Probabilistic Graphical Models **Fall 2013**

Lecture 1 — October 9, 2013

Lecturer: Guillaume Obozinski *Scribe: Huu Dien Khue Le, Robin Bénesse*

The web page of the course: http://www.di.ens.fr/~fbach/courses/fall2013/

1.1 Introduction

1.1.1 Problem

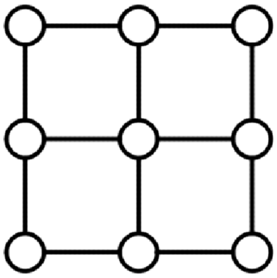
To model complex data, one is confronted with two main questions:

- How to manage the complexity of the data to be processed?
- How to infer global properties from local models?

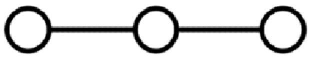
These questions lead to 3 types of problems: the representation of data (or how to obtain a global model from a local model), the inference of the distributions (how to use the model), and the learning of the models (what are the parameters of the models?).

1.1.2 Examples

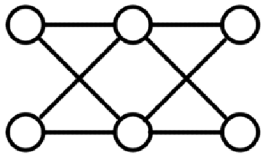
- **Image:** consider a 100×100 (pixels) monochromatic image. If each pixel is modelled by a discrete random variable (so there are 10000 of them), then the image can be modelled using a grid of the form:



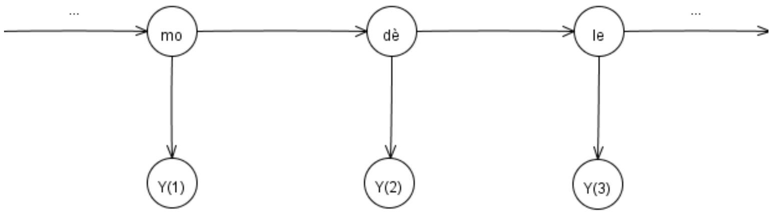
- **Bioinformatics:** consider a long sequence of 10000 ADN bases. If each base of this sequence is modelled by a discrete random variable (that, in general, can take values in $\{A, C, G, T\}$), then the sequence can be modelled by a Markov chain:



- **Finance:** consider the evolution of stock prices in discrete time, where we have values at time n. It is reasonable to postulate that the change of price of a stock at time n only depends only on its price (or the price of other stocks) at time $n-1$. For only two stocks, a possible simplified model is the following dependency graph:



- **Speech processing:** consider the syllables of a word and the way they are interpreted by a human ear or by a computer. Each syllable can be represented by a random sound. The objective is then to retrieve the word from the sequence of sounds heard or recorded. In this case, we can use a hidden Markov model:



- **Text:** consider a text with 1000000 words. The text is modelled by a vector such that each of its components equals to the number of times each keyword appears. This is usually called the “bag of words” model. This model seems to be weak, as it does not take the order of the words into account. However, it works quite well in practice. A so-called *naive Bayes classifier* can be used for classification (for example spam *vs* non spam).

It is clear that models which ignore the dependence among variables are too simple for real-world problems. On the other hand, models in which every random variable is dependent all or too many other ones are doomed both for statistical (lack of data) and computational reasons. Therefore, in practice, one has to make suitable assumptions to design models with the right level of complexity, so that the models obtained are able to *generalize* well from a statistical point of view and lead to tractable computations from an algorithmic perspective.

1.2 Basic notations and properties

In this section we recall some basic notations and properties of random variables.

Convention: Mathematically, the probability that a random variable X takes the value x is denoted $p(X = x)$. In this document, we simply write $p(X)$ to denote a distribution over the random variable X, or $p(x)$ to denote the distribution evaluated for the particular value x. It is similar for more variables.

Fundamental rules. For two random variables X, Y we have

- Sum rule:
$$p(X) = \sum_Y p(X, Y).$$

- Product rule:
$$p(X, Y) = p(Y|X)p(X).$$

Independence. Two random variables X and Y are said to be independent if and only if

$$P(X, Y) = P(X)P(Y).$$

Conditional independence. Let X, Y, Z be random variables. We define X and Y to be conditionally independent given Z if and only if

$$P(X, Y|Z) = P(X|Z)P(Y|Z).$$

Property: If X and Y are conditionally independent given Z, then

$$P(X|Y, Z) = P(X|Z).$$

Independent and identically distributed. A set of random variables is independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent.

Bayes formula. For two random variables X, Y we have

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}.$$

(Note that Bayes formula is not a Bayesian formula in the sense of Bayesian statistics).

1.3 Statistical models

Definition 1.1 (Statistical model) *A (parametric) statistical model $\mathcal{P}_\Theta$ is a collection of probability distributions (or a collection of probability density functions*[1]*) defined on the same space and parameterized by parameters θ belonging to a set $\Theta \subset \mathbb{R}^p$. Formally:*

$$\mathcal{P}_\Theta = \{p_\theta(\cdot) \mid \theta \in \Theta\}.$$

1.3.1 Bernoulli model

Consider a binary random variable X that can take the value 0 or 1. If $p(X = 1)$ is parametrized by $\theta \in [0, 1]$:

$$\begin{cases} \mathbb{P}(X = 1) = \theta \\ \mathbb{P}(X = 0) = 1 - \theta \end{cases}$$

then a probability distribution of the Bernoulli model can be written as

$$p(X = x; \theta) = \theta^x (1 - \theta)^{1-x} \tag{1.1}$$

and we can write

$$X \sim \mathrm{Ber}(\theta). \tag{1.2}$$

The Bernoulli model is the collection of these distributions for $\theta \in \Theta = [0, 1]$.

1.3.2 Binomial model

A binomial random variable $\mathrm{Bin}(\theta, N)$ is defined as the value of the sum of n i.i.d. Bernoulli r.v. with parameter θ. The distribution of a binomial random variable N is

$$\mathbb{P}(N = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

The set Θ is the same as for the Bernoulli model.

1.3.3 Multinomial model

Consider a discrete random variable C that can take one of K possible values $\{1, 2, \ldots, K\}$. The random variable C can be represented by a K-dimensional random variable $X = (X_1, X_2, \ldots, X_K)^T$ for which the event $\{C = k\}$ corresponds to the event

$$\{X_k = 1 \text{ and } X_l = 0, \forall l \neq k\}.$$

[1] In which case, they are all defined with respect to the same base measure, such as the Lebesgue measure in $\mathbb{R}^d$

If we parametrize $\mathbb{P}(C = k)$ by a parameter $\pi_k \in [0, 1]$, then by definition we also have

$$\mathbb{P}(X_k = 1) = \pi_k \ \ \forall k = 1, 2, \dots, K,$$

with $\sum_{k=1}^{K} \pi_k = 1$. The probability distribution over $\mathbf{x} = (x_1, \dots, x_k)$ can be written as

$$p(\mathbf{x}; \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{x_k} \tag{1.3}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)^T$. We will denote $\mathcal{M}(1, \pi_1, \dots, \pi_K)$ such a discrete distribution. The corresponding set of parameters is $\Theta = \{\boldsymbol{\pi} \in \mathbb{R}^+ \mid \sum_{k=1}^{K} \pi = 1\}$.

Now if we consider n independent observations of a $\mathcal{M}(1, \boldsymbol{\pi})$ multinomial random variable X, and we denote by N_k the number of observations for which $x_k = 1$, then the joint distribution of $N_1, N_2, \dots, N_K$ is called a multinomial $\mathcal{M}(n, \boldsymbol{\pi})$ distribution. It takes the form:

$$p(n_1, n_2, \dots, n_K; \boldsymbol{\pi}, n) = \frac{n!}{n_1! n_2! \dots n_K!} \prod_{k=1}^{K} \pi_k^{n_k} \tag{1.4}$$

and we can write

$$(N_1, \dots, N_K) \sim \mathcal{M}(N, \pi_1, \pi_2, \dots, \pi_K). \tag{1.5}$$

The multinomial $\mathcal{M}(n, \boldsymbol{\pi})$ is to the $\mathcal{M}(1, \boldsymbol{\pi})$ distribution, as the binomial distribution is to the Bernoulli distribution. In the rest of this course, when we will talk about multinomial distributions, we will always refer to a $\mathcal{M}(1, \boldsymbol{\pi})$ distribution.

1.3.4 Gaussian models

The Gaussian distribution is also known as the normal distribution. In the case of a scalar variable X, the Gaussian distribution can be written in the form

$$\mathcal{N}(x;\, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.6}$$

where μ is the mean and σ^2 is the variance. For a d-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{1.7}$$

where $\boldsymbol{\mu}$ is a d-dimensional vector, $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric positive definite matrix , and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. It is a well-known property that the parameter $\boldsymbol{\mu}$ is equal to the expectation of X and that the matrix $\boldsymbol{\Sigma}$ is the covariance matrix of X, which means that $\boldsymbol{\Sigma}_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$.

1.4 Parameter estimation by maximum likelihood

1.4.1 Definition

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. Suppose we have a sample $x_1, x_2, \ldots, x_n$ of n independent and identically distributed observations, coming from a distribution $p(x_1, x_2, \ldots, x_n; \theta)$ where θ is an unknown parameter (both x_i and θ can be vectors). As the name suggests, the MLE finds the parameter $\hat{\theta}$ under which the data $x_1, x_2, \ldots, x_n$ are most likely:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x_1, x_2, \ldots, x_n;\, \theta) \tag{1.8}$$

The probability on the right-hand side in the above equation can be seen as a function of θ and can be denoted by $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = p(x_1, x_2, \ldots, x_n;\, \theta) \tag{1.9}$$

This function is called the *likelihood*.
As $x_1, x_2, \ldots, x_n$ are independent and identically distributed, we have

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i;\, \theta) \tag{1.10}$$

In practice it is often more convenient to work with the logarithm of the likelihood function, called the *log-likelihood*:

$$\begin{aligned} \ell(\theta) = \log \mathcal{L}(\theta) &= \log \prod_{i=1}^{n} p(x_i;\, \theta) && (1.11) \\ &= \sum_{i=1}^{n} \log p(x_i;\, \theta) && (1.12) \end{aligned}$$

Next, we will apply this method for the models presented previously. **We assume that all the observations are independent and identically distributed** in all of the remainder of this lecture.

1.4.2 MLE for the Bernoulli model

Consider n observations $x_1, x_2, \ldots, x_n$ of a binary random variable X following a Bernoulli distribution $\mathrm{Ber}(\theta)$. From (1.12) and (1.1) we have

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log p(x_i;\, \theta) \\ &= \sum_{i=1}^{n} \log \theta^{x_i}(1-\theta)^{1-x_i} \\ &= N \log(\theta) + (n-N)\log(1-\theta) \end{aligned}$$

where $N = \sum_{i=1}^n x_i$.
As $\ell(\theta)$ is strictly concave, it has a unique maximizer, and since the function is in addition differentiable, its maximizer $\hat{\theta}$ is the zero of its gradient $\nabla\ell(\theta)$:

$$\nabla\ell(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{N}{\theta} - \frac{n-N}{1-\theta}.$$

It is easy to show that $\nabla\ell(\theta) = 0 \iff \theta = \frac{N}{n}$. Therefore we have

$$\hat{\theta} = \frac{N}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{1.13}$$

1.4.3 MLE for the multinomial model

Consider N observations $X_1, X_2, \ldots, X_N$ of a discrete random variable X following a multinomial distribution $\mathcal{M}(1, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)^T$. We denote $\mathbf{x}_i$ $(i = 1, 2, \ldots, N)$ the K-dimensional vectors of 0s and 1s representing X_i, as presented in Section 1.3.3.
From (1.12) and (1.3) we have

$$\begin{aligned}
\ell(\boldsymbol{\pi}) &= \sum_{i=1}^N \log p(\mathbf{x}_i;\, \boldsymbol{\pi}) \\
&= \sum_{i=1}^N \log\left(\prod_{k=1}^K \pi_k^{x_{ik}}\right) \\
&= \sum_{i=1}^N \sum_{k=1}^K x_{ik} \log \pi_k \\
&= \sum_{k=1}^K n_k \log \pi_k
\end{aligned}$$

where $n_k = \sum_{i=1}^N x_{ik}$ (n_k is therefore the number of observations of $x_k = 1$).
We need to maximize this quantity subject to the constraint $\sum_{k=1}^K \pi_k = 1$.

Brief review on Lagrange duality

Lagrangian. Consider the following convex optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.14}$$

where f is a convex function, $\mathcal{X} \subset \mathbb{R}^p$ is a convex set included in the domain[2] of f, $\mathbf{A} \in \mathbb{R}^{n\times p}, \mathbf{b} \in \mathbb{R}^n$.

The *Lagrangian* associated with this optimization problem is defined as

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{1.15}$$

[2] The domain of a function is the set on which the function is finite.

The vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ is called the *Lagrange multiplier vector*.

Lagrange dual function. The *Lagrange dual function* is defined as

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \tag{1.16}$$

The problem of maximizing $g(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is known as the *Lagrange dual problem.*

Max-min inequality. For any $f : \mathbb{R}^n \times \mathbb{R}^m$ and any $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, we have

$$f(w,z) \leq \max_{z \in Z} f(w,z) \Longrightarrow \min_{w \in W} f(w,z) \leq \min_{w \in W} \max_{z \in Z} f(w,z) \tag{1.17}$$

$$\Longrightarrow \max_{z \in Z} \min_{w \in W} f(w,z) \leq \min_{w \in W} \max_{z \in Z} f(w,z). \tag{1.18}$$

The last inequality is known as the *max-min inequality.*

Duality. It is easy to show that

$$\max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{A}\mathbf{x} = \mathbf{b} \\ +\infty & \text{otherwise.} \end{cases}$$

Which gives us

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) \tag{1.19}$$

Now from (1.16), (1.18) and (1.19) we have

$$\max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} f(\mathbf{x}) \tag{1.20}$$

This inequality says that the optimal value d^* of the Lagrange dual problem always lower-bounds the optimal value p^* of the original problem. This property is called the *weak duality.* If the equality $d^* = p^*$ holds, then we say that the *strong duality* holds. Strong duality means that the order of the minimization over $\mathbf{x}$ and the maximization over $\boldsymbol{\lambda}$ can be switched without affecting the result.

Slater's constraint qualification lemma. If there exists an $\mathbf{x}$ in the relative interior of $\mathcal{X} \cap \{\mathbf{A}\mathbf{x} = \mathbf{b}\}$ then strong duality holds. (Note that by definition $\mathcal{X}$ is included in the domain of f so that if $\mathbf{x} \in \mathbf{X}$ then $f(\mathbf{x}) < \infty$.)

Note that all the above notions and results are stated for the problem (1.14) only. For a more general problem and more details about Lagrange duality, please refer to [9] (chapter 5).

Back to our problem

We need to minimize

$$f(\boldsymbol{\pi}) = -\ell(\boldsymbol{\pi}) = -\sum_{k=1}^{K} n_k \log \pi_k \tag{1.21}$$

subject to the constraint $\mathbf{1}^T\boldsymbol{\pi} = 1$.
The Lagrangian of this problem is

$$L(\boldsymbol{\pi}, \lambda) = -\sum_{k=1}^{K} n_k \log \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{1.22}$$

Clearly, as $n_k \geq 0$ $(k = 1, 2, \ldots, K)$, f is convex and this problem is a convex optimization problem. Moreover, it is trivial that there exist $\pi_1, \pi_2, \ldots, \pi_K$ such that $\pi_k > 0$ $(k = 1, 2, \ldots, K)$ and $\sum_{k=1}^{K} \pi_k = 1$, so by *Slater's constraint qualification*, the problem has strong duality property. Therefore, we have

$$\min_{\boldsymbol{\pi}} f(\boldsymbol{\pi}) = \max_{\lambda} \min_{\boldsymbol{\pi}} L(\boldsymbol{\pi}, \lambda) \tag{1.23}$$

As $L(\boldsymbol{\pi}, \lambda)$ is convex with respect to $\boldsymbol{\pi}$, to find $\min_{\boldsymbol{\pi}} L(\boldsymbol{\pi}, \lambda)$, it suffices to take derivatives with respect to π_k. This yields

$$\frac{\partial L}{\partial \pi_k} = -\frac{n_k}{\pi_k} + \lambda = 0, \ k = 1, 2, \ldots, K.$$

or

$$\pi_k = \frac{n_k}{\lambda}, \ k = 1, 2, \ldots, K. \tag{1.24}$$

Substituting these into the constraint $\sum_{k=1}^{K} \pi_k = 1$ we get $\sum_{k=1}^{K} n_k = \lambda$, yielding $\lambda = N$. From this and (1.24) we get finally

$$\hat{\pi}_k = \frac{n_k}{N}, \ k = 1, 2, \ldots, K. \tag{1.25}$$

Remark: $\hat{\pi}_k$ is the fraction of the N observations for which $x_k = 1$.

1.4.4 MLE for the univariate Gaussian model

Consider n observations $x_1, x_2, \ldots, x_n$ of a random variable X following a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. From (1.12) and (1.6) we have

$$\begin{aligned} \ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log p(x_i;\, \mu, \sigma^2) \\ &= \sum_{i=1}^{n} \log \left[\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right] \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}. \end{aligned}$$

We need to maximize this quantity with respect to μ and σ^2. By taking derivative with respect to μ and then σ^2, it is easy to obtain that the pair $(\hat{\mu}, \hat{\sigma}^2)$, defined by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.26}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{1.27}$$

is the only stationary point of the likelihood. One can actually check (for example computing the Hessian w.r.t. (μ, σ^2) that this actually a maximum. We will have a confirmation of this in the lecture on exponential families.

1.4.5 MLE for the multivariate Gaussian model

Let $X \in \mathbb{R}^d$ be a Gaussian random vector, with mean vector $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ (positive definite):

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}} \frac{1}{\sqrt{\det \Sigma}} \exp\left(\frac{-(x-\mu)^\top \Sigma^{-1} (x-\mu)}{2} \right)$$

Let $x_1, \ldots, x_n$ be a i.i.d. sample. The log-likelihood is given by:

$$\begin{aligned}
\ell(\mu, \Sigma) &= \log p(x_1, \ldots, x_n;\, \mu, \Sigma) \\
&= \log \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma) \\
&= -\left(\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(\det \Sigma) + \frac{1}{2}\sum_{i=1}^{n} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right)
\end{aligned}$$

In this case, one should be careful that these log-likelihoods are not concave with respect to the pair of parameters (μ, Σ). They are concave w.r.t. μ when Σ is fixed but they are not even concave with respect to Σ when μ is fixed.

1.4.6 Digression: review on differentials

Differentiable function. A function f is differentiable at $x \in \mathbb{R}^d$ if there exists a linear form df such that:

$$f(x+h) - f(x) = df(h) + o(||h||)$$

Since $\mathbb{R}^d$ is a Hilbert space, we know that there exists $g \in \mathbb{R}^d$ such that $df_x(h) = \langle g, h \rangle$. We call g the gradient of f : $g = \nabla f(x)$.

Example 1 : if $f \mapsto a^\top x + b$ then we have :

$$f(x+h) - f(x) = a^\top h$$

and thus

$$\nabla f(x) = a.$$

Example 2 : if $f \mapsto \frac{1}{2} x^\top A x$ then we have :

$$\begin{aligned} f(x+h) - f(x) &= \frac{1}{2}(x+h)^T A(x+h) - \frac{1}{2} x^\top A x \\ &= \frac{1}{2}\left(x^\top A h + h^\top A x\right) + o\left(||h||\right) \end{aligned}$$

The gradient is then :

$$\nabla f(x) = \frac{1}{2}\left(Ax + A^\top x\right)$$

Let us first differentiate $\ell\left(\mu, \Sigma\right)$ w.r.t. μ.

We need to differentiate :

$$\mu \mapsto (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)$$

Which is equal to $f \circ g$ where :

$$\begin{array}{ccccc} f & : & \mathbb{R}^d & \rightarrow & \mathbb{R} \\ & & y & \mapsto & y^\top \Sigma^{-1} y \end{array}$$

and

$$\begin{array}{ccccc} g & : & \mathbb{R}^d & \rightarrow & \mathbb{R}^d \\ & & \mu & \mapsto & \mu - x_i \end{array}$$

Reminder : Composition of differentials

$$\begin{aligned} d(f \circ g) &= df_{g(x)}\left(dg_x(h)\right) \\ &= df_{g(x)} \circ dg_x(h) \end{aligned}$$

The differential of f is : $df_y(h) = \langle \nabla f(y), h \rangle = \langle \frac{1}{2}\left(\Sigma^{-1}y + \left(\Sigma^{-1}\right)^\top y\right), h\rangle = \langle \Sigma^{-1}y, h\rangle$ as Σ^{-1} is symmetric.

The function $\ell : \mu \mapsto (x_i - \mu)\Sigma^{-1}(x_i - \mu)$. We have :

$$d\ell_\mu(h) = \langle \Sigma^{-1}(\mu - x_i), h\rangle$$

We deduce the gradient of ℓ :

$$\nabla \ell(\mu) = \Sigma^{-1}(\mu - x_i)$$

1.4.7 Back to the MLE for the multivariate Gaussian

Remember that the function we want to differentiate is :

$$\ell\left(\mu, \Sigma\right) = -\left(\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log\left(\det \Sigma\right) + \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^\top \Sigma^{-1}\left(x_i - \mu\right)\right)$$

According to the results above, we have :

$$\begin{aligned}\nabla_\mu \ell\left(\mu, \Sigma^{-1}\right) &= \sum_{i=1}^{n}\Sigma^{-1}(\mu - x_i)\\ &= \Sigma^{-1}\left(n\mu - \sum_{i=1}^{n} x_i\right)\\ &= \Sigma^{-1}\left(n\mu - n\overline{x}\right)\end{aligned}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The gradient is equal to 0 iff :

$$\boxed{\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Let us now differentiate ℓ w.r.t. Σ^{-1}. Let $A = \Sigma^{-1}$. We have :

$$\ell\left(\mu, \Sigma\right) = -\left(\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(\det A\right) + \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^\top A\left(x_i - \mu\right)\right)$$

The last term is a real number, so it equal to its trace. Thus :

$$\begin{aligned}\ell\left(\mu,\Sigma\right) &= -\left(\frac{nd}{2}\log(2\pi)-\frac{n}{2}\log\left(\det A\right)+\frac{1}{2}\sum_{i=1}^{n}\mathrm{Trace}\left((x_i-\mu)^\top A(x_i-\mu)\right)\right)\\ &= -\left(\frac{nd}{2}\log(2\pi)-\frac{n}{2}\log(\det A)+\frac{n}{2}\mathrm{Trace}(A\widetilde{\Sigma})\right)\end{aligned}$$

where

$$\widetilde{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top$$

is the empirical covariance matrix.

Let $f : A \mapsto \frac{n}{2}\mathrm{Trace}\left(A\widetilde{\Sigma}\right)$.

We have :

$$f(A+H)-f(A)=\frac{n}{2}\mathrm{Trace}\left(H\widetilde{\Sigma}\right)$$

The gradient of the last term of ℓ is then :

$$\nabla f(A)=\frac{n}{2}\widetilde{\Sigma}$$

and :

$$\begin{aligned}df_A(H) &= \langle\nabla f(A),H\rangle\\ &= \mathrm{Trace}\left(\nabla f(A)^\top H\right)\end{aligned}$$

Let us focus on the second term. We have :

$$\begin{aligned}\log\left(\det(A+H)\right) &= \log\left(|A^{\frac{1}{2}}\left(I+A^{-\frac{1}{2}}HA^{-\frac{1}{2}}\right)A^{-\frac{1}{2}}|\right)\\ &= \log\left(|A|\right)+\log\left(\det(I+\widetilde{H})\right)\end{aligned}$$

where $\widetilde{H}=\left(A^{-\frac{1}{2}}\right)HA^{-\frac{1}{2}}$.

Let $g : A \mapsto \log\left(\det(A)\right)$ where $A=I+\widetilde{H}$.

We have :

$$\begin{aligned}\log\left(\det(I+\widetilde{H})\right)-\log\left(\det(I)\right) &= \sum_{i=1}^{d}\log(1+\lambda_j)\\ &\simeq \sum_{i=1}^{d}\lambda_j+o\left(||\widetilde{H}||\right)\\ &= \text{Trace}\left(\widetilde{H}\right)+o\left(||\widetilde{H}||\right)\end{aligned}$$

$\widetilde{H}$ is symmetric, so it can be written as :

$$\widetilde{H}=U\Lambda U^\top$$

where U is an orthogonal matrix and $\Lambda=\text{diag}(\lambda_1,...,\lambda_d)$.

$$\text{d}\left(\log\left(\det{}_A(H)\right)\right)=\text{Trace}\left(A^{-\frac{1}{2}}HA^{-\frac{1}{2}}\right)=\text{Trace}(HA^{-1})$$

We deduce the gradient of $\log\left(\det(A)\right)$:

$$\nabla\log\left(\det(A)\right)=A^{-1}$$

And the gradient of ℓ w.r.t. A is :

$$\nabla_A(\ell)=-\frac{n}{2}A^{-1}+\frac{n}{2}\widetilde{\Sigma}$$

It is equal to zero iff :

$$\widehat{\Sigma}=\widetilde{\Sigma}$$

when $\widetilde{\Sigma}$ is invertible.
Finally we have shown that the pair

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\quad\text{and}\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^\top$$

is the only stationary point of the likelihood. One can actually check (for example computing the Hessian w.r.t. (μ,Σ) that this actually a maximum. We will have a confirmation of this in the lecture on exponential families.

Bibliography

[1] J.M Amigo and M.B. Kennel. Variance estimators for the Lempel-Ziv entropy rate estimator. *Chaos*, 16:043102, 2006.

[2] Aim Fuchs Dominique Foata. *Calcul des probabilités.* 2ème édition. Dunod, 2003.

[3] Gilbert Saporta. *Probabilités, analyses des données et statistiques.* Technip, 1990.

[4] Frédéric Bonnans. *Optimisation continue, Cours et problèmes corrigés.* Dunod, 2003.

[5] Michael Jordan. *An introduction to graphical models.* In preparation.

[6] http://fr.wikipedia.org/wiki/multiplicateur_de_lagrange.

[7] S.J.D. Prince. *Computer Vision: Models Learning and Inference.* Cambridge University Press, 2012.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.