KERNEL INDEPENDENT COMPONENT ANALYSIS

Francis R. Bach

Computer Science Division
University of California
Berkeley, CA 94720, USA
fbach@cs.berkeley.edu

Michael I. Jordan

Computer Science Division
and Department of Statistics
University of California
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

ABSTRACT

We present a class of algorithms for independent component analysis (ICA) which use contrast functions based on canonical correlations in a reproducing kernel Hilbert space. On the one hand, we show that our contrast functions are related to mutual information and have desirable mathematical properties as measures of statistical dependence. On the other hand, building on recent developments in kernel methods, we show that these criteria can be computed efficiently. Minimizing these criteria leads to flexible and robust algorithms for ICA. We illustrate with simulations involving a wide variety of source distributions, showing that our algorithms outperform many of the presently known algorithms.

1. INTRODUCTION

Recent research on kernel methods has yielded important new computational tools for solving large-scale, nonparametric classification and regression problems [10]. While some forays have also been made into unsupervised learning, there is still much unexplored terrain in problems involving large collections of mutually interacting variables, problems in which Markovian or general graphical models have excelled. These latter models in fact have several limitations that invite kernel-based initiatives; in particular, they are almost entirely based on strong parametric assumptions, and lack the nonparametric flexibility of the kernel approaches.

Independent component analysis (ICA) [8] is an interesting unsupervised learning problem in which to explore these issues. On the one hand, ICA is heavily based on structural assumptions—viewed as a graphical model it is a directed bipartite graph linking a set of "source nodes" to a set of "observation nodes," in which the lack of edges between the source nodes encodes an assumption of mutual independence. On the other hand, the ICA problem is also strongly nonparametric—the distribution of the source variables is left unspecified. This is difficult to accommodate within the (current) graphical model formalism, in which all nodes must be endowed with a probability distribution. It is here that we will find kernel methods to be useful. We will show how kernel methods can be used to define a "contrast function" that can be used to estimate the parametric part of the ICA model (the source-to-observation edges), despite the absence of a specific distribution on the source nodes. As we will see, compared to current ICA algorithms, the new kernel-based approach is notable for its robustness.

We refer to our new approach to ICA as "KERNELICA." It is important to emphasize at the outset that KERNELICA is not the "kernelization" of an extant ICA algorithm. Rather, it is a new approach to ICA based on novel kernel-based measures of dependence. We introduce two such measures. In Section 3, we define a kernel-based contrast function in terms of the first eigenvalue of a certain generalized eigenvector problem, and show how this function relates to probabilistic independence. In Section 4.3, we introduce an alternative kernel-based contrast function based on the entire spectrum of the generalized eigenvector problem, and show how this function can be related to mutual information.

2. BACKGROUND ON ICA

Independent component analysis (ICA) is the problem of recovering a latent random vector $x = (x_1, \ldots, x_m)^\top$ from observations of m unknown linear functions of that vector. The components of x are assumed to be mutually independent. Thus, an observation $y = (y_1, \ldots, y_m)^\top$ is modeled as $y = Ax$, where x is a latent random vector with independent components, and where A is an $m \times m$ matrix of parameters. Given N independently, identically distributed observations of y, we hope to estimate A and thereby to recover the latent vector x corresponding to any particular y by solving a linear system.

By specifying distributions for the components x_i, one obtains a parametric model that can be estimated via maximum likelihood [5]. Working with $W = A^{-1}$ as the parameterization, one readily obtains a gradient or fixed-point algorithm that yields an estimate $\hat{W}$ and provides estimates of the latent components via $\hat{x} = \hat{W}y$ [8].

In practical applications, however, one does not generally know the distributions of the components x_i, and it is preferable to view the ICA model as a *semiparametric model* in which the distributions of the components of x are left unspecified [6]. Maximizing the likelihood in the semiparametric ICA model is essentially equivalent to minimizing the mutual information between the components of the estimate $\hat{x} = \hat{W}y$ [7]. Thus it is natural to view mutual information as a *contrast function* to be minimized in estimating the ICA model.

Unfortunately, the mutual information for real-valued variables is difficult to approximate and optimize on the basis of a finite sample, and much research on ICA has focused on alternative contrast functions [8, 7, 1]. These have either been derived as expansion-based approximations to the mutual information, or have had a looser relationship to the mutual information, essentially borrowing its key property of being equal to zero if and only if the arguments to the function are independent. In this paper, we define two novel contrast functions. Minimizing them will lead to two KERNELICA algorithms.

3. MEASURING STATISTICAL DEPENDENCE WITH KERNELS

In this section, we define the $\mathcal{F}$-correlation, a measure of statistical dependence among random variables $x_1, \ldots, x_m$. For simplicity, we restrict ourselves initially to the case of two real random variables, x_1 and x_2, treating the general case of m variables in Section 3.4. (It is also worth noting that the restriction to real random variables is again for simplicity; a similar measure of dependence can be defined for any type of data for which Mercer kernels can be defined).

We assume that we are given a reproducing-kernel Hilbert space (RKHS) $\mathcal{F}$ on $\mathbb{R}$, with kernel $K(x,y)$ and feature map $\Phi(x)$. In this paper, our focus is the Gaussian kernel, $K(x,y) = \exp(-(x-y)^2/2\sigma^2)$, which corresponds to an infinite-dimensional RKHS of smooth functions [10].

3.1. The $\mathcal{F}$-correlation

Given an RKHS $\mathcal{F}$, we define the $\mathcal{F}$-correlation as the maximal correlation between the random variables $f_1(x_1)$ and $f_2(x_2)$, where f_1 and f_2 range over $\mathcal{F}$:

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}(f_1(x_1), f_2(x_2)) \quad (1)$$

$$= \max_{f_1, f_2 \in \mathcal{F}} \frac{\operatorname{cov}(f_1(x_1), f_2(x_2))}{(\operatorname{var} f_1(x_1))^{1/2} (\operatorname{var} f_2(x_2))^{1/2}}. \quad (2)$$

Clearly, if the variables x_1 and x_2 are independent, then the $\mathcal{F}$-correlation is equal to zero. Moreover, if the set $\mathcal{F}$ is large enough, the converse is also true. For example, it is well known that if $\mathcal{F}$ contains the Fourier basis (all functions of the form $x \mapsto e^{i\omega x}$ where $\omega \in \mathbb{R}$), then $\rho_{\mathcal{F}} = 0$ implies that x_1 and x_2 are independent. In [3], we show that the converse is also true for the reproducing kernel Hilbert spaces based on Gaussian kernels.

For reasons that will become clear in Section 4.3, it is useful to work on a logarithmic scale; in particular, we define our first contrast function as $I_{\rho_{\mathcal{F}}} = -\frac{1}{2}\log(1-\rho_{\mathcal{F}})$. Our converse result implies that $I_{\rho_{\mathcal{F}}}$ is a valid contrast function; a function that is always nonnegative and equal to zero if and only if the variables x_1 and x_2 are independent.

The ability to restrict the maximization in Eq. (1) to an RKHS has an important computational consequence. In particular, we can exploit the *reproducing property*, $f(x) = \langle \Phi(x), f \rangle$, to obtain an interpretation of $\rho_{\mathcal{F}}$ in terms of linear projections. Indeed, the reproducing property implies that $\operatorname{corr}(f_1(x_1), f_2(x_2)) = \operatorname{corr}\left(\langle \Phi(x_1), f_1 \rangle, \langle \Phi(x_2), f_2 \rangle\right)$. Consequently, the $\mathcal{F}$-correlation is the maximal possible correlation between one-dimensional linear projections of $\Phi(x_1)$ and $\Phi(x_2)$. This is exactly the definition of the first *canonical correlation* [2] between $\Phi(x_1)$ and $\Phi(x_2)$. This interpretation will enable us to derive a computationally efficient algorithm.

3.2. Canonical correlation analysis

Canonical correlation analysis (CCA) is a multivariate statistical technique similar in spirit to principal component analysis (PCA). While PCA works with a single random vector and maximizes the variance of projections of the data, CCA works with a pair of random vectors (or in general with a set of m random vectors) and maximizes correlation between sets of projections. While PCA leads to an eigenvector problem, CCA leads to a generalized eigenvector problem. More precisely, given two random vectors, x_1 and x_2, the first canonical correlation between x_1 and x_2 can be defined as the maximum possible correlation between the two projections $\xi_1^\top x_1$ and $\xi_2^\top x_2$ of x_1 and x_2:

$$\rho(x_1, x_2) = \max_{\xi_1, \xi_2} \operatorname{corr}(\xi_1^\top x_1, \xi_2^\top x_2) \quad (3)$$

$$= \max_{\xi_1, \xi_2} \frac{\xi_1^\top C_{12} \xi_2}{\left(\xi_1^\top C_{11} \xi_1\right)^{1/2} \left(\xi_2^\top C_{22} \xi_2\right)^{1/2}}, \quad (4)$$

where C_{ij} denotes the covariance matrix $\operatorname{cov}(x_i, x_j)$. By taking derivatives with respect to ξ_1 and ξ_2, this problem is easily seen to reduce to the following generalized eigenvalue problem [2]:

$$\begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \rho \begin{pmatrix} C_{11} & 0 \\ 0 & C_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}. \quad (5)$$

We need to be able to solve this problem in feature space, and thus we need to consider a "kernelized" version of CCA.

3.3. Estimating the $\mathcal{F}$-correlation

Let $\{x_1^1, \ldots, x_1^N\}$ and $\{x_2^1, \ldots, x_2^N\}$ denote sets of N empirical observations of x_1 and x_2. The observations generate *Gram matrices* L_1 and L_2, defined as $(L_i)_{ab} = K(x_i^a, x_i^b)$. The *centered Gram matrices* [10] K_1 and K_2 are defined as the Gram matrices of the centered (in feature space) data points and are equal to $K_i = P L_i P$ where $P = I - \frac{1}{N}\mathbf{1}$ is a constant singular matrix ($\mathbf{1}$ is the $N \times N$ matrix composed of ones).

Following the spirit of the derivation of kernel PCA [10], it is straightforward to derive a "kernelization" of CCA, which turns out to involve substituting products of Gram matrices for the covariance matrices in Eq. (3), and maximizing

$$\frac{\alpha_1^\top K_1 K_2 \alpha_2}{(\alpha_1^\top (K_1 + N\kappa I/2)^2 \alpha_1)^{1/2} (\alpha_2^\top (K_2 + N\kappa I/2)^2 \alpha_2)^{1/2}},$$

where κ is a small positive regularization parameter. As for CCA in. Eq. (3), the solution is obtained by solving the following generalized eigenvalue problem (cf. Eq. (3) and (5)):

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + \frac{N\kappa}{2} I)^2 & 0 \\ 0 & (K_2 + \frac{N\kappa}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad (6)$$

Since $(K_i + \kappa I)^2$ is necessarily invertible, classical methods can be invoked to solve the generalized eigenvalue problem in Eq. (6). Thus kernel CCA reduces to finding the largest eigenvalue of $\widetilde{\mathcal{K}}_\kappa = \begin{pmatrix} 0 & r_\kappa(K_1) r_\kappa(K_2) \\ r_\kappa(K_2) r_\kappa(K_1) & 0 \end{pmatrix}$, with $r_\kappa(K_i) = K_i (K_i + \kappa I)^{-1}$.

3.4. Generalization to more than two variables

It is straightforward to extend CCA, and its kernelized counterpart, to the case of m variables [3]. The problem becomes that of finding the smallest eigenvalue of the generalized eigenvalue problem $\mathcal{K}\alpha = \lambda \mathcal{D}\alpha$, where $\mathcal{K}$ is defined by blocks $\mathcal{K}_{ij} = K_i K_j$ for $i \neq j$ and $\mathcal{K}_{ii} = (K_i + \kappa I)^2$, and $\mathcal{D}$ is block diagonal with blocks $\mathcal{D}_{ii} = (K_i + \kappa I)^2$. We still refer to this eigenvalue as the $\mathcal{F}$-correlation.[1]

[1] See [3] for a detailed explanation of why we use the *smallest* generalized eigenvalue in our general definition, and how this accords with our earlier definition. In brief, the definitions are equivalent because of a symmetry property of the eigenvalues for the CCA problem.

It is worth noting that the general version of the $\mathcal{F}$-correlation that we have defined does not characterize *mutual* dependence among m variables, but only characterizes *pairwise* independence. Empirically, this does not appear to be a limitation in the ICA setting, as we show in Section 5. However, in situations in which a measure of mutual independence is required, one can form such a measure by exploiting the general fact that mutual independence can be expressed in terms of pairwise mutual information terms involving sets of variables. (Thus, for example, in the three-variable case we have the expansion $I(x, y, z) = I((x, y), z) + I(x, y)$).

4. KERNEL INDEPENDENT COMPONENT ANALYSIS

Having defined a contrast function in terms of the solution of a generalized eigenvalue problem, we now obtain a KERNELICA algorithm by *minimizing* this contrast function with respect to the parameter matrix W.

4.1. Outline of algorithm

Given a set of data vectors $y^1, y^2, \ldots, y^N$, and given a parameter matrix W, we set $x^i = Wy^i$, for each i, and thereby form a set of estimated source vectors $\{x^1, x^2, \ldots, x^N\}$. The m components of these vectors yield a set of m centered Gram matrices, $K_1, K_2, \ldots, K_m$. These Gram matrices (which depend on W) define the contrast function, $C(W) = \hat{I}_{\rho_\mathcal{F}}(K_1, \ldots, K_m)$, as the solution to a generalized eigenvalue problem, $\mathcal{K}\alpha = \lambda\mathcal{D}\alpha$, where $\mathcal{K}$ and $\mathcal{D}$ are block matrices constructed from the Gram matrices K_i. The KERNELICA-KCCA algorithm involves minimizing this function $C(W)$ with respect to W.

4.2. Computational issues

In order to turn this sketch into a practical ICA algorithm, several computational issues have to be addressed, as we now discuss.

Numerical linear algebra. The $\mathcal{F}$-correlation involves computing the smallest generalized eigenvalue of matrices of size mN. Thus a naive implementation would scale as $O(N^3)$, a computational complexity whose cubic growth in the number of data points would be a serious liability in applications to large data sets. However, Gram matrices have a spectrum that tends to show rapid decay, and low-rank approximations of Gram matrices can therefore often provide sufficient fidelity for the needs of kernel-based algorithms [10]. In [3], we show theoretically that for a regularization parameter κ that is linear in N, we require low-rank approximations of size M, where M is a constant that is independent of the number N of samples. Since the Gram matrix K_i is positive semidefinite, the low-rank approximation can be found through incomplete Cholesky decomposition in time $O(M^2N)$, which gives a $M \times N$ matrix G_i such that $K_i \approx G_iG_i^\top$. We perform a singular value decomposition of G_i, in time $O(M^2N)$, to obtain an $N \times M$ matrix U_i with orthogonal columns (i.e., such that $U_i^\top U_i = I$), and an $M \times M$ diagonal matrix Λ_i such that $K_i \approx G_iG_i^\top = U_i\Lambda_iU_i^\top$.

We then have $r_\kappa(K_i) = (K_i + \kappa I)^{-1}K_i = U_iD_iU_i^\top$, where D_i is the diagonal matrix obtained from the diagonal matrix Λ_i by applying the function $\lambda \mapsto \lambda/(\lambda + \kappa)$ to its elements. Finally, in the two-dimensional case, our problem reduces to finding the largest eigenvalue of $\widetilde{\mathcal{R}}_\kappa = \begin{pmatrix} 0 & D_1U_1^\top U_2D_2 \\ D_2U_2^\top U_1D_1 & 0 \end{pmatrix}$, with the obvious extension to the m-dimensional case. This problem can be solved in time linear in N.

Gradient descent on the Stiefel manifold. Since decorrelation implies independence, it is common to enforce decorrelation of the estimated sources. This is done by *whitening* the data and subsequently restricting the minimization to orthogonal matrices W [8]. The set of orthogonal matrices, which is commonly referred to as the Stiefel manifold, can be equipped with a natural Riemannian metric, which implies that gradient algorithms can be used. In our simulations we used steepest descent with line search along geodesics. The algorithm necessarily converges to a local minimum of $C(W)$, from any starting point.

The ICA contrast functions have multiple local minima, however, and restarts are generally necessary if we are to find the global optimum. Empirically, the number of restarts that were needed was found to be small when the number of samples is sufficiently large so as to make the problem well-defined. We have also developed two initialization heuristics that have been found to be particularly useful in practice for large-scale problems,"one-unit contrast functions", and Hermite polynomial kernels. These are detailed in [3].

4.3. Kernel generalized variance

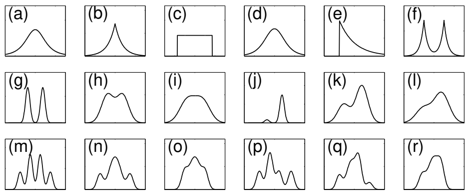
The $\mathcal{F}$-correlation is defined as the first eigenvalue of the kernelized CCA problem. It is obviously of interest to consider the other eigenvalues as well. Indeed, there is a classical relationship between the full CCA spectrum and the mutual information of Gaussian variables x_1 and x_2 [2]: the mutual information $I(x_1, x_2)$ is equal to $-\frac{1}{2}\log\prod_i(1 - \rho_i^2)$. The product $\prod_i(1 - \rho_i^2)$ is usually referred to as the *generalized variance*.

This suggests defining a corresponding quantity for kernelized CCA. In the case of two variables, we define the *kernel generalized variance (KGV)* as the product $\hat{\delta}_\mathcal{F} = \prod_i(1 - \rho_i^2)$, where ρ_i are the (positive) kernel canonical correlations. In the general case of m variables, we define $\hat{\delta}_\mathcal{F} = \det\mathcal{K}/\det\mathcal{D}$. Finally, by analogy with the mutual information for the Gaussian case, we also define a contrast function $\hat{I}_{\delta_\mathcal{F}} = -\frac{1}{2}\log\hat{\delta}_\mathcal{F}$. It turns out that $\hat{I}_{\delta_\mathcal{F}}(K_1, \ldots, K_m)$ has as its population counterpart a function $I_{\delta_\mathcal{F}}(x_1, \ldots, x_m)$ that is an approximation of the mutual information between the original non-Gaussian variables in the input space [3].

5. SIMULATION RESULTS

We have conducted an extensive set of simulation experiments using data obtained from a variety of source distributions. The sources that we used (Figure 1, Top) included subgaussian and supergaussian distributions, as well as distributions that are nearly Gaussian. We studied unimodal, multimodal, symmetric, and nonsymmetric distributions. We also varied the number of components, from 2 to 16, the number of training samples, from 250 to 4000, and studied the robustness of the algorithms to varying numbers of outliers (see [3] for details).

Comparisons were made with three existing ICA algorithms: the FastICA algorithm [8], the Jade algorithm [7], and the extended Infomax algorithm [9]. All simulations were performed in the situation when the true demixing matrix W_0 is known. We measure the performance of the algorithm in terms of the difference between W and W_0, via the standard ICA metric introduced by [1]. This measure is invariant to permutation and scaling of its arguments, lies between 0 and $100(m - 1)$, and is equal to zero for perfect demixing.



pdfs	F-ica	Jade	Imax	Kcca	Kgv
a	4.4	3.7	**1.8**	3.7	3.0
b	5.8	4.1	3.4	3.7	**2.9**
c	2.3	**1.9**	2.0	2.7	2.4
d	6.4	6.1	6.9	7.1	**5.7**
e	4.9	3.9	3.2	1.7	**1.5**
f	3.6	2.7	**1.0**	1.7	1.5
g	1.7	1.4	**0.5**	1.4	1.3
h	5.5	3.9	**3.2**	4.3	3.6
i	8.7	7.2	6.8	7.8	**6.5**
j	6.7	4.6	57.6	1.4	**1.3**
k	5.7	4.0	3.5	3.2	**2.6**
l	12.1	7.2	10.4	4.8	**4.2**
m	3.6	**2.9**	4.2	6.3	4.6
n	5.4	3.5	30.6	7.6	**3.0**
o	4.7	**3.3**	4.4	5.1	4.3
p	4.1	3.1	7.4	3.8	**3.0**
q	22.9	15.8	40.9	5.1	**3.9**
r	6.6	4.4	4.9	4.3	**3.6**
mean	6.4	4.6	10.7	4.2	**3.3**

Fig. 1. (Top) Source density functions. (Bottom) Performance of ICA algorithms for $m = 2$. The best performance in each row is indicated in bold font.

The results in Figure 1 (Bottom) show that the KERNELICA algorithms are competitive with current algorithms, and are particularly successful at handling asymmetric sources (see, e.g., the performance for sources j, l and q). In Figure 2 (Top), which reports results for random choices of source distributions, we see that the KERNELICA algorithms perform well for larger numbers of components. Finally, in Figure 2 (Bottom), we report the results of an experiment in which we added random outliers to the source data. We see that our algorithms are particularly resistant to outliers.

6. CONCLUSIONS

We have presented two novel, kernel-based measures of statistical dependence. These measures can be optimized with respect to a parameter matrix, yielding new algorithms for ICA. These algorithms are competitive with current algorithms, and are particularly notable for their resistance to outliers.

Our approach to ICA is more flexible and more demanding computationally than current algorithms, involving a search in a reproducing kernel Hilbert space—an inner loop which is not present in other algorithms. But the problem of measuring (and minimizing) departure from independence over all possible non-Gaussian source distributions is a difficult one, and we feel that the flexibility provided by our approach is appropriately targeted.

Many other problems at the intersection of graphical models and nonparametric estimation can also be addressed using these tools. In particular, in recent work [4], we have generalized ICA

m	N	F-ica	Jade	Imax	Kcca	Kgv
2	250	11	9	30	7	**5**
	1000	5	4	7	3	**2**
4	1000	18	13	25	12	**11**
	4000	8	7	11	6	**4**
8	2000	26	22	123	30	**20**
	4000	18	16	41	16	**8**
16	4000	42	38	130	31	**19**

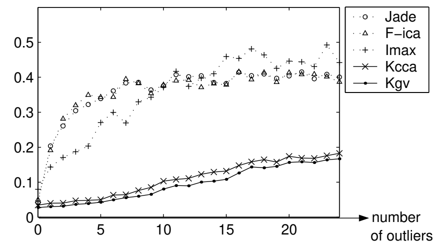


Fig. 2. (Top) Performance for larger number of components m. (Bottom) Performance as a function of the number of outliers.

to a model that no longer requires the sources to be independent, but requires them only to factorize according to a tree. The departure from a tree distribution can be measured in terms of a sum of mutual information terms, and approximated using the KGV.

7. REFERENCES

[1] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Adv. in NIPS, 8*, 1996.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley & Sons, 1984.

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

[4] F. R. Bach and M. I. Jordan. Tree-dependent component analysis. In *Proc. UAI*, 2002.

[5] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[6] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

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

[8] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley & Sons, 2001.

[9] T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended Infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[10] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.