CS 189 Introduction to Machine Learning
Spring 2018 Note 11

# 1 Canonical Correlation Analysis

The **Pearson Correlation Coefficient** $\rho(X, Y)$ is 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 $\mu$,

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

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

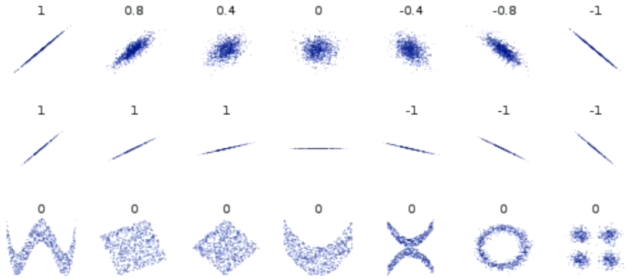
$$\text{Var}(X) = \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

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).

## 1.1 Correlation and Gaussians

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

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

then we can define a distribution on *normalized* $X$ and $Y$ and have their 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

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

## 1.2 Canonical Correlation Analysis

**Canonical Correlation Analysis (CCA)** is a method of modeling the relationship between two point sets by making use of the correlation coefficient. Formally, given zero-mean random vectors $X_{\text{rv}} \in \mathbb{R}^p$ and $Y_{\text{rv}} \in \mathbb{R}^q$, we want to find projection vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ that maximizes the correlation between ${X_{\text{rv}}}^\top u$ and ${Y_{\text{rv}}}^\top v$:

$$\max_{u,v} \rho({X_{\text{rv}}}^\top u, {Y_{\text{rv}}}^\top v) = \max_{u,v} \frac{\operatorname{Cov}({X_{\text{rv}}}^\top u, {Y_{\text{rv}}}^\top v)}{\sqrt{\operatorname{Var}({X_{\text{rv}}}^\top u)\operatorname{Var}({Y_{\text{rv}}}^\top v)}}$$

Observe that

$$\begin{aligned} \operatorname{Cov}({X_{\text{rv}}}^\top u, {Y_{\text{rv}}}^\top v) &= \mathbb{E}[({X_{\text{rv}}}^\top u - \mathbb{E}[{X_{\text{rv}}}^\top u])({Y_{\text{rv}}}^\top v - \mathbb{E}[{Y_{\text{rv}}}^\top v])] \\ &= \mathbb{E}[u^\top (X_{\text{rv}} - \mathbb{E}[X_{\text{rv}}])(Y_{\text{rv}} - \mathbb{E}[Y_{\text{rv}}])^\top v] \\ &= u^\top \mathbb{E}[(X_{\text{rv}} - \mathbb{E}[X_{\text{rv}}])(Y_{\text{rv}} - \mathbb{E}[Y_{\text{rv}}])^\top] v \\ &= u^\top \operatorname{Cov}(X_{\text{rv}}, Y_{\text{rv}}) v \end{aligned}$$

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

$$\operatorname{Var}({X_{\text{rv}}}^\top u) = u^\top \operatorname{Cov}(X_{\text{rv}}, X_{\text{rv}}) u$$
$$\operatorname{Var}({Y_{\text{rv}}}^\top v) = v^\top \operatorname{Cov}(Y_{\text{rv}}, Y_{\text{rv}}) v$$

so the correlation writes

$$\rho({X_{\text{rv}}}^\top u, {Y_{\text{rv}}}^\top v) = \frac{u^\top \operatorname{Cov}(X_{\text{rv}}, Y_{\text{rv}}) v}{\sqrt{u^\top \operatorname{Cov}(X_{\text{rv}}, X_{\text{rv}}) u \cdot v^\top \operatorname{Cov}(Y_{\text{rv}}, Y_{\text{rv}}) v}}$$

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

$$\operatorname{Cov}(X_{\text{rv}}, Y_{\text{rv}}) = \mathbb{E}[(X_{\text{rv}} - \underbrace{\mathbb{E}[X_{\text{rv}}]}_{0})(Y_{\text{rv}} - \underbrace{\mathbb{E}[Y_{\text{rv}}]}_{0})^\top] = \mathbb{E}[X_{\text{rv}} {Y_{\text{rv}}}^\top] \approx \frac{1}{n}\sum_{i=1}^{n} x_i {y_i}^\top = \frac{1}{n} X^\top Y$$

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

$$\mathrm{Cov}(X_{\text{rv}}, X_{\text{rv}}) = \mathbb{E}[X_{\text{rv}} X_{\text{rv}}^\top] \approx \frac{1}{n}\sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X^\top X$$

$$\mathrm{Cov}(Y_{\text{rv}}, Y_{\text{rv}}) = \mathbb{E}[Y_{\text{rv}} Y_{\text{rv}}^\top] \approx \frac{1}{n}\sum_{i=1}^n y_i y_i^\top = \frac{1}{n} Y^\top Y$$

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

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

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 $X$ and $Y$ such that the maximization problem is equivalent but easier to understand.

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

   $$X^\top X = U_x S_x U_x^\top$$

   Since

   $$S_x = \mathrm{diag}(\lambda_1(X^\top X), \ldots, \lambda_d(X^\top X))$$

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

   $$S_x^{1/2} = \mathrm{diag}\left(\sqrt{\lambda_1(X^\top X)}, \ldots, \sqrt{\lambda_d(X^\top X)}\right)$$

   Then, defining $W_x = U_x S_x^{-1/2} U_x^\top$, we have

   $$\begin{aligned}
   (XW_x)^\top (XW_x) &= W_x^\top X^\top X W_x \\
   &= U_x S_x^{-1/2} U_x^\top U_x S_x U_x^\top U_x S_x^{-1/2} U_x^\top \\
   &= U_x S_x^{-1/2} S_x S_x^{-1/2} U_x^\top \\
   &= U_x U_x^\top \\
   &= I
   \end{aligned}$$

   which shows that $W_x$ is a whitening matrix for $X$. The same process can be repeated to produce a whitening matrix $W_y = U_y S_y^{-1/2} U_y^\top$ for $Y$.

   Let's denote the whitened data $X_w = XW_x$ and $Y_w = YW_y$. Then by the change of variables $u_w = W_x^{-1} u, v_w = W_y^{-1} v$,

   $$\max_{u,v} \hat{\rho}(Xu, Yv) = \max_{u,v} \frac{(Xu)^\top Y v}{\sqrt{(Xu)^\top X u (Yv)^\top Y v}}$$

$$
\begin{aligned}
&= \max_{u,v} \frac{(XW_xW_x^{-1}u)^\top YW_yW_y^{-1}v}{\sqrt{(XW_xW_x^{-1}u)^\top XW_xW_x^{-1}u(YW_yW_y^{-1}v)^\top YW_yW_y^{-1}v}} \\
&= \max_{u_w,v_w} \frac{(X_wu_w)^\top Y_wv_w}{\sqrt{(X_wu_w)^\top X_wu_w(Y_wv_w)^\top Y_wv_w}} \\
&= \max_{u_w,v_w} \frac{{u_w}^\top {X_w}^\top Y_wv_w}{\sqrt{{u_w}^\top {X_w}^\top X_wu_w \cdot {v_w}^\top {Y_w}^\top Y_wv_w}} \\
&= \max_{u_w,v_w} \underbrace{\frac{{u_w}^\top {X_w}^\top Y_wv_w}{\sqrt{{u_w}^\top u_w \cdot {v_w}^\top v_w}}}_{\hat{\rho}(X_wu_w, Y_wv_w)}
\end{aligned}
$$

Note we have used the fact that ${X_w}^\top X_w$ and ${Y_w}^\top Y_w$ are identity matrices by construction.

2. Second, let's choose matrices $D_x$, $D_y$ to **decorrelate** $X_w$ and $Y_w$. This will let us simplify the covariance matrix $(X_wD_x)^\top(Y_wD_y)$ into a **diagonal** matrix. To do this, we'll make use of the SVD:

$$
{X_w}^\top Y_w = USV^\top
$$

The choice of $U$ for $D_x$ and $V$ for $D_y$ accomplishes our goal, since

$$
(X_wU)^\top(Y_wV) = U^\top {X_w}^\top Y_wV = U^\top(USV^\top)V = S
$$

Let's denote the decorrelated data $X_d = X_wD_y$ and $Y_d = Y_wW_y$. Then by the change of variables $u_d = D_x^{-1}u_w = {D_x}^\top u_w, v_d = D_y^{-1}v_w = {D_y}^\top v_w$,

$$
\begin{aligned}
\max_{u_w,v_w} \hat{\rho}(X_wu_w, Y_wv_w) &= \max_{u_w,v_w} \frac{(X_wu_w)^\top Y_wv_w}{\sqrt{{u_w}^\top u_w \cdot {v_w}^\top v_w}} \\
&= \max_{u_w,v_w} \frac{(X_wD_xD_x^{-1}u_w)^\top Y_wD_yD_y^{-1}v_w}{\sqrt{(D_xu_w)^\top D_xu_w \cdot (D_yv_w)^\top D_yv_w}} \\
&= \max_{u_d,v_d} \frac{(X_du_d)^\top Y_dv_d}{\sqrt{{u_d}^\top u_d \cdot {v_d}^\top v_d}} \\
&= \max_{u_d,v_d} \underbrace{\frac{{u_d}^\top X_dY_dv_d}{\sqrt{{u_d}^\top u_d \cdot {v_d}^\top v_d}}}_{\hat{\rho}(X_du_d, Y_dv_d)} \\
&= \max_{u_d,v_d} \frac{{u_d}^\top Sv_d}{\sqrt{{u_d}^\top u_d \cdot {v_d}^\top v_d}}
\end{aligned}
$$

Without loss of generality, suppose $u_d$ and $v_d$ are unit vectors[1] so that the denominator becomes 1, and we can ignore it:

$$
\max_{u_d,v_d} \frac{{u_d}^\top Sv_d}{\sqrt{{u_d}^\top u_d \cdot {v_d}^\top v_d}} = \max_{\substack{\|u_d\|=1 \\ \|v_d\|=1}} \frac{{u_d}^\top Sv_d}{\|u_d\|\|v_d\|} = \max_{\substack{\|u_d\|=1 \\ \|v_d\|=1}} {u_d}^\top Sv_d
$$

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

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

$$u_d^\top S v_d = \sum_i \sum_j (u_d)_i S_{ij} (v_d)_j = \sum_i S_{ii} (u_d)_i (v_d)_i$$

where $S_{ii}$, the singular values of ${X_w}^\top Y_w$, are arranged in descending order. Thus we have a weighted sum of these singular values, where the weights are given by the entries of $u_d$ and $v_d$, 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 $u_d$ and $v_d$ to 1, and the rest to 0:

$$u_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^p \qquad\qquad v_d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^q$$

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, $u_d$ and $v_d$ 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 ($u$ and $v$), we must invert the changes of variables we made:

$$u = W_x u_w = W_x D_x u_d \qquad\qquad v = W_y v_w = W_y D_y v_d$$

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

$$U_d = \begin{bmatrix} I_k \\ 0_{p-k,k} \end{bmatrix} \in \mathbb{R}^{p \times k} \qquad\qquad V_d = \begin{bmatrix} I_k \\ 0_{q-k,k} \end{bmatrix} \in \mathbb{R}^{q \times k}$$

where $I_k$ denotes the $k$-dimensional identity matrix. Then

$$U = W_x D_x U_d \qquad\qquad V = W_y D_y V_d$$

Note that $U_d$ and $V_d$ have orthogonal columns. The columns of $U$ and $V$, 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 $U_d, V_d$).

## 1.3 Comparison with PCA

An advantage of CCA over PCA is that it is invariant to scalings and affine transformations of $X$ and $Y$. Consider a simplified scenario in which two matrix-valued random variables $X, Y$ satisfy $Y = X + \epsilon$ where the noise $\epsilon$ 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 $\epsilon$! 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.

## 1.4 CCA regression

Once we've computed the CCA coefficients, one application is to use them for regression tasks, predicting $Y$ from $X$ (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 $A$ that linearly relates $XU$ and $YV$. We can accomplish this with ordinary least squares.

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

$$Y_c \approx X_c A$$

The least-squares solution is given by

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

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

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

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

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

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

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