EE445 Section: Principal Component Analysis

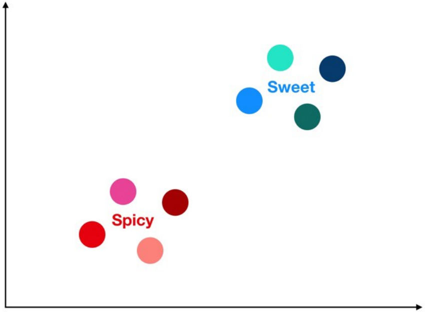
Curse of dimensionality

From Wiki:

When the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

In order to obtain a reliable result, the amount of data needed often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data, however, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.

Dimensionality Reduction - Toy Example



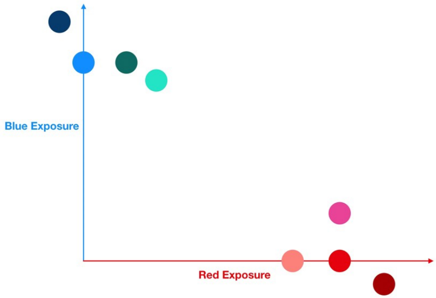
	Reddish	Bluish
	1	0
	1	0
	1	0
	1	0
	0	1
	0	1
	0	1
	0	1

Curse of dimensionality manifests

	Red	Maroon	Pink	Flamingo	Blue	Turquoise	Seaweed	Ocean
	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	0	0	0	1	0	0	0	0
	0	0	0	0	1	0	0	0
	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	1	0
	0	0	0	0	0	0	0	1

Dim reduction

	Red	Blue
Red	1.00	0
Maroon	1.20	-0.10
Pink	1.00	0.20
Flamingo	0.80	0
Blue	0	1.00
Turquoise	0.25	0.90
Seaweed	0.15	1.00
Ocean	-0.10	1.20



1. Visualization
2. Bias-variance tradeoff (remember the polynomial fitting?)

Concept Review

Eigenvalues, Spectral Decomposition

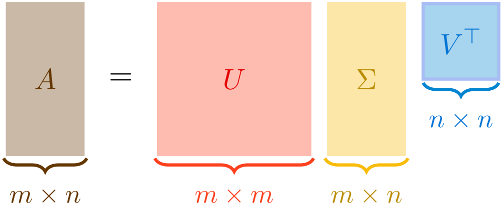
Eigenvalue: A nonzero vector x satisfying $Ax = \lambda x$ is a (right) **eigenvector** for the eigenvalue λ.

Spectral Decomposition Every symmetric matrix A can be **diagonalized** as $A = V \Lambda V^\top$ with V formed by the orthonormal eigenvectors of A and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ a diagonal matrix of the eigenvalues of A

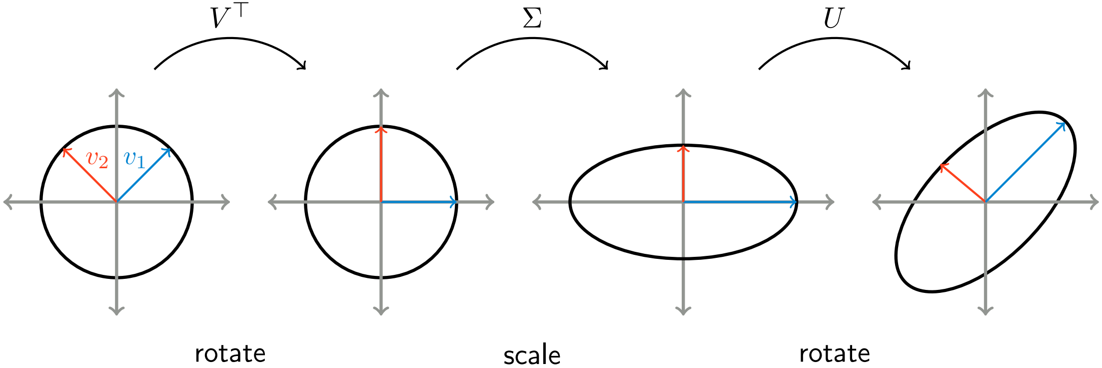
$$A = \underbrace{\begin{bmatrix} | & \cdots & | \\ v_1 & \cdots & v_n \\ | & \cdots & | \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \lambda_{n-1} & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} - & v_1^\top & - \\ \vdots & \vdots & \vdots \\ - & v_n^\top & - \end{bmatrix}}_{V^\top}$$

Singular Value Decomposition

For an arbitrary (non-symmetric) matrix A, we cannot compute the spectral decomposition.
But there are two related symmetric matrices: $AA^\top$ and $A^\top A$.



Geometric View of SVD



PCA: Data Preprocessing

Let $(z^{(1)}, \ldots, z^{(m)})$ be the original raw data, then preprocessing goes as follows:

1. Let $\mu = \frac{1}{m} \sum_{i=1}^{m} z^{(i)}$
2. Define $\tilde{x}^{(i)} = z^{(i)} - \mu$
3. Let $\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (\tilde{x}_j^{(i)})^2$
4. Define $x^{(i)} = (\tilde{x}_1^{(i)}/\sigma_1, \ldots, \tilde{x}_n^{(i)}/\sigma_n)$

- Steps 1-2 zero out the mean of the data
- Steps 3-4 rescale each coordinate to have unit variance, which ensures that different attributes are all treated on the same “scale."

PCA Optimization

Recall The length of the projection of x onto u is given by $x^\top u$.

Optimization To maximize the variance of the projections, we choose a unit-length u to maximize

$$\frac{1}{m}\sum_{i=1}^{m}((x^{(i)})^\top u)^2 = \frac{1}{m}\sum_{i=1}^{m} u^\top x^{(i)}(x^{(i)})^\top u = u^\top \underbrace{\left(\sum_{i=1}^{m} x^{(i)}(x^{(i)})^\top\right)}_{=:\Sigma = X^\top X} u$$

Covariance matrix Note that $\Sigma = X^\top X$ where

$$X = \begin{bmatrix} — & (x^{(1)})^\top & — \\ & \cdots & \\ — & (x^{(m)})^\top & — \end{bmatrix}$$

PCA Solution

Rewrite optimization

$$\max_u \|Xu\|^2 \text{ subject to } \|u\|^2 - 1 = 0 \quad \text{(note that } \|Xu\|^2 = u^\top \Sigma u\text{)}$$

Solve via Lagrangian

To solve, we write out the "Lagrangian".

$$\mathcal{L}(u, \lambda) = \|Xu\|^2 - \lambda(\|u\|^2 - 1) = u^\top \Sigma u - \lambda(u^\top u - 1)$$

$$\nabla_u \mathcal{L} = 2\Sigma u - 2\lambda u = 0 \implies \Sigma u = \lambda u$$

Interpretation

Hence, we choose an eigenvector u of Σ that chooses the largest eigenvalue. This is called the **principal eigenvector**, and is also the first right singular vector of X!!!

Practice Problem: PCA preserves inner products

Part (a):

Question What is the ij_{th} entry of the matrices $XX^\top$ and $X^\top X$? Express the matrix $XX^\top$ in terms of U and Σ, and, express the matrix $X^\top X$ in terms of Σ and V.

Part (b):

Define $\psi_{\text{PCA}}(x) = (v_1^\top x, \ldots v_k^\top x)$.
Question Show that

$$\psi_{\text{PCA}}\left(\mathbf{x}_i\right)^\top \psi_{\text{PCA}}\left(\mathbf{x}_j\right) = \mathbf{x}_i^\top \mathbf{V}_k \mathbf{V}_k^\top \mathbf{x}_j \quad \text{where} \quad \mathbf{V}_k = \left[\begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_k \end{array}\right].$$

Also show that $\mathbf{V}_k \mathbf{V}_k^\top = \mathbf{V}\mathbf{I}^k\mathbf{V}^\top$, where the matrix $\mathbf{I}^k$ denotes a $d \times d$ diagonal matrix with first k diagonal entries as 1 and all other entries as zero.

Part (c)

Suppose that we know the first k singular values are the dominant singular values. In particular, we are given that

$$\frac{\sum_{i=1}^{k} \sigma_i^4}{\sum_{i=1}^{d} \sigma_i^4} \geq 1 - \epsilon,$$

for some $\epsilon \in (0, 1)$. Then show that the PCA projection to the first k-right singular vectors preserves the inner products on average:

$$\frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(\mathbf{x}_i^\top \mathbf{x}_j\right)^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \left(\mathbf{x}_i^\top \mathbf{x}_j\right) - \left(\psi_{\text{PCA}}\left(\mathbf{x}_i\right)^\top \psi_{\text{PCA}}\left(\mathbf{x}_j\right)\right) \right|^2 \leq \epsilon.$$

Thus, we find that if there are dominant singular values, PCA projection can preserve the inner products on average. Hint: Using previous two parts and the definition of Frobenius norm might be useful.

PCA Numerical Example