University of BRISTOL

Kernel ICA

(it's ICA with a kernel)

Kieran Morris

University of bristol

23/02/2000

bristol.ac.uk

Review of ICA

ICA Model

We assume we have $\{x_i^0\}_{i=1}^n$ which are all p-dimensional realizations of a random variable X^0, then define $X = X^0 - \mathbb{E}[X^0]$. The ICA model assumes that

$$X = BS,\ B \in \mathbb{R}^{p \times p},\ \mathbb{E}[S] = 0,\ \text{Var}(S) = I_p \tag{1}$$

and all components of S are independent, and B is invertible.

We aim to recover S, think of these as 3 signals which are mixed when recorded on a microphone which we wish to retrieve the original signal back.

We have some useful results for our model.

Covariance assumption

We assumed in the model that $\text{Var}(S) = I_p$. We can do this without loss of generality. As we can always reduce a general $\text{Var}(S) = \Phi$ to the simple case.

Uniqueness

We can only obtain S up to an orthogonal transformation, i.e for any $G \in O(p)$, GS is a viable solution.

Orthogonal Solution

If we assume that $\text{Var}(X) = I_p$ then $BB^T = \text{Var}(BS) = \text{Var}(X) = I_p$ so our required B is orthogonal. This reduces our search space substantially!

For a measure of independence we can use the Mutual Information:

The Mutual Information

The Mutual Information of a random vector variable $Z = (Z_1, ...Z_p)$ is defined as

$$I(Z) = KL(Z \parallel \prod_{j=1}^{p} Z_j)) \tag{2}$$

where KL stands for the KL-divergence.

The mutual information is positive and $I(Z) =$ if and only if all Z_i are independent.

Objective

We find the B which minimises the following objective function, i.e

$$B \in \text{argmin}_{G \in O(p)} I(G^T X) \tag{3}$$

In practice there is not always a simple way to solve this, so we instead solve an approximate problem.

New Objective

Instead of minimising I, we solve the following simpler problem:

$$\hat{B} \in \text{argmax}_{G \in O(p)} \sum_{i=1}^{p} (\mathbb{E}[\phi(g_j^T X] - \mathbb{E}[\phi(Z)]) \tag{4}$$

for $Z \sim \mathcal{N}(0, 1)$ and $\phi : \mathbb{R} \to \mathbb{R}$. The solution to this problem is an approximate solution to the previous problem.

A few choices for ϕ are $\phi(u) = \frac{1}{a} \log \cosh au$ or $\phi(u) = -\exp(\frac{-u^2}{2})$. To minimise this objective function we simply approximate the first expectation with the same mean of our data and the second with numerical methods.

We have $\hat{B}$, now what?

Obtain S

Now that we have our orthogonal $\hat{B}$, we obtain back $\hat{S} = \hat{B}^T X$.

Whitening

Suppose we have instead that $\text{Var}(X) = \Sigma$ where Σ is full rank and hence $\Sigma = \Gamma L \Gamma^T$ for L diagonal and $\Gamma, \Gamma^T \in O(p)$. Then if we define $K = L^{-1/2}\Gamma^T$, $B_W = KB$ and $W = KX$ we have that $W = KX = (KB)S = B_W S$. It can additionally be shown that

$$\text{Var}(W) = I_p \tag{5}$$

and we have our initial setup! Giving us a way to find S even in the general case. We call the process of creating W whitening.

Kernel ICA

We introduce a new objective function for measuring independence of $X = (X_1, X_2)$ which lends itself well to kernel methods. We restrict to the 2-D case for simplicity.

$\mathcal{F}$-Correlation

For some vector (specifically Hilbert) space of functions $\mathcal{F}$, the $\mathcal{F}$-correlation is a generalised correlation in which we apply a family of functions to our random variables and use the optimal f which correlation is maximised, i.e

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \operatorname{corr}(f_1(X_1), f_2(X_2)) = \max_{f_1, f_2 \in \mathcal{F}} \frac{\operatorname{cov}(f_1(X_1), f_2(X_2))}{\sqrt{\operatorname{var}(f_1(X_1))\operatorname{var}(f_2(X_2))}} \tag{6}$$

This has the property that $\rho_{\mathcal{F}} = 0$ if and only if X_1 and X_2 are independent, so long as $\mathcal{F}$ is large enough.

Reproducing Kernel Hilbert Space

We say a Hilbert Space $\mathcal{H}$ consisting of functions on a set $\mathcal{X}$, we say $\mathcal{H}$ is a Reproducing Hilbert Space if there is some kernel $K : \mathcal{X}^2 \to \mathcal{H}$ such that for any $f \in \mathcal{H}$:

$$f(x) = \langle K(\cdot, x), f \rangle \tag{7}$$

Check that it makes sense that $K(\cdot, x) \in \mathcal{H}$

Moore-Aronszajn Theorem

For some $\mathcal{H}$, there is a unique kernel K for which $(\mathcal{H}, K)$ is an RKHS. Conversely, for any kernel on $\mathcal{X}$ we have a unique Hilbert space to create a RKHS.

This means that to choose a kernel on our dataset $X = \mathcal{X}$, means to choose a hilbert space to work over. Additionally for

Why do we care?

This allows us to perform the kernel trick!

The Gram Matrix for K

If K is a kernel on $\{x_i\}_{i=1}^n$, then define $K_{ij} = K(x_i, x_j)$ as the Gram Matrix for K.

Approximation for Covariance

If we assume the $\{f_1(x_i)\}_{i=1}^n$ is centered, then the sample covariance is:

$$\hat{\text{cov}}(f_1(x_i), f_2(x_j)) = \frac{1}{N}\sum_{i=1}^{N} f_1(x_i) f_2(x_j) \tag{8}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \langle K_1(\cdot, x_i), f_1\rangle \langle K_2(\cdot, x_j), f_2\rangle \tag{9}$$

By substituting in the reproducing property of the kernel K.

The final step verifying why the kernel trick works involves making a definition.

Basis expansion for f_1, f_2

Since $K_1(\cdot, x_i)$ and $K_2(\cdot, x_i)$ span a subspace of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, denote $f_1 = \alpha_1^T K_1(\cdot, x) + f_1^{\perp}$ and $f_2 = \alpha_2^T K_2(\cdot, x) + f_2^{\perp}$ as its decomposition, where $f_1^{\perp}$,$f_2^{\perp}$ denote the remainder terms.

This allows us to rewrite our inner products in terms of α_1,α_2:

$$\hat{\text{cov}}(f_1(x_i), f_2(x_j)) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_1^i K_1(x_1^i, x_2^k) K_2(x_2^j, x_2^k) \alpha_2^j \tag{10}$$

$$= \frac{1}{n} \alpha_1^T K_1 K_2 \alpha_2 \tag{11}$$

The Final Form

After applying the same argument to $\hat{\text{var}}$ we get the following expression to maximise over f_1, f_2:

$$\hat{\text{corr}}(f_1(x_1), f_2(x_2)) = \frac{\alpha_1^T K_1 K_2 \alpha_2}{\sqrt{\alpha_1^T K_1 K_2 \alpha_1 \alpha_2^T K_2 K_2 \alpha_2}} \tag{12}$$

Although notice that the the right hand side does not depend on f_1, f_2 directly. Recall that f_1 and f_2 can be defined in terms of α_1 and α_2. So we optimise the above function with respect to $\alpha = (\alpha_1, \alpha_2)$. Which in fact corresponds to a generalised eigenvalue problem.

In Practice

We have seen all the machinery that builds up to the alternate form of corr(), but how do we do it in practice?

Kernel ICA Algorithm

1. Whiten the Data to W
2. Choose K_1, ..., K_n kernels (potentially the same)
3. Perform a min-max optimisation to find the optimal $\hat{B}$, where we minimisation the $\mathcal{F}$-correlation of $G^T W (= S)$.