Kernel Independent Component Analysis

Jake Spiteri

University of Bristol

2020

Table of Contents

Summary of Independent Component Analysis

Problem:

- We want to recover a latent random vector $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ from observations $\boldsymbol{y} = (y_1, \ldots, y_m)$ which are unknown linear functions of $\boldsymbol{x}$.
- The components of $\boldsymbol{x}$ are modeled as mutually independent.
- An observation $\boldsymbol{y}$ is modeled as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A}$ is an $m \times m$ matrix of parameters.

Given N observations of $\boldsymbol{y}$, we want to estimate $\boldsymbol{A}$ and thus recover the latent vector $\boldsymbol{x}$.

Seeking Independence

Our problem can be reduced to finding $\boldsymbol{W} := \boldsymbol{A}^{-1}$ such that the components of $\hat{\boldsymbol{x}} = \hat{\boldsymbol{W}}\boldsymbol{y}$ are *independent*.

We have previously performed ICA by maximizing the negentropy, which is a measure of non-Gaussianity.

To achieve independence we can estimate parameters by minimizing a *contrast function*, where a contrast function is defined to always be nonnegative and equal to zero if and only if variables x_1 and x_2 are independent.

Kernel ICA

Kernel ICA uses kernel-based measures of statistical dependence.

Definition ($\mathcal{F}$-correlation)

For a reproducing-kernel Hilbert space (RKHS) $\mathcal{F}$, the $\mathcal{F}$-correlation between the random variables $f_1(x_1)$ and $f_2(x_2)$, where $f_1, f_2 \in \mathcal{F}$ is:

$$\begin{aligned}\rho_{\mathcal{F}} &= \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}\left(f_1\left(x_1\right), f_2\left(x_2\right)\right), \\ &= \max_{f_1, f_2 \in \mathcal{F}} \frac{\operatorname{cov}\left(f_1\left(x_1\right), f_2\left(x_2\right)\right)}{\left(\operatorname{var} f_1\left(x_1\right)\right)^{1/2}\left(\operatorname{var} f_2\left(x_2\right)\right)^{1/2}}.\end{aligned}$$

Clearly if x_1 and x_2 are independent, then the $\mathcal{F}$-correlation is zero.

Contrast function

We will use the following contrast function based on the $\mathcal{F}$-correlation

$$I_{\rho_{\mathcal{F}}} = -\frac{1}{2}\log(1 - \rho_{\mathcal{F}})$$

Reproducing Property

Restricting the maximization in the $\mathcal{F}$-correlation to the RKHS allows us to exploit the *reproducing property*:

$$f(x) = \langle \Phi(x), f \rangle, \quad \forall f \in \mathcal{F},$$

where $\Phi : \mathcal{X} \to \mathcal{F}$ is a map from our input space into the RKHS. This allows us to write

$$\begin{aligned} \rho_{\mathcal{F}} &= \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}\left(f_1\left(x_1\right), f_2\left(x_2\right)\right) \\ &= \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}\left(\left\langle \Phi\left(x_1\right), f_1\right\rangle, \left\langle \Phi\left(x_2\right), f_2\right\rangle\right) \end{aligned}$$

That is, the $\mathcal{F}$-correlation is the maximum correlation between one-dimensional linear projections of $\Phi\left(x_1\right), \Phi\left(x_2\right)$. This is the definition of the first *canonical correlation* between $\Phi\left(x_1\right)$, and $\Phi\left(x_2\right)$.

Problem Setup

- To use the $\mathcal{F}$-correlation as a contrast function for ICA, we need to compute canonical correlations in our feature space.
- We need a kernelization of the canonical correlation. This will allow us to work with an empirical sample and work in the feature space.

Kernelization of CCA

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 $\boldsymbol{L}_1, \boldsymbol{L}_2$, where $\{\boldsymbol{L}_i\}_{r,k} := K(x_i^r, x_j^k)$. We then compute the centered Gram matrices $\boldsymbol{K}_1, \boldsymbol{K}_2$. Our kernelized CCA problem becomes

$$\begin{aligned}
\hat{\rho}_{\mathcal{F}}(\boldsymbol{K}_1, \boldsymbol{K}_2) &= \max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^N} \operatorname{corr}\left(\boldsymbol{\alpha}_1^\top \boldsymbol{x}_1, \boldsymbol{\alpha}_2^\top \boldsymbol{x}_2\right) \\
&= \max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^N} \frac{\boldsymbol{\alpha}_1^\top \boldsymbol{K}_1 \boldsymbol{K}_2 \boldsymbol{\alpha}_2}{\left(\boldsymbol{\alpha}_1^\top \boldsymbol{K}_1^2 \boldsymbol{\alpha}_1\right)^{1/2} \left(\boldsymbol{\alpha}_2^\top \boldsymbol{K}_2^2 \boldsymbol{\alpha}_2\right)^{1/2}}
\end{aligned}$$

Kernelization of CCA

Based on the previous slide, we can perform a kernelized version of CCA by solving the generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

The $\mathcal{F}$-correlation is defined as the first (largest) eigenvalue of the kernelized CCA problem.

Kernelization of CCA

Based on the previous slide, we can perform a kernelized version of CCA by solving the generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \rho \begin{pmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}$$

The $\mathcal{F}$-correlation is defined as the first (largest) eigenvalue of the kernelized CCA problem. We can rewrite this as

$$\begin{pmatrix} \boldsymbol{K}_1^2 & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix},$$

where $\lambda = 1 + \rho$. We can easily generalize this result to more than two variables.

Kernelization of CCA

Based on the previous slide, we can perform a kernelized version of CCA by solving the generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

The $\mathcal{F}$-correlation is defined as the first (largest) eigenvalue of the kernelized CCA problem. We can rewrite this as

$$\begin{pmatrix} \boldsymbol{K}_1^2 & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix},$$

where $\lambda = 1 + \rho$. We can easily generalize this result to more than two variables. We will write this as

$$\mathcal{K}\alpha = \lambda\mathcal{D}\alpha$$

Outline of the kernel ICA algorithm

Algorithm KernelICA-KCCA

Input: Data vectors $y^1, y^2, \ldots, y^N$
Kernel $K(x, y)$

1. Whiten the data
2. Minimize (with respect to W) the contrast function $C(W)$ defined as:
   a. Compute the centered Gram matrices $K_1, K_2, \ldots, K_m$ of the estimated sources $\{x^1, x^2, \ldots, x^N\}$, where $x^i = W y^i$
   b. Define $\hat{\lambda}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m)$ as the minimal eigenvalue of the generalized eigenvector equation $\mathcal{K}_\kappa \alpha = \lambda \mathcal{D}_\kappa \alpha$
   c. Define $C(W) = \hat{I}_{\lambda_{\mathcal{F}}}(K_1, \ldots, K_m) = -\frac{1}{2} \log \hat{\lambda}_{\mathcal{F}}^{\kappa}(K_1, \ldots, K_m)$

Output: W