (2014).

X Chen, H Liu, and JG Carbonell. 2012. Structured sparse canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics*. 199–207.

A Cichonska, J Rousu, P Marttinen, AJ Kangas, P Soininen, T Lehtimäki, OT Raitakari, M-R Järvelin, V Salomaa, M Ala-Korpela, and others. 2016. metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* (2016), btw052.

N Cristianini, J Shawe-Taylor, and H Lodhi. 2002. Latent semantic kernels. *Journal of Intelligent Information Systems* 18, 2-3 (2002), 127–152.

R Cruz-Cano and MLT Lee. 2014. Fast regularized canonical correlation analysis. *Computational Statistics & Data Analysis* 70 (2014), 88–100.

J Dauxois and GM Nkiet. 1997. Canonical analysis of two Euclidean subspaces and its applications. *Linear Algebra Appl.* 264 (1997), 355–388.

DL Donoho and IM Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association* 90, 432 (1995), 1200–1224.

RB Dunham and DJ Kravetz. 1975. Canonical correlation analysis in a predictive system. *The Journal of Experimental Education* 43, 4 (1975), 35–42.

C Eckart and G Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1, 3 (1936), 211–218.

B Efron. 1979. Computers and the theory of statistics: thinking the unthinkable. *SIAM review* 21, 4 (1979), 460–480.

LM Ewerbring and FT Luk. 1989. Canonical correlations and generalized SVD: applications and new algorithms. In *32nd Annual Technical Symposium*. International Society for Optics and Photonics, 206–222.

J Fang, D Lin, SC Schulz, Z Xu, VD Calhoun, and Y-P Wang. 2016. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. *Bioinformatics* 32, 22 (2016), 3480–3488.

Y Fujikoshi and LG Veitch. 1979. Estimation of dimensionality in canonical correlation analysis. *Biometrika* 66, 2 (1979), 345–351.

K Fukumizu, FR Bach, and A Gretton. 2007. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research* 8, Feb (2007), 361–383.

C Fyfe and PL Lai. 2000. Canonical correlation analysis neural networks. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, Vol. 2. IEEE, 977–980.

GH Golub and CF Van Loan. 2012. *Matrix computations*. Vol. 3. JHU Press.

GH Golub and H Zha. 1995. The canonical correlations of matrix pairs and their numerical computation. In *Linear algebra for signal processing*. Springer, 27–49.

I González, S Déjean, PGP Martin, O Gonçalves, P Besse, and A Baccini. 2009. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems* 17, 02 (2009), 173–199.

BK Gunderson and RJ Muirhead. 1997. On estimating the dimensionality in canonical correlation analysis. *Journal of Multivariate Analysis* 62, 1 (1997), 121–136.

DR Hardoon, J Mourao-Miranda, M Brammer, and J Shawe-Taylor. 2007. Unsupervised analysis of fMRI data using kernel canonical correlation. *NeuroImage* 37, 4 (2007), 1250–1259.

DR Hardoon and J Shawe-Taylor. 2009. Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine learning* 74, 1 (2009), 23–38.

DR Hardoon and J Shawe-Taylor. 2011. Sparse canonical correlation analysis. *Machine Learning* 83, 3 (2011), 331–353.

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

MJR Healy. 1957. A rotation method for computing canonical correlations. *Math. Comp.* 11, 58 (1957), 83–86.

C Heij and B Roorda. 1991. A modified canonical correlation approach to approximate state space modelling. In *Decision and Control, 1991., Proceedings of the 30th IEEE Conference on*. IEEE, 1343–1348.

AE Hoerl and RW Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

JW Hooper. 1959. Simultaneous equations and canonical correlation theory. *Econometrica: Journal of the Econometric Society* (1959), 245–256.

CE Hopkins. 1969. Statistical analysis by canonical correlation: a computer application. *Health services research* 4, 4 (1969), 304.

P Horst. 1961. Relations among sets of measures. *Psychometrika* 26, 2 (1961), 129–149.

H Hotelling. 1935. The most predictable criterion. *Journal of educational Psychology* 26, 2 (1935), 139.

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

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

I Huopaniemi, T Suvitaival, J Nikkilä, M Orešič, and S Kaski. 2010. Multivariate multi-way analysis of multi-source data. *Bioinformatics* 26, 12 (2010), i391–i398.

A Kabir, RD Merrill, AA Shamim, RDW Klemn, AB Labrique, P Christian, KP West Jr, and M Nasser. 2014. Canonical correlation analysis of infant's size at birth and maternal factors: a study in rural Northwest Bangladesh. *PloS one* 9, 4 (2014), e94243.

M Kang, B Zhang, X Wu, C Liu, and J Gao. 2013. Sparse generalized canonical correlation analysis for biological model integration: a genetic study of psychiatric disorders. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 1490–1493.

JR Kettenring. 1971. Canonical analysis of several sets of variables. *Biometrika* (1971), 433–451.

A Kimura, M Sugiyama, T Nakano, H Kameoka, H Sakano, E Maeda, and K Ishiguro. 2013. SemiCCA: Efficient semi-supervised learning of canonical correlations. *Information and Media Technologies* 8, 2 (2013), 311–318.

A Klami and S Kaski. 2007. Local dependent components. In *Proceedings of the 24th international conference on Machine learning*. ACM, 425–432.

A Klami, S Virtanen, and S Kaski. 2012. Bayesian exponential family projections for coupled data sources. *arXiv preprint arXiv:1203.3489* (2012).

A Klami, S Virtanen, and S Kaski. 2013. Bayesian canonical correlation analysis. *Journal of Machine Learning Research* 14, Apr (2013), 965–1003.

D Krstajic, LJ Buturovic, DE Leahy, and S Thomas. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics* 6, 1 (2014), 1.

PL Lai and C Fyfe. 1999. A neural implementation of canonical correlation analysis. *Neural Networks* 12, 10 (1999), 1391–1397.

PL Lai and C Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* 10, 05 (2000), 365–377.

NB Larson, GD Jenkins, MC Larson, RA Vierkant, TA Sellers, CM Phelan, JM Schildkraut, R Sutphen, PPD Pharoah, S A Gayther, and others. 2014. Kernel canonical correlation analysis for assessing gene–gene interactions and application to ovarian cancer. *European Journal of Human Genetics* 22, 1 (2014), 126–131.

SC Larson. 1931. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology* 22, 1 (1931), 45.

H-S Lee. 2007. Canonical correlation analysis using small number of samples. *Communications in StatisticsSimulation and Computation®* 36, 5 (2007), 973–985.

SE Leurgans, RA Moyeed, and BW Silverman. 1993. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* (1993), 725–740.

H Lindsey, JT Webster, and S Halpern. 1985. Canonical Correlation as a Discriminant Tool in a Periodontal Problem. *Biometrical journal* 27, 3 (1985), 257–264.

P Marttinen, J Gillberg, A Havulinna, J Corander, and S Kaski. 2013. Genome-wide association studies with high-dimensional phenotypes. *Statistical applications in genetics and molecular biology* 12, 4 (2013), 413–431.

T Melzer, M Reiter, and H Bischof. 2001. Nonlinear feature extraction using generalized canonical correlation analysis. In *International Conference on Artificial Neural Networks*. Springer, 353–360.

T Melzer, M Reiter, and H Bischof. 2003. Appearance models based on kernel canonical correlation analysis. *Pattern recognition* 36, 9 (2003), 1961–1971.

W Meredith. 1964. Canonical correlations with fallible data. *Psychometrika* 29, 1 (1964), 55–65.

MS Monmonier and FE Finn. 1973. Improving the interpretation of geographical canonical correlation models. *The Professional Geographer* 25, 2 (1973), 140–142.

M Nakanishi, Y Wang, Y-T Wang, and T-P Jung. 2015. A Comparison Study of Canonical Correlation Analysis Based Methods for Detecting Steady-State Visual Evoked Potentials. *PloS one* 10, 10 (2015), e0140703.

RM Neal. 2012. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media.

T Ogura, Y Fujikoshi, and T Sugiyama. 2013. A variable selection criterion for two sets of principal component scores in principal canonical correlation analysis. *Communications in Statistics-Theory and Methods* 42, 12 (2013), 2118–2135.

E Parkhomenko, D Tritchler, and J Beyene. 2007. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, Vol. 1. BioMed Central Ltd, S119.

P Rai and H Daume. 2009. Multi-label prediction via sparse infinite CCA. In *Advances in Neural Information Processing Systems*. 1518–1526.

J Rousu, DD Agranoff, O Sodeinde, J Shawe-Taylor, and D Fernandez-Reyes. 2013. Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria. *PLoS Comput Biol* 9, 4 (2013), e1003018.

Y Saad. 2011. *Numerical methods for large eigenvalue problems*. Vol. 158. SIAM.

T Sakurai. 2009. Asymptotic expansions of test statistics for dimensionality and additional information in canonical correlation analysis when the dimension is large. *Journal of Multivariate Analysis* 100, 5 (2009), 888–901.

BK Sarkar and C Chakraborty. 2015. DNA pattern recognition using canonical correlation algorithm. *Journal of biosciences* 40, 4 (2015), 709–719.

SV Schell and WA Gardner. 1995. Programmable canonical correlation analysis: A flexible framework for blind adaptive spatial filtering. *IEEE transactions on signal processing* 43, 12 (1995), 2898–2908.

B Schölkopf, A Smola, and K-R Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10, 5 (1998), 1299–1319.

JA Seoane, C Campbell, INM Day, JP Casas, and TR Gaunt. 2014. Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Comput Biol* 10, 10 (2014), e1003876.

J Shawe-Taylor and N Cristianini. 2004. *Kernel methods for pattern analysis*. Cambridge university press.

X-B Shen, Q-S Sun, and Y-H Yuan. 2013. Orthogonal canonical correlation analysis and its application in feature fusion. In *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 151–157.

C Soneson, H Lilljebjörn, T Fioretos, and M Fontes. 2010. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC bioinformatics* 11, 1 (2010), 1.

L Song, B Boots, SM Siddiqi, GJ Gordon, and A Smola. 2010. Hilbert space embeddings of hidden Markov models. (2010).

Y Song, PJ Schreier, D Ramírez, and T Hasija. 2016. Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing* 128 (2016), 449–458.

M Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)* (1974), 111–147.

MJ Sullivan. 1982. Distribution of Edaphic Diatoms in a Missisippi Salt Marsh: A Canonical Correlation Analysis. *Journal of Phycology* 18, 1 (1982), 130–133.

A Tenenhaus, C Philippe, and V Frouin. 2015. Kernel generalized canonical correlation analysis. *Computational Statistics & Data Analysis* 90 (2015), 114–131.

A Tenenhaus, C Philippe, V Guillemot, K-A Le Cao, J Grill, and V Frouin. 2014. Variable selection for generalized canonical correlation analysis. *Biostatistics* (2014), kxu001.

A Tenenhaus and M Tenenhaus. 2011. Regularized generalized canonical correlation analysis. *Psychometrika* 76, 2 (2011), 257–284.

CJF Ter Braak. 1990. Interpreting canonical correlation analysis through biplots of structure correlations and weights. *Psychometrika* 55, 3 (1990), 519–531.

R Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

XM Tu. 1991. A bootstrap resampling scheme for using the canonical correlation technique in rank estimation. *Journal of chemometrics* 5, 4 (1991), 333–343.

XM Tu, DS Burdick, DW Millican, and LB McGown. 1989. Canonical correlation technique for rank estimation of excitation-emission matrixes. *Analytical Chemistry* 61, 19 (1989), 2219–2224.

V Uurtio, M Bomberg, K Nybo, M Itävaara, and J Rousu. 2015. Canonical correlation methods for exploring microbe-environment interactions in deep subsurface. In *International Conference on Discovery Science*. Springer, 299–307.

JP Van de Geer. 1984. Linear relations amongk sets of variables. *Psychometrika* 49, 1 (1984), 79–94.

T Van Gestel, JAK Suykens, J De Brabanter, B De Moor, and J Vandewalle. 2001. Kernel canonical correlation analysis and least squares support vector machines. In *International Conference on Artificial Neural Networks*. Springer, 384–389.

HD Vinod. 1976. Canonical ridge and econometrics of joint production. *Journal of Econometrics* 4, 2 (1976), 147–166.

S Waaijenborg, PC Verselewel de Witt Hamer, and AH Zwinderman. 2008. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology* 7, 1 (2008).

C Wang. 2007. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks* 18, 3 (2007), 905–910.

D Wang, L Shi, DS Yeung, and ECC Tsang. 2005. Nonlinear canonical correlation analysis of fMRI signals using HDR models. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 5896–5899.

GC Wang, N Lin, and B Zhang. 2013. Dimension reduction in functional regression using mixed data canonical correlation analysis. *Stat Interface* 6 (2013), 187–196.

DS Watkins. 2004. *Fundamentals of matrix computations*. Vol. 64. John Wiley & Sons.

FV Waugh. 1942. Regressions between sets of variables. *Econometrica, Journal of the Econometric Society* (1942), 290–310.

DM Witten, R Tibshirani, and T Hastie. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* (2009), kxp008.

KW Wong, PCW Fung, and CC Lau. 1980. Study of the mathematical approximations made in the basis-correlation method and those made in the canonical-transformation method for an interacting Bose gas. *Physical Review A* 22, 3 (1980), 1272.

T Yamada and T Sugiyama. 2006. On the permutation test in canonical correlation analysis. *Computational statistics & data analysis* 50, 8 (2006), 2111–2123.

H Yamamoto, H Yamaji, E Fukusaki, H Ohno, and H Fukuda. 2008. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochemical Engineering Journal* 40, 2 (2008), 199–204.

Y-H Yuan, Q-S Sun, and H-W Ge. 2014. Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition. *Pattern Recognition* 47, 3 (2014), 1411–1424.

Y-H Yuan, Q-S Sun, Q Zhou, and D-S Xia. 2011. A novel multiset integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recognition* 44, 5 (2011), 1031–1040.

B Zhang, J Hao, G Ma, J Yue, and Z Shi. 2014. Semi-paired probabilistic canonical correlation analysis. In *International Conference on Intelligent Information Processing*. Springer, 1–10.

H Zou and T Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.

Received February 2017; revised July 2017; accepted August 2017