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

(email: carlos.alzate@esat.kuleuven.be)

January 21, 2008

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 (Melzer, Reiter, & Bischof, 2001; Akaho, 2001; Lai & Fyfe, 2000; Bach & Jordan, 2002). 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 (Lai & Fyfe, 2000; Bach & Jordan, 2002; Gretton, Herbrich, Smola, Bousquet, & Schölkopf, 2005).

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 & Jordan, 2002). This link is called the $\mathcal{F}$-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 $\mathcal{F}$-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 & 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 (Hotelling, 1936; Gittins, 1985). Let $\{x_i^{(1)}\}_{i=1}^N$ and $\{x_i^{(2)}\}_{i=1}^N$ 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 $C_{11} = (1/N)\sum_{k=1}^{N} x_k^{(1)} x_k^{(1)^T}, C_{22} = (1/N)\sum_{k=1}^{N} x_k^{(2)} x_k^{(2)^T}, C_{12} = (1/N)\sum_{k=1}^{N} x_k^{(1)} x_k^{(2)^T}$. This is typically formulated as the optimization problem:

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

$$\text{such that} \quad \begin{cases} w^{(1)^T} C_{11} w^{(1)} = 1 \\ w^{(2)^T} C_{22} w^{(2)} = 1 \end{cases} \tag{3}$$

which leads to the generalized eigenvalue problem:

$$\left[\begin{array}{c|c} 0 & C_{12} \\ \hline C_{21} & 0 \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right] = \rho \left[\begin{array}{c|c} C_{11} & 0 \\ \hline 0 & C_{22} \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right]$$

where ρ 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

$$\left[\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right] = (1+\rho) \left[\begin{array}{c|c} C_{11} & 0 \\ \hline 0 & C_{22} \end{array}\right] \left[\begin{array}{c} w^{(1)} \\ \hline w^{(2)} \end{array}\right]$$

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 & Fyfe, 2000; Bach & Jordan, 2002). The data is 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 $\varphi^{(1)}(\cdot) : \mathbb{R}^{d_1} \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(\cdot) : \mathbb{R}^{d_2} \to \mathbb{R}^{n_2}$ to high dimensional feature spaces of dimension n_1 and n_2 respectively. The centered feature matrices $\Phi_c^{(1)} \in \mathbb{R}^{N \times n_1}, \Phi_c^{(2)} \in \mathbb{R}^{N \times n_2}$ become

$$\begin{aligned} \Phi_c^{(1)} &= [\varphi^{(1)}(x_1^{(1)})^T - \hat{\mu}_{\varphi^{(1)}}^T; \ldots; \varphi^{(1)}(x_N^{(1)})^T - \hat{\mu}_{\varphi^{(1)}}^T] \\ \Phi_c^{(2)} &= [\varphi^{(2)}(x_1^{(2)})^T - \hat{\mu}_{\varphi^{(2)}}^T; \ldots; \varphi^{(2)}(x_N^{(2)})^T - \hat{\mu}_{\varphi^{(2)}}^T], \end{aligned}$$

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

$$\max_{w_K^{(1)}, w_K^{(2)}} \quad w_K^{(1)^T}\Phi_c^{(1)^T}\Phi_c^{(2)}w_K^{(2)} \tag{4}$$

$$\text{such that} \quad \begin{cases} w_K^{(1)^T}\Phi_c^{(1)^T}\Phi_c^{(1)}w_K^{(1)} = 1 \\ w_K^{(2)^T}\Phi_c^{(2)^T}\Phi_c^{(2)}w_K^{(2)} = 1, \end{cases}$$

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

$$w_K^{(1)} = \Phi_c^{(1)^T}\beta^{(1)}$$
$$w_K^{(2)} = \Phi_c^{(2)^T}\beta^{(2)}.$$

Applying the kernel trick results in the following optimization problem:

$$\max_{\beta^{(1)}, \beta^{(2)}} \quad \beta^{(1)^T}\Omega_c^{(1)}\Omega_c^{(2)}\beta^{(2)} \tag{5}$$

$$\text{such that} \quad \begin{cases} \beta^{(1)^T}\Omega_c^{(1)}\Omega_c^{(1)}\beta^{(1)} = 1 \\ \beta^{(2)^T}\Omega_c^{(2)}\Omega_c^{(2)}\beta^{(2)} = 1 \end{cases}$$

where $\Omega_c^{(l)}$ denotes the l-th centered kernel matrix $\Omega_c^{(l)} = P\Omega^{(l)}P$, $P = I_N - (1/N)1_N 1_N^T$ is the centering matrix, I_N is the $N \times N$ identity matrix, 1_N is a vector of N ones, $\Omega^{(l)}$ is the kernel matrix with ij-entry: $\Omega_{ij}^{(l)} = K^{(l)}(x_i^{(l)}, x_j^{(l)})$ and $K^{(l)}(x_i, x_j) = \varphi^{(l)}(x_i)^T\varphi^{(l)}(x_j)$ is a Mercer kernel, for $l = 1, 2$. The problem in (5) can be written as the generalized eigenvalue problem:

$$\left[\begin{array}{c|c} 0 & \Omega_c^{(1)}\Omega_c^{(2)} \\ \hline \Omega_c^{(2)}\Omega_c^{(1)} & 0 \end{array}\right] \left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right] = \rho \left[\begin{array}{c|c} \Omega_c^{(1)}\Omega_c^{(1)} & 0 \\ \hline 0 & \Omega_c^{(2)}\Omega_c^{(2)} \end{array}\right] \left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right]. \tag{6}$$

The non-zero 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 (Gretton et al., 2005; Bach & Jordan, 2002).

The regularized method proposed in (Bach & Jordan, 2002) consists of finding the smallest eigenvalue of the following problem problem:

$$\left[\begin{array}{c|c} (\Omega_c^{(1)} + \frac{Nc}{2} I_N)^2 & \Omega_c^{(1)}\Omega_c^{(2)} \\ \hline \Omega_c^{(2)}\Omega_c^{(1)} & (\Omega_c^{(2)} + \frac{Nc}{2} I_N)^2 \end{array}\right] \left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right] = \xi \left[\begin{array}{c|c} (\Omega_c^{(1)} + \frac{Nc}{2} I_N)^2 & 0 \\ \hline 0 & (\Omega_c^{(2)} + \frac{Nc}{2} I_N)^2 \end{array}\right] \left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right]$$

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:

$$\max_{w,v,e,r} \gamma e^T r - \nu_1 \frac{1}{2} e^T e - \nu_2 \frac{1}{2} r^T r - \frac{1}{2} w^T w - \frac{1}{2} v^T v \quad \text{such that} \begin{cases} e = \Phi_c^{(1)} w \\ r = \Phi_c^{(2)} v. \end{cases} \tag{7}$$

Proposition 1. *(Suykens et al., 2002) Given a positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $K(x, z) = \varphi(x)^T \varphi(z)$ and regularization constants $\gamma, \nu_1, \nu_2 \in \mathbb{R}^+$, the dual problem to (7) is given by the following eigenvalue problem:*

$$\left[\begin{array}{c|c} 0 & \Omega_c^{(2)} \\ \hline \Omega_c^{(1)} & 0 \end{array}\right] \left[\begin{array}{c} \xi \\ \hline \tau \end{array}\right] = \lambda \left[\begin{array}{c|c} \nu_1 \Omega_c^{(1)} + I_N & 0 \\ \hline 0 & \nu_2 \Omega_c^{(2)} + I_N \end{array}\right] \left[\begin{array}{c} \xi \\ \hline \tau \end{array}\right]. \tag{8}$$

Proof. Consider the Lagrangian of (7):

$$\mathcal{L}(w, v, e, r; \xi, \tau) = \gamma e^T r - \nu_1 \frac{1}{2} e^T e - \nu_2 \frac{1}{2} r^T r - \frac{1}{2} w^T w - \frac{1}{2} v^T v - \xi^T (e - \Phi_c^{(1)} w) - \tau^T (r - \Phi_c^{(2)} v)$$

where ξ, τ are Lagrange multiplier vectors. The Karush-Kuhn-Tucker (KKT) conditions for optimality are:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow \quad w = {\Phi_c^{(1)}}^T \xi \\ \frac{\partial \mathcal{L}}{\partial v} = 0 & \rightarrow \quad v = {\Phi_c^{(2)}}^T \tau \\ \frac{\partial \mathcal{L}}{\partial e} = 0 & \rightarrow \quad \gamma r = \nu_1 e + \xi \\ \frac{\partial \mathcal{L}}{\partial r} = 0 & \rightarrow \quad \gamma e = \nu_2 r + \tau \\ \frac{\partial \mathcal{L}}{\partial \xi} = 0 & \rightarrow \quad e = \Phi_c^{(1)} w \\ \frac{\partial \mathcal{L}}{\partial \beta} = 0 & \rightarrow \quad r = \Phi_c^{(2)} v. \end{cases}$$

Eliminating the primal variables w, v, r, e gives:

$$\gamma \Phi_c^{(2)} {\Phi_c^{(2)}}^T \tau = \nu_1 \Phi_c^{(1)} {\Phi_c^{(1)}}^T \xi + \xi$$
$$\gamma \Phi_c^{(1)} {\Phi_c^{(1)}}^T \xi = \nu_2 \Phi_c^{(2)} {\Phi_c^{(2)}}^T \tau + \tau$$

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

$$\begin{cases} \Omega_c^{(2)} \tau = & \lambda(\nu_1 \Omega_c^{(1)} \xi + \xi) \\ \Omega_c^{(1)} \xi = & \lambda(\nu_2 \Omega_c^{(2)} \tau + \tau) \end{cases}$$

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

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*

$$z_e^{(i)} = \Phi_c^{(1)} w = \Omega_c^{(1)} \xi^{(i)}, \tag{9}$$
$$z_r^{(i)} = \Phi_c^{(2)} v = \Omega_c^{(2)} \tau^{(i)}, \tag{10}$$

$i = 1, \ldots, 2N$.

Remark 4. *The optimal $\lambda^\star, \xi^\star, \tau^\star$ can be selected such that the correlation coefficient is maximized*

$$\max_{i\in\{1,\ldots,2N\}} \frac{z_e^{(i)^T} z_r^{(i)}}{||z_e^{(i)}||_2 ||z_r^{(i)}||_2}.$$

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:

$$\begin{aligned} \max_{w^{(l)}, e^{(l)}} \frac{\gamma}{2} \sum_{l=1}^{m} \sum_{k\neq l}^{m} e^{(l)^T} e^{(k)} - \frac{1}{2} \sum_{l=1}^{m} \nu_l e^{(l)^T} e^{(l)} - \frac{1}{2} \sum_{l=1}^{m} w^{(l)^T} w^{(l)} \\ \text{such that } e^{(l)} = \Phi_c^{(l)} w^{(l)}, l = 1, \ldots, m \end{aligned} \tag{11}$$

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

$$e_i^{(l)} = w^{(l)^T} (\varphi^{(l)}(x_i^{(l)}) - \hat{\mu}_{\varphi^{(l)}}), i = 1, \ldots, N, l = 1, \ldots, m,$$

the l-th centered feature matrix $\Phi_c^{(l)} \in \mathbb{R}^{N\times n_h}$ is

$$\Phi_c^{(l)} = \begin{bmatrix} \varphi^{(l)}(x_1^{(l)})^T - \hat{\mu}_{\varphi^{(l)}}^T \\ \varphi^{(l)}(x_2^{(l)})^T - \hat{\mu}_{\varphi^{(l)}}^T \\ \vdots \\ \varphi^{(l)}(x_N^{(l)})^T - \hat{\mu}_{\varphi^{(l)}}^T \end{bmatrix},$$

γ, ν_l are regularization parameters, $\varphi^{(l)}(\cdot)$ are the mappings to high dimensional spaces and

$$\hat{\mu}_{\varphi^{(l)}} = (1/N) \sum_{i=1}^{N} \varphi^{(l)}(x_i^{(l)}), l = 1, \ldots, m.$$

Lemma 1. *Given m positive definite kernel functions $K^{(l)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $K^{(l)}(x, z) = \varphi^{(l)}(x)^T \varphi^{(l)}(z)$ 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:*

$$\mathcal{K}\alpha = \lambda \mathcal{R}\alpha \tag{12}$$

with $\mathcal{K}, \mathcal{R}$ and α defined as follows:

$$\mathcal{K} = \left[\begin{array}{c|c|c|c} 0 & \Omega_c^{(2)} & \dots & \Omega_c^{(m)} \\ \hline \Omega_c^{(1)} & 0 & \dots & \Omega_c^{(m)} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \Omega_c^{(1)} & \Omega_c^{(2)} & \dots & 0 \end{array}\right], \mathcal{R} = \left[\begin{array}{c|c|c|c} R^{(1)} & 0 & \dots & 0 \\ \hline 0 & R^{(2)} & \dots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \dots & R^{(m)} \end{array}\right], \alpha = \left[\begin{array}{c} \alpha^{(1)} \\ \hline \alpha^{(2)} \\ \hline \vdots \\ \hline \alpha^{(m)} \end{array}\right],$$

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

Proof. The Lagrangian for this constrained optimization problem is:

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

with conditions for optimality given by:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 & \rightarrow w^{(l)} = \Phi_c^{(l)^T} \alpha^{(l)} \\ \frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 & \rightarrow \gamma \sum_{k \neq l} e^{(k)} = \nu_l e^{(l)} + \alpha^{(l)} \\ \frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 & \rightarrow e^{(l)} = \Phi_c^{(l)} w^{(l)}, \end{cases}$$

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

$$\gamma \sum_{k \neq l}^{m} \Phi_c^{(k)^T} \Phi_c^{(k)} \alpha^{(k)} = (\nu_l \Phi_c^{(l)^T} \Phi_c^{(l)} + I_N) \alpha^{(l)}, l = 1, \dots, m$$

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

$$\sum_{k \neq l}^{m} \Omega_c^{(k)} \alpha^{(k)} = \lambda (\nu_l \Omega_c^{(l)} + I_N) \alpha^{(l)}, l = 1, \dots, m.$$

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

Remark 5. *Note that the regularized matrix $\mathcal{R}$ differs from the schemes proposed in (Bach & Jordan, 2002; Gretton et al., 2005; Yamanishi, Vert, Nakaya, & 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:*

$$z_i^{(l)} = \Phi_c^{(l)} w^{(l)} = \Omega^{(l)} \alpha_i^{(l)},$$

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* $\lambda^\star, \alpha^\star$ *can be selected such that the sum of the correlation coefficients is maximized:*

$$\max_{i \in \{1,\ldots,Nm\}} \sum_{l=1}^{m} \sum_{k \neq l}^{m} \frac{\left(z_i^{(l)}\right)^T z_i^{(k)}}{||z_i^{(l)}||_2 ||z_i^{(k)}||_2}.$$

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

$$z_{test}^{(l)} = \Omega_{test}^{(l)} \alpha^{(l)\star}, \tag{13}$$

where $\Omega_{test}^{(l)} \in \mathbb{R}^{N_{test} \times N}$ *is the kernel matrix evaluated using the test points with* ji*-entry* $\Omega_{test,ji}^{(l)} = K^{(l)}(\tilde{x}_j^{(l)}, x_i^{(l)}), j = 1, \ldots, N_{test}, i = 1, \ldots, N$.

4 ICA and Kernel CCA

4.1 The $\mathcal{F}$ 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 (Hyvärinen et al., 2001; Hyvärinen & Pajunen, 1999; Comon, 1994; Cardoso, 1999; Amari, Cichocki, & Yang, 1996) 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 & Jordan, 2002) with the $\mathcal{F}$-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:

$$\rho_{\mathcal{F}}\big(f^{(1)}(x^{(1)}), f^{(2)}(x^{(2)})\big) = \max_{f^{(1)}\in\mathcal{F}^{(1)}, f^{(2)}\in\mathcal{F}^{(2)}} \text{corr}\big(f^{(1)}(x^{(1)}), f^{(2)}(x^{(2)})\big)$$

where $f^{(l)}(x^{(l)}) = \langle \varphi^{(l)}(x^{(l)}), f^{(l)} \rangle, l = 1, 2$ 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 $\mathcal{F}$-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):

$$(\mathcal{K} + \mathcal{R})\alpha = \zeta \mathcal{R}\alpha, \tag{14}$$

where $\zeta = 1 + \lambda$. The largest eigenvalue of (12) is related to the smallest eigenvalue ζ_{min} of (14) and ζ_{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:

$$\hat{\rho}(\Omega^{(1)}, \ldots, \Omega^{(m)}) = 1 - \zeta_{\text{min}}. \tag{15}$$

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

$$\hat{\rho}(\Omega^{(1)}_{\text{test}}, \ldots, \Omega^{(m)}_{\text{test}}) = \sum_{l=1}^{m} \sum_{k\neq l}^{m} \frac{z^{(l)^T}_{\text{test}} z^{(k)}_{\text{test}}}{||z^{(l)}_{\text{test}}||_2 ||z^{(k)}_{\text{test}}||_2}. \tag{16}$$

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 Ω can be factorized as $\Omega = GG^T$ where $G \in \mathbb{R}^{N \times N}$ is lower triangular. If the eigenvalues of Ω 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 $M \ll N$. This method is known as the incomplete Cholesky decomposition with symmetric pivoting and was used in (Bach & Jordan, 2002; Gretton et al., 2005) to reduce the computational burden of the corresponding eigenvalue problem.

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

$$\tilde{\mathcal{K}}\alpha = \zeta \tilde{\mathcal{R}}\alpha \tag{17}$$

where

$$\tilde{\mathcal{K}} = \left[\begin{array}{c|c|c|c} U^{(1)}\Lambda_\nu^{(1)}U^{(1)^T} & U^{(2)}\Lambda^{(2)}U^{(2)^T} & \dots & U^{(m)}\Lambda^{(m)}U^{(m)^T} \\ \hline U^{(1)}\Lambda^{(1)}U^{(1)^T} & U^{(1)}\Lambda_\nu^{(2)}U^{(2)^T} & \dots & U^{(m)}\Lambda^{(m)}U^{(m)^T} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline U^{(1)}\Lambda^{(1)}U^{(1)^T} & U^{(2)}\Lambda^{(2)}U^{(2)^T} & \dots & U^{(m)}\Lambda_\nu^{(m)}U^{(m)^T} \end{array}\right]$$

,

$$\tilde{\mathcal{R}} = \left[\begin{array}{c|c|c|c} U^{(1)}\Lambda_\nu^{(1)}U^{(1)^T} & 0 & \dots & 0 \\ \hline 0 & U^{(1)}\Lambda_\nu^{(2)}U^{(2)^T} & \dots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \dots & U^{(m)}\Lambda_\nu^{(m)}U^{(m)^T} \end{array}\right],$$

$\Lambda_\nu^{(l)}$ is a diagonal matrix with nn-entry

$$\Lambda_{\nu,nn}^{(l)} = 1 + \nu_l \Lambda_{nn}^{(l)}, n = 1, \dots, M_l.$$

Substituting $\tilde{\alpha}^{(l)} = U^{(l)^T}\alpha^{(l)}, i = l, \dots, m$ and premultiplying both sides of (17) by blkdiag$(U^{(1)^T}, \dots, U^{(m)^T})$

leads to the following reduced problem:

$$\tilde{\mathcal{K}}_{\mathcal{R}}\tilde{\alpha} = \zeta\tilde{\alpha} \tag{18}$$

where

$$\tilde{\mathcal{K}}_{\mathcal{R}} = \left[\begin{array}{c|c|c|c} I_{M_1} & T_U^{(1)}U_\Lambda^{(2)} & \dots & T_U^{(1)}U_\Lambda^{(m)} \\ \hline T_U^{(2)}U_\Lambda^{(1)} & I_{M_2} & \dots & T_U^{(1)}U_\Lambda^{(m)} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline T_U^{(m)}U_\Lambda^{(1)} & T_U^{(m)}U_\Lambda^{(2)} & \dots & I_{M_m} \end{array}\right], \tilde{\alpha} = \left[\begin{array}{c} \tilde{\alpha}^{(1)} \\ \hline \tilde{\alpha}^{(2)} \\ \hline \vdots \\ \hline \tilde{\alpha}^{(m)} \end{array}\right],$$

$T_U^{(l)} = T^{(l)}U^{(l)^T}$, $T^{(l)}$ is a diagonal matrix $T_{nn}^{(l)} = 1/\Lambda_{nn}^{(l)}$, $U_\Lambda^{(l)} = U^{(l)}\Lambda^{(l)}, l = 1, \dots, m, n = 1, \dots, M_l$. Note that the size of $\tilde{\mathcal{K}}_{\mathcal{R}}$ is $(\sum_{l=1}^m M_l) \times (\sum_{l=1}^m M_l)$ which is much smaller than the original $mN \times mN$ problem in (14).

5.2 The Stiefel Manifold

The Stiefel manifold is described by the set of all $m \times q$ orthogonal matrices W (i.e. $W^TW = I_q$). The geometry of this special manifold was studied in (Edelman, Arias, & Smith, 1999) together with gradient descent algorithms for optimization on the manifold. The tangent space corresponds to the space of all matrices A such that W^TA is a skew-symmetric. A geodesic is defined as the shortest path between two points in the manifold and is determined as $G_W(t) = W\exp(tW^TA)$ where W is the starting point, A is the direction and t is the displacement. The direction A can be determined as $A = D - WD^TW$ where D is the derivative of the contrast function with respect to W. Gradient descent can then be used along the geodesic in the direction A to obtain orthogonal matrices W such that the contrast function is minimized. The kernel-based ICA approaches described in (Bach & Jordan, 2002; Gretton et al., 2005) apply gradient descent methods with golden search on the Stiefel manifold in order to obtain demixing matrices leading to independent components. We also use this approach to minimize the KRC. The computation of the KRC and the corresponding minimization on the Stiefel manifold are shown in the following algorithms:

Algorithm 1 Kernel Regularized Correlation (KRC)

Input: Mixture set of m sources $\{x_i^{(l)}\}_{i=1}^N, l = 1, \dots, m$, RBF kernel parameter σ^2, regularization parameter ν, Cholesky error threshold η.
Output: $\hat{\rho}(\Omega_c^{(1)}, \dots, \Omega_c^{(m)})$.

1: Compute the Cholesky factor $G_c^{(l)} \in \mathbb{R}^{N \times M_l}$ such that $||\Omega_c^{(l)} - G_c^{(l)}G_c^{(l)^T}||_2 \leq \eta, l = 1, \dots, m$.
2: Compute the SVD $G_c^{(l)} = U^{(l)}\Lambda^{(l)}U^{(l)^T}, l = 1, \dots, m$ and build the matrix $\tilde{\mathcal{K}}_{\mathcal{R}}$ (18).
3: Compute the smallest eigenvalue ζ_{min} of the problem $\tilde{\mathcal{K}}_{\mathcal{R}}\tilde{\alpha} = \zeta\tilde{\alpha}$.
4: **Return** $1 - \zeta_{\text{min}}$.

The computational complexity of the KRC is $\mathcal{O}((mM)^2N)$, where M is the maximal rank considered by the

Algorithm 2 Gradient descent on the Stiefel manifold

Input: Whitened mixture set of m sources $\{x_i^{(l)}\}_{i=1}^N, l = 1, \ldots, m$, RBF kernel parameter σ^2, regularization parameter ν, Cholesky error threshold η, initial demixing matrix W_0.
Output: Demixing matrix W.

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

incomplete Cholesky decompositions . However, as noted by (Bach & Jordan, 2002), the bottleneck of the computation is the incomplete Cholesky decomposition. Therefore, the empirical computational complexity of the KRC is $\mathcal{O}(mM^2N)$.

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 $\nu_l = \nu$ and fixed to some value. Given validation data $\{\check{x}_h^{(l)}\}, h = 1, \ldots, N_{\text{val}}, l = 1, \ldots, m$, the following model selection criterion is proposed:

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

where $z_{\text{val}}^{(l)} = \Omega_{\text{val}}^{(l)} \alpha^{(l)\star}$, $\Omega_{\text{val}}^{(l)}$ is the kernel matrix evaluated using the validation points with hi-entry $\Omega_{\text{val},hi}^{(l)} = K^{(l)}(\check{x}_h^{(l)}, x_i^{(l)}), h = 1, \ldots, N_{\text{val}}, i = 1, \ldots, N$ 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 σ^2 increases (Bach & Jordan, 2002). Therefore, $||\Omega_{\text{val}}||_F$ in (19) fixes the scaling

introduced by σ^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 seconds for $N = 28,000$, $m = 3$, $\eta = 10^{-4}$ on a Pentium 4, 2.8 GHz, 1GB RAM using Lanczos algorithm in Matlab. In the case of test and validation data, the computation time increased to 23.56 seconds 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 3 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. Figure 1 shows the contour plots of the Amari divergence, the KCCA and KGV criteria described in (Bach & 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 $\sigma = 5, \nu = 2$. 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 σ^2, ν on the optimization surface can be seen in Figure 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 6 different source distributions. The sample size was $N = 900$. The distributions include a multimodal asymmetric mixture of two Gaussians, a Student's t with 3 degrees of freedom, an 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. Table 1 and Figure 3 show the results. In all cases the KRC obtained the smallest Amari error. Mean computation times for the KCCA, the KGV and the KRC are shown in Figure 4.

7.3 Speech signals

We used three fragments of 6.25 seconds 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. Figure 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 140×200 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). Figure 6 shows the model selection curve. The regularization parameter was fixed to $\nu = 1.00$. Figure 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 56×56 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 favorably 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.

Acknowledgment

This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD / Postdocs & fellow grants), the Flemish Government FWO: PhD / 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:PhD 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

Akaho, S. (2001). A kernel method for canonical correlation analysis. In *Proceedings of the international meeting of the psychometric society (IMPS2001)*. Tokyo: Springer-Verlag.

Alzate, C., & Suykens, J. A. K. (2007). ICA through an LS-SVM based kernel CCA measure for independence. In *Proceedings of the 2007 international joint conference on neural networks (IJCNN'07)* (pp. 2920–2925).

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 757–763). Cambridge, MA: MIT Press.

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, *3*, 1–48.

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11*(1), 157–192.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.

Edelman, A., Arias, T. A., & Smith, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, *20*(2), 303–353.

Gittins, R. (1985). *Canonical analysis - a review with applications in ecology*. Berlin: Springer Verlag.

Gretton, A., Herbrich, R., Smola, A. J., Bousquet, O., & Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, *6*, 2075–2129.

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

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: John Wiley and Sons.

Hyvärinen, A., & Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492.

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: existence and uniqueness results. *Neural Networks*, *12*(3), 429–439.

Jutten, C., & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, *24*, 1–10.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, *58*(3), 433–451.

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

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its

application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th int'l conf. computer vision* (Vol. 2, pp. 416–423).

Melzer, T., Reiter, M., & Bischof, H. (2001). Nonlinear feature extraction using generalized canonical correlation analysis. In *Proceedings of the international conference on artificial neural networks (ICANN 2001)*. New York: Springer-Verlag.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Yamanishi, Y., Vert, J., Nakaya, A., & Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, *19*, 323i–330i.

Captions

Figure 1. Contour plots of the optimization surfaces for several kernel-based contrast functions for ICA. Three sources were mixed using a product of known Jacobi rotations. The optimal set of angles is shown as a black dot. *Top left:* Amari divergence. *Top right:* Proposed KRC using $\sigma^2 = 5, \nu = 2$. *Bottom left:* KCCA criterion (Bach & Jordan, 2002) using $\sigma^2 = 1, \kappa = 2 \times 10^{-2}$. *Bottom right:* KGV criterion (Bach & Jordan, 2002) using the same set of parameters as KCCA. Note that all contrast functions are smooth and the black dot corresponds to a local minimum.

Table 1. Mean Amari error (multiplied by 100) of the source distributions experiment after 20 runs. The KCCA and KGV parameters were set to the values recommended by the authors ($\sigma^2 = 1, \kappa = 2 \times 10^{-2}$) in (Bach & 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.

Figure 2. Effect of the KRC parameters on the optimization surface. *Top row:* $\sigma^2 = 0.1$. *Middle row:* $\sigma^2 = 1$. *Bottom row:* $\sigma^2 = 10$. *Left column:* $\nu = 0.1$. *Middle column:* $\nu = 1$. *Right column:* $\nu = 10$. The black dot indicates the optimal set of angles. Note that for most parameter values the black dot corresponds to a local minimum in the KRC.

Figure 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:* 3 sources. *Top right:* 4 sources. *Bottom left:* 5 sources. *Bottom right:* 6 sources. The KRC performs better with respect to the KCCA and KGV.

Figure 4. Mean computation times of the source distributions experiment. The KRC compares favorably with respect the KCCA contrast.

Figure 5. Speech demixing results after 10 runs with different segments. *Left:* Amari errors. *Right:* Computation times. The KRC outperforms the KCCA and KGV in terms of Amari error with comparable computation times with respect to the KGV.

Figure 6. Image demixing experiment. Model selection curve for a fixed regularization parameter $\nu = 1.00$. The plot shows the proposed contrast function evaluated on validation data. The obtained tuned RBF kernel parameter is $\sigma^2 = 0.57$.

Figure 7. Image demixing experiment. *Top row:* Original independent components. *Middle row:* Observed mixtures. The black squares show the regions used for training. Validation regions are shown as white squares. All remaining pixels are considered as test points. *Bottom row:* Estimated independent components using the KRC with tuned RBF kernel width $\sigma^2 = 0.57$. The regularization parameter was fixed to $\nu = 1.00$. Note that the estimated components are visually appealing with an Amari error of 0.15.

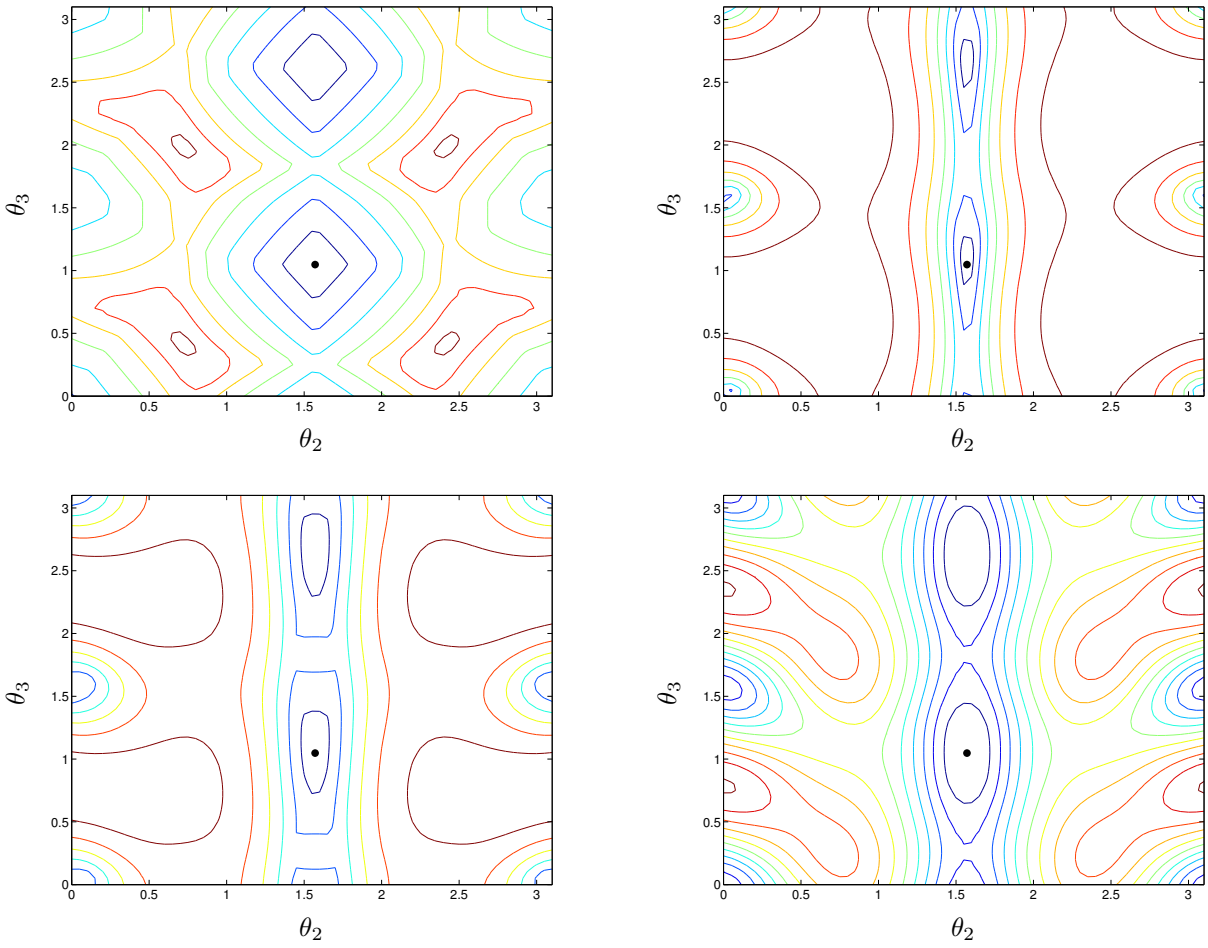


Figure 1:

# sources	KCCA	KGV	KRC	FastICA	Jade
2	4.02	2.77	**2.75**	12.92	4.95
3	7.43	2.89	**2.36**	9.37	6.98
4	6.12	3.55	**2.19**	9.88	8.30
5	6.39	5.75	**4.48**	11.03	8.88
6	8.42	5.57	**4.06**	9.23	7.49

Table 1:

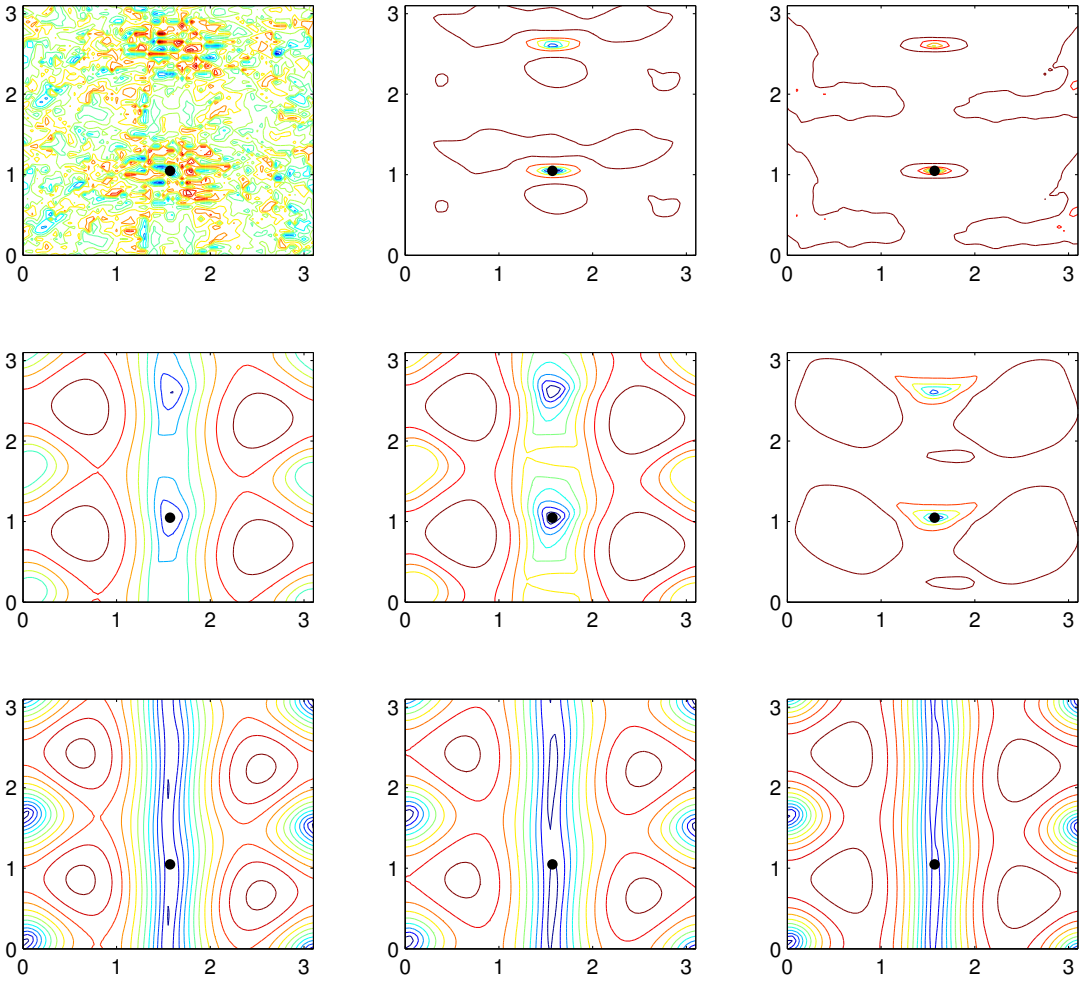


Figure 2:

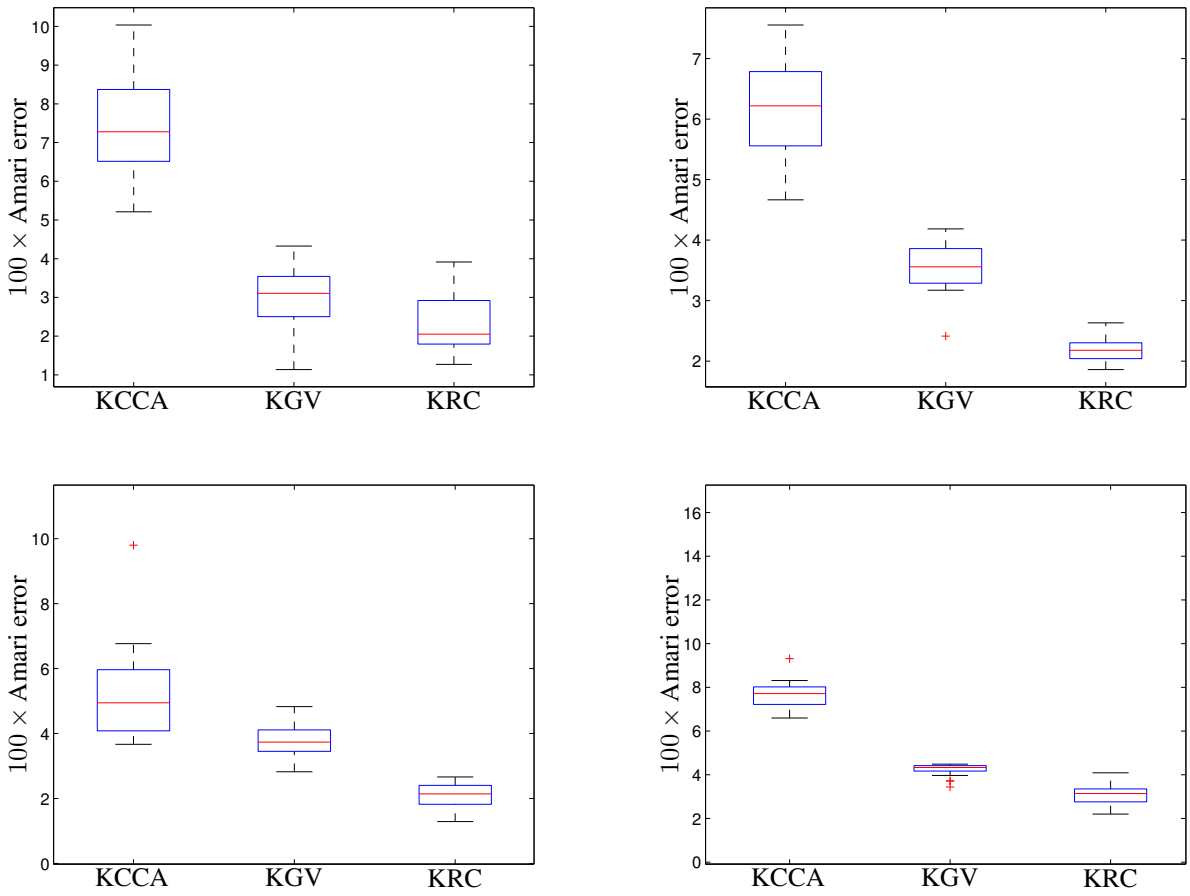

100 × Amari error
KCCA
KGV
KRC
100 × Amari error
KCCA
KGV
KRC
100 × Amari error
KCCA
KGV
KRC
100 × Amari error
KCCA
KGV
KRC

Figure 3:

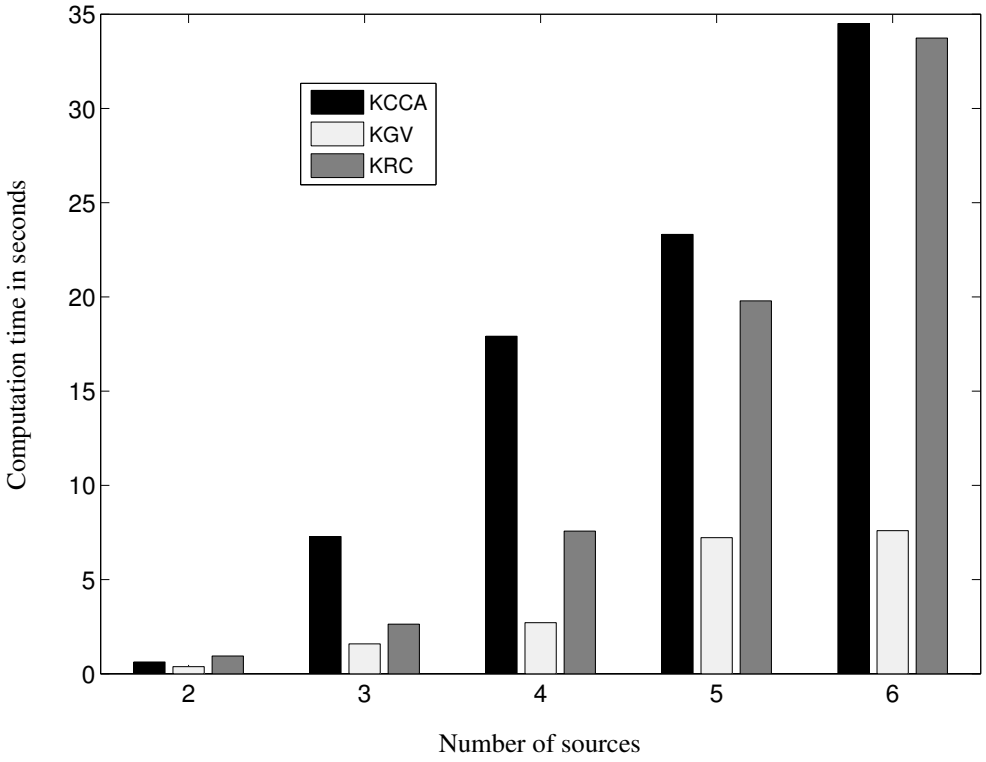


Figure 4:

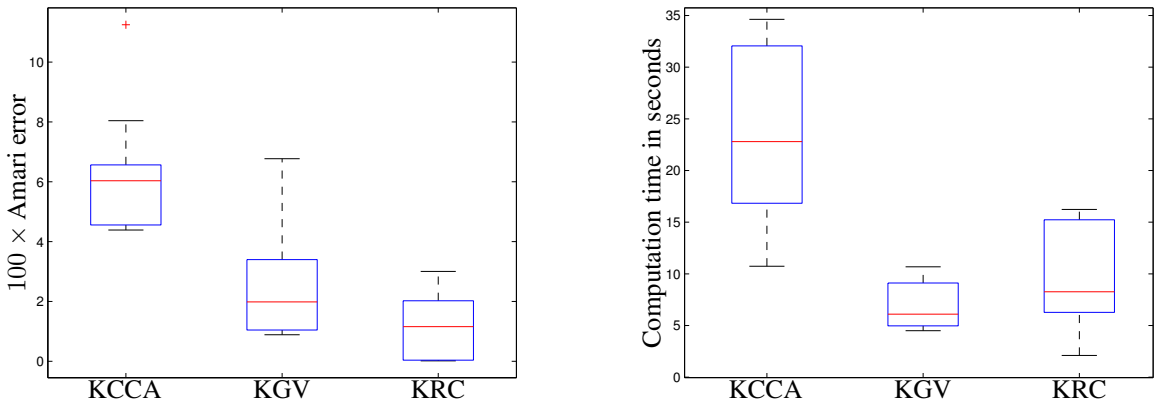


Figure 5:

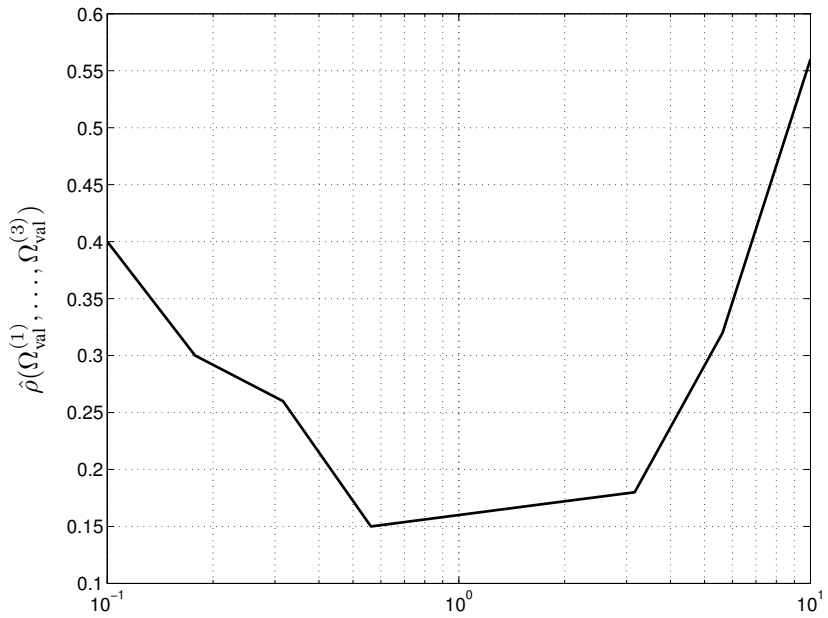


Figure 6:

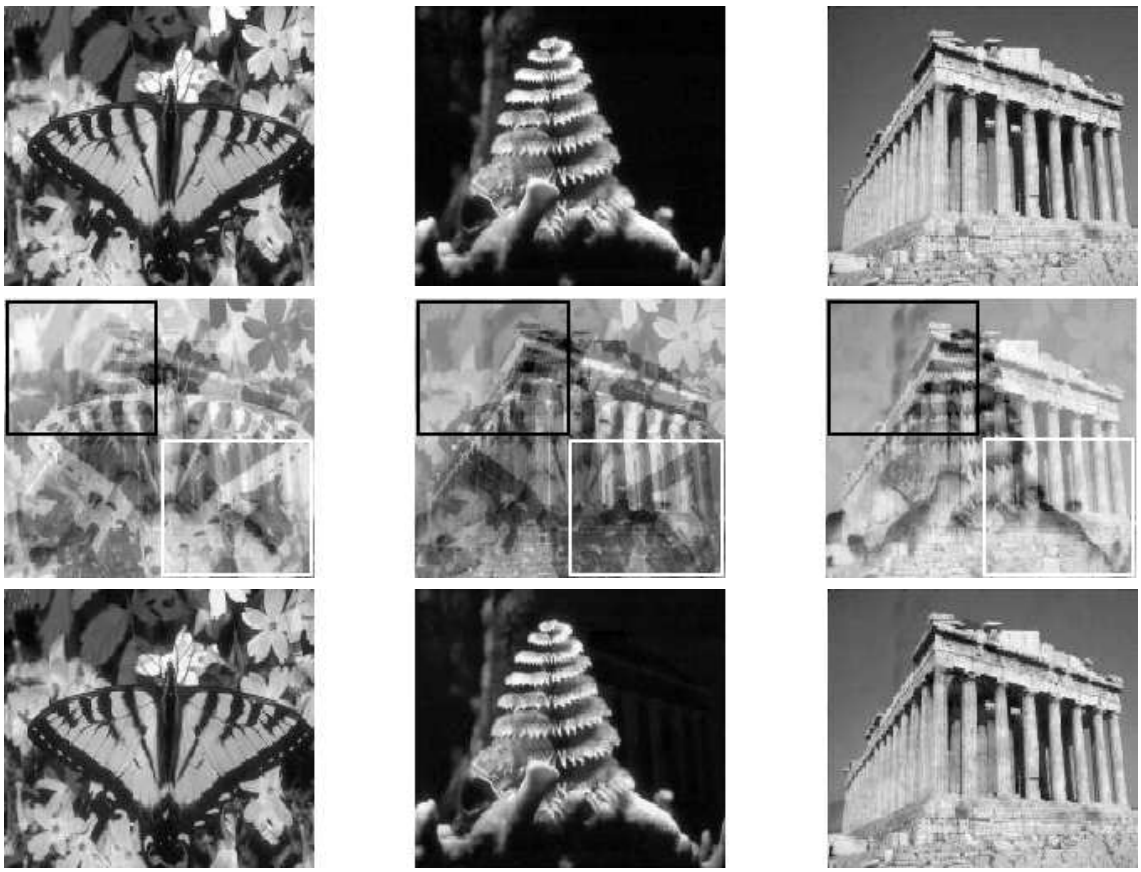


Figure 7: