EECS 189 Introduction to Machine Learning
Fall 2020

Note 11

1 Canonical Correlation Analysis

PCA provided us with a dimensionality-reduction approach that didn't use the labels y in any way. In that way, it was fundamentally unsupervised by nature. However, we can imagine that there can be situations in which the most relevant directions in $\mathbf{x}$ for understanding $\mathbf{y}$ are not necessarily the directions of greatest variation in $\mathbf{x}$. For example, what if the $\mathbf{x}$ data by nature was contaminated with a strong correlated noise signal? PCA would find the noise dimensions to be those that have the greatest variation and keep them, throwing away those dimensions where we could actually hope to get information relevant for predicting $\mathbf{y}$!

The other potentially troublesome aspect of PCA is that it is not invariant to a change of units or scaling. If we changed the units of some feature from meters to millimeters, then all the values for that feature would increase by a factor of a thousand, and suddenly, this direction might be favored by PCA. This is unavoidable because there is no natural reference point that would allow us to treat units as arbitrary.

Consequently, it is important to have an approach to dimensionality reduction and the discovery of linear structure from data that does take advantage of paired $(\mathbf{x}, \mathbf{y})$ data, preferably in a way that is robust to linear transformations of both $\mathbf{x}$ and $\mathbf{y}$ individually.

1.1 A latent space view with Gaussian random variables

What does it mean to extract the linear structure establishing the underlying relationship between $\mathbf{X}$ and $\mathbf{Y}$, two vector-valued quantites of which we have many paired samples. To understand what this should mean, we need to construct a model. The first thing that we do is assume we have a joint distribution for X and Y as random variables. In practice, we won't have the random variables in distribution, just paired samples of them. But it is easier to start understanding what we want by assuming that we have the entire distribution. This corresponds to how well we think we can do given infinite amounts of training data. The next we do is assume a particular form for the random variables. Since we are interested in linear structure, jointly Gaussian random variables are a useful model.

Our goal is to extract the underlying relationship or commonality between $\mathbf{X}$ and $\mathbf{Y}$. To do this, we assume that we have three underlying iid standard Gaussian random vectors $\mathbf{Z}_J$ (representing the common/joint part), $\mathbf{Z}_X$ (representing the randomness that is purely in $\mathbf{X}$ and not shared by $\mathbf{Y}$), and $\mathbf{Z}_Y$ (representing the randomness that is purely in $\mathbf{Y}$ and not shared by $\mathbf{X}$). Then we can assume

that they are related by an underlying linear relationship:

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B} & \mathbf{0}\\ \mathbf{0} & \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{Z}_X\\ \mathbf{Z}_J\\ \mathbf{Z}_Y\end{bmatrix} \tag{1}$$

As is typical in these situations, there is going to be some ambiguity in choosing the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ matrices. But the important thing is that somehow the $\mathbf{B}$ and $\mathbf{C}$ matrices together capture the joint relationship between $\mathbf{X}$ and $\mathbf{Y}$.

How will such a joint relationship manifest in the joint distributions for $\mathbf{X}$ and $\mathbf{Y}$? To understand that, we should first consider the scalar case.

1.2 Correlation and Scalar Gaussians

For the scalar case, A, B, C, D are just real numbers. So, the joint distribution of X, Y is $\mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma = \begin{bmatrix}A^2 + B^2 & BC\\ BC & C^2 + D^2\end{bmatrix}$. The first thing that we notice is that we cannot disentangle B and C. The second is that the information about the joint relationship (which we know is encoded by B and C) is all in the cross-covariance term, not in the individual variance term. Recall that we want to pull out the relationship in a way that does not depend on any individual scaling or linear transformation that we apply to X and Y.

Here's a neat fact: if X and Y are jointly Gaussian, i.e.

$$\begin{bmatrix}X\\ Y\end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

then we can define a distribution on *individually normalized* X and Y and have their joint inter-relationship entirely captured by $\rho(X, Y)$. First write

$$\rho(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Then

$$\Sigma = \begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma y^2\end{bmatrix} = \begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}$$

so

$$\begin{aligned}\begin{bmatrix}\sigma_x^{-1} & 0\\ 0 & \sigma_y^{-1}\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix}\sigma_x^{-1} & 0\\ 0 & \sigma_y^{-1}\end{bmatrix}\Sigma\begin{bmatrix}\sigma_x^{-1} & 0\\ 0 & \sigma_y^{-1}\end{bmatrix}^\top\right)\\ &\sim \mathcal{N}\left(0, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)\end{aligned}$$

This ρ quantity is the signature of the joint inter-relationship of the X and Y random variables.

To make things explicit, once we have the $\rho = \frac{BC}{\sqrt{(A^2+B^2)(C^2+D^2)}}$, we can come up with many possible backstories for the latent picture behind the observed random variables. Here is one that splits the influence of the latent space proportionately.

$$A = \sigma_x \sqrt{1 - |\rho|} \tag{2}$$

$$B = \sigma_x \sqrt{|\rho|} \tag{3}$$

$$C = \sigma_y \text{sign}(\rho) \sqrt{|\rho|} \tag{4}$$

$$D = \sigma_y \sqrt{1 - |\rho|} \tag{5}$$

Because $A^2 + B^2 = \sigma_x^2$, $C^2 + D^2 = \sigma_y^2$, and $\rho = \frac{BC}{\sqrt{(A^2+B^2)(C^2+D^2)}}$, this works.

1.3 Pearson Correlation

Although we defined this ρ above for a pair of jointly Gaussian random variables, it is really about linear structure. The **Pearson Correlation Coefficient** $\rho(X, Y)$ is effectively a way to measure how linearly related (in other words, how well a linear model captures the relationship between) random variables X and Y.

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Here are some important facts about it:

- It is commutative: $\rho(X, Y) = \rho(Y, X)$
- It always lies between -1 and 1: $-1 \leq \rho(X, Y) \leq 1$
- It is completely invariant to affine transformations: for any $a, b, c, d \in \mathbb{R}$,

$$\begin{aligned}
\rho(aX + b, cY + d) &= \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{Var}(aX + b)\,\text{Var}(cY + d)}} \\
&= \frac{\text{Cov}(aX, cY)}{\sqrt{\text{Var}(aX)\,\text{Var}(cY)}} \\
&= \frac{a \cdot c \cdot \text{Cov}(X, Y)}{\sqrt{a^2\,\text{Var}(X) \cdot c^2\,\text{Var}(Y)}} \\
&= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} \\
&= \rho(X, Y)
\end{aligned}$$

The correlation is defined in terms of random variables rather than observed data. Assume now that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are vectors containing n independent observations of X and Y, respectively. Recall the **law of large numbers**, which states that for i.i.d. X_i with mean μ,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu \quad \text{as } n \to \infty$$

We can use this law to justify a sample-based approximation to the mean:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \approx \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where the bar indicates the sample average, i.e. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then as a special case we have

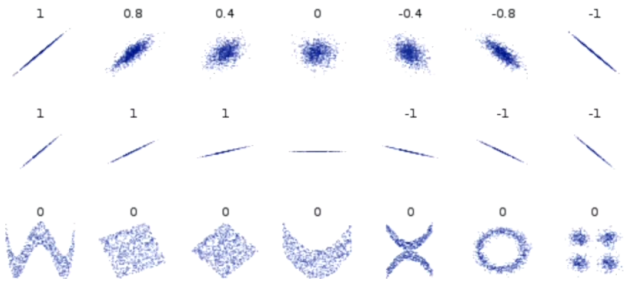
$$\text{Var}(X) = \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Var}(Y) = \text{Cov}(Y, Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Plugging these estimates into the definition for correlation and canceling the factor of $1/n$ leads us to the **Sample Pearson Correlation Coefficient** $\hat{\rho}$:

$$\begin{aligned}\hat{\rho}(x, y) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\tilde{x}^\top \tilde{y}}{\sqrt{\tilde{x}^\top \tilde{x} \cdot \tilde{y}^\top \tilde{y}}} \quad \text{where } \tilde{x} = x - \bar{x}, \tilde{y} = y - \bar{y}\end{aligned}$$

Here are some 2-D scatterplots and their corresponding correlation coefficients:



You should notice that:

- The magnitude of $\hat{\rho}$ increases as X and Y become more linearly correlated.
- The sign of $\hat{\rho}$ tells whether X and Y have a positive or negative relationship.
- The correlation coefficient is undefined if either X or Y has 0 variance (horizontal line).

1.4 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a method of modeling the relationship between two point sets by making use of the correlation coefficients.

As in PCA, it is useful to start with trying to find the directions that represent the most correlation. You can think of this as finding the parts of $\mathbf{X}_{\text{rv}}$ and $\mathbf{Y}_{\text{rv}}$ that depend on the first coordinate of $\mathbf{Z}_J$, where we choose the convention that the first coordinate represents the most shared dimension. We will then see how to move on to get the rest.

Formally, given zero-mean random vectors $\mathbf{X}_{\text{rv}} \in \mathbb{R}^p$ and $\mathbf{Y}_{\text{rv}} \in \mathbb{R}^q$, we want to find projection vectors $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$ that maximizes the correlation between $\mathbf{X}_{\text{rv}}{}^\top \mathbf{u}$ and $\mathbf{Y}_{\text{rv}}{}^\top \mathbf{v}$:

$$\max_{\mathbf{u},\mathbf{v}} \rho(\mathbf{X}_{\text{rv}}{}^\top \mathbf{u}, \mathbf{Y}_{\text{rv}}{}^\top \mathbf{v}) = \max_{\mathbf{u},\mathbf{v}} \frac{\text{Cov}(\mathbf{X}_{\text{rv}}{}^\top \mathbf{u}, \mathbf{Y}_{\text{rv}}{}^\top \mathbf{v})}{\sqrt{\text{Var}(\mathbf{X}_{\text{rv}}{}^\top \mathbf{u})\,\text{Var}(\mathbf{Y}_{\text{rv}}{}^\top \mathbf{v})}}$$

Observe that

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}{}^\top\mathbf{u}, \mathbf{Y}_{\mathrm{rv}}{}^\top\mathbf{v}) &= \mathbb{E}[(\mathbf{X}_{\mathrm{rv}}{}^\top\mathbf{u} - \mathbb{E}[\mathbf{X}_{\mathrm{rv}}{}^\top\mathbf{u}])(\mathbf{Y}_{\mathrm{rv}}{}^\top\mathbf{v} - \mathbb{E}[\mathbf{Y}_{\mathrm{rv}}{}^\top\mathbf{v}])] \\ &= \mathbb{E}[\mathbf{u}^\top(\mathbf{X}_{\mathrm{rv}} - \mathbb{E}[\mathbf{X}_{\mathrm{rv}}])(\mathbf{Y}_{\mathrm{rv}} - \mathbb{E}[\mathbf{Y}_{\mathrm{rv}}])^\top\mathbf{v}] \\ &= \mathbf{u}^\top\mathbb{E}[(\mathbf{X}_{\mathrm{rv}} - \mathbb{E}[\mathbf{X}_{\mathrm{rv}}])(\mathbf{Y}_{\mathrm{rv}} - \mathbb{E}[\mathbf{Y}_{\mathrm{rv}}])^\top]\mathbf{v} \\ &= \mathbf{u}^\top\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}})\mathbf{v}\end{aligned}$$

which also implies (since $\mathrm{Var}(Z) = \mathrm{Cov}(Z, Z)$ for any random variable Z) that

$$\begin{aligned}\mathrm{Var}(\mathbf{X}_{\mathrm{rv}}{}^\top\mathbf{u}) &= \mathbf{u}^\top\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{X}_{\mathrm{rv}})\mathbf{u} \\ \mathrm{Var}(\mathbf{Y}_{\mathrm{rv}}{}^\top\mathbf{v}) &= \mathbf{v}^\top\mathrm{Cov}(\mathbf{Y}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}})\mathbf{v}\end{aligned}$$

so the correlation can be written

$$\rho(\mathbf{X}_{\mathrm{rv}}{}^\top\mathbf{u}, \mathbf{Y}_{\mathrm{rv}}{}^\top\mathbf{v}) = \frac{\mathbf{u}^\top\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}})\mathbf{v}}{\sqrt{\mathbf{u}^\top\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{X}_{\mathrm{rv}})\mathbf{u} \cdot \mathbf{v}^\top\mathrm{Cov}(\mathbf{Y}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}})\mathbf{v}}}$$

Unfortunately, we do not have access to the true distributions of $\mathbf{X}_{\mathrm{rv}}$ and $\mathbf{Y}_{\mathrm{rv}}$, so we cannot compute these covariances matrices. However, we can estimate them from data. Assume now that we are given zero-mean data matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times q}$, where the rows of the matrix $\mathbf{X}$ are i.i.d. samples $\mathbf{x}_i \in \mathbb{R}^p$ from the random variable $\mathbf{X}_{\mathrm{rv}}$, and correspondingly for $\mathbf{Y}_{\mathrm{rv}}$. Then

$$\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}}) = \mathbb{E}[(\mathbf{X}_{\mathrm{rv}} - \underbrace{\mathbb{E}[\mathbf{X}_{\mathrm{rv}}]}_{0})(\mathbf{Y}_{\mathrm{rv}} - \underbrace{\mathbb{E}[\mathbf{Y}_{\mathrm{rv}}]}_{0})^\top] = \mathbb{E}[\mathbf{X}_{\mathrm{rv}}\mathbf{Y}_{\mathrm{rv}}{}^\top] \approx \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i{}^\top = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}$$

where again the sample-based approximation is justified by the law of large numbers. Similarly,

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}_{\mathrm{rv}}, \mathbf{X}_{\mathrm{rv}}) = \mathbb{E}[\mathbf{X}_{\mathrm{rv}}\mathbf{X}_{\mathrm{rv}}{}^\top] &\approx \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i{}^\top = \frac{1}{n}\mathbf{X}^\top\mathbf{X} \\ \mathrm{Cov}(\mathbf{Y}_{\mathrm{rv}}, \mathbf{Y}_{\mathrm{rv}}) = \mathbb{E}[\mathbf{Y}_{\mathrm{rv}}\mathbf{Y}_{\mathrm{rv}}{}^\top] &\approx \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i{}^\top = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}\end{aligned}$$

Plugging these estimates in for the true covariance matrices, we arrive at the problem

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{Y}\right)\mathbf{u}}{\sqrt{\mathbf{u}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{u} \cdot \mathbf{v}^\top\left(\frac{1}{n}\mathbf{Y}^\top\mathbf{Y}\right)\mathbf{v}}} = \max_{\mathbf{u},\mathbf{v}} \underbrace{\frac{\mathbf{u}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{v}}{\sqrt{\mathbf{u}^\top\mathbf{X}^\top\mathbf{X}\mathbf{u} \cdot \mathbf{v}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{v}}}}_{\hat{\rho}(\mathbf{X}\mathbf{u}, \mathbf{Y}\mathbf{v})}$$

Let's try to massage the maximization problem into a form that we can reason with more easily. Our strategy is to choose matrices to transform $\mathbf{X}$ and $\mathbf{Y}$ such that the maximization problem is equivalent but easier to understand.

1. First, let's choose matrices $\mathbf{W}_x, \mathbf{W}_y$ to **whiten** $\mathbf{X}$ and $\mathbf{Y}$. This will make the (co)variance matrices $(\mathbf{X}\mathbf{W}_x)^\top(\mathbf{X}\mathbf{W}_x)$ and $(\mathbf{Y}\mathbf{W}_y)^\top(\mathbf{Y}\mathbf{W}_y)$ become identity matrices and simplify our expression. To do this, note that $\mathbf{X}^\top\mathbf{X}$ is positive definite (and hence symmetric), so we can employ the eigendecomposition

$$\mathbf{X}^\top\mathbf{X} = \mathbf{U}_x\mathbf{S}_x\mathbf{U}_x{}^\top$$

Since

$$\mathbf{S}_x = \text{diag}(\lambda_1(\mathbf{X}^\top\mathbf{X}), \ldots, \lambda_d(\mathbf{X}^\top\mathbf{X}))$$

where all the eigenvalues are positive, we can define the "square root" of this matrix by taking the square root of every diagonal entry:

$$\mathbf{S}_x^{1/2} = \text{diag}\left(\sqrt{\lambda_1(\mathbf{X}^\top\mathbf{X})}, \ldots, \sqrt{\lambda_d(\mathbf{X}^\top\mathbf{X})}\right)$$

Then, defining $\mathbf{W}_x = \mathbf{U}_x\mathbf{S}_x^{-1/2}\mathbf{U}_x{}^\top$, we have

$$\begin{aligned}
(\mathbf{X}\mathbf{W}_x)^\top(\mathbf{X}\mathbf{W}_x) &= \mathbf{W}_x{}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}_x \\
&= \mathbf{U}_x\mathbf{S}_x^{-1/2}\mathbf{U}_x{}^\top\mathbf{U}_x\mathbf{S}_x\mathbf{U}_x{}^\top\mathbf{U}_x\mathbf{S}_x^{-1/2}\mathbf{U}_x{}^\top \\
&= \mathbf{U}_x\mathbf{S}_x^{-1/2}\mathbf{S}_x\mathbf{S}_x^{-1/2}\mathbf{U}_x{}^\top \\
&= \mathbf{U}_x\mathbf{U}_x{}^\top \\
&= \mathbf{I}
\end{aligned}$$

which shows that $\mathbf{W}_x$ is a whitening matrix for $\mathbf{X}$. The same process can be repeated to produce a whitening matrix $\mathbf{W}_y = \mathbf{U}_y\mathbf{S}_y^{-1/2}\mathbf{U}_y{}^\top$ for $\mathbf{Y}$.

Let's denote the whitened data $\mathbf{X}_w = \mathbf{X}\mathbf{W}_x$ and $\mathbf{Y}_w = \mathbf{Y}\mathbf{W}_y$. Then by the change of variables $\mathbf{u}_w = \mathbf{W}_x^{-1}\mathbf{u}, \mathbf{v}_w = \mathbf{W}_y^{-1}\mathbf{v}$,

$$\begin{aligned}
\max_{\mathbf{u},\mathbf{v}} \hat{\rho}(\mathbf{X}\mathbf{u}, \mathbf{Y}\mathbf{v}) &= \max_{\mathbf{u},\mathbf{v}} \frac{(\mathbf{X}\mathbf{u})^\top\mathbf{Y}\mathbf{v}}{\sqrt{(\mathbf{X}\mathbf{u})^\top\mathbf{X}\mathbf{u}(\mathbf{Y}\mathbf{v})^\top\mathbf{Y}\mathbf{v}}} \\
&= \max_{\mathbf{u},\mathbf{v}} \frac{(\mathbf{X}\mathbf{W}_x\mathbf{W}_x^{-1}\mathbf{u})^\top Y\mathbf{W}_y\mathbf{W}_y^{-1}\mathbf{v}}{\sqrt{(\mathbf{X}\mathbf{W}_x\mathbf{W}_x^{-1}\mathbf{u})^\top\mathbf{X}\mathbf{W}_x\mathbf{W}_x^{-1}\mathbf{u}(Y\mathbf{W}_y\mathbf{W}_y^{-1}\mathbf{v})^\top\mathbf{Y}\mathbf{W}_y\mathbf{W}_y^{-1}\mathbf{v}}} \\
&= \max_{\mathbf{u}_w,\mathbf{v}_w} \frac{(\mathbf{X}_w\mathbf{u}_w)^\top\mathbf{Y}_w\mathbf{v}_w}{\sqrt{(\mathbf{X}_w\mathbf{u}_w)^\top\mathbf{X}_w\mathbf{u}_w(\mathbf{Y}_w\mathbf{v}_w)^\top\mathbf{Y}_w\mathbf{v}_w}} \\
&= \max_{\mathbf{u}_w,\mathbf{v}_w} \frac{\mathbf{u}_w{}^\top\mathbf{X}_w{}^\top\mathbf{Y}_w\mathbf{v}_w}{\sqrt{\mathbf{u}_w{}^\top\mathbf{X}_w{}^\top\mathbf{X}_w\mathbf{u}_w \cdot \mathbf{v}_w{}^\top\mathbf{Y}_w{}^\top\mathbf{Y}_w\mathbf{v}_w}} \\
&= \max_{\mathbf{u}_w,\mathbf{v}_w} \underbrace{\frac{\mathbf{u}_w{}^\top\mathbf{X}_w{}^\top\mathbf{Y}_w\mathbf{v}_w}{\sqrt{\mathbf{u}_w{}^\top\mathbf{u}_w \cdot \mathbf{v}_w{}^\top\mathbf{v}_w}}}_{\hat{\rho}(\mathbf{X}_w\mathbf{u}_w, \mathbf{Y}_w\mathbf{v}_w)}
\end{aligned}$$

Note we have used the fact that $\mathbf{X}_w{}^\top\mathbf{X}_w$ and $\mathbf{Y}_w{}^\top\mathbf{Y}_w$ are identity matrices by construction.

2. Second, let's choose matrices $\mathbf{D}_x, \mathbf{D}_y$ to **decorrelate** $\mathbf{X}_w$ and $\mathbf{Y}_w$. This will let us simplify the covariance matrix $(\mathbf{X}_w\mathbf{D}_x)^\top(\mathbf{Y}_w\mathbf{D}_y)$ into a **diagonal** matrix.

   Recall that our ultimate goal is to understand the underlying latent structure behind $\mathbf{X}_{\text{rv}}$ and $\mathbf{Y}_{\text{rv}}$. The whitening was a normalizing change of coordinates. The decorrelation is there so that we can pick out independent underlying components of $\mathbf{Z}_J$. (Since jointly Gaussian random variables are independent if they are uncorrelated.) Alternatively, you can consider decorrelation as reducing the problem to a sequence of scalar problems.

To do this, we'll make use of the SVD:

$$\mathbf{X}_w{}^\top \mathbf{Y}_w = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$

The choice of $\mathbf{U}$ for $\mathbf{D}_x$ and $\mathbf{V}$ for $\mathbf{D}_y$ accomplishes our goal, since

$$(\mathbf{X}_w\mathbf{U})^\top(\mathbf{Y}_w\mathbf{V}) = \mathbf{U}^\top\mathbf{X}_w{}^\top\mathbf{Y}_w\mathbf{V} = \mathbf{U}^\top(\mathbf{U}\mathbf{S}\mathbf{V}^\top)\mathbf{V} = \mathbf{S}$$

Let's denote the decorrelated data $\mathbf{X}_d = \mathbf{X}_w\mathbf{D}_x$ and $\mathbf{Y}_d = \mathbf{Y}_w\mathbf{D}_y$. Then by the change of variables $\mathbf{u}_d = \mathbf{D}_x^{-1}\mathbf{u}_w = \mathbf{D}_x{}^\top\mathbf{u}_w$, $\mathbf{v}_d = \mathbf{D}_y^{-1}\mathbf{v}_w = \mathbf{D}_y{}^\top\mathbf{v}_w$,

$$\begin{aligned}
\max_{\mathbf{u}_w,\mathbf{v}_w} \hat{\rho}(\mathbf{X}_w\mathbf{u}_w, \mathbf{Y}_w\mathbf{v}_w) &= \max_{\mathbf{u}_w,\mathbf{v}_w} \frac{(\mathbf{X}_w\mathbf{u}_w)^\top\mathbf{Y}_w\mathbf{v}_w}{\sqrt{\mathbf{u}_w{}^\top\mathbf{u}_w \cdot \mathbf{v}_w{}^\top\mathbf{v}_w}} \\
&= \max_{\mathbf{u}_w,\mathbf{v}_w} \frac{(\mathbf{X}_w\mathbf{D}_x\mathbf{D}_x^{-1}\mathbf{u}_w)^\top\mathbf{Y}_w\mathbf{D}_y\mathbf{D}_y^{-1}\mathbf{v}_w}{\sqrt{(\mathbf{D}_x\mathbf{u}_w)^\top\mathbf{D}_x\mathbf{u}_w \cdot (\mathbf{D}_y\mathbf{v}_w)^\top\mathbf{D}_y\mathbf{v}_w}} \\
&= \max_{\mathbf{u}_d,\mathbf{v}_d} \frac{(\mathbf{X}_d\mathbf{u}_d)^\top\mathbf{Y}_d\mathbf{v}_d}{\sqrt{\mathbf{u}_d{}^\top\mathbf{u}_d \cdot \mathbf{v}_d{}^\top\mathbf{v}_d}} \\
&= \max_{\mathbf{u}_d,\mathbf{v}_d} \underbrace{\frac{\mathbf{u}_d{}^\top\mathbf{X}_d\mathbf{Y}_d\mathbf{v}_d}{\sqrt{\mathbf{u}_d{}^\top\mathbf{u}_d \cdot \mathbf{v}_d{}^\top\mathbf{v}_d}}}_{\hat{\rho}(\mathbf{X}_d\mathbf{u}_d, \mathbf{Y}_d\mathbf{v}_d)} \\
&= \max_{\mathbf{u}_d,\mathbf{v}_d} \frac{\mathbf{u}_d{}^\top\mathbf{S}\mathbf{v}_d}{\sqrt{\mathbf{u}_d{}^\top\mathbf{u}_d \cdot \mathbf{v}_d{}^\top\mathbf{v}_d}}
\end{aligned}$$

Without loss of generality, suppose $\mathbf{u}_d$ and $\mathbf{v}_d$ are unit vectors[1] so that the denominator becomes 1, and we can ignore it:

$$\max_{\mathbf{u}_d,\mathbf{v}_d} \frac{\mathbf{u}_d{}^\top S\,\mathbf{v}_d}{\sqrt{\mathbf{u}_d{}^\top\mathbf{u}_d \cdot \mathbf{v}_d{}^\top\mathbf{v}_d}} = \max_{\substack{\|\mathbf{u}_d\|=1 \\ \|\mathbf{v}_d\|=1}} \frac{\mathbf{u}_d{}^\top\mathbf{S}\mathbf{v}_d}{\|\mathbf{u}_d\|\|\mathbf{v}_d\|} = \max_{\substack{\|\mathbf{u}_d\|=1 \\ \|\mathbf{v}_d\|=1}} \mathbf{u}_d{}^\top\mathbf{S}\mathbf{v}_d$$

The diagonal nature of $\mathbf{S}$ implies $S_{ij} = 0$ for $i \neq j$, so our simplified objective expands as

$$\mathbf{u}_d{}^\top\mathbf{S}\mathbf{v}_d = \sum_i \sum_j (\mathbf{u}_d)_i S_{ij} (\mathbf{v}_d)_j = \sum_i S_{ii} (\mathbf{u}_d)_i (\mathbf{v}_d)_i$$

where S_{ii}, the singular values of $\mathbf{X}_w{}^\top\mathbf{Y}_w$, are arranged in descending order. Thus we have a weighted sum of these singular values, where the weights are given by the entries of $\mathbf{u}_d$ and $\mathbf{v}_d$, which are constrained to have unit norm. To maximize the sum, we "put all our eggs in one basket" and extract S_{11} by setting the first components of $\mathbf{u}_d$ and $\mathbf{v}_d$ to 1, and the rest to 0:

$$\mathbf{u}_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^p \qquad\qquad \mathbf{v}_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^q$$

[1] Why can we assume this? Observe that the value of the objective does not change if we replace $\mathbf{u}_d$ by $\alpha\mathbf{u}_d$ and $\mathbf{v}_d$ by $\beta\mathbf{v}_d$, where α and β are any positive constants. Thus if there are maximizers $\mathbf{u}_d, \mathbf{v}_d$ which are not unit vectors, then $\mathbf{u}_d/\|\mathbf{u}_d\|$ and $\mathbf{v}_d/\|\mathbf{v}_d\|$ (which are unit vectors) are also maximizers.

Any other arrangement would put weight on S_{ii} at the expense of taking that weight away from S_{11}, which is the largest, thus reducing the value of the sum.

Finally we have an analytical solution, but it is in a different coordinate system than our original problem! In particular, $\mathbf{u}_d$ and $\mathbf{v}_d$ are the best weights in a coordinate system where the data has been whitened and decorrelated. To bring it back to our original coordinate system and find the vectors we actually care about ($\mathbf{u}$ and $\mathbf{v}$), we must invert the changes of variables we made:

$$\mathbf{u} = \mathbf{W}_x \mathbf{u}_w = \mathbf{W}_x \mathbf{D}_x \mathbf{u}_d \qquad \mathbf{v} = \mathbf{W}_y \mathbf{v}_w = \mathbf{W}_y \mathbf{D}_y \mathbf{v}_d$$

More generally, to get the best k directions, we choose

$$\mathbf{U}_d = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{p-k,k} \end{bmatrix} \in \mathbb{R}^{p \times k} \qquad \mathbf{V}_d = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{q-k,k} \end{bmatrix} \in \mathbb{R}^{q \times k}$$

where I_k denotes the k-dimensional identity matrix. Then

$$\mathbf{U} = \mathbf{W}_x \mathbf{D}_x \mathbf{U}_d \qquad \mathbf{V} = \mathbf{W}_y \mathbf{D}_y \mathbf{V}_d$$

Note that $\mathbf{U}_d$ and $\mathbf{V}_d$ have orthogonal columns. The columns of $\mathbf{U}$ and $\mathbf{V}$, which are the projection directions we seek, will in general not be orthogonal, but they will be linearly independent (since they come from the application of invertible matrices to the columns of $\mathbf{U}_d, \mathbf{V}_d$).

Following (2), (3), (4), and (5), it is also possible to use what we have calculated to give an explicit learned latent-space realization for the $\mathbf{X}_{\text{rv}}, \mathbf{Y}_{\text{rv}}$ in terms of standard Gaussian random variables $\mathbf{Z}_X$, $\mathbf{Z}_J$, $\mathbf{Z}_Y$. In particular, matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ of the appropriate sizes. This is left as an exercise to the reader once you realize that after whitening and decorrelating (both invertible transformations), we are left with a collection of scalar problems that would represent independent random variables if all the variables were indeed jointly Gaussian.

CCA thus illustrates how it is possible to learn a latent representation for common (linear) structure given paired data. This is a powerful idea not limited to the specific case of CCA. In effect, CCA shows how we can discover (synthesize) features that distill what aspects of input data is relevant for understanding output data.

This is subtly different from what happens in ordinary least squares because in ordinary least squares, each individual element of $\mathbf{y}$ is predicted independently. In OLS, the different output variables are not used collectively to distill the most relevant dimensions of the input. By contrast, in CCA, the different output variables do vote collectively to determine relevant dimensions in the input.

1.5 Comparison with PCA

An advantage of CCA over PCA is that it is invariant to scalings and affine transformations of $\mathbf{X}$ and $\mathbf{Y}$. Consider a simplified scenario in which two matrix-valued random variables $\mathbf{X}, \mathbf{Y}$ satisfy $\mathbf{Y} = \mathbf{X} + \boldsymbol{\epsilon}$ where the noise $\boldsymbol{\epsilon}$ has huge variance. What happens when we run PCA on $\mathbf{Y}$? Since PCA maximizes variance, it will actually project $\mathbf{Y}$ (largely) into the column space of $\boldsymbol{\epsilon}$! However, we're interested in $\mathbf{Y}$'s relationship to $\mathbf{X}$, not its dependence on noise. How can we fix this? As it

turns out, CCA solves this issue. Instead of maximizing variance of $\mathbf{Y}$, we maximize correlation between $\mathbf{X}$ and $\mathbf{Y}$. In some sense, we want the maximize "predictive power" of information we have.

1.6 CCA regression

Once we've computed the CCA coefficients, one application is to use them for regression tasks, predicting $\mathbf{Y}$ from $\mathbf{X}$ (or vice-versa). Recall that the correlation coefficient attains a greater value when the two sets of data are *more linearly correlated*. Thus, it makes sense to find the $k \times k$ weight matrix $\mathbf{A}$ that linearly relates $\mathbf{XU}$ and $\mathbf{YV}$. We can accomplish this with ordinary least squares.

Denote the projected data matrices by $\mathbf{X}_c = \mathbf{XU}$ and $\mathbf{Y}_c = \mathbf{YV}$. Observe that $\mathbf{X}_c$ and $\mathbf{Y}_c$ are zero-mean because they are linear transformations of $\mathbf{X}$ and $\mathbf{Y}$, which are zero-mean. Thus we can fit a linear model relating the two:

$$\mathbf{Y}_c \approx \mathbf{X}_c\mathbf{A}$$

The least-squares solution is given by

$$\begin{aligned}\mathbf{A} &= (\mathbf{X}_c{}^\top\mathbf{X}_c)^{-1}\mathbf{X}_c{}^\top\mathbf{Y}_c \\ &= (\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{V}\end{aligned}$$

However, since what we *really* want is an estimate of $\mathbf{Y}$ given new (zero-mean) observations $\tilde{\mathbf{X}}$ (or vice-versa), it's useful to have the entire series of transformations that relates the two. The predicted canonical variables are given by

$$\hat{\mathbf{Y}}_c = \tilde{\mathbf{X}}_c\mathbf{A} = \tilde{\mathbf{X}}\mathbf{U}(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{V}$$

Then we use the canonical variables to compute the actual values:

$$\begin{aligned}\hat{\mathbf{Y}} &= \hat{\mathbf{Y}}_c(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top \\ &= \tilde{\mathbf{X}}\mathbf{U}(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U})^{-1}(\mathbf{U}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{V})(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\end{aligned}$$

We can collapse all these terms into a single matrix $\mathbf{A}_{\text{eq}}$ that gives the prediction $\hat{\mathbf{Y}}$ from $\tilde{\mathbf{X}}$:

$$\mathbf{A}_{\text{eq}} = \underbrace{\mathbf{U}}_{\text{projection}}\underbrace{(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U})^{-1}}_{\text{whitening}}\underbrace{(\mathbf{U}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{V})}_{\text{decorrelation}}\underbrace{(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top}_{\text{projection back}}$$