Fundamentals of Probabilistic Data Mining

Jean-Baptiste Durand

Duration: 3 hours. Allowed documents: handwritten notes only. No computer nor calculator nor mobile phone allowed.
Both exercises are independent. If you cannot prove a statement, you can admit it in the next questions. (not in the previous questions!) The approximate weight of each exercise in the final mark is given as a %.

Exercise 1 [55%]:

We recall the definition and main features of the probabilistic principal component analysis (PPCA). This is a model for n independent random variables $X_1^n = X_1, \ldots, X_n$, assumed as independent and satisfying

$$X_i = WZ_i + \mu + \varepsilon_i \text{ with } Z_i \sim \mathcal{N}(0, I_M),$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_D) \perp\!\!\!\perp Z_i \text{ and } W \in \mathbb{R}^{D\times M}.$$

where $M < D$.
Each $X_i \in \mathbb{R}^D$ is thus a noisy affine transformation of some latent (i.e. unobserved) $Z_i \in \mathbb{R}^M$. The unknown quantities μ, σ^2 and W are deterministic parameters.
1) Show that he marginal distribution of X_i is: $X_i \sim \mathcal{N}(\mu, WW^T + \sigma^2 I_D)$.
As a consequence, every information related to the distribution of X_1^n is contained in μ, σ^2 and WW^T.
2) Show that if the actual generative model for each X_i actually is $X_i = (WR)Z_i + \mu + \varepsilon_i$ for some orthogonal matrix R, we still have

$$X_i \sim \mathcal{N}(\mu, WW^T + \sigma^2 I_D). \tag{0.1}$$

What are the practical consequences of such statement?
3) Using the formula of conditioned Gaussian vectors (0.2) in question 1 from exercise 2, show that for every $i = 1, \ldots, n$,

$$Z_i | X_i = x_i \sim \mathcal{N}(-W^T C^{-1}(x_i - \mu), I_M - W^T C^{-1} W).$$

where $C = WW^T + \sigma^2 I_D$.

Also show that

$$Z_i | X_i = x_i \sim \mathcal{N}(-\mathcal{M}^{-1} W^T (x_i - \mu), \sigma^2 \mathcal{M}^{-1})$$

where $\mathcal{M} = W^T W + \sigma^2 I_M$.

In practice, which of both formulas would you use to perform computations? Why?

4) Compute and simplify the expression of the log-likelihood function $\ln p_{\mu,W,\sigma^2}(x_1^n)$.

Compute the maximum likelihood estimator (MLE) $\hat{\mu}_n$ of μ (as a function of x_1^n).

Let S denote the sample covariance matrix. In what follows, we admit that the MLEs of the remaining parameters satisfy:

Proposition 1

- $\hat{W}_n = \hat{U}_n(\hat{\Lambda}_n - \hat{\sigma}_n^2 I_M)^{\frac{1}{2}}$ *(up to some rotation), where* $\hat{U}_n \in \mathbb{R}^{D \times M}$ *is composed by the* M *eigenvectors of* S *with* M *largest eigenvalues and* $\hat{\Lambda}_n$ *is the diagonal matrix of the* M *largest eigenvalues.*
- $\hat{\sigma}_n^2$ *is the mean of the* $D - M$ *smallest eigenvalues.*

We will then compare direct computation of the MLE with some EM algorithm.

5) Let $\lambda = (\mu, W, \sigma)$ denote the parameter. Show that the completed loglikelihood is

$$\ln p_\lambda(x_1^n, z_1^n) = -\frac{Dn}{2} \ln(2\pi\sigma^2) - \frac{Mn}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{1}{\sigma^2} \|x_i - \mu - W z_i\|^2 + \|z_i\|^2 \right].$$

6) Deduce that, up to some quantity that does not depend from λ, the EM algorithm consists at iteration m in maximizing

$$\begin{aligned} Q(\lambda, \lambda^{(m)}) = & -\frac{Dn}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \Big\{ E_{\lambda^{(m)}}[Z_i^T Z_i | x_i] + \frac{1}{\sigma^2} \|x_i - \mu\|^2 \\ & - \frac{2}{\sigma^2} E_{\lambda^{(m)}}[Z_i | x_i]^T W^T (x_i - \mu) + \frac{1}{\sigma^2} E_{\lambda^{(m)}}[Z_i^T W^T W Z_i | x_i] \Big\}. \end{aligned}$$

with respect to λ, where $\lambda^{(m)}$ is the current value of the parameter at iteration m.

7) Explain why we can fix μ to the MLE $\hat{\mu}_n$ in every iteration.

8) Provide explicit formulas to explain how to deduce $E_{\lambda^{(m)}}[Z_i | x_i]$ and $E_{\lambda^{(m)}}[Z_i Z_i^T | x_i]$ from the data and $\lambda^{(m)}$.

9) Show that $E_{\lambda^{(m)}}[Z_i^T W^T W Z_i | x_i] = \text{tr}(E_{\lambda^{(m)}}[Z_i Z_i^T | x_i] W^T W)$.

10) Using $\nabla_A \text{tr}(AB) = B^T$, show that the reestimation step of the algorithm is given by:

$$\begin{aligned} W^{(m+1)} &= \left[\sum_{i=1}^{n} (x_i - \hat{\mu}_n) E_{\lambda^{(m)}}[Z_i | x_i]^T \right] \left[\sum_{i=1}^{n} E_{\lambda^{(m)}}[Z_i Z_i^T | x_i] \right]^{-1} \\ (\sigma^{(m+1)})^2 &= \frac{1}{Dn} \sum_{i=1}^{n} \Big\{ \|x_i - \hat{\mu}_n\|^2 - 2 E_{\lambda^{(m)}}[Z_i | x_i] [W^{(m+1)}]^T (x_i - \hat{\mu}_n) \\ & \quad + \text{tr}\left(E_{\lambda^{(m)}}[Z_i Z_i^T | x_i] [W^{(m+1)}]^T W^{(m+1)} \right) \Big\}. \end{aligned}$$

11) Compute the per iteration time-complexity of the EM algorithm. Compare it with the time-complexity of direct computation of the MLE (see Proposition 1). In which case(s) would you prefer the EM algorithm?

Exercise 2 [45%]: probabilistic graphical models

Multivariate Gaussians

1) We consider a Gaussian vector $(X_1, \ldots, X_4)$ with 0 mean and covariance matrix Σ defined as

$$\Sigma = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 \\ 2 & 0 & 4 & 1 \\ 0 & 1 & 1 & 4 \end{pmatrix}.$$

We give the upper triangular part of the symmetric matrix Σ^{-1}:

$$\begin{pmatrix} \frac{11}{12} & -\frac{1}{5} & -\frac{3}{5} & \frac{1}{5} \\ & \frac{7}{5} & \frac{1}{5} & -\frac{2}{5} \\ & & \frac{3}{5} & -\frac{1}{5} \\ & & & \frac{2}{5} \end{pmatrix}$$

We remind the formulas of conditioned Gaussian vectors (A and B being two disjoint subsets of indices): let $X_A = (X_i)_{i\in A}, \Sigma_{A,B} = (\Sigma_{i,j})_{i\in A, j\in B}$, then $X_A | X_B = x_B \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$ with

$$\begin{cases} \mu_{A|B} = \mu_A - \Sigma_{A,B}\Sigma_B^{-1}(x_B - \mu_B) \\ \Sigma_{A|B} = \Sigma_A - \Sigma_{A,B}\Sigma_B^{-1}\Sigma_{B,A}. \end{cases} \tag{0.2}$$

Using (0.2):

1. Decide whether $X_1 \perp\!\!\!\perp X_4 | X_2$ or not.

2. Decide whether $X_2 \perp\!\!\!\perp X_3 | X_1$ or not.

2) Find five marginal or conditional independence relationships between pairs of variables. Ensure that none of them can be deduced from the four other.

3) Draw some minimal undirected I-MAP for $P(X_1, \ldots, X_4)$. Is it a perfect map? Provide some detailed justification for your answers (several lines of comments required).

4) Draw some minimal directed I-MAP for $P(X_1, \ldots, X_4)$. Is it a perfect map? Provide some detailed justification for your answers (several lines of comments required).

I-equivalence PDAGs

We now consider two distributions P_1 and P_2 for a 5-dimensional random vector $(X_1, \ldots, X_5)$, having respectively DAG $\mathcal{G}_1$ and DAG $\mathcal{G}_2$ as perfect maps (Fig. 1).

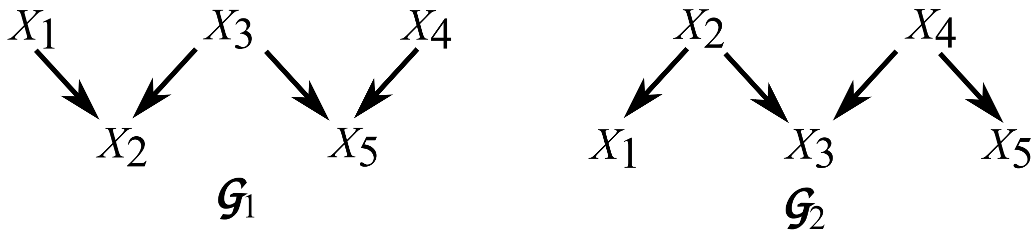

Figure 1: *Perfect independence maps $\mathcal{G}_1$ and $\mathcal{G}_2$.*

4) Draw a minimal directed I-MAP for the marginal $P_1(X_1, X_2, X_4, X_5)$. Is it a perfect I-MAP? What practical conclusions are to be drawn from your statements?

5) Draw a minimal directed I-MAP for the marginal $P_2(X_1, X_3, X_5)$. Is it a perfect I-MAP? What practical conclusions are to be drawn from your statements?