Statistical Inference and Machine Learning | Homework 3

Team members : Xia Shengzhao, Chica Linares Andrés, Gloria Dal Santo, Loïc Piccot, Elio Ovide Sanchez

December 19, 2020

1 Principal Component Analysis (PCA)

This Exercise aims to perform dimensionality reduction on a given dataset (leaf.csv). The original dataset has 14 features, we have to use PCA to perform Data Compression. Note that before applying PCA, we took off the two first columns of the data which represent the plant species and the number of specimens available – they are not representative attributes of the leaves.

1.1 Eigenvalues

First of all, we center the data and we scale it by its standard deviation. Then we compute its covariance matrix to then use MATLAB function 'eig' which gives us the eigenvectors and eigenvalues matrices.

The next step is to order the eigenvalues (they are the diagonal scalars from the eigenvalues matrix) in descending order with the aim of selecting the most informative ones, it is to say the ones with the biggest values. Figure 1 shows the eigenvalues in ascending order.

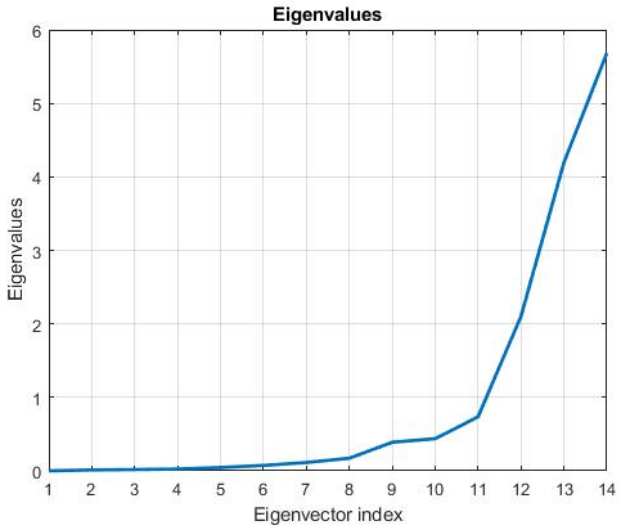

Figure 1: Eigenvalues

1.2 Cut off

Only the eigenvectors corresponding to this selected eigenvalues will be extracted from the eigenvector matrix. Generally when computing PCA, we set the desired percentage of data variability we want to keep. For example, let's say we want to keep 90% of data variability. To visualize the necessary number

of eigenvectors to reach this percentage, we compute the cumulative variance explained which is showed by the red dash line in Figure 2. We can deduce that the first four eigenvectors are sufficient.

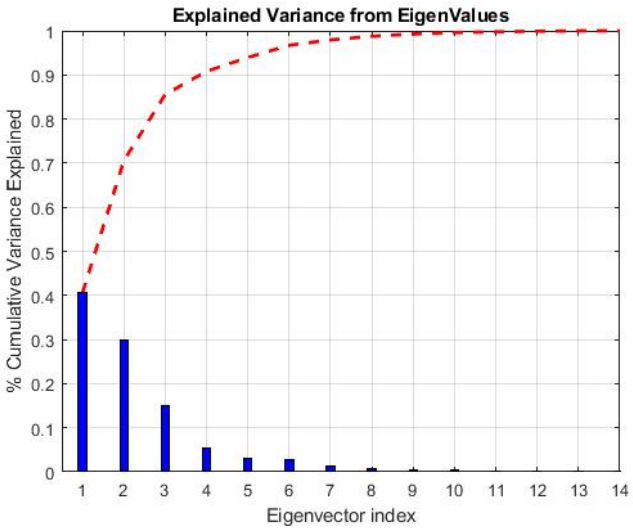

Figure 2: Cumulative Variance Explained evolution with each Eigenvectors

1.3 Data projection

According to the assignment, we consider k=2 and use the first two eigenvectors. The following figure shows a 2-D scatterplot representing all the leaves in terms of the discovered 2 principle components.

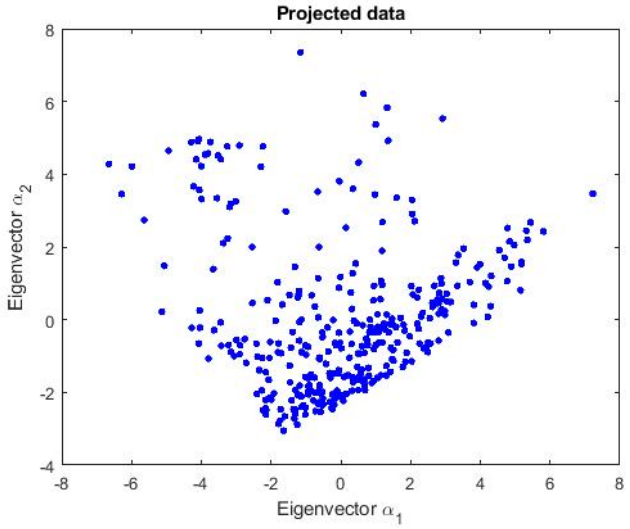

Figure 3: Leaves visualization in terms of the 2 principle components

1.4 Mean Squared reconstruction Error (MSE)

Finally, we compute the mean squared reconstruction error between the original and the reconstructed datasets.

$$MSE = \|x_i - \tilde{x}_i\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 = 0.2936$$

2 EM algorithm

To derive the Estimation and Maximization steps of the EM algorithm we applied the Jensen's inequality and the concept of Evidence Lower BOund (ELBO) as described below.

2.1 Estimation

We define $Q(z^{(i)})$, a possible distribution over $z^{(i)}$ such that $\sum_i Q(z^{(i)}) = 1$ and $Q(z^{(i)}) \geq 0$, as the conditional probability $p(z^{(i)} = k|x^{(i)}; \theta)$. In this setting, the latent variable $z^{(i)}$ corresponds to the topic c, $x^{(i)}$ represent the document D_i, assumed to be known, and θ is the set of parameters of both x and z, $\theta : \{\mu,\ \pi\}$. Then, $Q(z^{(i)})$ can be written as

$$Q(z^{(i)}) = p(z^{(i)} = k|x^{(i)}; \theta) = \frac{p(D_i|c = k; \mu)p(c = k; \pi)}{\sum_l^{n_c} p(D_i|c = l; \mu)p(c = l; \pi)}$$

recalling that $p(D_i|c = k; \mu_{jk}) = \prod_{j=1}^{n_w} \mu_{jk}^{T_{ij}}$ and $p(c = k; \pi) = \pi_k$, we obtain

$$\begin{aligned} Q(z^{(i)}) &= \frac{p(D_i|c = k; \mu)p(c = k; \pi)}{\sum_l^{n_c} p(D_i|c = l; \mu)p(c = l; \pi)} \\ &= \frac{\pi_k \prod_{j=1}^{n_w} \mu_{jk}^{T_{ij}}}{\sum_{l=1}^{n_c} \pi_l \prod_{j=1}^{n_w} \mu_{jl}^{T_{ij}}} = \gamma_{ik} \end{aligned}$$

2.2 Maximization

In order to update the mixture parameters, we introduced the ELBO proceeding as follows:

$$\begin{aligned} \theta &= \arg\max_\theta \sum_{i=1}^{n_d} \text{ELBO}(x^i; Q(z^{(i)}), \theta) \\ &= \arg\max_\theta \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} Q(z^{(i)}) \log \frac{p(z^{(i)}; \theta)p(x^{(i)}|z^{(i)}; \theta)}{Q(z^{(i)})} \\ &= \arg\max_\theta \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \log \frac{p(c = k; \theta)p(x^{(i)}|c = k; \theta)}{\gamma_{ik}} \\ &= \arg\max_\theta \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \left(\log \pi_k \prod_{l=1}^{n_w} \mu_{lk}^{T_{il}} - \log \gamma_{ik} \right) \end{aligned}$$

In the following, the term $\gamma_{ik} \log \gamma_{ik}$ has been neglected since it is irrelevant for the estimation of the parameters.
To find the update formulas, we compute the arg max with respect to the parameters π_c and μ_{jc} separately

$$\begin{aligned} \pi_c &= \arg\max_{\pi_c} \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \log \left(\pi_k \prod_{l=1}^{n_w} \mu_{lk}^{T_{il}} \right) \\ &= \arg\max_{\pi_c} \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \log \pi_k + \gamma_{ik} \sum_{l=1}^{n_w} T_{il} \log \mu_{lk} \end{aligned}$$

to find the arg max we set the gradient of the function to zero and we solve it for π_c

$$\nabla_{\pi_c} \left(\sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \log \pi_k + \lambda \left(1 - \sum_{k=1}^{n_c} \pi_k \right) \right) = 0 \tag{1}$$

The Lagrange multiplier λ ha been introduced to take into account the constraint $\sum_{k=1}^{n_c} \pi_k = 1$.
Solving eq.(1) we obtain

$$\pi_c = \frac{\sum_{i=1}^{n_d} \gamma_{ic}}{\lambda} = \frac{\sum_{i=1}^{n_d} \gamma_{ic}}{\sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik}} = \frac{1}{n_d} \sum_{i=1}^{n_d} \gamma_{ic}$$

Similarly, for μ_{jc} we have

$$\mu_{jc} = \arg\max_{\mu_{jc}} \sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \log \pi_k + \gamma_{ik} \sum_{l=1}^{n_w} T_{il} \log \mu_{lk}$$

Introducing the Lagrange multiplier λ_k relative to the constraint $\sum_{l=1}^{n_w} \mu_{lk} = 1, \forall k = 1, \dots n_c$, we finally obtain

$$\nabla_{\mu_{jc}} \left(\sum_{i=1}^{n_d} \sum_{k=1}^{n_c} \gamma_{ik} \sum_{l=1}^{n_w} T_{il} \log \mu_{lk} + \sum_{k=1}^{n_c} \lambda_k \left(1 - \sum_{l=1}^{n_w} \mu_{lk} \right) \right) = 0$$

$$\Rightarrow \mu_{jc} = \frac{1}{\lambda_c} \sum_{i=1}^{n_d} \gamma_{ic} T_{ij} = \frac{\gamma_{ic} T_{ij}}{\sum_{i=1}^{n_d} \sum_{l=1}^{n_w} \gamma_{ic} T_{il}}$$

2.3 Implementation

The accuracy of our implementation varies with the parameter initialization, we repeat the algorithm for 100 times and report its mean and standard deviation as follows:

	Mean	Standard Deviation
Accuracy	0.7612	0.0785

Table 1: Mean and Standard Deviation of Accuracy

3 MCMC algorithm

3.1 Task 1

We verify derived $\delta_k^*, a_k^*, b_k^*, \alpha_k^*, m_k^*$ respectively as follows:

- For $\delta_k^* = \delta_k + n_k$:

  We can derive it with the posterior distribution of ρ and the Bayes rule:

  $$p(\rho \mid x, z) = \frac{p(x, z \mid \rho) p(\rho)}{p(x, z)}$$

  Since $\rho \sim \text{Dirichlet}(\delta_1, \dots, \delta_K) \propto \prod_{k=1}^K \rho_k^{\delta_k - 1}$, we can obtain that:

  $$\begin{aligned} p(\rho \mid x, z) &\propto p(x, z \mid \rho) p(\rho) \\ &\propto \prod_{k=1}^K \rho_k^{n_k} \prod_{k=1}^K \rho_k^{\delta_k - 1} \\ &\propto \prod_{k=1}^K \rho_k^{n_k + \delta_k - 1} \end{aligned}$$

  As $\rho \mid x, z \sim \text{Dirichlet}(\delta_1^*, \dots, \delta_K^*) \propto \prod_{k=1}^K \rho_k^{\delta_k^* - 1}$, we finally get $\prod_{k=1}^K \rho_k^{n_k + \delta_k - 1} \propto \prod_{k=1}^K \rho_k^{\delta_k^* - 1}$, which should hold in any situation and then require $\delta_k^* = \delta_k + n_k$.

- For $a_k^* = a_k + n_k$ and $b_k^* = b + \sum_{i:z_i=k} (x_i - \mu_k)^2$:

  We can derive them similarly by the posterior distribution of ϕ and the Bayes rule. As we know:

  $$X \sim \text{Gamma}(a, \lambda) \Rightarrow p(x) = \begin{cases} \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \Rightarrow p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

, so $\phi_k \sim \text{Gamma}(a/2, b/2) \propto \phi_k^{\frac{a}{2}-1} e^{\frac{-b}{2}\phi_k}$ and $p(x, z \mid \phi_k) \propto \prod_{i:z_i=k} (\phi_k)^{\frac{1}{2}} e^{-\frac{\phi_k}{2}(x_i-\mu_k)^2}$. Then we have:

$$\begin{aligned}
p(\phi_k \mid x, z) &\propto p(x, z \mid \phi_k) p(\phi_k) \\
&\propto \left(\prod_{i:z_i=k} (\phi_k)^{\frac{1}{2}} e^{-\frac{\phi_k}{2}(x_i-\mu_k)^2}\right) \cdot (\phi_k^{\frac{a}{2}-1} e^{\frac{-b}{2}\phi_k}) \\
&= \left((\phi_k)^{\frac{n_k}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_k}(x_i-\mu_k)^2\phi_k}\right) \cdot (\phi_k^{\frac{a}{2}-1} e^{\frac{-b}{2}\phi_k}) \\
&\propto \phi_k^{\frac{a+n_k}{2}-1} e^{-\frac{1}{2}\phi_k\left(b+\sum_{i:z_i=k}(x_i-\mu_k)^2\right)} \\
&\sim \text{Gamma}\left(\frac{a+n_k}{2}, \frac{b+\sum_{i:z_i=k}(x_i-\mu_k)^2}{2}\right)
\end{aligned}$$

, which means $\phi_k \mid x, z \sim \text{Gamma}\left(\frac{a_k^*}{2}, \frac{b_k^*}{2}\right) \propto \text{Gamma}\left(\frac{a+n_k}{2}, \frac{b+\sum_{i:z_i=k}(x_i-\mu_k)^2}{2}\right)$. To make it hold in any situation, we have $a_k^* = a_k + n_k$ and $b_k^* = b + \sum_{i:z_i=k}(x_i-\mu_k)^2$.

- For $\alpha_k^* = \alpha_k + n_k$ and $m_k^* = \frac{\alpha_k m_k + n_k \bar{x}_k}{a_k + n_k}$ where $\bar{x}_k = \frac{1}{n_k}\sum_{i:z_i=k} x_i$:

  We can derive them starting from a reformulation of the likelihood function as follows

$$\begin{aligned}
\prod_{iz_i=k} p(x_i|\mu_k, \phi_k) &= \prod_{i:z_i=k} \sqrt{\phi_k} \exp\left\{-\frac{\phi_k}{2}(x_i-\mu_k)^2\right\} \\
&\propto \exp\left\{-\frac{\phi_k}{2}\sum_{i:z_i=k}(x_i-\mu_k)^2\right\} \\
&= \exp\left\{-\frac{\phi_k}{2}\sum_{i:z_i=k}(x_i^2 - 2\mu_k x_i + \mu_k^2)\right\} \\
&= \exp\left\{-\frac{\phi_k}{2}\left(\sum_{i:z_i=k} x_i^2 - 2n_k\mu_k\bar{x}_k + n_k\mu_k^2\right)\right\} \\
&= \exp\left\{-\frac{\phi_k n_k}{2}\left(\mu_k^2 - 2\mu_k\bar{x}_k + \frac{1}{n_k}\sum_{i:z_i=k} x_i^2\right)\right\} \\
&= \exp\left\{-\frac{\phi_k n_k}{2}(\mu_k - \bar{x}_k)^2\right\}
\end{aligned}$$

  with this reformulation the likelihood of the normally distributed data $\mathcal{N}_k(\mu_k, \phi_k^{-1})$ can be expressed as $\mathcal{N}_k(\bar{x}, (\phi_k n_k)^{-1})$.

Applying Bayes rule we can obtain the following expression for the posterior distribution of μ_k

$$\begin{aligned}
p(\mu_k|x,z,\phi_k) &\propto \exp\left\{\frac{\phi_k\alpha_k}{2}(\mu_k-m_k)^2\right\}\cdot\exp\left\{-\frac{\phi_k n_k}{2}\left(\mu_k-\bar{x}_k\right)^2\right\}\\
&= \exp\left\{\frac{\phi_k\alpha_k}{2}(\mu_k^2-2\mu_k m_k+m_k^2)-\frac{\phi_k n_k}{2}\left(\mu_k^2-2\mu_k\bar{x}_k+\bar{x}_k^2\right)\right\}\\
&= \exp\left\{-\frac{\phi_k}{2}\left(\mu_k^2\left(\alpha_k+n_k\right)-2\mu_k\left(\alpha_k m_k+n_k\bar{x}_k\right)+\alpha_k m_k^2+n_k\bar{x}_k^2\right)\right\}\\
&= \exp\left\{-\frac{\phi_k(\alpha_k+n_k)}{2}\left(\mu_k^2-2\mu_k\frac{(\alpha_k m_k+n_k\bar{x}_k)}{(\alpha_k+n_k)}+\frac{\alpha_k m_k^2+n_k\bar{x}_k^2}{(\alpha_k+n_k)}\right)\right\}\\
&\propto \exp\left\{-\frac{\phi_k(\alpha_k+n_k)}{2}\left(\mu_k-\frac{\alpha_k m_k+n_k\bar{x}_k}{a_k+n_k}\right)^2\right\}
\end{aligned}$$

We can see that the distribution is still Normal:

$$\mu_k|x,z,\phi_k \sim \mathcal{N}\left(\frac{\alpha_k m_k+n_k\bar{x}_k}{a_k+n_k},\frac{1}{\phi_k(\alpha_k+n_k)}\right)$$

Comparing the last equation with the expression that we expected

$$\mu_k|x,z,\phi_k \sim \mathcal{N}\left(m_k^*,\frac{1}{\phi_k\alpha_k^*}\right)$$

that is

$$\begin{aligned}
p(\mu_k|x,z,\phi_k) &\propto \exp\left\{-\frac{\phi_k\alpha_k^*}{2}\left(\mu_k-m_k^*\right)^2\right\}\\
&= \exp\left\{-\frac{\phi_k}{2}\left(\mu_k^2\alpha_k^*-2\alpha_k^* m_k^*\mu_k+\alpha_k^* m_k^{*2}\right)\right\}
\end{aligned}$$

we can see that $\alpha_k^*=\alpha_k+n_k$ and $m_k^*=\frac{\alpha_k m_k+n_k\bar{x}_k}{a_k+n_k}$ make the two distributions equivalent with respect to μ_k, concluding the proof.

3.2 Task 2

Figure 4 shows the posterior distribution of the unknown parameters evaluated on 1000 samples. Note that the variance in the distribution of μ is particularly low, hence the histogram bins are tighter.
The means of the obtained parameters, summarized in Table 2, have been used to parametrize the Gaussian mixture model whose distribution on 1000 samples is shown in Figure 5. In the figure the original samples x are compared with the distribution of $\tilde{x}$, sampled from the estimated model. It can be seen that the estimated parameters were able to describe the distribution of the original data even though not perfectly, especially around 0.5.

	$\bar{\rho}_k$	$\bar{\phi}_k$	$\bar{\mu}_k$
k_1	0.42	194.94	0.29
k_2	0.58	82.63	0.57

Table 2: Mean values of the estimated parameters

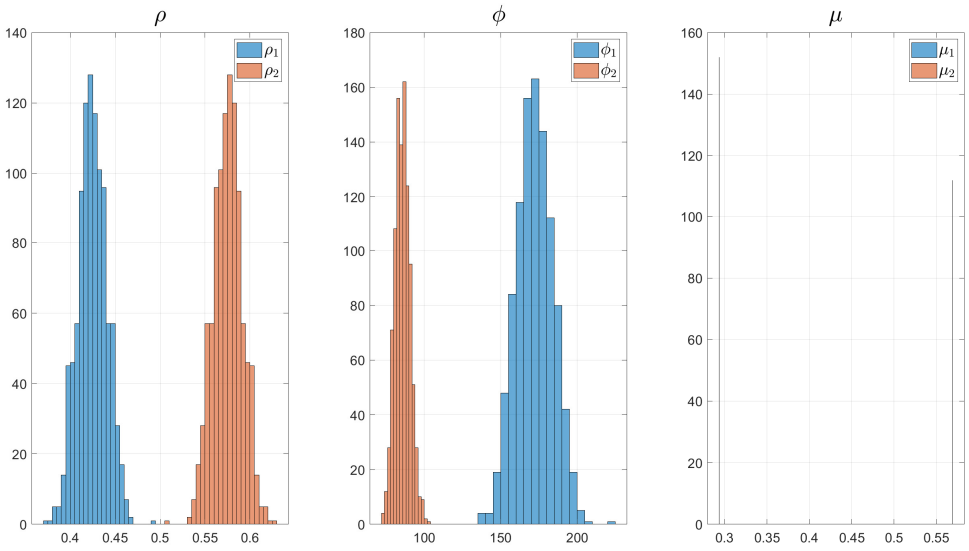

Figure 4: Posterior distribution of the unknown parameters

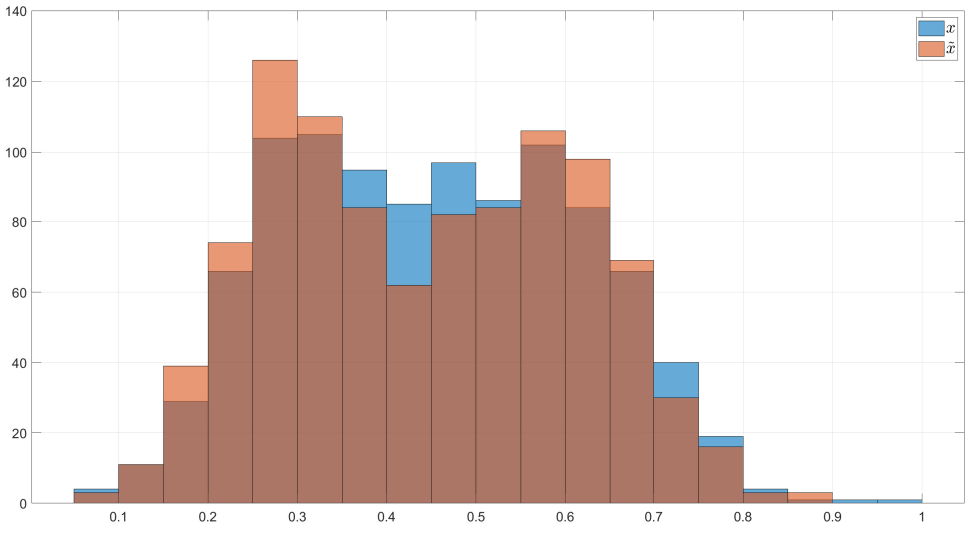

Figure 5: Original dataset samples x and estimated samples $\tilde{x}$