Canonical Correlation Analysis

What is CCA?

by Alvin Wan . alvinwan.com/cs189/fa17

First, the definition: CCA determines a subspace for us to project onto, that maximizes normalized covariance. That seems like a mouthful, so let's break this down. Consider this formula:

$$\rho(a,b) = \frac{\text{cov}(a,b)}{\sigma_a \sigma_b} = \frac{\text{cov}(a,b)}{\sqrt{\text{var}(a)\,\text{var}(b)}}$$

for random variables a, b. We want this quantity to be large, because the higher this number, the more information we have about a when we have b. This notion of "information" is hand-wavy, however. Let's make the motivation more concrete, first.

The formula $\rho(a,b)$ is called the **Pearson correlation coefficient (PCC)**. The expression has values between 1 and -1, where:

- 1 means positively correlated. Change in one random variable a means proportional change in b.
- 0 means no correlation.
- -1 means negative correlated. Change in one random variable a means a proportional change in the reverse direction for b.

More importantly, PCC tells us whether or not a, b are linearly related. Note that below, when I say "correlation", I really mean PCC.

1 Motivation

We need intuition for the fundamental question: why is maximizing correlation important? We can do this by showing where other approaches, like PCA, fail. Take

a simplified scenario, with two matrix-valued random variables X, Y. For simplicity, say $Y = X + \epsilon$ where noise ϵ has huge variance. What happens when we run PCA on Y? Since PCA maximizes variance, it will actually project Y (largely) into the column space of ϵ. However, we're interested in Y's relationship to X, not its dependence on noise. How can we fix this? As it turns out, CCA solves this issue. Instead of maximizing variance of Y, we maximize correlation between X and Y. In some sense, we want the maximize "predictive power" of information we have.

Although, how does correlation translate into "predictive power"? We can consider a perspective using linear algebra and another using independence, albeit weakly:

1. Take a, b to be vector-valued random variables, neither of which have zero mean. Then, we know that these vectors are orthogonal if and only if $E[ab] = 0$. Otherwise, the two vectors are linearly-related, and as a result, we can potentially model one random variable using the other. We expand on this below.

2. Take the extreme cases. When a, b are independent, we know $\text{cov}(a, b) = \rho(a, b) = 0$. In general, the converse does not necessarily hold, but *if* a, b *are jointly Gaussian*, $\text{cov}(a, b) = \rho(a, b) = 0 \iff a, b$ independent. Now, take $\text{cor}(a, b) = 1$. Then, we know that each change in a will see a corresponding change in b in the same direction with proportional magnitude.

The motivation should make both intuitive and mathematical sense now, so let's consider this a maximization problem. However, what are we maximizing over?

2 Optimization

Take vector-valued random variable $\vec{x}$. We want to project this vector onto a subspace. We'll characterize this new subspace using just one vector $\vec{u}$, so our projection is $\vec{u}^T\vec{x}$. Likewise, we can project y onto a subspace and we get $v^T y$. Then, our goal is to maximize the correlation between these projected vectors. We control the subspaces u, v, and we want to adjust them so that $u^T x, v^T y$ are correlated as much as possible. Formulated more explicitly, we have

$$\max_{u,v \in \mathbb{R}^d} \text{cor}(u^T x, v^T y)$$

This is a simple case where we consider just one vector, for each subspace. However, we can characterize each of these subspaces with more than one vector. We simply stack each new basis vector as columns in a matrix U, to obtain

$$\max_{U,V \in \mathbb{R}^{nxd}} \text{cor}(U^T x, V^T y)$$

3 Solve

Here, we have a simplified scenario, but the gist is the following:

1. We project x, y into subspaces.
2. Adjust those subspaces so we maximize correlation between the projected x, y.
3. Pick some vector(s) that characterizes those subspaces.

For simplicity, we are characterizing these subspaces by single vectors, as we can solve this optimization problem, one basis vector at a time. How is that? First, recall the iterative algorithm for PCA. For some matrix A:

1. Solve $x_i = \text{argmax}_{x:\|x\|=1} x^T A x$.
2. Find new A by removing components of A in the x_i direction. For each a_i, replace with $a_i - x_i \langle a_i, x_i \rangle$.
3. Repeat step 1 until you have k of the x_i vectors.

In discussion, we showed that CCA can be rewritten in a quadratic form.

$$\max_{u,v} \begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} 0 & X^T Y \\ Y^T X & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$
$$\text{subject to} \| \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \| = \sqrt{2}$$

Simply apply the same algorithm specified above, as we are again interested in finding the basis for a new subspace. Let's close with a question, to gauge your intuition for CCA.

Question

Take the following system, as proposed in class, where $x, y \in \mathbb{R}^{n \times d}$. For simplicity, take $U, V \in \mathbb{R}^{a \times d}, C \in \mathbb{R}^{b \times d}$. The remaining A, B, D, E are size-conforming matrices.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A & B & 0 \\ 0 & D & E \end{bmatrix} \begin{bmatrix} U \\ C \\ V \end{bmatrix}$$

Given this system, when would CCA demonstrate the most significant advantage over PCA, if we would like to predict y from x? Consider cases where $a >> b, a = b,$ and $a << b$.

Answer

When $a >> b$. We can start by rewriting this:

$$x = AU + BC, y = DC + EV$$

If it isn't apparent from the formulation above, it should be apparent here, that x, y share C. In this sense, the "coefficients" B, D contain the information we desire. A, E are simply "noise".

What does $a >> b$ mean? This means that U, V are much larger than C. By construction, the range of U, V are much larger than C's. In a stochastic interpretation, U, V then account for more variance. As a result, PCA would more likely project into spaces spanned by U, V. To maximize predictive power, however, we want to consider x, y in the range of C.