Mathematics for Data Sciences **September 6, 2011**

Lecture 1. Introduction, Sample Mean, Variance, and PCA

Instructor: Yuan Yao, Peking University *Scribe: Bowei Yan*

Introduction

In this very first lecture, we talk about data representation as vectors, matrices (*esp.* graphs, networks), and tensors, *etc.* Data are mappings of real world based on sensory measurements, whence the real world puts constraints on the variations of data. Data science is the study of laws in real world which shapes the data.

We start the first topic on sample mean and variance in high dimensional Euclidean spaces, and show they are maximal likelihood estimators based on multivariate Gaussian assumption. Principle Component Analysis (PCA) is the projection of high dimensional data on its top singular vectors.

Sample Mean, Sample Variance, Principal Component Analysis

Let $x_1, ..., x_n \in \mathbb{R}^p$ are sampled from a distribution $p(X)$ on $\mathbb{R}^p$,

$$\mathbb{E}[X] = \mu, \quad Cov[X] = \mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma \in \mathbb{R}^{p\times p}$$

Take normal distribution as an example:

$$\mathcal{N}(\mu,\Sigma) \sim \frac{1}{\sqrt{2\pi|\Sigma|}} \exp[-(X-\mu)^T\Sigma^{-1}(X-\mu)],$$

where $|\Sigma|$ is the determinant of covariance matrix Σ. The maximum likelihood estimators (MLE) of μ and Σ are[1]

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T.$$

The Law of Large Numbers (LLN) tells us that for fixed p, when $n \to \infty$, $\hat{\mu}_n \to \mu$ and $\hat{\Sigma}_n \to \Sigma$. However as we can see in the following classes, they are not the best estimators when the dimension of the data p gets large.

To get the MLE of normal distribution, we need to

$$\max_{\mu,\Sigma} P(x_1, ..., x_n|\mu,\Sigma) = \max_{\mu,\Sigma} \prod_{i=1}^n \frac{1}{\sqrt{2\pi|\Sigma|}} \exp[-(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)]$$

It is equivalent to maximize the log-likelihood

$$I = \log P(x_1, ..., x_n|\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log|\Sigma| + C$$

[1] There is a constant difference in variance-covariance matrix as will be shown later.

Let μ^* is the MLE of μ, we have

$$0 = \frac{\partial I}{\partial \mu^*} = -\sum_{i=1}^{n} \Sigma^{-1}(X_i - \mu^*)$$

$$\Rightarrow \mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}_n$$

To get the estimation of Σ, we need to maximize

$$I(\Sigma) = tr(I) = -\frac{1}{2} tr \sum_{i=1}^{n}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu) - \frac{n}{2} tr \log|\Sigma| + C$$

$$\begin{aligned}
-\frac{1}{2} tr \sum_{i=1}^{n}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu) &= -\frac{1}{2}\sum_{i=1}^{n} tr[\Sigma^{-1}(X_i - \mu)(X_i - \mu)^T] \\
&= -\frac{1}{2}(tr \Sigma^{-1}\hat{\Sigma}_n)(n-1) \\
&= -\frac{n-1}{2} tr(\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}\hat{\Sigma}_n^{\frac{1}{2}}) \\
&= -\frac{n-1}{2} tr(\hat{\Sigma}_n^{\frac{1}{2}}\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}) \\
&= -\frac{n-1}{2} tr(S)
\end{aligned}$$

where $S = \hat{\Sigma}_n^{\frac{1}{2}}\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}$ is symmetric and positive definite. Above we repeatedly use $tr(AB) = tr(BA)$. Then we have

$$\Sigma = \hat{\Sigma}_n^{-\frac{1}{2}} S^{-1} \hat{\Sigma}_n^{-\frac{1}{2}}$$

$$-\frac{n}{2}\log|\Sigma| = \frac{n}{2}\log|S| + \frac{n}{2}\log|\hat{\Sigma}_n| = f(\hat{\Sigma}_n)$$

Therefore,

$$\max I(\Sigma) \Leftrightarrow \min \frac{n-1}{2} tr(S) - \frac{n}{2}\log|S| + Const(\hat{\Sigma}_n, 1)$$

Suppose $S = U\Lambda U$ is the eigenvalue decomposition of S, $\Lambda = diag(\lambda_i)$

$$J = \frac{n-1}{2}\sum_{i=1}^{p}\lambda_i - \frac{n}{2}\sum_{i=1}^{p}\log(\lambda_i) + Const$$

$$\frac{\partial J}{\partial \lambda_i} = \frac{n-1}{2} - \frac{n}{2}\frac{1}{\lambda_i} \Rightarrow \lambda_i = \frac{n}{n-1}$$

$$S = \frac{n}{n-1} I_p$$

This gives the MLE solution

$$\Sigma^* = \frac{n-1}{n}\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T,$$

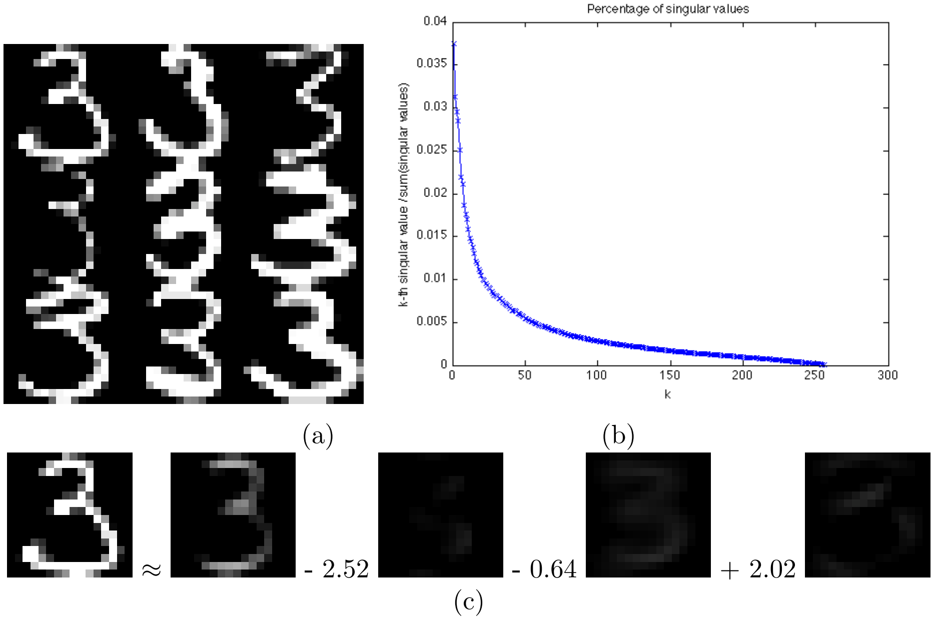
which differs to $\hat{\Sigma}_n$ only in that the denominator $(n-1)$ is replaced by n. In covariance matrix, $(n-1)$ is used because for a single sample $n = 1$, there is no variance at all.

PCA takes the eigenvector decomposition of $\hat{\Sigma}_n = \hat{V}\hat{\Lambda}\hat{V}^T$ and studies its top k eigenvectors $\hat{v}_1, \ldots, \hat{v}_k$ as the principle components. This is equivalent to the singular value decomposition (SVD) of $\hat{X} = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$ in the following sense,

$$\tilde{X} = \hat{X} - \frac{1}{n} ee^T \hat{X} = \tilde{U}\tilde{S}\tilde{V}^T, \quad e = (1, \ldots, 1)^T \in \mathbb{R}^n$$

where top right singular vectors $\tilde{v}_1, \ldots, \tilde{v}_k$ gives the same principle components.

Example. Take the dataset of hand written digit '3', $\hat{X} \in \mathbb{R}^{658 \times 256}$ contains 658 images, each of which is of 16-by-16 grayscale image as hand written digit 3. Figure 1 shows a random selection of 9 images, the sorted singular values divided by total sum of singular values, and an approximation of x_1 by top 3 principle components: $x_1 = \hat{\mu}_n - 2.5184\tilde{v}_1 - 0.6385\tilde{v}_2 + 2.0223\tilde{v}_3$.



(a) (b)

(c)

Figure 1: (a) random 9 images. (b) percentage of singular values over total sum. (c) approximation of the first image by top 3 principle components (singular vectors).

Although sample mean $\hat{\mu}_n$ and sample variance $\hat{\Sigma}_n$ are the most common statistics widely used in data analysis, they may suffer some problems in high dimensional settings, *e.g.* for large p and small n scenario. In 1956, Stein shows that the sample mean is not the best estimator in terms of the mean square error, for $p > 2$; moreover in 2006, Jonestone shows by random matrix theory that the sample variance is far from consistency for fixed ratio p/n. Among other works, these two pieces of excellent works inspired a long pursuit toward modern high dimensional statistics with a large unexplored field ahead.