IFT 6760A - Lecture 6
Dimensionality Reduction

Scribe(s): Charles Onu, Mohamed Abdelsalam **Instructor:** Guillaume Rabusseau

1 Summary

In the previous lecture, we reviewed the different matrix norms, and we covered the low rank approximation, including the Ecart-Young-Mirsky theorem. We introduced as well the Rayleigh-Ritz theorem, the Courant-Fischer (Min-max) theorem and the trace maximization/minimization.

In this lecture we review dimensionality reduction, mainly using Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Locally Linear Embedding (LLE). We show as well the relationship between these methods, and the theorems introduced during previous lectures.

2 Principal Component Analysis (PCA)

PCA is concerned with projecting some data onto a low dimensional subspace. We will present PCA from 2 perspectives which were separately developed by Hotelling, 1933 [5] and Pearson, 1901 [6]:

1. Maximization of the variance [5]
2. Minimization of the reconstruction error [6]

Definition 1 (Correlation)**.** *The Pearson* correlation *between two random variables X and Y is defined as their covariance, divided by the product of their standard deviations:*

$$\rho_{X,Y} = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}} = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{(\mathbb{E}[X^2]-\mathbb{E}[X]^2)(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)}}$$

Remark 2. *This means that:*

1. *if two random variables X and Y are independent, then $Cor(X,Y) = 0$*
2. *if X is a positive multiple of Y ($X = \alpha Y$ for $\alpha > 0$), then $Cor(X,Y) = 1$*

Proof.

$$
\begin{aligned}
1.\ Cor(X,Y) &= \frac{Cov(X,Y)}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}} \\
Cov(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[XY - X\,\mathbb{E}[Y] - Y\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] && \text{(for independent variables, } \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]) \\
&= \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0
\end{aligned}
$$

$$
\begin{aligned}
2.\ Cor(X,Y) &= \frac{Cov(X,Y)}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}} \\
&= \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{(\mathbb{E}[X^2]-\mathbb{E}[X]^2)(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)}} \\
&= \frac{\mathbb{E}[(\alpha Y-\mathbb{E}[\alpha Y])(Y-\mathbb{E}[Y])]}{\sqrt{(\mathbb{E}[\alpha^2 Y^2]-\mathbb{E}[\alpha Y]^2)(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)}} && \text{(using the linearity of expectation)} \\
&= \frac{\alpha\,\mathbb{E}[(Y-\mathbb{E}[Y])(Y-\mathbb{E}[Y])]}{\alpha\sqrt{(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)}} \\
&= \frac{\mathbb{E}[(Y-\mathbb{E}[Y])^2]}{\sqrt{(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)^2}} && \text{(using } \mathbb{V}[Z] = \mathbb{E}[(Z-\mathbb{E}[Z])^2] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2) \\
&= \frac{\mathbb{V}[Y]}{\sqrt{\mathbb{V}[Y]^2}} = 1
\end{aligned}
$$

□

Note however that although independence implies uncorrelation, uncorrelation doesn't necessarily imply independence. As the correlation coefficient defined above measures only the linear association between the random variables. For more elaboration, see [9] and [1].

2.1 Maximizing the variance

Let $\mathbf{x} \in \mathbb{R}^d$ be a random variable, which we would like to represent by a lower dimensional variable $\mathbf{y} \in \mathbb{R}^p$ that tries to keep as much information about $\mathbf{x}$ as possible. PCA achieves that by choosing each component $y_i = \mathbf{a}_i^T\mathbf{x}$ ($y_i \in \mathbb{R}$) so as to maximize the variance while being uncorrelated to the previous components. Thus for $i = 1, ..., p$, we need to find the $\mathbf{a}_i$ that satisfies:

$$\mathbf{a}_i = \underset{\mathbf{a}\in\mathbb{R}^d}{\arg\max}\, \mathbb{V}[\mathbf{a}^T\mathbf{x}] \quad \text{such that} \quad Cor(\mathbf{a}^T\mathbf{x}, \mathbf{a}_j^T\mathbf{x}) = 0 \quad \text{for all} \quad j < i$$

Note that for $\mathbb{V}[\mathbf{a}^T\mathbf{x}]$ to be bounded, the length of $\mathbf{a}$ needs to be constrained, so we impose an additional constraint $||\mathbf{a}||^2 = 1$, leading to:

$$\mathbf{a}_i = \underset{\substack{\mathbf{a}\in\mathbb{R}^d,\ ||\mathbf{a}||^2=1 \\ Cor(\mathbf{a}^T\mathbf{x},\mathbf{a}_j^T\mathbf{x})=0\ \forall\ j<i}}{\arg\max}\ \mathbb{V}[\mathbf{a}^T\mathbf{x}]$$

We have (note that $\mathbf{a}$ is not a random variable):

$$
\begin{aligned}
\mathbb{V}[\mathbf{a}^T\mathbf{x}] &= \mathbb{E}[(\mathbf{a}^T\mathbf{x} - \mathbb{E}[\mathbf{a}^T\mathbf{x}])^2] && \text{(using the linearity of expectation)} \\
&= \mathbb{E}[(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\,\mathbb{E}[\mathbf{x}])^2] \\
&= \mathbb{E}[\mathbf{a}^T(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\mathbf{a}] && \text{(again, using the linearity of expectation)} \\
&= \mathbf{a}^T\,\mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T]\mathbf{a} \\
&= \mathbf{a}^T\mathbf{C}_{\mathbf{xx}}\mathbf{a} && (\mathbf{C}_{\mathbf{xx}} = Cov(\mathbf{x},\mathbf{x}))
\end{aligned}
$$

hence we have:

$$\mathbf{a}_1 = \underset{\mathbf{a}\in\mathbb{R}^d,\ ||\mathbf{a}||^2=1}{\arg\max}\ \mathbf{a}^T\mathbf{C_{xx}}\mathbf{a}$$

Therefore, $\mathbf{a}_1$ is the unit eigenvector associated with the top eigenvalue of $\mathbf{C_{xx}}$

Theorem 3. *$\mathbf{a}_i$ is the unit eigenvector associated with the i_{th} eigenvalue of $\mathbf{C_{xx}}$*

Proof. For the base case of $i = 1$, we can rewrite

$$\mathbf{a}_1 = \underset{\mathbf{a}\in\mathbb{R}^d,\ ||\mathbf{a}||^2=1}{\arg\max}\ \mathbf{a}^T\mathbf{C_{xx}}\mathbf{a}$$

using Lagrange multipliers as: (see [4] for a review on Lagrange multipliers)

$$\mathbf{a}_1 = \underset{\mathbf{a}\in\mathbb{R}^d}{\arg\max}\, L = \underset{\mathbf{a}\in\mathbb{R}^d}{\arg\max}\, \mathbf{a}^T\mathbf{C_{xx}}\mathbf{a} + \alpha(1 - \mathbf{a}^T\mathbf{a})$$

we can get the maximum by differentiating L with respect to a and setting the result to zero

$$\begin{aligned}
\frac{dL}{d\mathbf{a}} &= 2\mathbf{a}^T\mathbf{C_{xx}} - 2\alpha\mathbf{a}^T = 0 \\
\mathbf{a}^T\mathbf{C_{xx}} &= \alpha\mathbf{a}^T \quad \text{(taking the transpose, and as } \mathbf{C_{xx}} \text{ is symmetric)} \\
\mathbf{C_{xx}}\mathbf{a} &= \alpha\mathbf{a}
\end{aligned}$$

hence substituting the (unit) eigenvector/eigenvalue pairs for $\mathbf{a}$ and α satisfy the stationary points, with the top eigenvalue and its corresponding eigenvector giving the maximum. This gave the result for $\mathbf{a}_1$, for the subsequent $\mathbf{a}_i$, we have to satisfy another condition:

$$\begin{aligned}
Cor(\mathbf{a}_i^T\mathbf{x}, \mathbf{a}_j^T\mathbf{x}) &= 0 \quad \text{for all} \quad j < i \\
\frac{Cov(\mathbf{a}_i^T\mathbf{x}, \mathbf{a}_j^T\mathbf{x})}{\sqrt{\mathbb{V}[\mathbf{a}_i^T\mathbf{x}]\,\mathbb{V}[\mathbf{a}_j^T\mathbf{x}]}} &= 0 \\
Cov(\mathbf{a}_i^T\mathbf{x}, \mathbf{a}_j^T\mathbf{x}) &= 0 \\
\mathbf{a}_i^T\mathbf{C_{xx}}\mathbf{a}_j &= 0 \quad \text{(using the result obtained above, } \mathbf{C_{xx}}\mathbf{a}_j = \lambda_j\mathbf{a}_j) \\
\lambda_j\mathbf{a}_i^T\mathbf{a}_j &= 0 \\
\mathbf{a}_i^T\mathbf{a}_j &= 0
\end{aligned}$$

which can be satisfied by choosing $\mathbf{a}_i$ to be the i_{th} unit eigenvector of $\mathbf{C_{xx}}$ (remember that $\mathbf{C_{xx}}$ is a symmetric matrix with d orthogonal eigenvectors). For more details, see [8] □

Another way to look at the previous proof is by using the Rayleigh-Ritz theorem, which states that:

$$\max_{||\mathbf{x}||_2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_1\ with\ \mathbf{x} = \mathbf{e}_1$$

$$\max_{\substack{||\mathbf{x}||_2=1 \\ \mathbf{x}\in span(\mathbf{e}_1,..,\mathbf{e}_{i-1})^T}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_i\ with\ \mathbf{x} = \mathbf{e}_i$$

Where $\mathbf{e}_i$ is the unit eigenvector corresponding to the i_{th} eigenvalue.

PCA in Practice

Let's say we have an input $\mathbf{X} \in \mathbb{R}^{d\times N}$, where $\mathbf{X} = \begin{pmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{pmatrix}$, with $\mathbf{x}_i \in \mathbb{R}^d$. We would like to have an output $\mathbf{Y} = \begin{pmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_N\end{pmatrix}$, with $\mathbf{y}_i \in \mathbb{R}^p$, capturing as much information about $\mathbf{X}$ as we can while keeping $p << d$.

First, we would like to make sure that our data is centered (i.e. $\sum_{i=1}^{N} = \mathbf{0}$), which can easily be done during preprocessing. This way, the empirical estimate of the covariance $\hat{\mathbf{C}}_{XX}$ matrix could be calculated as $\hat{\mathbf{C}}_{XX} = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$.
As $\hat{\mathbf{C}}_{XX}$ is a symmetric matrix, we can compute an eigenvalue decomposition for it:

$$\hat{\mathbf{C}}_{XX} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

$$\mathbf{U} = \begin{pmatrix}\mathbf{e}_1 \cdots \mathbf{e}_p \cdots \mathbf{e}_d\end{pmatrix} = \begin{pmatrix}\mathbf{U}_p & \tilde{\mathbf{U}}\end{pmatrix}$$

Where $\mathbf{U}_p$ contains the first p eigenvectors as its columns. Therefore $\mathbf{Y}$ can be obtained using $\mathbf{Y} = \mathbf{U}_p^T\mathbf{X}$.

Another way for applying PCA is through using Singular Value Decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{pmatrix}\mathbf{U}_p & \tilde{\mathbf{U}}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_p & \\ & \tilde{\mathbf{\Sigma}}\end{pmatrix}\begin{pmatrix}\mathbf{V}_p^T \\ \tilde{\mathbf{V}}^T\end{pmatrix}$$

$$\hat{\mathbf{C}}_{XX} = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \frac{1}{N}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \frac{1}{N}\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

Which means that the columns of $\mathbf{U}$ are still the principal directions, and the relationship between the eigenvalues of $\hat{\mathbf{C}}_{XX}$ and its singular values is $\mathbf{D} = \frac{1}{N}\mathbf{\Sigma}$. $\mathbf{Y}$ can be obtained through:

$$\begin{aligned}
\mathbf{Y} = \mathbf{U}_p^T\mathbf{X} &= \mathbf{U}_p^T\begin{pmatrix}\mathbf{U}_p & \tilde{\mathbf{U}}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_p & \\ & \tilde{\mathbf{\Sigma}}\end{pmatrix}\begin{pmatrix}\mathbf{V}_p^T \\ \tilde{\mathbf{V}}^T\end{pmatrix} \\
&= \begin{pmatrix}\mathbf{U}_p^T\mathbf{U}_p & \mathbf{U}_p^T\tilde{\mathbf{U}}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_p & \\ & \tilde{\mathbf{\Sigma}}\end{pmatrix}\begin{pmatrix}\mathbf{V}_p^T \\ \tilde{\mathbf{V}}^T\end{pmatrix} \\
&= \begin{pmatrix}\mathbf{I}_p & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_p & \\ & \tilde{\mathbf{\Sigma}}\end{pmatrix}\begin{pmatrix}\mathbf{V}_p^T \\ \tilde{\mathbf{V}}^T\end{pmatrix} \\
&= \mathbf{\Sigma}_p\mathbf{V}_p^T
\end{aligned}$$

2.2 PCA as a Minimization Problem

PCA also can be thought of as a minimization of the reconstruction error. We would like to find a projection $\mathbf{\Pi} = \mathbf{U}\mathbf{U}^T$ that projects our data in $\mathbb{R}^d$ onto a p-dimensional subspace in $\mathbb{R}^d$, where $\mathbf{U} \in \mathbb{R}^{d\times p}$ is an orthogonal matrix. We want to minimize the following loss:

$$\begin{aligned}
l &= \sum_{i=1}^{N}||\mathbf{\Pi}(\mathbf{x}_i) - \mathbf{x}_i||^2 \\
&= ||\mathbf{\Pi}\mathbf{X} - \mathbf{X}||_F^2 && \text{(remember, } \mathbf{X} \in \mathbb{R}^{d\times N}\text{)} \\
&= Tr((\mathbf{\Pi}\mathbf{X} - \mathbf{X})^T(\mathbf{\Pi}\mathbf{X} - \mathbf{X})) \\
&= Tr(\mathbf{X}^T\mathbf{\Pi}^T\mathbf{\Pi}\mathbf{X} - \mathbf{X}^T\mathbf{\Pi}^T\mathbf{X} - \mathbf{X}^T\mathbf{\Pi}\mathbf{X} + \mathbf{X}^T\mathbf{X}) && (\mathbf{\Pi}^T = \mathbf{\Pi},\ \mathbf{\Pi}^T\mathbf{\Pi} = \mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{\Pi}) \\
&= Tr(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{\Pi}\mathbf{X}) && \text{(Trace is a linear mapping)} \\
&= Tr(\mathbf{X}^T\mathbf{X}) - Tr(\mathbf{X}^T\mathbf{\Pi}\mathbf{X}) \\
&= Tr(\mathbf{X}^T\mathbf{X}) - Tr(\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X})
\end{aligned}$$

We can now re-express the minimization problem as follows:

$$\begin{aligned}
\underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\min}\, l &= \underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\min}(Tr(\mathbf{X}^T\mathbf{X}) - Tr(\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X})) && (\mathbf{X}^T\mathbf{X} \text{ is not a function of } \mathbf{U}) \\
&= \underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\min}(-Tr(\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X})) \\
&= \underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\max} Tr(\mathbf{X}^T\mathbf{U}\mathbf{U}^T\mathbf{X}) && (\text{using cyclic property of trace}) \\
&= \underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\max} Tr(\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U}) \\
&= \underset{\substack{\mathbf{U}\in\mathbb{R}^{d\times p}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}}{\arg\max} NTr(\mathbf{U}^T\mathbf{C}_{\mathbf{xx}}\mathbf{U})
\end{aligned}$$

From the Trace maximization theorem we saw in the previous lecture, we can get that the columns of the resulting $\mathbf{U}$ consists of the unit eigenvectors corresponding to the p highest eigenvalues of $\mathbf{C}_{\mathbf{xx}}$.

3 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) provides another framework for dimensionality reduction. Here we have 2 random variables $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{y} \in \mathbb{R}^{d_2}$ for which we want to derive 2 low-dimensional representations $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathbb{R}^p$ whose correlation is maximized. One can think of $\mathbf{x}$ and $\mathbf{y}$ as two views of the same object; for example CCA can be used to combine information from both audio and lip features in a speaker identification task [3]. More examples of practical problems where this framework applies to can be found for example in [2]. The idea behind CCA is very close to PCA: we want to find uncorrelated components for each view, so as to maximize the correlation between the components from the two views (instead of maximizing the variance in PCA). Formally,

$$(\mathbf{u}_i, \mathbf{v}_i) = \underset{\substack{\mathbf{u}\in\mathbb{R}^{d_1},\, \mathbf{v}\in\mathbb{R}^{d_2}\\ Cor(\mathbf{u}^T\mathbf{x}, \mathbf{u}_j^T\mathbf{x})=0 \forall j<i\\ Cor(\mathbf{v}^T\mathbf{y}, \mathbf{v}_j^T\mathbf{y})=0 \forall j<i}}{\arg\max} Cor(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

Let us assume, without loss of generality, that $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{y}] = 0$, then

$$\begin{aligned}
Cor(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) &= \frac{\mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})]}{\sqrt{\mathbb{E}[(\mathbf{u}^T\mathbf{x})^2]\mathbb{E}[(\mathbf{v}^T\mathbf{y})^2]}} \\
&= \frac{\mathbf{u}^T[\mathbf{x}\mathbf{y}^T]\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbb{E}[\mathbf{x}\mathbf{x}^T]\mathbf{u}}\sqrt{\mathbf{v}^T\mathbb{E}[\mathbf{y}\mathbf{y}^T]\mathbf{v}}} \\
&= \frac{\mathbf{u}^T\mathbf{C}_{\mathbf{xy}}\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{C}_{\mathbf{xx}}\mathbf{u}}\sqrt{\mathbf{v}^T\mathbf{C}_{\mathbf{yy}}\mathbf{v}}}
\end{aligned}$$

We can re-write our objective as:

$$(\mathbf{u}_i, \mathbf{v}_i) = \underset{\substack{\mathbf{u}\in\mathbb{R}^{d_1},\, \mathbf{v}\in\mathbb{R}^{d_2}\\ \mathbf{u}^T\mathbf{C}_{\mathbf{xx}}\mathbf{u}=1\\ \mathbf{v}^T\mathbf{C}_{\mathbf{yy}}\mathbf{v}=1}}{\arg\max}\, \mathbf{u}^T\mathbf{C}_{\mathbf{xy}}\mathbf{v}$$

Let $\mathbf{C}_{\mathbf{xx}}{}^{\frac{1}{2}} \in \mathbb{R}^{d_1\times d_1}$ and $\mathbf{C}_{\mathbf{yy}}{}^{\frac{1}{2}} \in \mathbb{R}^{d_2\times d_2}$ be such that $\mathbf{C}_{\mathbf{xx}} = (\mathbf{C}_{\mathbf{xx}}{}^{\frac{1}{2}})^T\mathbf{C}_{\mathbf{xx}}{}^{\frac{1}{2}}$ and $\mathbf{C}_{\mathbf{yy}} = (\mathbf{C}_{\mathbf{yy}}{}^{\frac{1}{2}})^T\mathbf{C}_{\mathbf{yy}}{}^{\frac{1}{2}}$. Observe that the existence of such matrices follows from the fact that $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{C}_{\mathbf{yy}}$ are positive semi-definite. Now let us set $\mathbf{a} = \mathbf{C}_{\mathbf{xx}}{}^{\frac{1}{2}}\mathbf{u}$ and $\mathbf{b} = \mathbf{C}_{\mathbf{yy}}{}^{\frac{1}{2}}\mathbf{v}$. Then, we have that:

$$(\mathbf{a}_i, \mathbf{b}_i) = \underset{\substack{\mathbf{a}\in\mathbb{R}^{d_1},\ \mathbf{b}\in\mathbb{R}^{d_2} \\ ||\mathbf{a}||=1,\ ||\mathbf{b}||=1}}{\arg\max} \mathbf{a}^T \underbrace{\mathbf{C_{xx}}^{-\frac{1}{2}} \mathbf{C_{xy}} \mathbf{C_{yy}}^{-\frac{1}{2}}}_{\mathbf{\Omega}} \mathbf{b}$$

We assumed here that $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ are invertible/full rank, which is a valid assumption as we can always get rid of any redundant variables beforehand (for example using PCA and discarding components associated with zero eigenvalues of the covariance matrix). Then we have:

$$(\mathbf{a}_i, \mathbf{b}_i) = \underset{\substack{\mathbf{a}\in\mathbb{R}^{d_1},\ \mathbf{b}\in\mathbb{R}^{d_2} \\ ||\mathbf{a}||=1,\ ||\mathbf{b}||=1}}{\arg\max} \mathbf{a}^T \mathbf{\Omega} \mathbf{b}$$

The solution $(\mathbf{a}_1, \mathbf{b}_1)$ is given by the top left/right singular vectors of $\mathbf{\Omega}$. More generally, The solution $(\mathbf{a}_i, \mathbf{b}_i)$ is given by the left/right singular vectors corresponding to the top i_{th} singular value of $\mathbf{\Omega}$.

Proof. Using the Lagrangian form of the maximization problem, we have:

$$L = \mathbf{a}^T \mathbf{\Omega} \mathbf{b} + \alpha(1 - \mathbf{a}^T \mathbf{a}) + \beta(1 - \mathbf{b}^T \mathbf{b})$$

By differentiation with respect to $\mathbf{a}$ and $\mathbf{b}$ and setting the result to zero, we get:

$$\mathbf{\Omega} \mathbf{b} = \alpha \mathbf{a}$$

$$\mathbf{\Omega}^T \mathbf{a} = \beta \mathbf{b}$$

Multiplying both by $\mathbf{a}^T$ and $\mathbf{b}^T$, we get:

$$\mathbf{a}^T \mathbf{\Omega} \mathbf{b} = \alpha \mathbf{a}^T \mathbf{a} = \alpha$$

$$\mathbf{b}^T \mathbf{\Omega}^T \mathbf{a} = \beta \mathbf{b}^T \mathbf{b} = \beta$$

from which it follows that $\alpha = \beta$ is a singular value of ω with corresponding left and right singular vectors $\mathbf{a}$ and $\mathbf{b}$. Consequently, the solution to the maximization problem $(\mathbf{a}_1, \mathbf{b}_1)$ is given by the top left/right singular vectors.
For the subsequent $(\mathbf{a}_i, \mathbf{b}_i)$, we have:

$$Cor(\mathbf{u}_i^T \mathbf{x}, \mathbf{u}_j^T \mathbf{x}) = Cor((\mathbf{C_{xx}}^{\frac{-1}{2}} \mathbf{a}_i)^T \mathbf{x}, (\mathbf{C_{xx}}^{\frac{-1}{2}} \mathbf{a}_j)^T \mathbf{x}) = \mathbf{a}_i^T \mathbf{a}_j = 0 \quad \text{for all} \quad j < i$$

$$Cor(\mathbf{v}_i^T \mathbf{y}, \mathbf{v}_j^T \mathbf{y}) = Cor((\mathbf{C_{yy}}^{\frac{-1}{2}} \mathbf{b}_i)^T \mathbf{y}, (\mathbf{C_{yy}}^{\frac{-1}{2}} \mathbf{b}_j)^T \mathbf{y}) = \mathbf{b}_i^T \mathbf{b}_j = 0 \quad \text{for all} \quad j < i$$

Which can be satisfied (along with the maximization problem) by choosing $(\mathbf{a}_i, \mathbf{b}_i)$ as the left/right singular vectors corresponding to the top i_{th} singular value. □

4 Locally Linear Embedding (LLE)

Whereas PCA and CCA are linear dimensionality reduction techniques, LLE or locally linear embedding [7] is a framework for non-linear dimensionality reduction. The principal idea is, given a set of points in high dimension, we reconstruct each point as a linear combination of its k neighbours.
Given N points $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^d$, we want to find low-dimensional embeddings $\mathbf{y}_1, ..., \mathbf{y}_N \in \mathbb{R}^p$ (where $p << d$) that preserve the local structure of the original points. To do this, we will carry out 2 main steps:

1. Compute the weights w_{ij} which best reconstruct each point $\mathbf{x}_i$ from its k neighbours.

$$E(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{x}_i - \sum_{j=1}^{N} w_{ij} \mathbf{x}_j||^2 \tag{1}$$

   subject to $w_{ij} \neq 0 \iff \mathbf{x}_j$ is a neighbour of $\mathbf{x}_i$ and $\forall i, \sum_{j=1}^{N} w_{ij} = 1$

   Remark: It can be shown mathematically from the objective function that the solution of $\min_{\mathbf{W}} E(\mathbf{W})$ is invariant to scaling, rotation, and translation.

2. Then, find the low-dimensional representations $\mathbf{y}_1, ..., \mathbf{y}_N \in \mathbb{R}^p$ which minimize

$$\min_{\mathbf{y}_1,...,\mathbf{y}_N \in \mathbb{R}^p} \sum_{i=1}^{N} ||\mathbf{y}_i - \sum_{j=1}^{N} w_{ij}\mathbf{y}_j||^2 \tag{2}$$

We will now show how both optimization problem can be solved easily using tools from linear algebra we saw in previous lectures. To solve for (2), let us express the objective with matrices:

$$\min_{\mathbf{Y} \in \mathbb{R}^{N \times p}} ||\mathbf{Y} - \mathbf{W}\mathbf{Y}||_F^2$$

where the rows of $\mathbf{Y}$ are the low-dimensional embeddings $\mathbf{y}_1, ..., \mathbf{y}_N$ we are seeking.
This objective is not well-posed because we can choose the trivial embedding of all zeros $\mathbf{Y} = \mathbf{0}$ which solves it. However observe that, similarly to the first optimization function, the cost function in Eq. (2) is invariant to translations, rotations and scaling, we can thus without loss of generality enforce that the embeddings are centered ($\sum_{i=1}^{N} \mathbf{y}_i = 0$) and that they have unit covariance ($\sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}$). Therefore, we consider the constrained optimization problem

$$\min_{\substack{\mathbf{Y} \in \mathbb{R}^{N \times p} \\ \mathbf{Y^T Y = I} \\ \mathbf{Y^T 1 = 0}}} ||\mathbf{Y} - \mathbf{W}\mathbf{Y}||_F^2 = \min_{\substack{\mathbf{Y} \in \mathbb{R}^{N \times p} \\ \mathbf{Y^T Y = I} \\ \mathbf{Y^T 1 = 0}}} Tr(\mathbf{Y^T} \underbrace{(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})}_{\mathbf{M}} \mathbf{Y}) = \min_{\substack{\mathbf{Y} \in \mathbb{R}^{N \times p} \\ \mathbf{Y^T Y = I} \\ \mathbf{Y^T 1 = 0}}} Tr(\mathbf{Y^T M Y})$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector of all ones. Using again the trace maximization/minimization result from the previous lecture, the solution of this problem will be given by choosing the bottom eigenvectors of the matrix $\mathbf{M}$. To conclude, first observe that $\mathbf{1}$ is an eigenvector of $\mathbf{M}$ for the eigenvalue 0 since $\mathbf{W1} = \mathbf{1}$. We can thus take the $p + 1$ bottom eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_{p+1} \in \mathbb{R}^N$ of $\mathbf{M}$ and discard the first one: since all the remaining eigenvectors will be orthogonal to $\mathbf{v}_1 = \mathbf{1}$ we have that the constraint $\mathbf{Y^T 1} = \mathbf{0}$ is satisfied as well.
To solve for (1), let $\mathbf{x}$ be one of the points, and let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_k \in \mathbb{R}^d$ be its k neighbours. We want to solve

$$\min_{\substack{w_1,..,w_k \in \mathbb{R} \\ \sum_{j=1}^{k} w_j = 1}} ||\mathbf{x} - \sum_{j=1}^{N} w_j \boldsymbol{\eta}_j||^2 = \min_{\substack{w_1,..,w_k \in \mathbb{R} \\ \sum_{j=1}^{k} w_j = 1}} ||\sum_{j=1}^{N} w_j(\mathbf{x} - \boldsymbol{\eta}_j)||^2 = \min_{\substack{w_1,..,w_k \in \mathbb{R} \\ \sum_{j=1}^{k} w_j = 1}} \mathbf{w}^T \mathbf{C} \mathbf{w}$$

where $\mathbf{w} \in \mathbb{R}^k$ and the matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$ is defined by $\mathbf{C}_{ij} = (\mathbf{x} - \boldsymbol{\eta}_i)^T(\mathbf{x} - \boldsymbol{\eta}_j)$. We first form the lagrangian

$$L = \mathbf{w}^T \mathbf{C} \mathbf{w} + \lambda(\mathbf{1}^T \mathbf{w} - 1)$$

By setting $\nabla_w L = 0$ and $\frac{\partial}{\partial \lambda} L = 0$ we obtain

$$\mathbf{w} = \frac{\mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}$$

References

[1] Uncorrelated vs. independent. URL `https://www.stat.cmu.edu/~cshalizi/uADA/13/reminders/uncorrelated-vs-independent.pdf`.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[3] R. Arora and K. Livescu. Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7135–7139. IEEE, 2013.

[4] C. Bishop. *Pattern Recognition and Machine Learning*, chapter Appendix E. Springer, 2011. URL `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/CourseBios362/LagrangeMultipliers-Bishop-PatternRecognitionMachineLearning.pdf`.

[5] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[6] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11): 559–572, 1901. URL http://www.stats.org.uk/pca/Pearson1901.pdf.

[7] L. K. Saul and S. T. Roweis. An introduction to locally linear embedding. *unpublished. Available at: http://www. cs. toronto. edu/~ roweis/lle/publications. html*, 2000.

[8] K. Stratos. A hitchhikers guide to pca and cca. URL http://karlstratos.com/notes/pca_cca.pdf.

[9] M. Ventura. Why zero correlation does not necessarily imply independence. URL https://stats.stackexchange.com/q/179537.