COS513: FOUNDATIONS OF PROBABILISTIC MODELS LECTURE 15

ALEX LORBERT, GUNGOR POLATKAN

1. Dimensionality Reduction

The goal of dimensionality reduction is to compute a reduced representation of our data. The benefits of such a reduction include visualization of data, storage of data, and the *possible* extraction of systematic structures. In general, if we have a p-dimensional vector $(x_1, x_2, \ldots, x_p)$ we wish to find a way to represent this vector with q-dimensions as $(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_q)$ with $p > q$. In this lecture we assume only real valued vectors.

2. Principal Component Analysis (PCA)

The main idea of PCA is to project our data to a lower dimensional manifold. For example, if $p = 2$ and our data "seem" linear ($q = 1$) then we wish to project the data points onto a "suitable" line (see Figure 1). This projection is not without cost since our data do not really live on a line. In PCA our free parameter is the selection of q.

There are at least three ways to think about our lower dimensional subspace:

(1) We can maximize the variance of the projection along $\mathbb{R}^q$ [1]. In the previous example, a selection of a horizontal line results in the projected data points being "squashed".
(2) We can minimize the <u>reconstruction error</u>, i.e. the distance between the the original data and the projected data [2]. [Note: this is <u>not</u> the same as regression where we minimize the RSS].
(3) We can view PCA via an MLE of a parameter in a latent variable model [3].

3. The Multivariate Gaussian Distribution

The probability density function of a Gaussian random vector $X \in \mathbb{R}^p$ is

$$p\left(x|\mu, \Sigma\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right\}$$

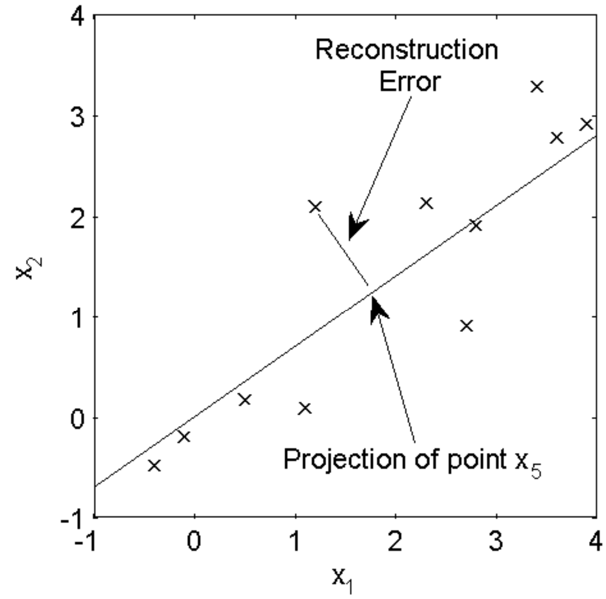

FIGURE 1. Example of dimensionality reduction with $p = 2$ and $q = 1$.

with mean $\mu \in \mathbb{R}^p$ and symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. If we let X_i denote the i^{th} element of X and σ_{ij} denote the ij^{th} element of Σ then we have the following relationships:

$$\begin{aligned} \mu_i &= \mathbb{E}\left[X_i\right] && \text{(mean)} \\ \sigma_{ij} &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] && \text{(covariance)} \\ \sigma_{ii} &= \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2 && \text{(variance)} \end{aligned}$$

Letting $f(x) = -\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)$ defines contours of equal probability (see Figure 2). When Σ is diagonal the elements of X are uncorrelated implying statistical independence in the case of Gaussian random vectors.

3.1. **MLE of the Multivariate Gaussian.** Let $X_1, \ldots, X_N \in \mathbb{R}^p$ denote iid Gaussian random vectors. The MLE is given by

$$\left(\widehat{\mu}, \widehat{\Sigma}\right) = \arg\max_{(\mu,\Sigma)} \sum_{n=1}^{N} \log p\left(x_n | \mu, \Sigma\right)$$

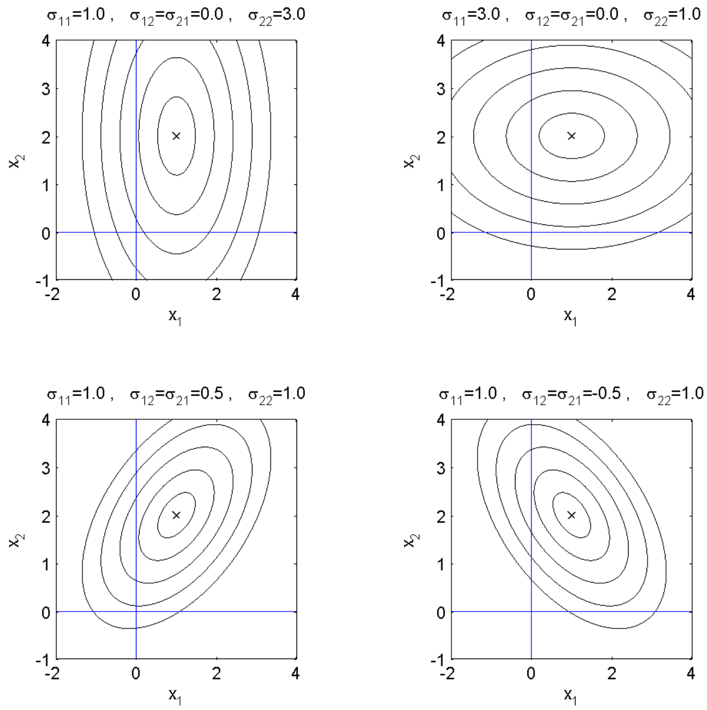

FIGURE 2. Some examples of multivariate Gaussian equiprobable contours in 2 dimensions with $\mu = [1\ 2]^T$

and its solution is given by

$$\begin{aligned} \widehat{\mu} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\ \widehat{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N} (x_n - \widehat{\mu})(x_n - \widehat{\mu})^T \end{aligned}$$

3.2. **Subvectors of Multivariate Gaussian Random Vectors.** For p-dimensional Gaussian random vector $X = \langle X_1, X_2, \ldots, X_p\rangle$ we can write $X = \left\langle \widetilde{X}_1, \widetilde{X}_2 \right\rangle$ where $\widetilde{X}_1 = \langle X_1, \ldots, X_k\rangle$ and $\widetilde{X}_1 = \langle X_{k+1}, \ldots, X_p\rangle$. It follows that $\widetilde{X}_1$ and $\widetilde{X}_2$ are jointly normal and for X having mean μ and covariance matrix Σ, we partition our parameters as follows:

$$\mu = \langle \mu_1, \mu_2 \rangle \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

with

$$\mu_i = \mathbb{E}\left[\widetilde{X}_i\right], \text{ and } \Sigma_{ij} = \mathbb{E}\left[\widetilde{X}_i \widetilde{X}_j^T\right].$$

Using the chain rule we obtain $p(\widetilde{x}_1, \widetilde{x}_2) = p(\widetilde{x}_2)\, p(\widetilde{x}_1|\widetilde{x}_2)$. To obtain the marginal of $\widetilde{X}_2$, denoted X_m, we marginalize over $\widetilde{X}_1$ and obtain the Gaussian random vector with mean $\mu_m = \mu_2$ and covariance matrix $\Sigma_m = \Sigma_{22}$. The conditional variable $\widetilde{X}_1|\widetilde{X}_2$, denoted X_c, is also Gaussian with parameters

$$\begin{aligned} \mu_c &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\widetilde{x}_2 - \mu_2) \\ \Sigma_c &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad \text{(Schur complement)} \end{aligned}$$

4. PCA AND FACTOR ANALYSIS (FA)

We begin with the latent variable graphical model (see Figure 3): the latent random variable $z_n \sim \mathcal{N}(\mathbf{0}, I_q)$, observed variable $x_n \sim \mathcal{N}(\mu + \Lambda z_n, \Psi)$, and with parameters $\Lambda \in \mathbb{R}^{p\times q}$ and positive definite diagonal matrix $\Psi \in \mathbb{R}^{q\times q}$ (I_k is the $k \times k$ identity matrix). Without loss of generality, we can assume $\mu = \mathbf{0}$ since we can always center the observed data.

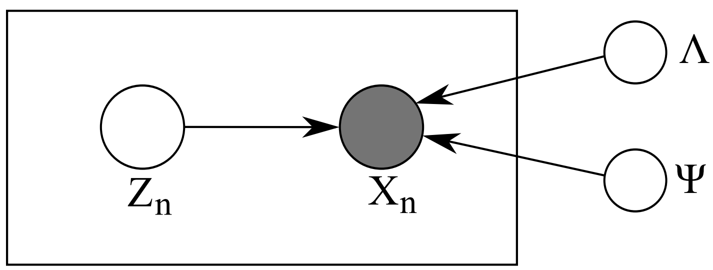

FIGURE 3. Latent variable graphical model

We can view this as a generative process as also seen by samples at Figure 4: for each $n = 1, \ldots, N$

(1) Select a random point on the q-manifold with distribution $\mathcal{N}(\mathbf{0}, I_q) \implies z_n$
(2) Use Λ to map this random point to $\mathbb{R}^p \implies \Lambda z_n$
(3) Select a random point in $\mathbb{R}^p$ with distribution $\mathcal{N}(\Lambda z_n, \Psi) \implies x_n$

The difference between PCA and FA is the structure of Ψ:

$$\begin{array}{ll} \text{PCA} & \Psi = diag(\sigma^2 \mathbf{1}_p) \\ \text{FA} & \Psi = diag\left(\langle \sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2 \rangle\right) \end{array}$$

where $\mathbf{1}_k$ is the k-dimensional vector of all ones. The solution to PCA is exact and involves selecting the eigenvectors of $[x_1|x_2|\cdots|x_n]\times[x_1|x_2|\cdots|x_n]^T$ corresponding to the largest p eigenvalues (in magnitude). FA, on the other hand, does not have an explicit solution and so we rely on the EM algorithm. The graphical model above has a strong resemblance to the regression model (z_n resemble the covariates and x_n resemble the response). The

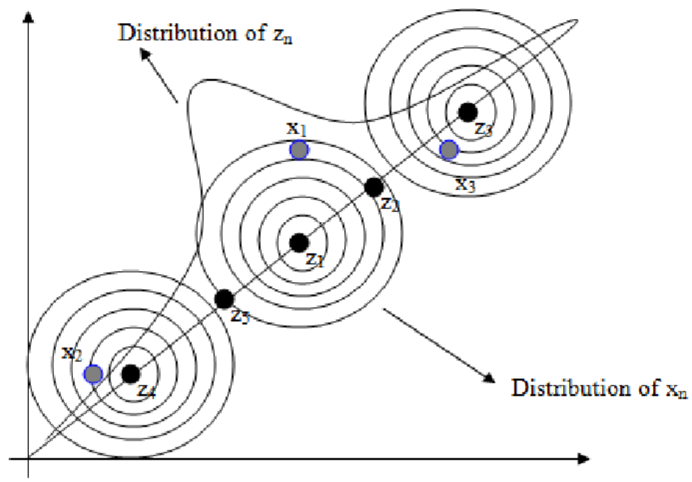

FIGURE 4. A few samples from the generative process given at Figure 3

steps of the EM algorithm are

$$\begin{array}{ll} \text{E step} & z|x \\ \text{M step} & \widehat{\Lambda} \end{array}$$

where

$$\widehat{\Lambda}^{(t+1)} = \left(\sum_{n=1}^{N} \mathbb{E}\left[z_n z_n^T | x \right] \right)^{-1} \left(\sum_{n=1}^{N} \mathbb{E}\left[z_n | x_n \right]^T x_n \right)$$

which resembles the normal equations.

REFERENCES

[1] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," J. Educational Psychology, vol. 24, pp. 417–441, 1933.
[2] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philosophical Magazine, vol. 2, no. 6, pp. 559–572, 1901.
[3] C. M. Bishop and C. K. I. Williams, "Em optimization of latent-variable density models," in Advances in Neural Information Processing Systems 8, pp. 465–471, MIT Press, 1996.