A regularized kernel CCA contrast function for ICA

Carlos Alzate*, Johan A.K. Suykens

Department of Electrical Engineering ESAT-SCD-SISTA, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium

Received 27 August 2007; received in revised form 30 November 2007; accepted 28 December 2007

Abstract

A new kernel based contrast function for independent component analysis (ICA) is proposed. This criterion corresponds to a regularized correlation measure in high dimensional feature spaces induced by kernels. The formulation is a multivariate extension of the least squares support vector machine (LS-SVM) formulation to kernel canonical correlation analysis (CCA). The regularization is incorporated naturally in the primal problem leading to a dual generalized eigenvalue problem. The smallest generalized eigenvalue is a measure of correlation in the feature space and a measure of independence in the input space. Due to the primal-dual nature of the proposed approach, the measure of independence can also be extended to out-of-sample points which is important for model selection ensuring statistical reliability of the proposed measure. Computational issues due to the large size of the matrices involved in the eigendecomposition are tackled via the incomplete Cholesky factorization. Simulations with toy data, images and speech signals show improved performance on the estimation of independent components compared with existing kernel-based contrast functions.

Keywords: Independent component analysis; Kernel canonical correlation analysis; Least squares support vector machines

1. Introduction

Canonical correlation analysis (CCA) finds linear relationships between two sets of variables (Hotelling, 1936). The objective is to find basis vectors on which the projected variables are maximally correlated. The CCA problem can be solved by means of a generalized eigenvalue problem. Generalized CCA consists of a generalization of CCA to more than two sets of variables (Kettenring, 1971). Kernel CCA as a nonlinear extension of CCA first maps the data into a high dimensional feature space induced by a kernel and then performs linear CCA. In this way, nonlinear relationships can be found (Akaho, 2001; Bach & Jordan, 2002; Lai & Fyfe, 2000; Melzer, Reiter, & Bischof, 2001). Without regularization, kernel CCA does not characterize the canonical correlation of the variables leading to a generalized eigenvalue problem with all eigenvalues equal to ±1 and ill-conditioning. Several ad-hoc regularization schemes have been proposed in order to obtain useful estimates of the correlation measure (Bach & Jordan, 2002; Gretton, Herbrich, Smola, Bousquet, & Schölkopf, 2005; Lai & Fyfe, 2000).

Independent component analysis (ICA) corresponds to a class of methods with the objective of recovering underlying latent factors present on the data. The observed variables are linear mixtures of the components which are assumed to be mutually independent (Cardoso, 1999; Comon, 1994; Hyvärinen, Karhunen, & Oja, 2001; Hyvärinen & Pajunen, 1999). Different approaches to ICA have been proposed in the context of neural networks (Jutten & Herault, 1991), tensor-based algorithms (Comon, 1994) and contrast functions (Hyvärinen et al., 2001). There is a wide range of applications of ICA such as exploratory data analysis, signal processing, feature extraction, noise reduction and blind source separation. A link between independence in the input space and kernel canonical correlation using universal kernels was presented in Bach and Jordan (2002). This link is called the $F$-correlation and establishes that if two variables in the feature space induced by an RBF kernel are completely uncorrelated then the original variables are independent in the input space. In the case of more than two variables, the $F$-correlation characterizes pairwise independence. Therefore, independent components can be estimated by minimizing...
a measure of correlation in the feature space. Different links have also been studied in the literature such as the constrained covariance (COCO) (Gretton et al., 2005), the kernel generalized variance (KGV) (Bach & Jordan, 2002) and the kernel mutual information (KMI) (Gretton et al., 2005).

In this paper, a kernel regularized correlation measure (KRC) is proposed. The KRC is an extended version of the method introduced in Alzate and Suykens (2007). This measure can be used as a contrast function for ICA. The formulation arises from a clear optimization framework where regularization is incorporated naturally in the primal problem leading to a generalized eigenvalue problem as the dual. The proposed method is a multivariate extension of the least squares support vector machines (LS-SVM) approach to kernel CCA (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002). The KRC can also be extended to out-of-sample data which is important for parameter selection.

This paper is organized as follows: In Section 2, we review linear CCA, kernel CCA and the LS-SVM interpretation of kernel CCA. In Section 3, we present a multivariate kernel CCA extension to more than two sets of variables. Section 4 contains a review of the link between kernel CCA and ICA together with the proposed contrast function. Section 5 discusses an approximation of the contrast function using the incomplete Cholesky decomposition together with an algorithm. In Section 6, we propose a model selection criterion. Section 7 contains the empirical results and in Section 8 we give conclusions.

2. LS-SVMs and kernel CCA

Least squares support vector machines (LS-SVM) formulations to different problems were discussed in Suykens et al. (2002). This class of kernel machines emphasizes primal-dual interpretations in the context of constrained optimization problems. LS-SVMs have been studied with respect to kernel Fisher discriminant analysis, kernel ridge regression, kernel PCA, kernel canonical correlation analysis, kernel partial least-squares, recurrent networks and control (Suykens et al., 2002). In this Section we discuss LS-SVM formulations to kernel CCA which are relevant for the sequel of the paper.

2.1. Linear CCA

The problem of CCA consists of measuring the linear relationship between two sets of variables (Gittins, 1985; Hotelling, 1936). Let \( \{x_1^{(1)}\}_{i=1}^N \) and \( \{x_1^{(2)}\}_{i=1}^N \) denote \( N \) observations of \( x^{(1)} \) and \( x^{(2)} \). Assume that the observations have zero mean. Then, the objective of CCA is to find basis vectors \( w^{(1)} \) and \( w^{(2)} \) such that the projected variables \( w^{(1)T}x^{(1)} \) and \( w^{(2)T}x^{(2)} \) are maximally correlated:

\[
\max_{w^{(1)},w^{(2)}} \rho = \frac{w^{(1)T}C_{12}w^{(2)}}{\sqrt{w^{(1)T}C_{11}w^{(1)}}\sqrt{w^{(2)T}C_{22}w^{(2)}}} \tag{1}
\]

where \( C_{11} = (1/N)\sum_{k=1}^N x_k^{(1)}x_k^{(1)T} \), \( C_{22} = (1/N)\sum_{k=1}^N x_k^{(2)}x_k^{(2)T} \). This is typically formulated as the optimization problem:

\[
\max_{w^{(1)},w^{(2)}} \quad w^{(1)T}C_{12}w^{(2)} \tag{2}
\]

such that \[
\begin{bmatrix}
  w^{(1)T}C_{11}w^{(1)} = 1 \\
  w^{(2)T}C_{22}w^{(2)} = 1 
\end{bmatrix} \tag{3}
\]

which leads to the generalized eigenvalue problem:

\[
\begin{bmatrix}
  0 & C_{12} \\
  C_{21} & 0
\end{bmatrix} \begin{bmatrix}
  w^{(1)} \\
  w^{(2)}
\end{bmatrix} = \rho \begin{bmatrix}
  C_{11} & 0 \\
  0 & C_{22}
\end{bmatrix} \begin{bmatrix}
  w^{(1)} \\
  w^{(2)}
\end{bmatrix}
\]

where \( \rho \) is the correlation coefficient. This problem has \( d_1 + d_2 \) eigenvalues \( \{\rho_1, -\rho_1, \ldots, \rho_p, -\rho_p, 0, \ldots, 0\} \), where \( p = \min(d_1, d_2) \).

An alternative scheme is to find the smallest eigenvalue of the following generalized eigenvalue problem:

\[
\begin{bmatrix}
  C_{11} & C_{12} \\
  C_{21} & C_{22}
\end{bmatrix} \begin{bmatrix}
  w^{(1)} \\
  w^{(2)}
\end{bmatrix} = (1 + \rho) \begin{bmatrix}
  C_{11} & 0 \\
  0 & C_{22}
\end{bmatrix} \begin{bmatrix}
  w^{(1)} \\
  w^{(2)}
\end{bmatrix}
\]

which provides a bounded correlation measure ranging from zero to one.

2.2. Kernel CCA

A nonlinear extension of CCA using kernels was introduced in Lai and Fyfe (2000) and Bach and Jordan (2002). The data are first embedded into a high dimensional feature space induced by a kernel and then linear CCA is applied. This way, nonlinear relations between variables can be found. Consider the feature maps \( \phi^{(1)}(\cdot) : \mathbb{R}^{d_1} \to \mathbb{R}^{n_1} \) and \( \phi^{(2)}(\cdot) : \mathbb{R}^{d_2} \to \mathbb{R}^{n_2} \) to high dimensional feature spaces of dimension \( n_1 \) and \( n_2 \) respectively. The centred feature matrices \( \Phi^{(1)} = \{\phi^{(1)}(x_i^{(1)})\} \in \mathbb{R}^{N \times n_1} \), \( \Phi^{(2)} = \{\phi^{(2)}(x_i^{(2)})\} \in \mathbb{R}^{N \times n_2} \) become

\[
\Phi^{(1)}_c = \{\phi^{(1)}(x_1^{(1)})^T - \mu^{(1)}; \ldots; \phi^{(1)}(x_N^{(1)})^T - \mu^{(1)}\},
\]

\[
\Phi^{(2)}_c = \{\phi^{(2)}(x_1^{(2)})^T - \mu^{(2)}; \ldots; \phi^{(2)}(x_N^{(2)})^T - \mu^{(2)}\},
\]

where \( \mu^{(l)} = (1/N)\sum_{i=1}^N \phi^{(l)}(x_i^{(l)}) \), \( l = 1, 2 \). The canonical correlation in the feature space can then be written as:

\[
\max_{w^{(1)}_K, w^{(2)}_K} \quad w^{(1)T}_K \Phi^{(1)T}_c \Phi^{(2)} w^{(2)}_K \tag{4}
\]

such that \[
\begin{bmatrix}
  w^{(1)T}_K \phi^{(1)}_c \phi^{(1)}_c w^{(1)}_K = 1 \\
  w^{(2)T}_K \phi^{(2)}_c \phi^{(2)}_c w^{(2)}_K = 1,
\end{bmatrix}
\]

the projection vectors \( w^{(1)}_K \) and \( w^{(2)}_K \) are assumed to lie in the span of the mapped data:

\[
\begin{bmatrix}
  w^{(1)}_K \\
  w^{(2)}_K
\end{bmatrix} = \Phi^{(1)T}_c \beta^{(1)} \tag{5}
\]

\[
\begin{bmatrix}
  w^{(1)}_K \\
  w^{(2)}_K
\end{bmatrix} = \Phi^{(2)T}_c \beta^{(2)}.
\]

Applying the kernel trick results in the following optimization problem:

\[
\max_{\beta^{(1)}, \beta^{(2)}} \quad \beta^{(1)T} \Omega^{(1)} \Omega^{(2)} \beta^{(2)}
\]
such that
\[ \begin{align*}
\beta^{(1)T} \Omega_c^{(1)} \beta^{(1)} &= 1 \\
\beta^{(2)T} \Omega_c^{(2)} \beta^{(2)} &= 1 
\end{align*} \]

where \( \Omega_c^{(i)} \) denotes the \( i \)-th centred kernel matrix \( \Omega_c^{(i)} = P \Omega^{(i)} P, \ P = I_N - (1/N) 1_N 1_N^T \) is the centring matrix, \( I_N \) is the \( N \times N \) identity matrix, \( 1_N \) is a vector of \( N \) ones, \( \Omega^{(i)} \) is the kernel matrix with \( ij \)-entry: \( \Omega^{(i)}_{ij} = K^{(i)}(x_i^T, x_j^T) \) and \( K^{(i)}(x_i, x_j) = \varphi^{(i)}(x_i)^T \varphi^{(i)}(x_j) \) is a Mercer kernel, for \( i = 1, 2 \). The problem in (5) can be written as the generalized eigenvalue problem:
\[ \begin{bmatrix}
0 & \Omega_c^{(1)} \\
\Omega_c^{(2)} & 0
\end{bmatrix} \begin{bmatrix}
\beta^{(1)} \\
\beta^{(2)}
\end{bmatrix} = \rho \begin{bmatrix}
\Omega_c^{(1)} \\
\Omega_c^{(2)}
\end{bmatrix} \begin{bmatrix}
\beta^{(1)} \\
\beta^{(2)}
\end{bmatrix}. \]

The nonzero solutions to this generalized eigenvalue problem are always \( \rho = \pm 1 \). This is due to the fact that the angle between the subspaces spanned by the columns of the kernel matrices \( \Omega_c^{(1)} \) and \( \Omega_c^{(2)} \) is always equal to zero independent of the data. Hence, the canonical correlation estimate is always equal to one (Bach & Jordan, 2002). This issue can also be interpreted as ill-conditioning since the kernel matrices can be singular or near-singular. Several ad-hoc regularization schemes have been proposed in order to obtain useful estimators of the canonical correlation in high dimensional feature spaces (Bach & Jordan, 2002; Gretton et al., 2005).

The regularized method proposed in Bach and Jordan (2002) consists of finding the smallest eigenvalue of the following problem:
\[ \begin{bmatrix}
\Omega_c^{(1)} + \frac{N_c}{N} I_N \\
\Omega_c^{(2)}
\end{bmatrix} \begin{bmatrix}
\beta^{(1)} \\
\beta^{(2)}
\end{bmatrix} = \lambda \begin{bmatrix}
\Omega_c^{(1)} \\
\Omega_c^{(2)}
\end{bmatrix} \begin{bmatrix}
\beta^{(1)} \\
\beta^{(2)}
\end{bmatrix} \]

where \( c \) is small positive regularization constant depending of the sample size.

### 2.3. LS-SVM approach to kernel CCA

An LS-SVM formulation to kernel CCA was introduced in Suykens et al. (2002) with the following primal form:
\[ \max_{w,v,\gamma} \gamma e^T r - v_1 \frac{1}{2} e^T e - v_2 \frac{1}{2} r^T r - \frac{1}{2} w^T w - \frac{1}{2} v^T v \]

such that
\[ \begin{align*}
\gamma &\geq \varphi^{(1)} w \\
\gamma &\geq \varphi^{(2)} v,
\end{align*} \]

Proposition 1 (Suykens et al., 2002). Given a positive definite kernel function \( K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) with \( K(x, z) = \varphi(x)^T \varphi(z) \) and regularization constants \( \gamma, v_1, v_2 \in \mathbb{R}^+ \), the dual problem to (7) is given by the following eigenvalue problem:
\[ \begin{bmatrix}
0 & \Omega_c^{(2)} \\
\Omega_c^{(1)} & 0
\end{bmatrix} \begin{bmatrix}
\xi^{(1)} \\
\xi^{(2)}
\end{bmatrix} = \lambda \begin{bmatrix}
\Omega_c^{(1)} \\
\Omega_c^{(2)}
\end{bmatrix} \begin{bmatrix}
\xi^{(1)} \\
\xi^{(2)}
\end{bmatrix}. \]

**Proof.** Consider the Lagrangian of (7):
\[ \mathcal{L}(w, v, e; \gamma, \tau) = \gamma e^T r - v_1 \frac{1}{2} e^T e - v_2 \frac{1}{2} r^T r - \frac{1}{2} w^T w - \frac{1}{2} v^T v - \gamma e^T \xi + \gamma r^T \xi - \tau e^T v - \tau r^T w \]

where \( \xi, \tau \) are Lagrange multiplier vectors. The Karush-Kuhn-Tucker (KKT) conditions for optimality are
\[ \begin{align*}
\frac{\partial \mathcal{L}}{\partial w} &= 0 & \rightarrow & w &= \varphi^{(1)T} \xi \\
\frac{\partial \mathcal{L}}{\partial v} &= 0 & \rightarrow & v &= \varphi^{(2)T} \tau \\
0 &= 0 & \rightarrow & \gamma r &= v_1 e + \xi \\
0 &= 0 & \rightarrow & \gamma e &= v_2 r + \tau \\
0 &= 0 & \rightarrow & e &= \varphi^{(1)} w \\
0 &= 0 & \rightarrow & r &= \varphi^{(2)} v.
\end{align*} \]

Eliminating the primal variables \( w, v, e, r \) gives:
\[ \gamma v_1 \varphi^{(1)T} \varphi^{(1)} \xi + \xi = v_2 \varphi^{(2)T} \varphi^{(2)} \tau + \tau \]
defining \( \lambda = 1/\gamma \) and applying the definition of a positive definite kernel function leads to the following set of equations:
\[ \begin{align*}
\Omega_c^{(1)} \xi &= \lambda (v_1 \varphi^{(1)}(x_1)^T \xi + \xi) \\
\Omega_c^{(2)} \xi &= \lambda (v_2 \varphi^{(2)}(x_1)^T \tau + \tau)
\end{align*} \]

which can be written as the generalized eigenvalue problem in (8). \( \square \)

**Remark 1.** Note that the primal objective function (7) does not have the same expression as the correlation coefficient but takes the numerator with a positive sign and the denominator with a negative sign. Moreover, it can be considered as a generalization of the problem \( \min_{w,v} \| e - r \|_2^2 \) which is known in CCA (Gittins, 1985).

**Remark 2.** Note that the regularization is incorporated in the primal optimization problem in a natural way.

**Remark 3.** The projections of the mapped training data onto the \( i \)-th eigenvector (also called the score variables) become
\[ \begin{align*}
z_c^{(i)} &= \varphi^{(1)} w = \Omega_c^{(1)} \xi^{(i)}, \\
z_c^{(i)} &= \varphi^{(2)} v = \Omega_c^{(2)} \tau^{(i)}, \quad i = 1, \ldots, 2N.
\]
Remark 4. The optimal $\lambda^*$, $\xi^*$, $\tau^*$ can be selected such that the correlation coefficient is maximized:

$$\max_{i \in \{1, \ldots, 2N\}} \frac{z_i^{(2)} z_i^{(3)}}{\|z_i^{(2)}\|_2 \|z_i^{(3)}\|_2}.$$  

3. Multivariate kernel CCA

In this section, we extend the LS-SVM formulation to kernel CCA to $m$ sets of variables. Given training data $\{x_i^{(l)}\}$, $i = 1, \ldots, N$, $l = 1, \ldots, m$, the multivariate kernel CCA problem can be formulated in the primal as:

$$\max_{w^{(l)}, e^{(l)}} \frac{\gamma}{2} \sum_{l=1}^{m} \sum_{k \neq l} e^{(l)^T} e^{(k)} - \frac{1}{2} \sum_{l=1}^{m} \sum_{j \neq l} e^{(j)} e^{(l)^T} e^{(l)}$$

$$- \frac{1}{2} \sum_{l=1}^{m} w^{(l)^T} w^{(l)}$$

such that $e^{(l)} = \Phi_{c}^{(l)} w^{(l)}$, $l = 1, \ldots, m$

where $e^{(l)}$ is the compact form of the $l$-th projected variables

$$e^{(l)} = w^{(l)^T} (\Phi^{(l)} (x_i^{(l)}) - \mu_{\psi^{(l)}}),$$

$i = 1, \ldots, N$, $l = 1, \ldots, m$.

the $l$-th centred feature matrix $\Phi_{c}^{(l)} \in \mathbb{R}^{N \times nh}$ is

$$\Phi_{c}^{(l)} = \begin{bmatrix}
\phi^{(l)} (x_{1}^{(l)})^T - \mu_{\psi^{(l)}}
\phi^{(l)} (x_{2}^{(l)})^T - \mu_{\psi^{(l)}}
\vdots
\phi^{(l)} (x_{N}^{(l)})^T - \mu_{\psi^{(l)}}
\end{bmatrix},$$

where $\gamma$, $\nu_l$ are regularization parameters, $\Phi^{(l)}(\cdot)$ are the mappings to high dimensional spaces and

$$\mu_{\psi^{(l)}} = (1/N) \sum_{i=1}^{N} \phi^{(l)} (x_i^{(l)}), \quad l = 1, \ldots, m.$$

Lemma 1. Given $m$ positive definite kernel functions $K^{(l)} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ with $K^{(l)}(x, z) = \phi^{(l)}(x)^T \phi^{(l)}(z)$ and regularization constants $\gamma, \nu_l \in \mathbb{R}^+$, $l = 1, \ldots, m$, the dual problem to (11) is given by the following generalized eigenvalue problem:

$$K_\alpha = \gamma R_\alpha$$

with $K$, $R$ and $\alpha$ defined as follows:

$$K = \begin{bmatrix}
0 & \Omega^{(2)} & \cdots & \Omega^{(m)} \\
\Omega^{(1)} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\Omega^{(1)} & \Omega^{(2)} & \cdots & 0 \\
\end{bmatrix},$$

$$R = \begin{bmatrix}
R^{(1)} & 0 & \cdots & 0 \\
0 & R^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & R^{(m)} \\
\end{bmatrix},$$

$$R^{(l)} = (I_N + \nu_l \Omega^{(l)}), l = 1, \ldots, m$$ and $\lambda = 1/\gamma$.

Proof. The Lagrangian for this constrained optimization problem is:

$$\mathcal{L}(w^{(l)}, e^{(l)}; \alpha^{(l)}) = \frac{\gamma}{2} \sum_{l=1}^{m} \sum_{k \neq l} e^{(l)^T} e^{(k)} - \frac{1}{2} \sum_{l=1}^{m} \sum_{j \neq l} e^{(j)} e^{(l)^T} e^{(l)}$$

$$- \frac{1}{2} \sum_{l=1}^{m} w^{(l)^T} w^{(l)}$$

with conditions for optimality given by

$$\frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \rightarrow w^{(l)} = \Phi_{c}^{(l)} \alpha^{(l)}$$

$$\frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \rightarrow \gamma \sum_{k \neq l} e^{(k)} = \nu_l e^{(l)} + \alpha^{(l)}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \rightarrow e^{(l)} = \Phi_{c}^{(l)} w^{(l)},$$

for $l = 1, \ldots, m$. Eliminating the primal variables $w^{(l)}, e^{(l)}$ gives:

$$\gamma \sum_{k \neq l} \Phi_{c}^{(k)^T} \Phi_{c}^{(l)} \alpha^{(k)} = (\nu_l \Phi_{c}^{(l)^T} \Phi_{c}^{(l)} + I_N) \alpha^{(l)}, \quad l = 1, \ldots, m$$

defining $\lambda = 1/\gamma$ and applying the definition of a positive definite kernel function leads to the following set of equations:

$$\sum_{k \neq l} \Phi_{c}^{(k)^T} \Phi_{c}^{(l)} \alpha^{(k)} = \lambda (\nu_l \Phi_{c}^{(l)^T} \Phi_{c}^{(l)} + I_N) \alpha^{(l)}, \quad l = 1, \ldots, m$$

which can be written as the generalized eigenvalue problem in (12).

Remark 5. Note that the regularized matrix $R$ differs from the schemes proposed in Bach and Jordan (2002), Gretton et al. (2005) and Yamanishi, Vert, Nakaya, and Kanehisa (2003). These schemes start from an unregularized kernel CCA problem and perform regularization afterwards in an ad-hoc manner. As already discussed, the proposed regularization arises naturally from the primal optimization problem providing insights and leading to better numerically conditioned eigenvalue problems.

Remark 6. The score variables for training data can be expressed as

$$z_i^{(l)} = \Phi_{c}^{(l)} w_i^{(l)} = \Omega^{(l)} \alpha_i^{(l)},$$

where $\alpha_i^{(l)}$ is the $\alpha^{(l)}$ corresponding to the $i$-th eigenvector, $i = 1, \ldots, Nm$, $l = 1, \ldots, m$.

Remark 7. The optimal $\lambda^*$, $\alpha^*$ can be selected such that the sum of the correlation coefficients is maximized:

$$\max_{i \in \{1, \ldots, N\}} \sum_{l=1}^{m} \sum_{k \neq l} \|z_i^{(l)}\|_2 \|z_i^{(k)}\|_2$$

Remark 8. Given new test data $\{\tilde{x}_j^{(l)}\}$, $j = 1, \ldots, N_{\text{test}}$, $l = 1, \ldots, m$ the projections $z_{\text{test}}^{(l)}$ can be calculated as:

$$z_{\text{test}}^{(l)} = \tilde{z}_{\text{test}}^{(l)} \alpha^{(l)},$$
where $\Omega^{(l)}_{\text{test},ij} \in \mathbb{R}^{N_{\text{test}} \times N}$ is the kernel matrix evaluated using the test points with $ji$-entry $\Omega^{(l)}_{\text{test},ji} = K^{(l)}(\mathbf{x}^{(l)}_j, \mathbf{x}^{(l)}_i)$, $j = 1, \ldots, N_{\text{test}}, i = 1, \ldots, N$.

4. ICA and kernel CCA

4.1. The $F$ correlation

Consider the following statistical model:

$$Y = MS,$$

where $S$ is a $p \times N$ matrix representing $N$ identically distributed observations of $p$ independent components, $Y$ is the $p \times N$ matrix of mixtures and $M$ is a $p \times p$ mixing matrix assumed to be invertible. The ICA problem (Amari, Cichocki, & Yang, 1996; Cardoso, 1999; Comon, 1994; Hyvärinen et al., 2001; Hyvärinen & Pajunen, 1999) consists of estimating $S$ based on $Y$. If the distributions of the components are specified, maximum likelihood can be used to estimate the ICA parametric model (Cardoso, 1999). However, in practice, the distributions of the components are usually not known and the ICA problem is tackled using an estimate of the demixing matrix $W = M^{-1}$ allowing the estimation of independent components via $\hat{S} = \hat{W}Y$ (Hyvärinen et al., 2001). A typical approach to ICA consists of an approximation of the mutual information between the components of $\hat{S}$ (Amari et al., 1996; Cardoso, 1999). However, alternative contrast functions based on higher-order moments have also been proposed.

The matrix of mixtures $Y$ is often preprocessed by making its covariance matrix equal to the identity matrix. This preprocessing step is called whitening and reduces the number of parameters to be estimated because the demixing matrix $\hat{W}$ is then necessarily orthogonal (Hyvärinen et al., 2001). Therefore, the ICA problem can be reduced to finding an orthogonal matrix $\hat{W}$ such that the components of $\hat{S}$ are independent. A measure of independence based on kernel canonical correlation was introduced in Bach and Jordan (2002) with the $F$-correlation functional as contrast function. This functional establishes a link between canonical correlation in feature spaces induced by RBF kernels and independence in the original input space and it is defined as:

$$\rho_F \left( f^{(1)}(\mathbf{x}^{(1)}), f^{(2)}(\mathbf{x}^{(2)}) \right) = \max_{f^{(1)} \in \mathcal{F}^{(1)}, f^{(2)} \in \mathcal{F}^{(2)}} \text{corr} \left( f^{(1)}(\mathbf{x}^{(1)}), f^{(2)}(\mathbf{x}^{(2)}) \right),$$

where $f^{(l)}(\mathbf{x}^{(l)}) = (\varphi^{(l)}(\mathbf{x}^{(l)}), f^{(l)}), l = 1, 2$ is the reproducing property of RKHS. In the case of RBF kernels, this functional is zero if and only if $x^{(1)}$ and $x^{(2)}$ are independent (Bach & Jordan, 2002). In the multivariate case, the $F$-correlation characterizes pairwise independence. Different kernel-based measures of independence have been proposed such as the constrained covariance (COCO) (Gretton et al., 2005), the kernel mutual information (KMI) (Gretton et al., 2005) and the kernel generalized variance (KGV) (Bach & Jordan, 2002).

4.2. The kernel regularized correlation (KRC) contrast function

The largest generalized eigenvalue of the problem (12) provides a measure of correlation among the sets of data. Even though this measure may not correspond to the correlation coefficient, it can be used as a contrast function for ICA. Consider the following modification to (12):

$$(\mathbb{K} + \mathbb{R})\alpha = \zeta \mathbb{R}\alpha,$$

(14)

where $\zeta = 1 + \lambda$. The largest eigenvalue of (12) is related to the smallest eigenvalue $\xi_{\min}$ of (14) and $\xi_{\min}$ is bounded between zero and one. Therefore, the proposed kernel regularized correlation (KRC) contrast function for ICA is defined as follows for training data:

$$\hat{\rho}(\Omega^{(1)}, \ldots, \Omega^{(m)}) = 1 - \xi_{\min}.$$  

(15)

The KRC for test data is the sum of the correlation coefficients of the projections:

$$\hat{\rho}(\Omega^{(1)}_{\text{test}}, \ldots, \Omega^{(m)}_{\text{test}}) = \sum_{l=1}^{m} \sum_{k \neq \ell} \frac{\zeta^{(l)}_{\text{test}} \cdot \zeta^{(k)}_{\text{test}}}{\|\zeta^{(l)}_{\text{test}}\|_2 \|\zeta^{(k)}_{\text{test}}\|_2}.$$  

(16)

5. Optimization on the Stiefel manifold

5.1. Computational issues

For real-life problems, the cost of storing (14) and computing the smallest eigenvalue can be prohibitive due to the size of the matrices involved in the eigendecomposition. However, low-rank approximations of the kernel matrices can be found by means of the incomplete Cholesky decomposition. An $N \times N$ kernel matrix $\Omega$ can be factorized as $\Omega = GG^T$ where $G \in \mathbb{R}^{N \times M}$ is lower triangular. If the eigenvalues of $\Omega$ decay sufficiently rapidly then it is possible to make the approximation $\Omega \approx GG^T$ where $G \in \mathbb{R}^{N \times M}$ such that $\|\Omega - GG^T\|_2 \leq \eta$ and $M < N$. This method is known as the incomplete Cholesky decomposition with symmetric pivoting and was used in Bach and Jordan (2002) and Gretton et al. (2005) to reduce the computational burden of the corresponding eigenvalue problem.

Making use of the incomplete Cholesky decomposition, the centred kernel matrices in (14) can be approximated as $\Omega^{(l)}_c \approx G_c^{(l)} G_c^{(l)^T}$. The SVD of $G_c^{(l)}$ is $U^{(l)} A^{(l)} (U^{(l)^T})^T$ where $U^{(l)}$ is an $N \times M_l$ orthogonal matrix and $A^{(l)}$ is an $M_l \times M_l$ diagonal matrix. Therefore we can approximate (14) with

$$\tilde{K}\alpha = \zeta \tilde{R}\alpha,$$  

(17)

where

$$\tilde{K} = \begin{bmatrix}
U^{(1)}(U^{(1)}A^{(1))})^T & U^{(2)}(A^{(1))}U^{(2))})^T & \cdots & U^{(m)}(A^{(1))}U^{(m))})^T \\
U^{(1)}(A^{(1))}U^{(1))})^T & U^{(2)}(A^{(1))}U^{(2))})^T & \cdots & U^{(m)}(A^{(1))}U^{(m))})^T \\
\vdots & \vdots & \ddots & \vdots \\
U^{(1)}(A^{(1))}U^{(1))})^T & U^{(2)}(A^{(1))}U^{(2))})^T & \cdots & U^{(m)}(A^{(1))}U^{(m))})^T
\end{bmatrix},$$
\[\tilde{R} = \begin{bmatrix} U^{(1)^{T}}A^{(1)}U^{(1)^{T}} & 0 & \cdots & 0 \\ 0 & U^{(1)^{T}}A^{(2)}U^{(2)^{T}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & U^{(m)^{T}}A^{(m)}U^{(m)^{T}} \end{bmatrix}.\]

\[A^{(l)}_{\nu,n} = 1 + \nu A^{(l)}_{\nu,nn}, \quad \nu = 1, \ldots, M_l.\]

Substituting \(\tilde{\alpha}^{(l)} = U^{(l)^{T}}\alpha^{(l)}, \nu, l = 1, \ldots, m\) and premultiplying both sides of (17) by \(\text{blkdiag}(U^{(1)^{T}}, \ldots, U^{(m)^{T}})\) leads to the following reduced problem:

\[\tilde{K}_R\tilde{\alpha} = \xi\tilde{\alpha},\]

where

\[\tilde{K}_R = \begin{bmatrix} I_{M_1} & T^{(1)}_{U^{(1)}U^{(2)}} & \cdots & T^{(1)}_{U^{(1)}U^{(m)}} \\ T^{(2)}_{U^{(2)}U^{(1)}} & I_{M_2} & \cdots & T^{(2)}_{U^{(2)}U^{(m)}} \\ \vdots & \vdots & \ddots & \vdots \\ T^{(m)}_{U^{(m)}U^{(1)}} & T^{(m)}_{U^{(m)}U^{(2)}} & \cdots & I_{M_m} \end{bmatrix},\]

\[\tilde{\alpha} = \begin{bmatrix} \tilde{\alpha}^{(1)} \\ \tilde{\alpha}^{(2)} \\ \vdots \\ \tilde{\alpha}^{(m)} \end{bmatrix},\]

\[T^{(l)} = T^{(l)^{T}}T^{(l)}, \quad T^{(l)} \text{ is a diagonal matrix } T^{(l)}_{ll} = 1/A^{(l)}_{\nu,nn}.\]

Algorithm 1: Kernel Regularized Correlation (KRC)

**Input:** Mixture set of \(m\) sources \(\{x_i^{(l)}\}_{l=1}^{m}, l = 1, \ldots, m\), RBF kernel parameter \(\sigma^2\), regularization parameter \(\nu\), Cholesky error threshold \(\eta\).

**Output:** \(\hat{\rho}(\Omega_c^{(1)}, \ldots, \Omega_c^{(m)}).\)

1. Compute the Cholesky factor \(G_c^{(l)} \in \mathbb{R}^{N \times M_l}\) such that \(\|\tilde{\Omega}_c^{(l)} - G_c^{(l)}G_c^{(l)^{T}}\|_2 \leq \eta, l = 1, \ldots, m\).
2. Compute the SVD \(G_c^{(l)} = U^{(l)}A^{(l)}U^{(l)^{T}}, l = 1, \ldots, m\) and build the matrix \(\tilde{K}_R\) (18).
3. Compute the smallest eigenvalue \(\xi_{\text{min}}\) of the problem \(\tilde{K}_R\tilde{\alpha} = \xi\tilde{\alpha}\).
4. Return \(1 - \xi_{\text{min}}\).

Algorithm 2: Gradient descent on the Stiefel manifold

**Input:** Whitened mixture set of \(m\) sources \(\{x_i^{(l)}\}_{l=1}^{m}, l = 1, \ldots, m\), RBF kernel parameter \(\sigma^2\), regularization parameter \(\nu\), Cholesky error threshold \(\eta\), initial demixing matrix \(W_0\).

**Output:** Demixing matrix \(W\).

1. \(W \leftarrow W_0\)
2. **repeat**
3. Compute \(D\) the derivative of the KRC with respect to \(W\) using first-order differences.
4. Compute the direction \(A = D - W D^T W\).
5. Find a displacement \(t\) along the geodesic using golden search in the direction \(A\) such that the KRC decreases.
6. Update \(W \leftarrow W \exp(tW^T A)\)
7. **until** change of \(W\) in two consecutive iterations is less than some given threshold.

Cholesky decompositions. However, as noted by Bach and Jordan (2002), the bottleneck of the computation is the incomplete Cholesky decomposition. Therefore, the empirical computational complexity of the KRC is \(O(m M^2 N)\).

6. Model selection

Typically, the free parameters in kernel-based approaches to ICA are set heuristically to some values that show good performance on a number of experiments. However, overfitting may occur due to the fact that the estimation of the correlation coefficient is done solely on training data especially when few data are available. One of the main advantages of our proposed approach is a clear constrained optimization problem for kernel CCA and the corresponding measure of independence. This allows the computation of projected variables for out-of-sample (test) data points that can be used to select values for the free parameters. The parameters values are chosen such that the normalized correlation coefficient is minimal for validation data. This ensures statistical reliability of the estimated canonical correlation.

We assume that the RBF kernel parameters are identical \((\sigma^{(l)})^2 = \sigma^2, l = 1, \ldots, m\) which means considering one feature map \(\varphi^{(l)} = \varphi\). The regularization parameters are also considered to be identical \(\nu = \nu\) and fixed to some value.
Given validation data \( \{ \tilde{x}_h^{(l)} \}, h = 1, \ldots, N_{\text{val}}, l = 1, \ldots, m \), the following model selection criterion is proposed:

\[
\hat{\rho}(\Omega_{\text{val}}^{(1)}, \ldots, \Omega_{\text{val}}^{(m)}) = \sum_{l=1}^{m} \sum_{k \neq l} \frac{z_{\text{val}}^{(l)T}(k) z_{\text{val}}^{(k)}}{\| z_{\text{val}}^{(l)} \|_2 \| z_{\text{val}}^{(k)} \|_2 \| \Omega_{\text{val}} \|_F}
\] (19)

where \( z_{\text{val}}^{(l)} = \Omega_{\text{val}}^{(l)} \alpha^{(l)*} \), \( \Omega_{\text{val}}^{(l)} \) is the kernel matrix evaluated using the validation points with \( hi \)-entry \( \Omega_{\text{val}, hi}^{(l)} = K^{(l)}(x_h^{(l)}, x_i^{(l)}), h = 1, \ldots, N_{\text{val}}, i = 1, \ldots, N \) and \( \Omega_{\text{val}} = \text{blkdiag}(\Omega_{\text{val}}^{(1)}, \ldots, \Omega_{\text{val}}^{(m)}) \). Note that, for RBF kernels the correlation measure decreases as \( \sigma^2 \) increases (Bach & Jordan, 2002). Therefore, \( \| \Omega_{\text{val}} \|_F \) in (19) fixes the scaling introduced by \( \sigma^2 \).

7. Empirical results

In this section, we present empirical results. In all experiments, we assume that the number of sources \( m \) is given and the incomplete Cholesky parameter was fixed to \( \eta = 1 \times 10^{-4} \). We use the Amari divergence (Amari et al., 1996) to assess the performance of the proposed approach when the optimal demixing matrix \( W \) is known. This error measure compares two demixing matrices and is equal to zero if and only if the matrices represent the same components. Moreover, the Amari divergence is invariant to permutation and scaling which makes it suitable for ICA. We compared four different existing algorithms for ICA: FastICA (Hyvärinen & Oja, 1997), the KCCA contrast (Bach & Jordan, 2002), the KGV (Bach & Jordan, 2002), Jade (Cardoso, 1999) and the proposed KRC. The kernel-based algorithms were initialized with the solution of the KGV using a Hermite polynomial of degree \( d = 3 \) and width \( \sigma = 1.5 \). This solution does not characterize independence but it was shown to be a good starting point (Bach & Jordan, 2002). The computation time of the eigenvalue decomposition involved in the calculation of the KRC for training data was 0.98 s for \( N = 28,000, m = 3, \eta = 10^{-4} \) on a Pentium 4, 2.8 GHz, 1 GB RAM using Lanczos algorithm in Matlab. In the case of test and validation data, the computation time increased to 23.56 s because the projected variables should be calculated in order to compute the correlation coefficient.

7.1. Toy examples—simple signals

In this experiment we used three simple signals: a sine function, a sawtooth function and a sinc function. The sample
size was $N = 400$. The sources were mixed using a product of known Jacobi rotations $\theta_3 = -\pi/3, \theta_2 = -\pi/2, \theta_1 = -\pi/5$. In this way, the optimal demixing matrix is a 3-D rotation matrix with angles $-\theta_1, -\theta_2, -\theta_3$. The first angle was fixed to $\pi/5$ and an estimate of the demixing matrix was made using a grid in the range $[0, \pi]$ for the remaining two angles. Fig. 1 shows the contour plots of the Amari divergence, the KCCA and KGV criteria described in Bach and Jordan (2002) and the proposed KRC criterion. The KCCA and KGV parameters were set to the values recommended by the authors. The KRC parameters were set to $\sigma^2 = 5, \nu = 2$. The black dot shows the location of the optimal set of angles which corresponds to a local minimum for all criteria. The effect of the KRC parameters $\sigma^2, \nu$ on the optimization surface can be seen in Fig. 2. Note that the optimal set of angles corresponds to a local minimum in the KRC for a wide range of parameter values.

7.2. Toy examples—source distributions

This experiment consisted of demixing six different source distributions. The sample size was $N = 900$. The distributions include a multimodal asymmetrical mixture of two Gaussians, a Student’s $t$ with three degrees of freedom, a uniform distribution, a mixture of a Gaussian and uniform, a beta distribution and a gamma distribution. We performed 20 runs with different random mixing matrices and compute the mean Amari error for the KCCA, the KGV, the FastICA algorithm, Jade and the proposed KRC. As in the previous toy example, the parameters of the KCCA and KGV were set to the default values. The FastICA parameters were set to the default values and the kurtosis-based contrast function was used. In all cases the KRC obtained the smallest Amari error.

<table>
<thead>
<tr>
<th># sources</th>
<th>KCCA</th>
<th>KGV</th>
<th>KRC</th>
<th>FastICA</th>
<th>Jade</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4.02</td>
<td>2.77</td>
<td>2.75</td>
<td>12.92</td>
<td>4.95</td>
</tr>
<tr>
<td>3</td>
<td>7.43</td>
<td>2.89</td>
<td>2.36</td>
<td>9.37</td>
<td>6.98</td>
</tr>
<tr>
<td>4</td>
<td>6.12</td>
<td>3.55</td>
<td>2.19</td>
<td>9.88</td>
<td>8.30</td>
</tr>
<tr>
<td>5</td>
<td>6.39</td>
<td>5.75</td>
<td>4.48</td>
<td>11.03</td>
<td>8.88</td>
</tr>
<tr>
<td>6</td>
<td>8.42</td>
<td>5.57</td>
<td>4.06</td>
<td>9.23</td>
<td>7.49</td>
</tr>
</tbody>
</table>

The KCCA and KGV parameters were set to the values recommended by the authors ($\sigma^2 = 1, \kappa = 2 \times 10^{-2}$) in Bach and Jordan (2002). The KRC parameters were set to $\sigma^2 = 0.5, \nu = 1$. The FastICA algorithm was set to the kurtosis-based contrast and the default parameters. In all cases the KRC obtained the smallest Amari error.
Fig. 3. Amari errors of the source distributions experiment for the KCCA, KGV and the proposed KRC contrast after 20 runs with different mixing matrices. Top left: three sources. Top right: four sources. Bottom left: five sources. Bottom right: six sources. The KRC performs better with respect to the KCCA and KGV.

Fig. 4. Mean computation times of the source distributions experiment. The KRC compares favourably with respect the KCCA contrast. The smallest Amari error. Mean computation times for the KCCA, the KGV and the KRC are shown in Fig. 4.

7.3. Speech signals

We used three fragments of 6.25 s of speech sampled at 8 KHz with 8 bit quantization. We performed 10 runs with a random window of $N = 10,000$ consecutive data points. The fragments were combined with a randomly generated mixing matrix. The KCCA, KGV and the proposed KRC were initialized with Jade and the parameters were fixed to the defaults. Fig. 5 shows the performance and computation times. The KRC performs better than KCCA and KGV in terms of...
the Amari error with computation time comparable to the KGV. The Jade algorithm provides a fast and efficient way to initialize the KRC. Even though speech signals violate the i.i.d. data assumption, ICA has been applied to speech demixing with good results.

7.4. Image demixing

We used three $140 \times 200$ pixels images from the Berkeley image dataset (Martin, Fowlkes, Tal, & Malik, 2001). The mixed images were divided in 4 parts. The upper left part was used as training data and the lower right part as validation data. The two remaining regions were considered as test data. The RBF kernel parameter $\sigma^2 = 0.57$ was tuned using the proposed contrast function on validation data (19). Fig. 6 shows the model selection curve. The regularization parameter was fixed to $\nu = 1.00$. Fig. 7 shows the images together with the results. Note that the estimated demixing matrix computed for training data generalizes for unseen data. The use of validation points to tune the kernel parameter ensures reliability of the proposed measure of independence. The size of the matrix involved in the eigendecomposition was reduced from $84,000 \times 84,000$ to $56 \times 56$ using the incomplete Cholesky decomposition with $\eta = 1 \times 10^{-4}$. The mean Amari errors after 10 runs are 0.34 for the KCCA, 0.16 for the KGV, 0.30 for the FastICA algorithm and 0.155 for the proposed KRC.

8. Conclusions

A new kernel-based contrast function for ICA is proposed. This function called the kernel regularized correlation (KRC) is a measure of independence in the input space and a
measure of regularized correlation in the feature space induced by RBF kernels. This formulation is based on LS-SVMs and follows from a clear constrained optimization framework providing primal-dual insights and leading to a naturally regularized measure of independence. This approach also provides out-of-sample extensions for test points which is important for model selection ensuring statistical reliability of the measure for independence. Simulations with toy datasets show improved performance compared to existing ICA approaches. The computational complexity of the KRC compares favourably to similar kernel-based methods such as the KCCA and the KGV. Experiments with images and speech using the incomplete Cholesky decomposition to solve a reduced generalized eigenvalue problem also delivered good results.

Acknowledgments

This work was supported by grants and projects for the Research Council K.U. Leuven (GOA-Mefisto 666, GOA-Ambiorics, several Ph.D./Postdocs & fellow grants), the Flemish Government FWO: Ph.D./Postdocs grants, projects G.0240.99, G.0211.05, G.0407.02, G.0197.02, G.0080.01, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0226.06, G.0302.07, ICCoS, ANMMM; AWI-IWT;Ph.D. grants, GBOU (McKnow) Soft4s, the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/01/40), the EU(FP5-Quprodis, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (ISMC/IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). Johan Suykens is a professor at the K.U. Leuven, Belgium. The scientific responsibility is assumed by its authors.

References


