CSCI 5525 Machine Learning Fall 2019

Lecture 19: Principal Component Analysis

April 2020

Lecturer: Steven Wu *Scribe: Steven Wu*

We have thus far focused on supervised learning. The goal there is somewhat clear: we want to find a predictor $\hat{f}$ based on a training data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that $\hat{f}(x)$ matches y on most instances (x, y).

Now we will switch over to unsupervised learning, where the learner only observes a collection unlabeled examples $x_1, \ldots, x_n$. There are no target labels. The objectives for unsupervised learning can be quite diverse. For example, we might want to encode the data in some compact representation (e.g., lower-dimensional space), recover *latent* structure (e.g., clusters, subspace) in the data, or extract features for subsequent supervised learning tasks. We will study PCA as our first example of unsupervised learning.

Principal Component Analysis

SVD Let us first recall how *singular value decomposition (SVD)* works. In the lectures on linear regression, we study the "thin" version of SVD. Here we will state the "full" factorization of the SVD. Given any matrix $M \in \mathbb{R}^{n \times d}$, we want to factorize the matrix as $M = USV^\intercal$, such that

- r is the rank of the matrix M;
- $U \in \mathbb{R}^{n \times n}$ is orthonormal, that is $U^\intercal U = I_n$;
- $V \in \mathbb{R}^{d \times d}$ is orthonormal, that is $V^\intercal V = I_d$;
- $S \in \mathbb{R}^{n \times d}$ is a rectangular diagonal matrix with entries $S_{ii} = s_i$ and $S_{ij} = 0$ for all $i \neq j$, where $(s_1, \ldots, s_r, 0, \ldots, 0)$ such that $s_1 \geq s_2 \geq s_3 \ldots$ denote the ordered sequence of singular values.

We could also express the factorization as a sum $M = \sum_{i=1}^{r} s_i u_i v_i^\intercal$, where each u_i is a column vector for U and each v_i is a column vector for V. Note that $\{u_i\}_{i=1}^r$ spans the column space of M and $\{v_i\}_{i=1}^r$ spans the row space of M.

Truncated SVD For any $k \leq r$, let $U_k \in \mathbb{R}^{n \times k}$ be the matrix given by the first k columns of U, $V_k \in \mathbb{R}^{d \times k}$ be the matrix given by the first k columns of V, and $S_k = \text{diag}(s_1, \ldots, s_k)$. This gives the following factorization restricted to the top-k space:

$$M_k = U_k S_k V_k = \sum_{i=1}^{k} s_i u_i v_i^\intercal$$

Note that when $k = r$, this recovers the thin SVD.

PCA. The problem of *principal component analysis* aims to solve the following optimization problem: given as input a data matrix $X \in \mathbb{R}^{n\times d}$, find an encoder E and decoder D to minimize the following reconstruction error:

$$\min_{D\in\mathbb{R}^{k\times d}, E\in\mathbb{R}^{d\times k}} \|X - XED\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm: for any matrix A,

$$\|A\|_F = \sqrt{\sum_{(i,j)} A_{ij}^2} = \sqrt{\text{tr}(A^\intercal A)}$$

The PCA method solves the problem with the following procedure: compute $X = USV^\intercal$, then return encoder $E = V_k$, decoder $D = V_k^\intercal$, encoded data $XV_k = U_kS_k \in \mathbb{R}^{n\times k}$, and decoded data $XV_kV_k^\intercal$. Note that $V_kV_k^\intercal \in \mathbb{R}^{d\times d}$ performs orthogonal projection onto subspace spanned by V_k.

Analysis

Why does the PCA method solve the problem in (1)? We will prove the following fact, which will be useful for analyzing the problem in (1).

Fact 0.1. *Let $X \in \mathbb{R}^{n\times d}$ and $k \le r =$rank(X).*

$$\min_{M\in\mathbb{R}^{d\times d}\ \textit{rank}(M)=k} \|X - XM\|_F^2 = \min_{D\in\mathbb{R}^{k\times d}, E\in\mathbb{R}^{d\times k}} \|X - XED\|_F^2 = \min_{D\in\mathbb{R}^{d\times k}, D^\intercal D=I} \|X - XDD^\intercal\|_F^2$$

Proof. Note that feasible sets have the three minimization problems have the following relationship:

$$\begin{aligned}\{M \in \mathbb{R}^{d\times d} \colon \text{rank}(M) = k\} &\supseteq \{ED \colon E \in \mathbb{R}^{d\times k}, D \in \mathbb{R}^{k\times d}\} \\ &\supseteq \{DD^\intercal \colon D \in \mathbb{R}^{d\times k}, D^\intercal D = I_k\}\end{aligned}$$

Since the three minimization problems are increasingly constrained, we have

$$\min_{M\in\mathbb{R}^{d\times d}\ \text{rank}(M)=k} \|X - XM\|_F^2 \ge \min_{D\in\mathbb{R}^{k\times d}, E\in\mathbb{R}^{d\times k}} \|X - XED\|_F^2 \ge \min_{D\in\mathbb{R}^{d\times k}, D^\intercal D=I} \|X - XDD^\intercal\|_F^2$$

Thus, to establish the stated equality, it suffices to show

$$\min_{D\in\mathbb{R}^{k\times d}, D^\intercal D=I_k} \|X - XDD^\intercal\|_F^2 \le \min_{M\in\mathbb{R}^{d\times d}\ \text{rank}(M)=k} \|X - XM\|_F^2$$

Consider any $M \in \mathbb{R}^{d\times d}$. With slight abuse of notations, we will also write down the SVD of M as $USV^\intercal$ with rank$(M) \le k$. We can write

$$\begin{aligned}\|X - XM\|_F^2 &= \|X - XV_kV_k^\intercal + XV_kV_k^\intercal - XM\|_F^2 \\ &= \|X - XV_kV_k^\intercal\|_F^2 + \|XV_kV_k^\intercal - XM\|_F^2 + 2\text{tr}\left((X - XV_kV_k^\intercal)^\intercal(XV_kV_k^\intercal - XM)\right)\end{aligned}$$

If we can show that the trace term is zero, then the above implies

$$\|X - XM\|_F^2 \geq \|X - XV_kV_k^\intercal\|_F^2 \geq \min_{D\in\mathbb{R}^{d\times k}, D^\intercal D = I_k} \|X - XDD^\intercal\|_F^2$$

which will complete the proof.

Now we show that the trace term is zero.

$$\begin{aligned}
&\text{tr}\left((X - XV_kV_k^\intercal)^\intercal(XV_kV_k^\intercal - XM)\right)\\
=&\text{tr}\left((I - V_kV_k^\intercal)^\intercal X^\intercal(X - XU_kS_kV_k^\intercal)V_kV_k^\intercal\right)\\
=&\text{tr}\left(X^\intercal(X - XU_kS_kV_k^\intercal)V_kV_k^\intercal(I - V_kV_k^\intercal)^\intercal\right) \qquad \text{(cyclic property of trace)}
\end{aligned}$$

Let us focus on the last two factors:

$$(I - V_kV_k^\intercal)^\intercal V_kV_k^\intercal = \left(\sum_{j=1}^{d} v_jv_j^\intercal - \sum_{j=1}^{k} v_jv_j^\intercal\right)\sum_{j=1}^{k} v_jv_j^\intercal = 0$$

Thus, $\text{tr}\left((X - XV_kV_k^\intercal)^\intercal(XV_kV_k^\intercal - XM)\right) = 0$. □

The fact above suggests that it suffices to solve the following special case of the "encode-decode" problem with the same matrix D:

$$\min_{D\in\mathbb{R}^{d\times k}, D^\intercal D = I} \|X - XDD^\intercal\|_F^2 \tag{2}$$

Furthermore, since

$$\|XDD^\intercal\|_F^2 = \text{tr}((XDD^\intercal)^\intercal(XDD^\intercal)) = \text{tr}((XD)^\intercal(XDD^\intercal D)) = \text{tr}((XD)^\intercal(XD)) = \|XD\|_F^2$$

Then the objective can be further re-written as

$$\begin{aligned}
\|X - XDD^\intercal\|_F^2 &= \|X\|_F^2 - 2\text{tr}((XDD^\intercal)^\intercal X) + \|XDD^\intercal\|_F^2\\
&= \|X\|_F^2 - 2\text{tr}(D^\intercal X^\intercal XD) + \|XD\|_F^2\\
&= \|X\|_F^2 - \|XD\|_F^2
\end{aligned}$$

Thus, (2) is equivalent to

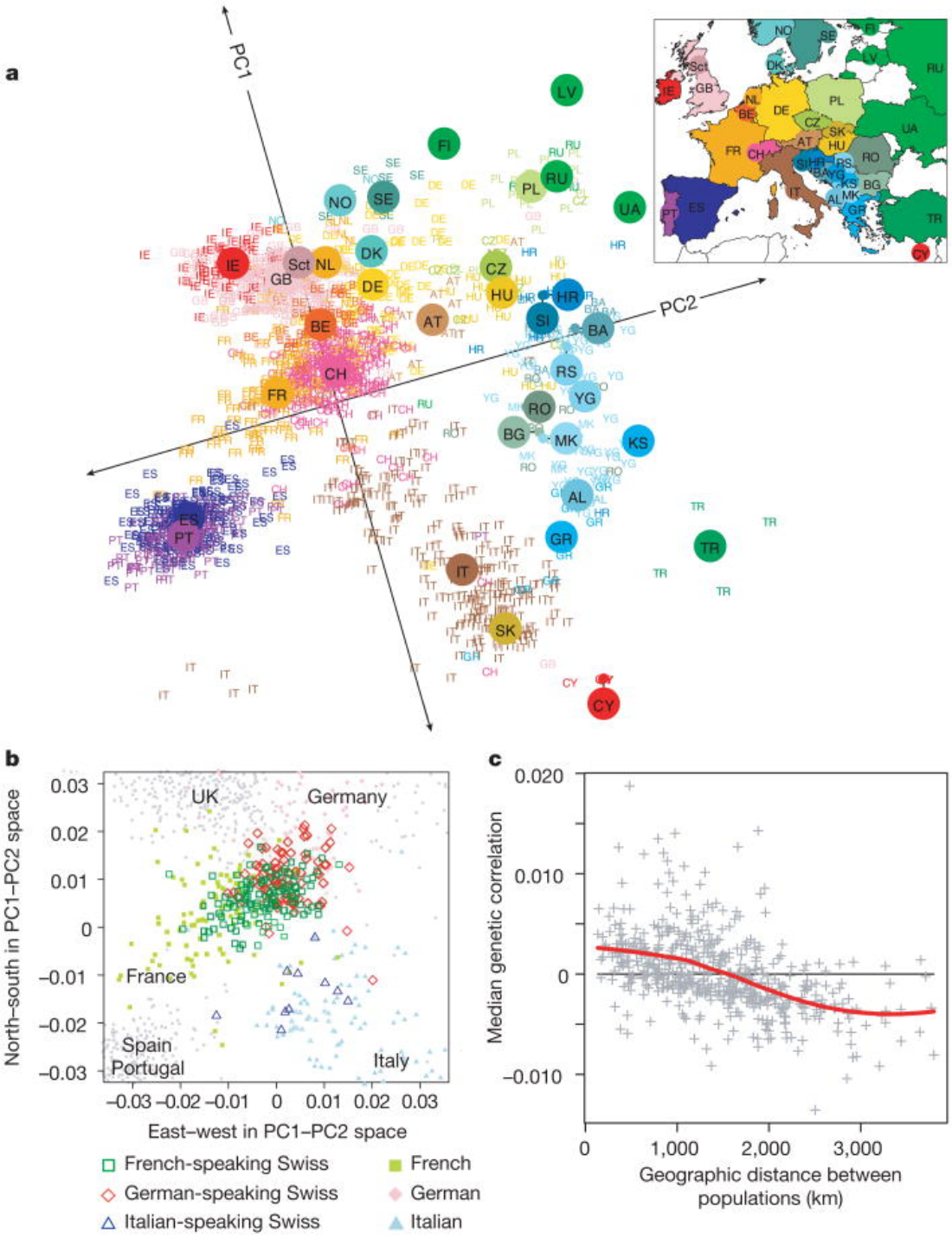
$$\min_{D\in\mathbb{R}^{d\times k}, D^\intercal D = I} \|X\|_F^2 - \|XD\|_F^2 \qquad \Leftrightarrow \qquad \max_{D\in\mathbb{R}^{d\times k}, D^\intercal D = I} \|XD\|_F^2$$

Finally, it can be shown that

$$\max_{D\in\mathbb{R}^{d\times k}, D^\intercal D = I} \|XD\|_F^2 = \|XV_k\|_F^2 = \sum_{i=1}^{k} s_i^2$$

where V_k from the truncated SVD of X.

PCA example. In this famous paper, the authors performed PCA over the genome data of 1,387 Europeans, and show that the structure of the projected data looks remarkably like the geographic map of Europe. See the Figure below.


a
PC1
PC2
b
North–south in PC1–PC2 space
East–west in PC1–PC2 space
UK
Germany
France
Spain
Portugal
Italy
French-speaking Swiss
German-speaking Swiss
Italian-speaking Swiss
French
German
Italian
c
Median genetic correlation
Geographic distance between populations (km)