ICA through an LS-SVM based Kernel CCA Measure for Independence

Carlos Alzate, Johan A. K. Suykens

***Abstract*— A new measure for independence based on canonical correlation in high dimensional feature spaces is presented. This measure can be used as a contrast function for independent component analysis (ICA). The formulation fits in the least squares support vector machines (LS-SVM) framework as a primal-dual interpretation of kernel canonical correlation analysis (CCA) in the context of constrained optimization problems. Regularization is incorporated naturally in the primal formulation leading to a dual generalized eigenvalue problem. Due to the primal-dual nature of the proposed approach, the measure for independence can be calculated for out-of-sample data points which is important for parameter selection ensuring statistical reliability of the estimated measure. Simulations results with small toy datasets performing model selection on a validation set showed good performance avoiding overfitting. Experiments with image demixing using approximated kernel matrices via incomplete Cholesky decomposition showed good results together with a reduced computational cost.**

I. Introduction

Independent component analysis (ICA) corresponds to a family of methods for finding underlying latent factors present in the data. The observed variables are represented as linear mixtures of the underlying components which are assumed to be mutually independent [3], [4], [10], [11]. These methods have been sucessfully applied in exploratory data analysis, signal processing, feature extraction, noise reduction and blind source separation. ICA is related to principal component analysis (PCA) [12], factor analysis and projection pursuit [6]. Several ICA approaches have been proposed from different points of view such as neural networks [13], tensor-based algorithms [4] and contrast functions [10].

Canonical correlation analysis (CCA) is a method useful for finding linear relations between two sets of variables [9], [7]. The objective is to find projected variables with maximal correlation. Kernel CCA as a nonlinear extension to CCA first maps the data into high dimensional features spaces via nonlinear feature maps induced by kernels and then performs linear CCA on the mapped data. A link between canonical correlation in the feature space induced by an RBF kernel and independence in the original input space was presented in [2]. If the projected variables are uncorrelated in this feature space then the corresponding input variables are pairwise independent. In this way, independent components can be obtained through minimization of a contrast function based on canonical correlation of the mapped data. Some kernel-based contrast functions that can be used for ICA include the KCCA, the kernel generalized variance (KGV) criteria [2], the constrained covariance (COCO) and the kernel mutual information (KMI) [8].

In this paper, we derive a new measure for independence based on an LS-SVM approach to kernel CCA. This formulation arises from a clear primal optimization problem where regularization is incorporated naturally leading to a dual generalized eigenvalue problem. This measure can also be extended to out-of-sample data points which become important for parameter selection.

This paper is organized as follows: Section II contains a review of linear and kernel CCA together with an LS-SVM formulation to kernel CCA for two sets of variables. In Section III, we derive an extension for LS-SVM based kernel CCA for m sets of variables. Section IV contains the proposed measure for independence based on canonical correlation. Section V describes an approximated eigenvalue problem based on the incomplete Cholesky decomposition in order to reduce the computational complexity. Section VI contains a model selection method. In Section VII we report the empirical results and in Section VIII we give conclusions.

II. LS-SVMs and Kernel CCA

A. Linear CCA

The problem of CCA consists of finding linear relations between two sets of variables [9], [7]. Given two random vectors with zero mean $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$, the objective is to determine vectors w and v such that the projected variables $w^T x$ and $v^T y$ are maximally correlated:

$$\max_{w,v} \rho = \frac{w^T C_{xy} v}{\sqrt{w^T C_{xx} w}\sqrt{v^T C_{yy} v}} \tag{1}$$

where $C_{xx} = E[xx^T], C_{yy} = E[yy^T], C_{xy} = E[xy^T]$. This is typically formulated as the optimization problem:

$$\max_{w,v} \quad w^T C_{xy} v \tag{2}$$

$$\text{such that} \quad \begin{cases} w^T C_{xx} w = 1 \\ v^T C_{yy} v = 1 \end{cases} \tag{3}$$

which leads to the generalized eigenvalue problem:

$$\left[\begin{array}{c|c} 0 & C_{xy} \\ \hline C_{yx} & 0 \end{array}\right] \left[\begin{array}{c} w \\ \hline v \end{array}\right] = \rho \left[\begin{array}{c|c} C_{xx} & 0 \\ \hline 0 & C_{yy} \end{array}\right] \left[\begin{array}{c} w \\ \hline v \end{array}\right]$$

where ρ is the correlation coefficient. This problem has $d_x + d_y$ eigenvalues $\{\rho_1, -\rho_1, \ldots, \rho_p, -\rho_p, 0, \ldots, 0\}$, where $p = \min(d_x, d_y)$.

The authors are with the Department of Electrical Engineering ESAT-SCD-SISTA, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium (email: carlos.alzate@esat.kuleuven.be; johan.suykens@esat.kuleuven.be).

B. Kernel CCA

Kernel CCA corresponds to a nonlinear extension of CCA using kernels. The data is first embedded into a high dimensional feature space induced by a kernel and then linear CCA is applied [14], [2]. This way, nonlinear relations between variables can be found. Let $\{x_k^{(1)}\}_{k=1}^N$ and $\{x_k^{(2)}\}_{k=1}^N$ denote N observations of $x^{(1)}$ and $x^{(2)}$ respectively and 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 spaces of dimension n_1 and n_2 respectively. The centered feature matrices become

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

the projection vectors w_K and v_K lie in the span of the mapped data

$$\begin{aligned} w_K &= \Phi_c^{(1)}\beta^{(1)} \\ v_K &= \Phi_c^{(2)}\beta^{(2)}\end{aligned}$$

resulting in the following optimization problem:

$$\begin{aligned}\max_{\beta^{(1)},\beta^{(2)}} \quad & \beta^{(1)^T}\Omega_c^{(1)}\Omega_c^{(2)}\beta^{(2)} \\ \text{such that} \quad & \begin{cases}\beta^{(1)^T}\Omega_c^{(1)}\Omega_c^{(1)}\beta^{(1)} = 1 \\ \beta^{(2)^T}\Omega_c^{(2)}\Omega_c^{(2)}\beta^{(2)} = 1\end{cases}\end{aligned} \tag{4}$$

where $\Omega_c^{(i)}$ denotes the i-th centered kernel matrix $\Omega_c^{(i)} = P\Omega^{(i)}P$, P is the centering matrix $P = I_N - (1/N)1_N 1_N^T$, I_N is the $N \times N$ identity matrix, 1_N is a vector of N ones, $\Omega^{(i)}$ is the kernel matrix with jk-entry: $\Omega_{jk}^{(i)} = K^{(i)}(x_k^{(i)}, x_j^{(i)})$, $K^{(i)}(\cdot,\cdot)$ is a Mercer kernel and $K^{(i)}(x_k, x_j) = \varphi^{(i)}(x_k)^T\varphi^{(i)}(x_j)$ is the kernel trick, for $i = 1, 2$. The problem in (4) can be written as the generalized eigenvalue problem:

$$\begin{aligned}&\left[\begin{array}{c|c} 0 & \Omega_c^{(1)}\Omega_c^{(2)} \\ \hline \Omega_c^{(2)}\Omega_c^{(1)} & 0 \end{array}\right]\left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right] = \\ &\rho\left[\begin{array}{c|c} \Omega_c^{(1)}\Omega_c^{(1)} & 0 \\ \hline 0 & \Omega_c^{(2)}\Omega_c^{(2)} \end{array}\right]\left[\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array}\right]\end{aligned} \tag{5}$$

The non-zero solutions to this generalized eigenvalue problem are $\rho = \pm 1$ [8], [2]. Hence, the eigenvalues of the kernel CCA problem do not characterize canonical correlation. This is due to the fact that the problem in (5) is ill-conditioned because the kernel matrices can be singular or near-singular. Therefore, regularization is needed in order to obtain useful estimators of the canonical correlation in high dimensional feature spaces. Several ad-hoc regularization schemes have been proposed based mainly in ridge-type regularization.

The method proposed in [2] results in the following regularized form:

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

where c is small positive regularization constant.

C. LS-SVM Approach to Kernel CCA

An LS-SVM formulation to kernel CCA was introduced in [15]. This formulation emphasizes primal-dual aspects in the context of constrained optimization problems and has the following primal form:

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

The primal objective function takes the numerator of the correlation coefficient with a positive sign and the denominator with a negative sign. Besides, it can be considered as a generalization of the problem $\min_{w,v} ||e - r||_2^2$ which is known in CCA [7].

The Lagrangian becomes:

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

where α, τ are Lagrange multipliers. The conditions for optimality are:

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

Eliminating w, v, r, e, defining $\lambda = 1/\gamma$ and applying the kernel trick, the following set of equations can be obtained:

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

which can be written as the following generalized eigenvalue problem:

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

Note that the regularization is incorporated in the primal optimization problem in natural way.

III. Multivariate Kernel CCA

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

$$\max_{w^{(i)},e^{(i)}} \frac{\gamma}{2}\sum_{i=1}^m\sum_{j\neq i}^m e^{(i)^T}e^{(j)} - \frac{1}{2}\sum_{i=1}^m \nu_i e^{(i)^T}e^{(i)} - \frac{1}{2}\sum_{i=1}^m w^{(i)^T}w^{(i)} \tag{6}$$
$$\text{such that } e^{(i)} = \Phi_c^{(i)} w^{(i)}, i = 1, \ldots, m$$

where $e^{(i)}$ is the compact form of the i-th projected variables $e_k^{(i)} = w^{(i)^T}(\varphi^{(i)}(x_k^{(i)}) - \hat{\mu}_{\varphi^{(i)}}), k = 1, \ldots, N, i = 1, \ldots, m,$ the i-th centered feature matrix $\Phi_c^{(i)} \in \mathbb{R}^{N \times n_h}$ is

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

γ, ν_i are regularization parameters, $\varphi^{(i)}(\cdot)$ are the mappings to high dimensional spaces and $\hat{\mu}_{\varphi^{(i)}} = (1/N)\sum_{k=1}^{N} \varphi^{(i)}(x_k^{(i)})$ for $i = 1, \ldots, m$.

The Lagrangian for this constrained optimization problem is:

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

with conditions for optimality given by:

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

for $i = 1, \ldots, m$. Eliminating $w^{(i)}, e^{(i)}$ and defining $\lambda = 1/\gamma$ leads to the following set of equations:

$$\sum_{j\neq i}^{m} \Omega_c^{(j)} \alpha^{(j)} = \lambda(I_N + \nu_i \Omega_c^{(i)})\alpha^{(i)}, i = 1, \ldots, m.$$

which can be written as the following generalized eigenvalue problem:

$$\mathcal{K}\alpha = \lambda \mathcal{R}\alpha \tag{7}$$

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

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

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

$$\alpha = \left[\begin{array}{c} \alpha^{(1)} \\ \hline \alpha^{(2)} \\ \hline \vdots \\ \hline \alpha^{(m)} \end{array}\right],$$

$R^{(i)} = (I_N + \nu_i \Omega^{(i)}), i = 1, \ldots, m$. Note that the regularized matrix $\mathcal{R}$ differs from the schemes proposed in [2], [8]. 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. Given new test data $\{\tilde{x}_l^{(i)}\}, l = 1, \ldots, N_{\text{test}}, i = 1, \ldots, m$ the projections onto the eigenvector solution can be calculated as:

$$e_{\text{test}} = \Omega_{\text{test}} \alpha^*, \tag{8}$$

where

$$\Omega_{\text{test}} = \left[\begin{array}{c|c|c|c} \Omega_{\text{test}}^{(1)} & 0 & \ldots & 0 \\ \hline 0 & \Omega_{\text{test}}^{(2)} & \ldots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \ldots & \Omega_{\text{test}}^{(m)} \end{array}\right],$$

$$e_{\text{test}} = \left[\begin{array}{c} e_{\text{test}}^{(1)} \\ \hline e_{\text{test}}^{(2)} \\ \hline \vdots \\ \hline e_{\text{test}}^{(m)} \end{array}\right],$$

$\Omega_{\text{test}}^{(i)}, i = 1, \ldots, m$ is the kernel matrix evaluated using the test points with lk-entry $\Omega_{\text{test},lk}^{(i)} = K_i(\tilde{x}_l^{(i)}, x_k^{(i)})$ and α^* is the eigenvector corresponding to the largest eigenvalue of (7). For given values of the parameters σ_i (RBF kernel), our proposed kernel canonical correlation estimator for training data is given by:

$$\hat{\rho}_K(\Omega^{(1)}, \ldots, \Omega^{(m)}) = \max_k \lambda_k \tag{9}$$

where $\lambda_k, k = 1, \ldots, N$ are the eigenvalues of (7). For test data, the normalized canonical correlation of the projected variables is used:

$$\hat{\rho}_K(\Omega_{\text{test}}^{(1)}, \ldots, \Omega_{\text{test}}^{(m)}) = \sum_{i=1}^{m}\sum_{j\neq i}^{m} \frac{e_{\text{test}}^{(i)^T} e_{\text{test}}^{(j)}}{||e_{\text{test}}^{(i)}||_2 ||e_{\text{test}}^{(j)}||_2} \tag{10}$$

IV. A Measure of Independence using LS-SVM based Kernel CCA

Consider the following statistical model:

$$Y = MS,$$

where Y is a $p \times N$ observed matrix, M is a $p \times p$ full rank mixing matrix and S is a $p \times N$ matrix representing p independent components. The ICA problem [10], [11], [4], [3], [1] 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 [3]. However, in practice, the distributions of the components are usually not known and the ICA problem is tackled with a semiparametric model $W = M^{-1}$ allowing the estimation of independent components via $\hat{S} = \hat{W}Y$ [10]. A common approach to ICA is to minimize approximations of the mutual information between the components of $\hat{S}$, but alternative contrasts function have also been proposed. If the covariance matrix of the data is equal to the identity matrix (also known

as *whitened* data), the demixing matrix W is necessarily orthogonal [10]. Therefore, the ICA problem can be reduced to finding an orthogonal matrix W such that the components of $\hat{S}$ are independent.

A measure of independence based on kernel canonical correlation was introduced in [2] 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:

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

where $f^{(i)} = \langle \varphi^{(i)}(x^{(i)}), f^{(i)} \rangle, i = 1, 2$ 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 [2]. This also extends to the multivariate case in which the $\mathcal{F}$-correlation characterizes pairwise independence. Therefore, the contrast function proposed in [2] corresponds to the smallest eigenvalue of the extension to m sources of (6). This extension was formulated in such a way that the smallest eigenvalue corresponds to the maximal canonical correlation.

Our proposed measure of independence is given by (9) for training data and (10) for test data. Given Y, the estimated independent components $\hat{S} = \hat{W}Y$ can be obtained by minimizing the contrast function with respect to the demixing matrix $\hat{W}$ so that the kernel canonical correlation estimator is as close to zero as possible.

V. Optimization on the Stiefel Manifold

A. Incomplete Cholesky Decomposition

For real-life problems, the cost of computing the largest eigenvalue of (7) 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 $\Omega = HH^T$ 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 $||\Omega - GG^T||_2$ is below some given threshold η and $M \ll N$. This method is known as the incomplete Cholesky decomposition and was used in [2], [8] to reduce the computational burden of the corresponding eigenvalue problem.

Making use of the incomplete Cholesky decomposition, the centered kernel matrices in (7) can approximated as $\Omega_c^{(i)} \approx G_c^{(i)} G_c^{(i)^T}$. The SVD of $G_c^{(i)}$ is $U^{(i)}\Lambda^{(i)}U^{(i)^T}$ where $U^{(i)}$ is an $N \times M_i$ orthogonal matrix and $\Lambda^{(i)}$ is an $M_i \times M_i$ diagonal matrix. Therefore we can approximate (7) with

$$\tilde{\mathcal{K}}\alpha = \lambda \tilde{\mathcal{R}}\alpha \tag{11}$$

where

$$\tilde{\mathcal{K}} = \begin{bmatrix} 0 & U^{(2)}\Lambda^{(2)}U^{(2)^T} & \dots & U^{(m)}\Lambda^{(m)}U^{(m)^T} \\ U^{(1)}\Lambda^{(1)}U^{(1)^T} & 0 & \dots & \Omega_c^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ U^{(1)}\Lambda^{(1)}U^{(1)^T} & U^{(2)}\Lambda^{(2)}U^{(2)^T} & \dots & 0 \end{bmatrix},$$

$$\tilde{\mathcal{R}} = \begin{bmatrix} U^{(1)}\Lambda_\nu^{(1)}U^{(1)^T} & 0 & \dots & 0 \\ 0 & U^{(1)}\Lambda_\nu^{(2)}U^{(2)^T} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U^{(m)}\Lambda_\nu^{(m)}U^{(m)^T} \end{bmatrix},$$

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

$$\Lambda_{\nu,nn}^{(i)} = 1 + \nu_i \Lambda_{nn}^{(i)}, n = 1, \dots, M_i$$

Substituting $\tilde{\alpha}^{(i)} = U^{(i)^T}\alpha^{(i)}, i = 1, \dots, m$ and premultiplying both sides of (11) by blkdiag$(U^{(1)^T}, \dots, U^{(m)^T})$ leads to the following reduced problem:

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

where

$$\tilde{\mathcal{K}}_{\mathcal{R}} = \begin{bmatrix} 0 & T_U^{(1)}U_\Lambda^{(2)} & \dots & T_U^{(1)}U_\Lambda^{(m)} \\ T_U^{(2)}U_\Lambda^{(1)} & 0 & \dots & T_U^{(1)}U_\Lambda^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ T_U^{(m)}U_\Lambda^{(1)} & T_U^{(m)}U_\Lambda^{(2)} & \dots & 0 \end{bmatrix},$$

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

$T_U^{(i)} = T^{(i)}U^{(i)^T}$, $T^{(i)}$ is a diagonal matrix $T_{nn}^{(i)} = 1/\Lambda_{nn}^{(i)}$, $U_\Lambda^{(h)} = U^{(h)}\Lambda^{(h)}, i, h = 1, \dots, m, n = 1, \dots, M_i$. Note that the size of $\tilde{\mathcal{K}}_{\mathcal{R}}$ is $(\sum_{i=1}^m M_i) \times (\sum_{i=1}^m M_i)$ which is much smaller than the original $mN \times mN$ problem in (7).

B. 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 [5] 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 $G_W(t) = \exp(tW^TA)$ 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 [2], [8] applied gradient descent methods on the Stiefel manifold in order to obtain demixing matrices leading to independent components. We also used this approach to minimize our proposed contrast function.

VI. 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 specially 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 $(\sigma^{(i)})^2 = \sigma^2, i = 1, \ldots, m$ which means considering one feature map $\varphi^{(i)} = \varphi$. The regularization parameters are also considered to be identical $\nu_i = \nu$ and fixed to some value. For model selection, the following criterion is proposed:

$$\hat{\rho}_K(\Omega_{\text{val}}^{(1)}, \ldots, \Omega_{\text{val}}^{(m)}) = \sum_{i=1}^{m} \sum_{j \neq i}^{m} \frac{e_{\text{val}}^{(i)^T} e_{\text{val}}^{(j)}}{||e_{\text{val}}^{(i)}||_2 ||e_{\text{val}}^{(j)}||_2 ||\Omega_{\text{val}}||_F} \tag{13}$$

where e_{val} and Ω_{val} are calculated in a similar way as in (8) but using validation data. Note that (13) is a normalized version of (10) to fix the scaling introduced by σ^2.

VII. Empirical Results

In this section, empirical results are presented. In all the experiments, we assume that the number of sources m is given and that $m = p$. The computation time of the eigenvalue decomposition involved in the calculation of the contrast function for training data was 0.5 seconds for $N = 1000$, $m = 5$, $\eta = 10^{-6}$ on a Pentium 4, 2.8 GHz, 1GB RAM using Lanczsos algorithm in Matlab. In the case of test and validation data, the computation time increased to 8 seconds because the projected variables should be calculated in order to compute the correlation coefficient.

A. Toy Examples

In this experiment, we used 5 simple sources but with a small number of data points $N = 50$. One of the sources is Gaussian. Since the size of the corresponding generalized eigenvalue problem is relatively small, $\mathcal{K}$ and $\mathcal{R}$ are 250×250 matrices, we solved the problem without using the incomplete Cholesky decomposition. The regularization parameter ν was fixed to 10 and the proposed contrast function was evaluated over validation data for several values of σ^2. Fig. 1 shows the model selection curve. The obtained independent components of the kernel ICA with KCCA contrast function [2] and our proposed approach are shown in Fig. 2. Note that our approach did not show overfitting due to the parameter selection using validation data.

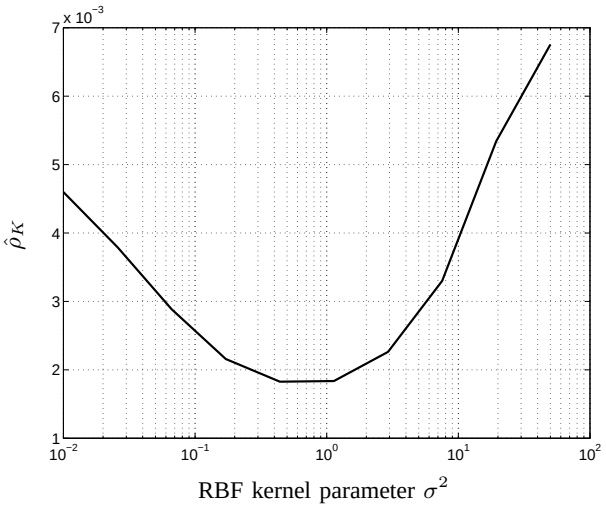

Fig. 1. Toy problem - Tuning the RBF kernel parameter σ^2. The plot shows the proposed contrast function evaluated on validation data with fixed regularization parameter ν set to 10. The tuned σ^2 is 0.7

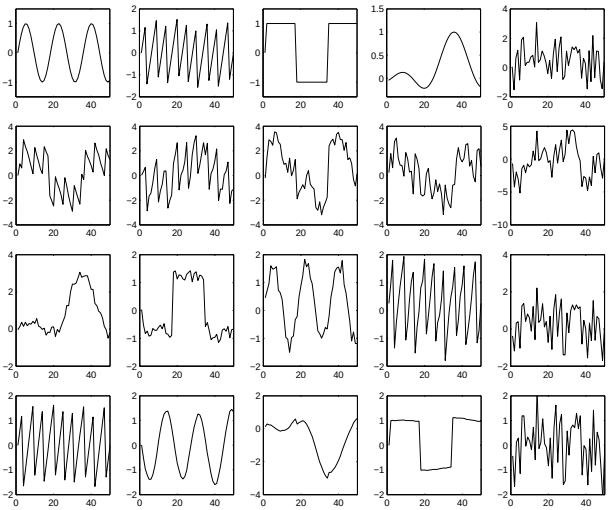

Fig. 2. Toy problem - Results. *First row:* Original sources. *Second row:* Some observed mixtures. *Third row:* Best results of the kernel ICA proposed in [2] using the KCCA contrast function showing some overfitting. The parameters were fixed to the recommended values in [2] and 20 restarts were performed. *Fourth row:* Best results of our proposed approach using the tuned $\sigma^2 = 0.7$.

B. Image demixing

The second experiment consisted of image demixing using 200×200 pixels images with 3 sources. Therefore, the matrices involved in the eigendecomposition have size 120000×120000 making the problem intractable if one tries to solve it directly. Applying the incomplete Cholesky decomposition with threshold $\eta = 10^{-6}$, we reduced the problem to a matrix of size 400×400. In this experiment we fixed the parameters to $\sigma^2 = 1, \nu = 10$. The best results after 20 restarts of the algorithm with different starting points are shown in Fig. 3.



Fig. 3. Image demixing - Results. *First row:* Original images. *Second and third rows:* Some observed mixtures. *Fourth row:* Estimated independent components using the proposed approximation with the incomplete Cholesky decomposition showing visually appealing results.

VIII. Conclusions

A new measure for independence based on an LS-SVM formulation to kernel CCA is proposed. This formulation follows from a clear constrained optimization framework providing primal-dual insights and leading to a naturally regularized kernel CCA approach. The kernel canonical correlation in high dimensional feature spaces corresponds to a measure of independence in the original input space which can be used to find independent components. 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 when few data is available. Simulations with small toy datasets showed better performance avoiding overfitting. Experiments with images using the incomplete Cholesky decomposition to reduce the size of the corresponding 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

[1] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763, Cambridge, MA, 1996. MIT Press.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

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

[4] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

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

[6] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. of Computers*, 23(9):881–890, 1974.

[7] R. Gittins. *Canonical Analysis - A Review with Applications in Ecology*. Berlin: Springer Verlag, 1985.

[8] A. Gretton, R. H., A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[9] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2001.

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

[12] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

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

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

[15] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.