Data Analysis

Linear Algebra Refresher

Thm. **0.0.1 Rank Theorem:**

- Let $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times k}$, with $\text{rank}(B) = n$. Then $\text{rank}(AB) = \text{rank}(A)$.
- Let $A \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{\ell\times m}$, with $\text{rank}(C) = n$. Then $\text{rank}(CA) = \text{rank}(A)$.

0.1 Lagrange Multipliers - Mine

Thm. **0.1.1 Single Constraint Lagrange Multiplier:** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function we wish to minimize under the constraint $g = 0$, with $g : \mathbb{R}^n \to \mathbb{R}$.

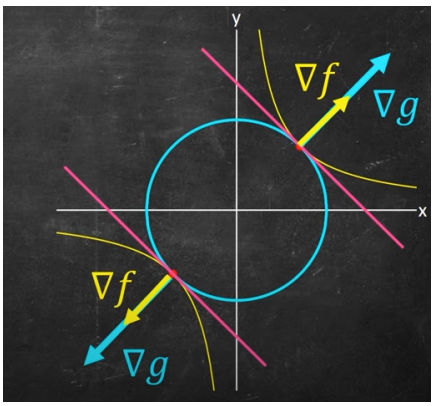


Figure 1: Lagrangian single constraint

At a critical point, we would have:

$$\boldsymbol{\nabla} f = \lambda \boldsymbol{\nabla} g$$

for some $\lambda \in \mathbb{R}$.

Equivalently, we can define the Lagrangian:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

and then look for stationary points, i.e. points that satisfy:

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \quad \textit{and} \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0$$

Thm. **0.1.2 Multiple Constraints Lagrange Multipliers:** Now assume we wish to minimize $f : \mathbb{R}^3 \to \mathbb{R}$, under constraints of both $g_1 = 0$ and $g_2 = 0$, with $g_1, g_2 : \mathbb{R}^3 \to \mathbb{R}$.

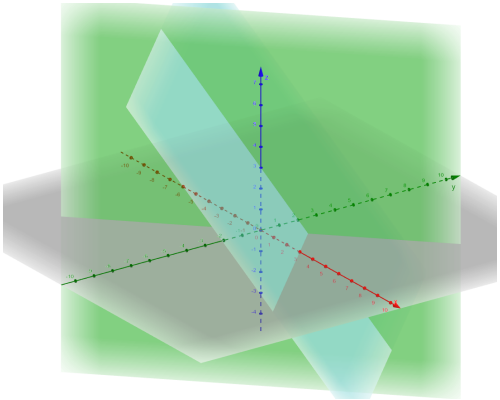


Figure 2: The constraints $g_1 = 0$ and $g_2 = 0$ visualized.
https://www.geogebra.org/3d/nxbqmx2g

Every point $\boldsymbol{x}$ on the constraint-surface of a constraint g_i has a space of allowable directions we can walk along. Note that this space is given by all vectors perpendicular to $\boldsymbol{\nabla} g_i(\boldsymbol{x})$. Therefore, the set of directions that are allowed by *all* constraints is the space of directions that are perpendicular to *all* the gradients. Let's denote the span of the gradients by S, and the set of allowable directions by A. Then $A = S^{\perp}$.

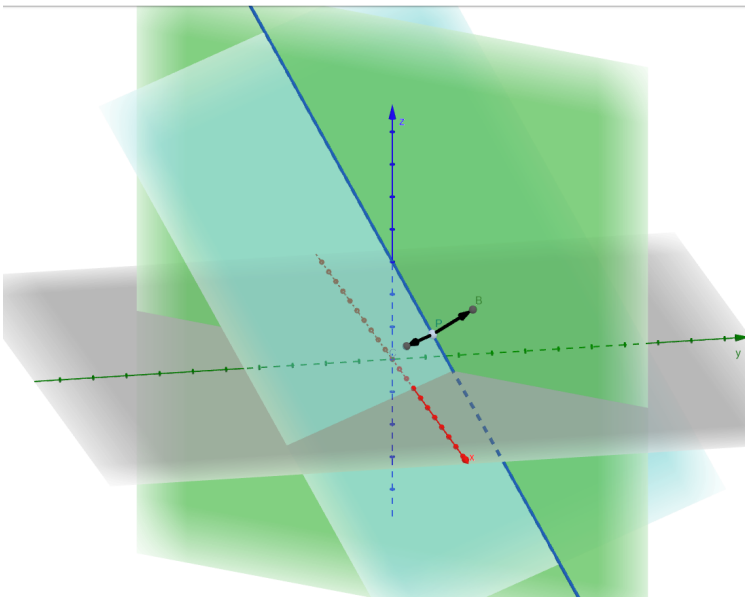


Figure 3: The blue line A represents all allowable directions. The plane perpendicular to it (not shown) is S.

We are interested in points where f does not change as we walk. I.e. in the set of our allowed directions A, f does not change. Specifically, this means that $\boldsymbol{\nabla} f \in A^{\perp} = S$. Thus, there are scalars λ_1, λ_2 such that:

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \sum_{i=1}^{2} \lambda_i \boldsymbol{\nabla} g_i(\boldsymbol{x})$$

More generally, for $f : \mathbb{R}^n \to \mathbb{R}$ with m constraints $g_i : \mathbb{R}^n \to \mathbb{R}$ we look for $\lambda_1, \dots \lambda_m$ s.t.:

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \sum_{k=1}^{m} \lambda_k \boldsymbol{\nabla} g_k(\boldsymbol{x}) \quad \text{and} \quad g_1(\boldsymbol{x}) = \ldots = g_m(\boldsymbol{x}) = 0$$

equivalently, we define:

$$\mathcal{L}(\boldsymbol{x}, \lambda_1, \dots \lambda_m) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k g_k(\boldsymbol{x})$$

and solve:

$$\boldsymbol{\nabla}_{\boldsymbol{x},\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = 0 \iff \begin{cases} \boldsymbol{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k \boldsymbol{\nabla}_{\boldsymbol{x}} g_k(\boldsymbol{x}) = 0 \\ g_1(\boldsymbol{x}) = \ldots = g_m(\boldsymbol{x}) = 0 \end{cases}$$

0.2 Lagrange Multipliers

Let $f : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$, for $i = 1, \dots, m$ be continuously differentiable functions.

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h_i(x) = 0, \quad i = 1, \dots m \end{aligned}$$

we denote $h : \mathbb{R}^n \to \mathbb{R}^m$ by:

$$h = (h_1, \dots h_m)$$

and the constraints are $h(x) = \mathbf{0}$.

\# Thm. **0.2.1 Lagrange Multiplier Theorem:** For a given local minimum $x^{\star}$, there exists scalars λ_1, λ_n, called *Lagrange Multipliers*, such that:

$$\boldsymbol{\nabla} f(x^{\star}) + \sum_{i=1}^{m} \lambda_i \boldsymbol{\nabla} h_i(x^{\star}) = \mathbf{0}$$

1 Week 1

1.1 Principal Component Analysis (PCA)

1.1.1 Statistical POV of PCA

We treat $\boldsymbol{x} \in \mathbb{R}^D$ as a random vector, aka a multivariate random variable. We assume features are scaled, i.e. $E[x_i] = 0$ and $\text{Var}[x_i] = 1$. Generally, $E[x_i x_j] \neq 0$, therefore the features are correlated. We wish to find a representation $\boldsymbol{y} \in \mathbb{R}^d$, $d << D$, given as a linear combination of $\boldsymbol{x}$, i.e. $\boldsymbol{y} = A\boldsymbol{x}, A \in \mathbb{R}^{d \times D}$, such that the features are uncorrelated, i.e.

$$\text{Cov}[y_i, y_j] = E[y_i y_j] - E[y_i]E[y_j] = 0$$

and since $E[y_i] = E[y_j] = 0$, this means we look for $E[y_i y_j] = 0$. In addition, we look for the most "meaningful" uncorrelated directions - those directions along which $\boldsymbol{x}$ varies the most. How do we capture this? Recall that with the random matrix:

$$P_x = \boldsymbol{x}\boldsymbol{x}^T$$

and $\text{Cov}[\boldsymbol{x}] = E[\boldsymbol{x}\boldsymbol{x}^T]$ we can measure the variance of the projection of $\boldsymbol{x}$ along a vector $\boldsymbol{u}$ by:

$$\text{Var}[\boldsymbol{u}^T\boldsymbol{x}] = \boldsymbol{u}^T E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{u} = E[\boldsymbol{u}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u}] = E[(\boldsymbol{u}^T\boldsymbol{x})^2]$$

This gives the objective function:

$$\boldsymbol{u}_i^\star = \arg\max_{\boldsymbol{u}_i} E[(\boldsymbol{u}_i^T\boldsymbol{x})^2]$$
$$\text{s.t. } \boldsymbol{u}_i^T\boldsymbol{u}_j = \delta_{ij} \text{ for } j \leq i$$

where the condition means that we require the directions to be uncorrelated, since

$$E[(y_2)(y_1)] = E[(\boldsymbol{u}_1^T\boldsymbol{x})(\boldsymbol{u}_2^T\boldsymbol{x})] = E(\boldsymbol{u}_1^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u}_2)] = \boldsymbol{u}_1 P_x \boldsymbol{u}_2 = \lambda_1 \boldsymbol{u}_1^T\boldsymbol{u}_2$$

so $E[(y_2)(y_1)] = 0$ iff $\boldsymbol{u}_1 \perp \boldsymbol{u}_2$.

Since $\text{Var}[\boldsymbol{u}^T\boldsymbol{x}] = \boldsymbol{u}^T P_x \boldsymbol{u}$, we have from Courant-Fischer that the largest eigenvectors of P_x are those that maximize it.

1.1.2 Sample POV of PCA

Assume we have n samples of dimension D, each sample denoted by $\boldsymbol{x}^{(i)} \in \mathbb{R}^D$. Denote $X \in \mathbb{R}^{n \times d}$, where each row contains a sample. Assume the samples are centered. PCA seeks to find the directions (principal components) that maximize the variance of the projected data. Let S denote the sample covariance matrix:

$$S \overset{\text{def}}{=} \frac{1}{n} X^T X = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)}(\boldsymbol{x}^{(i)})^T = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)} \otimes \boldsymbol{x}^{(i)}$$

Then $S \in \mathbb{R}^{D \times D}$.

Objective Function PCA solves the following optimization problem:

$$\max_{\boldsymbol{w}} \ \boldsymbol{w}^T S \boldsymbol{w} \quad \text{subject to } \|\boldsymbol{w}\| = 1$$

The principal components are the eigenvectors $\boldsymbol{w}$ corresponding to the largest eigenvalues of S. Since S is PSD, it has real non-negative eigenvalues, and we write:

$$S = V \Lambda V^T$$

with $V \in \mathbb{R}^{D \times D}$, and its columns, the eigenvectors of S are called the **principal directions**.

Limitations: PCA can only capture linear relations when the covariance matrix is computed in the original feature space. Hence for data with nonlinear structure, PCA may not reduce dimensionality effectively. The solution is to use Feature Mapping.

Remark (Summary): Let $X \in \mathbb{R}^{n \times D}$ be a matrix of n samples with D features each. Assume the data is centered. We perform PCA by computing $S = \frac{1}{n} X^T X = V \Lambda V^T$. We call the columns of V the **principal directions**. We can then embed each sample of X into a lower dimension k by projecting it into the first k principal directions, getting k **principal scores**. We can then reconstruct the sample by using the k principal scores to take a weighted sum of the k principal directions, producing a reconstruction. Note - we don't use the eigenvalues in Λ in the embedding/reconstruction process. They only provide measurement of the variance of the i-th principal direction.

Remark (Transpose Trick): Let $X \in \mathbb{R}^{m \times n}$, with $m >> n$, with m features. To find principal directions, the eigenvalues of XX^T, we need to compute XX^T, an $m \times m$ matrix, which is computationally expensive. Instead, we can use the following relation: Let v be an eigenvector XX^T, i.e.

$$\begin{aligned} & XX^T v = \lambda v \\ \Rightarrow\ & X^T X(X^T v) = \lambda X^T v \\ \Rightarrow\ & X^T X w = \lambda w \end{aligned}$$

I.e. $w := X^T v$ is an eigenvector of $X^T X$ with eigenvalue λ. Additionally, we have that:

$$\begin{aligned} & Xw = XX^T v = \lambda v \\ \Rightarrow\ & v = \lambda^{-1} X w \end{aligned}$$

Therefore, we can compute $X^T X$ eigenpairs (w, λ) and get XX^T eigenpairs (v, λ) by $v = \lambda^{-1} X w$.

1.2 Feature Mapping

Feature Space By mapping our data to higher feature space $\mathcal{F}$, we can separate non-linear data, as in Fig. 4. Denote by

$$\phi : \mathbb{R}^D \to \mathcal{F} \qquad \boldsymbol{x}^{(i)} \to \phi(\boldsymbol{x}^{(i)})$$

recall that $\boldsymbol{x}^{(i)} \in \mathbb{R}^D$.

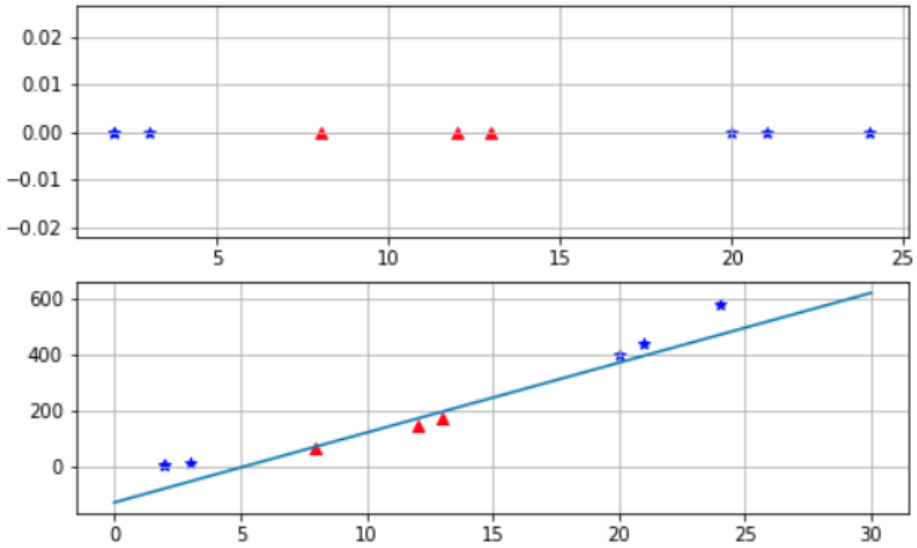


Figure 4: Non-linear data. Here, the function is $\phi : \mathbb{R} \to \mathbb{R}^2$, with $\phi(x) = (x, x^2)$.

In this new representation, we **can** linearly separate the samples. Note - we do need to recenter the projected samples. We will write $\phi(\boldsymbol{x}^{(i)})$ again assuming it is centered. Another assumption is that $\left\{\phi(\boldsymbol{x}^{(i)})\right\}_{i=1}^{n}$ are linearly independent.

Now we can use PCA, only we do so on the new covariance matrix:

$$S_\phi \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}^{(i)}) \phi(\boldsymbol{x}^{(i)})^T$$

Continuing similarly to PCA, we wish to find the (largest) eigenvectors of S_ϕ. If we denote $\mathcal{F}$ dimension by $|\mathcal{F}|$, then $S_\phi \in \mathbb{R}^{|\mathcal{F}| \times |\mathcal{F}|}$. This suggest a new problem - if $|\mathcal{F}| >> n$, then computations become expensive. Luckily, we notice that:

\# Thm. **1.2.1 The kernel theorem:**

Computing the (nonzero) eigenvectors of S_ϕ requires only knowing the inner products

$$k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) := \left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}) \right\rangle$$

Hence, if computing $k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ can be done easily, then we reduced the complexity of the problem

Proof. We will first prove a lemma.

Lemma 1.2.2: Range of S_ϕ: Recall that:

$$S_\phi = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^T$$

Let $\boldsymbol{w} \in \mathcal{F}$ be any vector. Then:

$$S_\phi \boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^T\boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\left\langle \phi(\boldsymbol{x}^{(i)}), \boldsymbol{w}\right\rangle$$

Therefore, $S_\phi \boldsymbol{w}$ is a linear combination of $\{\phi(\boldsymbol{x}^{(i)})\}_{i=1}^n$, with the coefficients being $\alpha_i = \langle \phi(\boldsymbol{x}^{(i)}), \boldsymbol{w}\rangle$. From the assumption that $\left\{\phi(\boldsymbol{x}^{(i)})\right\}_{i=1}^n$ are linearly independent, this representation is unique.

□

Continuing the proof: denote by $\boldsymbol{v}$ a (yet unknown) eigenvector of S_ϕ, i.e.:

$$S_\phi \boldsymbol{v} = \lambda \boldsymbol{v} \tag{1}$$

Using the lemma, we can write:

$$\boldsymbol{v} = \sum_{j=1}^{n}\alpha_j\phi(\boldsymbol{x}^{(j)}) \tag{2}$$

Where $\alpha_j \in \mathbb{R}, \alpha_j = \left\langle \phi(\boldsymbol{x}^{(j)}), \boldsymbol{v}\right\rangle$ are the coefficients we wish to determine. We write $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$. Plugging this into Eq. (1), we get:

$$\begin{aligned}
S_\phi \boldsymbol{v} &= \frac{1}{n}\Big(\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^T\Big)\boldsymbol{v} \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^T\Big)\sum_{j=1}^{n}\alpha_j\phi(\boldsymbol{x}^{(j)}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\sum_{j=1}^{n}\alpha_j\left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)})\right\rangle
\end{aligned}$$

Define a new matrix $K \in \mathbb{R}^{n\times n}$ by:

$$K_{ij} = \left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)})\right\rangle$$

[Note - we've assumed that ϕ sends such that $\{\phi(\boldsymbol{x}^{(i)})\}_{i=1}^n$ are centered. Another way to have this, is to define $\overline{K} = K - \frac{1}{n}\mathbf{1}^T K - \frac{1}{n}K\mathbf{1} + \frac{1}{n^2}\mathbf{1}^T K\mathbf{1}$, i.e. subtract row and col mean, and add total mean to remedy over correction.]

Then we get that:

$$S_\phi \boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})\sum_{j=1}^{n}\alpha_j\overline{K}_{ij} = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})(\overline{K}\boldsymbol{\alpha})_i$$

Developing the RHS of Eq. (1), we have:

$$\lambda\boldsymbol{v} = \lambda\sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}^{(i)})$$

Rewriting Eq. (1), we get:

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}^{(i)})(\overline{K}\boldsymbol{\alpha})_i = \lambda\sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}^{(i)})$$

From the linear independence of $\left\{\phi(\boldsymbol{x}^{(i)})\right\}_{i=1}^n$, we know the coefficient for each $\phi(\boldsymbol{x}^{(i)})$ must be equal, hence we get:

$$\begin{aligned}
&\frac{1}{n}(\overline{K}\boldsymbol{\alpha})_i = \lambda\alpha_i \\
\Rightarrow\ &\left(\overline{K}\boldsymbol{\alpha}\right)_i = n\lambda\alpha_i \\
\Rightarrow\ &K\boldsymbol{\alpha} = n\lambda\boldsymbol{\alpha}
\end{aligned}$$

Therefore, the original problem of finding eigenpairs $(\boldsymbol{v}, \lambda)$ for the $|\mathcal{F}| \times |\mathcal{F}|$ matrix S_ϕ, is actually equivalent to finding the eigenpairs $(\boldsymbol{\alpha}, n\lambda)$ of the $n \times n$ matrix $\overline{K}$. Since $|\mathcal{F}| >> n$, the second task is computationally easier. Using kernel function k, we compute (centered) matrix $\overline{K}$, find its eigenvalues $\boldsymbol{\alpha}_1, \dots \boldsymbol{\alpha}_n$ with eigenvalues $n\lambda_1, \dots, n\lambda_n$, and proceed. □

1.3 Kernel PCA

We finish by performing kernel PCA, projecting the lifted samples $\phi(\boldsymbol{x}^{(i)})$ onto the m-th largest eigenvector of the (centered) kernel matrix $\overline{K}$, using **only the kernel function**:

$$k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) := \langle \phi(\boldsymbol{x}^{(i)}, \phi(\boldsymbol{x}^{(j)}) \rangle$$

Let $\boldsymbol{\alpha}_m = [(\boldsymbol{\alpha}_m)_1, \dots, (\boldsymbol{\alpha}_m)_n]$ denote the m-th largest eigenvector of the centered kernel matrix $\overline{K}$. Denote by $\boldsymbol{y}^{(i)} \in \mathbb{R}^d$ the embedding of $\boldsymbol{x}^{(i)}$ by kernel PCA using the first d largest eigenvectors of $\overline{K}$. Let $\boldsymbol{v}_m$ the m-th largest eigenvector of S_ϕ. Remember that $\boldsymbol{v}_m = \sum_{j=1}^n \phi(\boldsymbol{x}^{(j)}) (\boldsymbol{\alpha}_m)_j$. Then we have that:

$$\begin{aligned}
(\boldsymbol{y}^{(i)})_m = \langle\, \phi(\boldsymbol{x}^{(i)})\,,\, \boldsymbol{v}_m \,\rangle &= \langle\, \phi(\boldsymbol{x}^{(i)})\,,\, \sum_{j=1}^n \phi(\boldsymbol{x}^{(j)})(\boldsymbol{\alpha}_m)_j \,\rangle \\
&= \sum_{j=1}^n (\boldsymbol{\alpha}_m)_j \,\langle\, \phi(\boldsymbol{x}^{(i)})\,,\, \phi(\boldsymbol{x}^{(j)}) \,\rangle \\
&= \sum_{j=1}^n (\boldsymbol{\alpha}_m)_j \overline{K}_{ij} \\
&= (\overline{K}\boldsymbol{\alpha}_m)_i
\end{aligned}$$

And therefore:

$$\begin{bmatrix} | & \dots & | \\ \boldsymbol{y}^{(1)} & \dots & \boldsymbol{y}^{(n)} \\ | & \dots & | \end{bmatrix} = \begin{bmatrix} - & \boldsymbol{\alpha}_1^T & - \\ & \dots & \\ - & \boldsymbol{\alpha}_d^T & - \end{bmatrix} K$$

We create embeddings using **only** the kernel function k.

@ Example: Consider a dataset consisting of points lying on a circle on $\mathbb{R}^2$:

$$\boldsymbol{x}^{(i)} = \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix}, \quad \theta_i = \frac{2\pi i}{n}, \quad i = 1, \dots, n$$

clearly this is a 1 dimensional data, however it currently in $\mathbb{R}^2$. Since this data is has a non-linear structure, classic PCA would not help in reducing its dimension. We use the RBF function:

$$k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = e^{-\frac{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|}{2\sigma^2}}$$

For simulation, The un-centered kernel matrix K is computed for $n = 100$ and $\sigma = 0.5$.

The kernel matrix is centered as:

$$K_{\text{centered}} = K - \frac{1}{n}\mathbf{1}K - K\frac{1}{n}\mathbf{1} + \frac{1}{n}\mathbf{1}K\frac{1}{n}\mathbf{1},$$

where $\mathbf{1}$ is the all-ones matrix.

The eigenvalues and eigenvectors of K_{centered} are computed. The eigenvalues in descending order are:

$$\lambda_1 = 17.8751, \quad \lambda_2 = 17.8751, \quad \lambda_3 = 11.7627, \dots$$

The projection onto the first principal component is given by:

$$\boldsymbol{z} = K_{\text{centered}}\boldsymbol{v}_1,$$

where $\boldsymbol{v}_1$ is the eigenvector corresponding to λ_1.

Below is a plot of the projection:

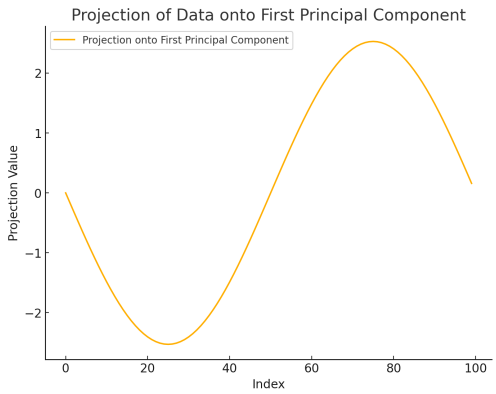


Figure 5: Projecting the points on the circle using the largest eigenvector of the centered kernel matrix. Clearly we are projecting $\boldsymbol{x}^{(i)}$, associated to θ_i, to a slightly scaled version of $-\sin(\theta_i)$

Remark ((Classic) PCA is linear): (Classic) PCA can only find linear relations in the data. Assume $\boldsymbol{x}^{(i)}$ is a sample, and $W \in \mathbb{R}^{k\times d}$ is the projection matrix containing the k largest eigenvectors of the sample covariance matrix $S = \frac{1}{n}X^TX$. Let $\boldsymbol{y}^{(i)}$ denote the projection of $\boldsymbol{x}^{(i)}$, i.e.:

$$\boldsymbol{y}^{(i)} = W\boldsymbol{x}^{(i)}$$

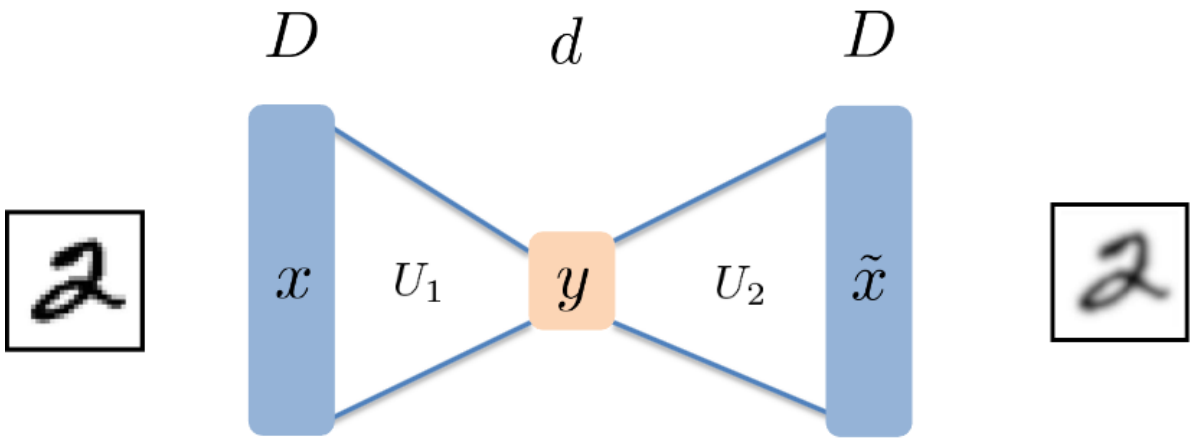
Notice we are only applying a linear transformation on $\boldsymbol{x}^{(i)}$, hence PCA is linear.

Kernel PCA is nonlinear.

2 Week 2

2.1 Encoder Decoder

Let $x \in \mathbb{R}^D$ be some sample. A common task is to do dimensionality reduction, to some representation $y \in \mathbb{R}^d$, and then do reconstruction to $\widetilde{x} \in \mathbb{R}^D$. This is an encoder-decoder framework.



- Setup: Let $X \in \mathbb{R}^{D\times n}$, where D is number of features and n is number of samples. Assume the data is centered, there are no redundant features, and $n \geq D$, i.e. $\text{rank}(X) = D$. Denote by $Y \in \mathbb{R}^{d\times n}$ an encoding of X, and denote by $\widetilde{X} \in \mathbb{R}^{D\times n}$ a reconstruction of X using Y.
- Objective: We wish, using linear operators, to find $\widetilde{X}$ s.t. :

$$\min_{\widetilde{X}} \|X - \widetilde{X}\|_F^2$$

Since both the encoding and decoding is done via linear operators, we can write:

$$\min_{\widetilde{X}} \|X - \widetilde{X}\|_F^2 \quad \text{s.t.} \quad \widetilde{X} = U_2Y = U_2U_1X \quad U_1 \in \mathbb{R}^{d\times D}, U_2 \in \mathbb{R}^{D\times d}$$

- Solution: PCA. Let $\widehat{P} = \frac{1}{n}XX^T$ be the covariance matrix, and let U be the first d largest eigenvectors of $\widehat{P}$. Then $U_1 = U^T, U_2 = U$.
- Geometric PCA is not-unique: Notice that for every orthogonal matrix $R \in \mathbb{R}^{d\times d}$, we have that $U_1 = (UR)^T, U_2 = (UR)$ is also a minimizer. Denote $\widetilde{U} = UR$. For example, if our data really only has entries in the first two entries, then we can encode it to $X - Y$ plane. However every rotation following this projection would also serve, provided the decoder also includes the inverse rotation.

  I.e. the encoding Y isn't unique. However, statistical PCA is unique - how can this be? The answer is that in statistical PCA, we add a requirement that there is no correlation between the features (i.e. they are orthogonal), i.e. every row (feature) of Y is uncorrelated with every other row. If $Y = U^T X \in \mathbb{R}^{d\times n}$, then we have that $E[YY^T] = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$, i.e. the encodings are not correlated. However, for $\widetilde{Y} = \widetilde{U}^T X = (UR)^T X$, we have that:

$$E[\widetilde{Y}\widetilde{Y}^T] = R^T \Lambda R$$

  i.e. there can be non-zero off-diagonal elements, and hence the features can be correlated, unlike (statistical) PCA.

2.2 Relating Statistical and Geometrical PCA

2.3 SVD, PCA

Thm.**2.3.1 SVD:** Let $X \in \mathbb{R}^{m\times n}$, with m samples and n features per sample, $q = \min\{m, n\}$ and $r = \text{rank}(M)$. SVD decomposes X to

$$X = U\Sigma V^T$$

with $U \in \mathbb{R}^{m\times m}, \Sigma \in \mathbb{R}^{m\times n}, V \in \mathbb{R}^{n\times n}$. Σ is a rectangular diagonal matrix, with non-increasing entries.

Thm. **2.3.2 PCA and SVD:** Recall that in PCA (assuming X is centered), we compute:

$$\widehat{P} = \frac{1}{n-1}X^T X = \frac{1}{n-1}V\Sigma^T U^T U\Sigma V^T = V\frac{\Sigma^2}{n-1}V^T$$

I.e. the right singular vectors in V (from the SVD) are the principal directions (eigenvectors) from PCA, and the PCA eigenvalues λ_i are related to σ_i via $\lambda_i = \frac{\sigma_i^2}{n-1}$.

Since the columns of V are the principal directions, we can project onto them to get principal scores, i.e. an optimal lower dimension embedding. The rank k embedding $Y \in \mathbb{R}^{m\times k}$ is given by:

$$Y = XV_{n\times k} = U\Sigma V^T V_{n\times k} = U\Sigma_{m\times k} = U_{m\times k}\Sigma_{k\times k}$$

where:

$$\Sigma_{k\times k} = \text{diag}(\sigma_1, \ldots, \sigma_k) = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_k \end{bmatrix}$$

Note that $Y = U_{m\times k}\Sigma_{k\times k}$ are the principal scores. Therefore we can produce a reconstruction $\widetilde{X} \in \mathbb{R}^{m\times n}$ by using the principal scores to take a weighted sum of the principal directions:

$$\widetilde{X} = Y(V^T)_{k\times n} = (U_{m\times k}\Sigma_{k\times k})(V^T)_{k\times n}$$

This can be written as:

$$\widetilde{X} = Y(V^T)_{k\times n} = U_{m\times k}\Sigma_{k\times k}(V^T)_{k\times n} = \tag{1}$$

$$= U_{m\times k}\begin{bmatrix} - & \sigma_1 v_1^T & - \\ - & \vdots & - \\ - & \sigma_k v_k^T & - \end{bmatrix} = \sum_{i=1}^{k}\sigma_i u_i v_i^T \tag{2}$$

Alternatively, we can think of the reconstruction $\widetilde{X}$ of rank k as taking the first k largest singular values, and zeroing the rest. Let $\Sigma^{(k)} \in \mathbb{R}^{m\times n}$ be defined by:

$$\Sigma^{(k)} = \begin{cases} \sigma_i, & i \le k \\ 0, & i \ge k \end{cases}$$

Then:

$$\widetilde{X} = U\Sigma^{(k)}V^T$$

This follows from the RHS of Eq. (1).

Thm. **2.3.3 Reconstruction Error:**

$$\epsilon(k) = \|X - \widetilde{X}\|_F^2 = \mathrm{Tr}\left((X - \widetilde{X})^T(X - \widetilde{X})\right) = \sum_{i=k+1}^{n} \sigma_i^2$$

Remark (Model Selection - How to choose embedding dimension k): knee point

2.4 Stochastic Neighbor Embedding

$*$ Def. **2.4.1 (Shannon) Entropy:** Let X be a discrete random variable which takes values in the set $\mathcal{X}$, with PMF $p : \mathcal{X} \to [0, 1]$ such that $p(x) := \mathbb{P}[X = x]$. Then:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x) \log p(x)$$

$*$ Def. **2.4.2 KL Divergence:** Let p, q be PMF over $\mathcal{X}$. Then:

$$D(p||q) := \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

$*$ Def. **2.4.3 Jensen-Shannon Divergence:** Symmetric version of the KL-Divergence:

$$D_{JS}(p; q) := \frac{1}{2}\big(D(p||\rho) + D(q||\rho)\big)$$

where:

$$\rho = \frac{1}{2}(p + q)$$

Let $X = \{x_1, \ldots, x_n\}$ be n samples, with $x_i \in \mathbb{R}^D$, and a distance measure $\|x_i - x_k\|$.

$*$ Def. **2.4.4 "Conditional Probability":** Define "conditional probability":

$$p_{j|i} := \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k\neq i} \exp(-\|x_i - x_k\|^2/2\sigma^2)}$$

This is the Heat-Kernel Affinity. We define $p_{i|i} := 0$ (convention). Let E be the symmetric matrix defined by $E_{ij} = e^{-\|x_i - x_j\|^2}$. Define the non-symmetric matrix $P_{ji} = p_{j|i}$ which is E_{ji} divided by entire i-th column of$E$. Therefore the columns sum to1, but the rows don't.

- Probability Measure This is a probability measure, since $\sum_j p_{j|i} = 1$
- Non-Symmetric $p_{j|i} \neq p_{i|j}$. Example: Imagine three points $a < b < c$ on a line, with b in the middle. Then $p_{b|a} > p_{a|b}$, since b has other close neighbors (more popular) than a.
- Distance $p_{i|j}$ transforms the euclidean distance $\|x_i - x_j\|$ into a probability measure. Low distance = High probability (of being neighbors). This allows us to turn the problem of finding similar embeddings into finding similar probability distributions P and Q.

Thm. **2.4.5 SNE Embedding:** We wish to find low dimensional representation $Y = \{y_1, \ldots y_n\} \in \mathbb{R}^d$ with $d << D$, where we measure affinity $q_{j|i}$ by:

$$q_{j|i} := \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i} \exp(-\|y_i - y_k\|^2)}$$

(some omit the σ in q), and:

$$P_i := p_{\cdot|i} = (p_{1|i}, \ldots, p_{n|i})$$
$$Q_i := q_{\cdot|i} = (q_{1|i}, \ldots, q_{n|i})$$

Thus P_i is the normalized similarity vector for x_i, and defines a conditional probability distribution.

We wish that y_i is similar to x_i.

∗ Def. **2.4.6 Perplexity:** σ_i is chosen using perplexity, with:

$$\text{Perp}(P_i) := 2^{H(P_i)} \quad \text{with } H(P_i) := -\sum_{j \neq i} p_{j|i} \log_2 p_{j|i}$$

$\text{Perp}(P_i)$ is a smooth measure of the effective number of neighbors of x_i.

∗ Def. **2.4.7 SNE Cost Function:** Define the Cost Function, that minimizing it aims to maintain neighborhood similarity:

$$C(Y) = C(\{y_i\}) = \sum_i KL(P_i|Q_i) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

We look for:

$$\min_Y C(Y)$$

- Goal: Goal of SNE is to minimize mismatches between $p_{j|i}$ and $q_{j|i}$. SNE minimizes the sum of KL divergences over all data points, using a gradient descent method. We denote the cost function C:
- Non-Symmetry: For **close** points ($p_{j|i}$ high) there is **high** cost for mapping them **far** (i.e. small $q_{j|i}$). However, for **far** points ($p_{j|i}$ low), there is **small** cost for mapping them close (i.e. assigning high $q_{j|i}$). This leads to asymmetry, where SNE preserves **local** structure, but not **global** structure.

Thm. **2.4.8 Cost Function Gradient:** in Tirgul.

$$\frac{\partial C}{\partial y_i} = 2 \sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}\right)(y_i - y_j)$$

Thm. **2.4.9 SNE Gradient Descent:**

- We initialize y by sampling map points randomly from an isotropic Gaussian with small variance that is centered around the origin.
- A relatively large momentum is added to gradient to avoid poor local minima. Alternatively, the current gradient is added to an exponentially decaying sum of previous gradients.

  $$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\partial C}{\partial y} + \alpha(t)(\mathcal{Y}^{(t-1)} + \mathcal{Y}^{(t-2)})$$

  where $\alpha(t)$ is the momentum at iteration t and η is the learning rate.

Remark (SNE Notes):

- t-SNE is good for visualization, not necessarily as input to a downstream . It does not preserve global structure.
- t-SNE is nonlinear.
- t-SNE does not preserve distances. Close points will behave similarly, however, more global structure is not preserved. E.g. let x be our base points, and x_1, x_2 two points with $d(x_1 - x) < d(x_2 - x)$. Then it's not necessarily true that $d(y_1 - y) < d(y_2 - y)$, where y, y_1, y_2 are the t-SNE embeddings. If that's a goal, then use MDS. Intuitively, PCA preserves large pairwise distances, but not small pairwise distances. t-SNE does not aim to preserve any (global) distances.
- T-SNE is designed to preserve local similarities between data points, which means it excels at revealing clusters and local patterns in high-dimensional data. However, this focus on local structure comes at the expense of accurately representing global relationships.
- Probability Distribution: T-SNE uses a heavy-tailed Student t-distribution in the low-dimensional space, which tends to push apart points that are not similar

2.5 t-SNE

Crowding Problem Consider a fixed point $x_i \in \mathbb{R}^D$. The set of neighbors $\{x_j \mid p_{j|i}\}$ is large may be "evenly spread" across the entire D-dimensional volume around x_i. In other words, x_i has many neighbors that are moderately close. Therefore, trying to project this relationship faithfully into a $d << D$-dimensional volume, preserving distances, is impossible - distortions and overlap are necessary. There is simply not enough "room" in $\mathbb{R}^d$, we have "overcrowding". Therefore we must either push points very close, or spread them out.

If we use Gaussian distribution for q_{ij}, this will lead to highly-distorted representation. Instead, tSNE uses a heavy-tailed distribution. The slower decay allows t-SNE to spread out points in the low-dimensional space without completely sacrificing neighborhood probabilities.

tSNE solves this by using:

1. Symmetry: uses a symmetrized version of the SNE cost function, with simpler gradients.
2. Student-t distribution: Uses Student-t distribution instead of a Gaussian to compute the *similarity* in the low-dimensional space. A heavy-tailed distribution goes to zero slower than an exponential. This will give outliers high values.

Solving Asymmetry We can do **symmetric SNE**, by taking the denominator with respect to all pairwise possibilities. We write:

$$p_{ij} := \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_k \sum_{\ell \neq k} \exp(-\|x_k - x_\ell\|^2/2\sigma^2)}$$

Recall that previously, $P_{ji} = p_{j|i}$ was given as E_{ji} divided by the i-th column of E. Now instead, P_{ji} is defined to be E_{ji} divided by the *entire* matrix E (which has a zero main diagonal), and therefore $P_{ji} = P_{ij}$. In the *low-dimensional* map $q_{i|j}$ as the denominator (Z_i) by pairwise similarities:

$$q_{ij} := \frac{e^{-\|y_i - y_j\|^2}}{\sum_k \sum_{\ell \neq k} e^{-\|y_k - y_\ell\|^2}}$$

However, this still does not solve the crowding problem. p_{ij} can still be arbitrarily small, which is problematic. Therefore, tSNE uses a different conditional probability:

$$p_{ij} := \frac{p_{i|j} + p_{j|i}}{2n}$$

This symmetric definition ensures that:

$$\sum_j p_{ij} > \frac{1}{2n}$$

i.e. p_{ij} is bounded below, regardless of how much of an outliers x_i and x_j are, and therefore y_i and y_j will have to be somewhat reasonable.

Heavy-Tail Additionally, to solve the crowding problem, we give more weight to outliers, by using for the mapping q_{ij} a heavy-tailed distribution instead of the Gaussian. Specifically we use the Student-t distribution:

$$q_{ij} := \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k \sum_{\ell \neq k} (1 + \|y_k - y_j\|^2)^{-1}}$$

The heavy tails "allows" the map to place points further apart.

2.6 Canonical Correlation Analysis (CCA)

The idea is to reduce the number of variables without sacrificing too much information. Whereas PCA deals with random variables that come from a single set, CCA assumes the variables come from **two** sets. We select an uncorrelated linear combination of the two sets of variables, which are pairwise highly correlated.

We assume wlog that $E[X] = E[Y] = 0$. Additionally, We assume both random vectors have no redundant features, i.e. no features that can be expressed as a linear combination of the others. This gives that XX^T and YY^T are (strictly) positive definite (and not *semi*-positive definite).

Objective: Given multivariate random vectors $X \in \mathbb{R}^n, Y \in \mathbb{R}^m$, find *linear* mappings, $a \in \mathbb{R}^n, b \in \mathbb{R}^m$, also known as the (first) **canonical directions**, such that for the random variables $U = a^T X$ and $V = b^T Y$, we have:

$$\begin{aligned} &\max_{a,b} \left[(a^T X)(b^T Y)\right] \\ &\text{subject to} \quad E\left[(a^T X)^2\right] = 1\,,\ E\left[(b^T Y)^2\right] = 1 \end{aligned} \tag{1}$$

Denoting $\Sigma_{XX} := E[XX^T]$ and $\Sigma_{YY} := E[YY^T]$ and $\Sigma_{XY} := E[XY^T], \Sigma_{YX} := E[YX^T]$ our objective becomes

$$\max_{a,b} a^T \Sigma_{XY} b \tag{2}$$

$$\text{s.t. } a^T \Sigma_{XX} a = 1 \ \ b^T \Sigma_{YY} b = 1 \tag{3}$$

Note that:

$$E[a^T X] = a^T E[X] = 0$$

and similarly $E[b^T Y] = 0$.

Define:

$$M \stackrel{\text{def}}{=} \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \in \mathbb{R}^{n \times m}$$

Then we have:

$$MM' = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^T \Sigma_{XX}^{-1/2} \in \mathbb{R}^{n \times n} \quad , \quad M'M = \Sigma_{YY}^{-1/2} \Sigma_{XY}^T \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_Y^{-1/2}$$

We've shown that $MM', M'M$ are PD and have the same eigenvalues.

Define:

$$B = \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{XY}^T \quad , \quad C = \Sigma_{YY}^{-1} \Sigma_{XY}^T \Sigma_{XX} \Sigma_{XY}$$

We note that

$$B = \Sigma_{XX}^{-1/2} M M^T \Sigma_{XX}^{1/2} \tag{4}$$

$$C = \Sigma_{YY}^{-1/2} M^T M \Sigma_{YY}^{1/2} \tag{5}$$

i.e. B is *similar* to MM^T and hence they have the same eigenvalues. Likewise, C is similar to $M^T M$, and hence $MM^T, M^T M, B, C$ all have the same eigenvalues $\lambda_1 > \ldots > \lambda_r > 0$, where r is the rank of Σ_{XY}.

We Solve Eq. (2) using Lagrange Multipliers. Define:

$$L(\rho_x, \rho_y, a, b) := a^T \Sigma_{XY} b - \frac{\rho_x}{2}(a^T \Sigma_{XX} a) - \frac{\rho_y}{2}(b^T \Sigma_{YY} b)$$

Differentiating L and looking for stationary points, we get:

$$\boldsymbol{\nabla}_a L = \Sigma_{XY} b - \rho_x \Sigma_{XX} a = \mathbf{0} \tag{6}$$

$$\boldsymbol{\nabla}_b L = \Sigma_{XY}^T a - \rho_y \Sigma_{YY} b = \mathbf{0} \tag{7}$$

We pre-multiply by a^T and b^T respectively, and we get:

$$a^T \Sigma_{XY} b = \rho_x a^T \Sigma_{XX} a = \rho_x \tag{8}$$

$$b^T \Sigma_{XY}^T a = b^T \rho_y \Sigma_{YY} b = \rho_y \tag{9}$$

where we've used Eq. (3). Therefore we get that $\rho_x = \rho_y = \rho$.

Since we've assumed Σ_{XX}, Σ_{YY} are (strictly) positive definite, they are invertible, and therefore from Eq. (6) and Eq. (7), we get:

$$a = \frac{\Sigma_{XX}^{-1} \Sigma_{XY} b}{\rho} \tag{10}$$

Plugging this into Eq. (7), we get:

$$\begin{aligned} & \frac{1}{\rho} \Sigma_{XY}^T \Sigma_{XX}^{-1} \Sigma_{XY} b = \rho \Sigma_{YY} b \\ \Rightarrow\ & \Sigma_{YY}^{-1} \Sigma_{XY}^T \Sigma_{XX}^{-1} \Sigma_{XY} b = \rho^2 b \\ \Rightarrow\ & C b = \rho^2 b \end{aligned}$$

In an analogous way we get:

$$B a = \rho^2 a$$

Therefore we get that the optimal $\widehat{a}, \widehat{b}$ would be the largest eigenvector of B and C, respectively. The reasoning is like in PCA.

Further, note that:

$$B\widehat{a} = \rho^2 \widehat{a} \tag{11}$$

$$\Rightarrow MM^T(\Sigma_{XX}^{1/2} \widehat{a}) = \rho^2 (\Sigma_{XX}^{1/2} \widehat{a}) \tag{12}$$

Denote $\widehat{u} = \Sigma_{XX}^{1/2}\widehat{a}$. Then $\widehat{u}$ is an eigenvector of B with eigenvalue ρ^2. Since we've shown that MM^T and B have the same eigenvalues, we get that $\widehat{u}$ is the largest eigenvector of MM^T. Therefore, In order to compute $\widehat{a}$, all we need to do is find largest eigenvector MM^T, i.e. $\widehat{u}$, and then compute $\widehat{a} = \Sigma_{XX}^{-1/2}\widehat{u}$. Similarly, if $\widehat{v}$ is largest eigenvector of M^TM, then $\widehat{b} = \Sigma_{YY}^{-1/2}\widehat{v}$.

Therefore, we can obtain the canonical directions using the eigenvalues of MM^T, M^TM.

3 Variational Auto Encoders

$*$ Def. **3.0.1 Modeling:** Modeling is unveiling the underlying ruling processes, by posing hypotheses and predictions, based on observations. For instance, physicists model how fluids flow, and biologists the structure of organisms.

Modeling often involves representation, where we describe a phenomena using specific qualities and quantities related to the process we are interested in. We describe an object by its shape, color, position, volume, etc. When looking at data - a large collection of samples - in some cases, it is reasonable to believe that those representations follow some distributions. For example, human height clearly follow some probability distribution. In such cases, we can think of the samples as being "generated" from those distributions. When such a hypothesis is true, we can generate new samples in the population, provided we estimated its probabilities.

Remark (): A complete probabilistic model captures both the distributions of its components and the relations and dependencies between them. Usually linear relations, given by covariances, are used.

$*$ Def. **3.0.2 Probabilistic Model:** Assume the observed variable $\boldsymbol{x}$ is a random sample from an unknown underlying process, whose true probability distribution is $p_\theta(\boldsymbol{x}$

$*$ Def. **3.0.3 Generative Latent-Variable Model:** A parametric family of joint probability distributions over observed variables x and unobserved latent variables z, defined by:

$$p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$$

where θ are the model parameters. The latent variables