Maximum Entropy and Bayesian inference: Where do we stand and where do we go?

Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes,
Unité mixte de recherche 8506 (CNRS-Supélec-UPS 11)
Supélec, Plateau de Moulon, 3 rue Juliot-Curie, 91192 Gif-sur-Yvette, France

Abstract. In this tutorial talk, we will first review the main established tools of probability and information theories. Then, we will consider the following main questions which arise in any inference method: i) Assigning a (prior) probability law to a quantity to represent our knowledge about it, ii) Updating the probability laws when there is new piece of information, and iii) Extracting quantitative estimates from a (posterior) probabilty law.

For the first, the main tool is the Maximum Entropy Principle (MEP). For the second, we have two tools: i) Minimising the relative entropy (the Kullbak-Leibler discrepency measure), and ii) The Bayes rule. We will make precise the appropriate situations to use them as well as their possible links. For the third problem, we will see that, even if it can be handeled through decision theory, the choice of an utility function may depend on the two previous tools used to arrive at that posterior probability. Finally, these points will be illustrated through examples of inference methods for some inverse problems such as image restoration or blind source separation.

Key Words: Information theory, Entropy, Relative Entropy, Assigning and updating probabilities, Likelihood, Bayesian inference, Bayesian computation, Variational Bayes.

NOTATIONS AND INTRODUCTION

In what follows, we will use the following notations:

A discrete valued quantity of interest:	$X \in \{\omega_1, \cdots, \omega_n\}$
Probabilities:	$\boldsymbol{p} = \{p_1, \cdots, p_n\}, \quad p_j = \mathrm{P}(X = \omega_j)$
Information quantities:	$\boldsymbol{I} = \{I_1, \cdots, I_n\}, \quad I_j = \ln \frac{1}{p_j} = -\ln p_j$
Entropy [1] :	$H(\boldsymbol{p}) = \mathrm{E}\{I_j\} = -\sum_{j=1}^n p_j \ln p_j$
Prior probabilities:	$\boldsymbol{q} = \{q_1, \cdots, q_n\}$
Relative Entropy (Kullbak-Leibler):	$K(\boldsymbol{p} : \boldsymbol{q}) = \sum_{j=1}^n p_j \ln p_j/q_j$
Data type 1: K expected values:	$d_k = \mathrm{E}\{\phi_k(X)\} = \sum_{j=1}^n p_j \, \phi_k(\omega_j)$, $k = 1, \cdots, K$
Data type 2: N direct samples:	$\boldsymbol{x} = \{x_1, \cdots, x_N\}$
Data type 3: N indirect samples:	$\boldsymbol{y} = \{y_1, \cdots, y_N\}$ with $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$
Data type 4: N indirect noisy samples:	$\boldsymbol{y} = \{y_1, \cdots, y_N\}$ with $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$

For a continuous valued quantity of interest $X \in \mathcal{C}$, where $\mathcal{C}$ is a compact, we note by $p(x)$ its probability density function (pdf). Then the entropy (rate) of $p(x)$ is defined as $H(p) = -\int p(x) \ln p(x) \, \mathrm{d}x$ and the relative entropy of $p(x)$ over $q(x)$ is defined as $K(p : q) = \int p(x) \ln p(x)/q(x) \, \mathrm{d}x$.

ASSIGNING PROBABILITIES

Assigning a probability distribution to a quantity X to represent our knowledge about it depends on the nature of that knowledge. We consider first two cases: i) a set of expected values and ii) a set of direct observations on X. The main tool for the first is the Maximum Entropy Principle (MEP) [2, 3, 4, 5] and for the second is the Maximum Likelihood (ML). We then will see the link between the two approaches.

Maximum Entropy Principle (MEP)

The mathematical problem is stated as:
Given a set of data type 1: $d_k = \mathrm{E}\{\phi_k(X)\} = \sum_{j=1}^{n} p_j\,\phi_k(\omega_j),\ k = 1, \cdots, K,$ assign the probabilities $\boldsymbol{p} = \{p_1, \cdots, p_n\}$.
This problem has, in general, an infinite number of possible solutions. The main tool here to choose one of them is the Maximum Entropy Principle (MEP):
Among all the possible solutions choose the one with maximum entropy

$$\text{maximize} \quad H(\boldsymbol{p}) = -\sum_j p_j \ln p_j \quad \text{s.t.} \quad \sum_j p_j\,\phi_k(\omega_j) = d_k, \quad k = 1, \cdots, K$$

The solution is obtained by defining the Lagrangian

$$\mathcal{L} = -\sum_{j=1}^{n} p_j \ln p_j + \sum_{k=0}^{K} \lambda_k \left(\sum_{j=1}^{n} p_j\,\phi_k(\omega_j) - d_k \right)$$

and finding its stationnary point: $\begin{cases} \frac{\partial \mathcal{L}}{\partial p_j} = 0 \longrightarrow p_j = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\omega_j)\right] \\ \frac{\partial \mathcal{L}}{\partial \lambda_k} = 0 \longrightarrow -\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} = d_k \longrightarrow \lambda^* \end{cases}$,
gives the ME solution:

$$p_j = \frac{1}{Z(\boldsymbol{\lambda}^*)} \exp\left[-\sum_{k=1}^{K} \lambda_k^* \phi_k(\omega_j)\right] = \exp\left[-\lambda_0 - \sum_{k=1}^{K} \lambda_k^* \phi_k(\omega_j)\right]$$

where $\quad Z(\boldsymbol{\lambda}) = \exp[\lambda_0] = \sum_{j=1}^{n} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\omega_j)\right]$

For the continuous case, by extension, we have:

$$\text{maximize} \quad H(\boldsymbol{p}) = -\int p(x) \ln p(x)\ \mathrm{d}x \quad \text{s.t.} \quad \int p(x)\phi_k(x)\ \mathrm{d}x = d_k,\ k = 1, \cdots, K$$

Again writing the expression of the Lagrangian

$$\mathcal{L} = -\int p(x) \ln p(x)\ \mathrm{d}x + \sum_{k=0}^{K} \lambda_k \left(\int p(x)\,\phi_k(x)\ \mathrm{d}x - d_k \right)$$

and finding its stationnary point, we obtain

$$p(x) = \frac{1}{Z(\boldsymbol{\lambda}^*)} \exp\left[-\sum_{k=1}^{K} \lambda_k^* \phi_k(x)\right]$$

where $Z(\boldsymbol{\lambda}) = \exp[\lambda_0] = \int \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\omega_j)\right] \mathrm{d}x$.

In both cases, this solution has the following properties:

$$\begin{aligned} -\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} &= -\frac{\partial \lambda_0(\boldsymbol{\lambda})}{\partial \lambda_k} = \mathrm{E}\{\phi_k(X)\}, \\ -\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k \partial \lambda_l} &= -\frac{\partial \lambda_0(\boldsymbol{\lambda})}{\partial \lambda_k \partial \lambda_l} = \mathrm{E}\{\phi_k(X)\phi_l(X)\}, \end{aligned}$$

$$H = \lambda_0 + \sum_k \lambda_k \mathrm{E}\{\phi_k(X)\} \quad \text{and} \quad H_{\max} = \lambda_0 + \sum_k \lambda_k d_k.$$

For more details see [6].

Maximum Likelihood (ML)

Considering the case where we have observed a set of direct samples $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ of X and we want to assign a probability distribution $\boldsymbol{p}$ to it to represent this knowledge. The main idea behind the Maximum Likelihood (ML) approach is to consider a parametric family $p(x|\theta)$ to represent this knowledge. Then, it is assumed that the samples x_j are obtained independently from this distribution thus defining the likelihood $\mathcal{L}(\theta) = p(\boldsymbol{x};\theta) = \prod_{j=1}^{N} p(x_j|\theta)$. Then, the Maximum Likelihood estimate is defined as $\widehat{\theta} = \arg\max_\theta \{\mathcal{L}(\boldsymbol{x}|\theta)\}$. Finally, $p(x|\widehat{\theta})$ will represent the state of knowledge of this model and those data.

A particular case of parametric family is the exponential family where $p(x|\theta)$ is the the following form

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\left[-\sum_{k=1}^{K} \theta_k \phi_k(x)\right]$$

for which we can see some link between ME and ML solutions.

We also may note that, even those methods called *non-parametric*, have a parametric form. For example in Kernel based method $p(x|\theta) = \sum_{j=1}^{N} \theta_j \, h(x - x_j)$ where h is the Kernel, depends on at least $N + 1$ parameters.

Link between MEP and Maximum Likelihood (ML)

Considering the continuous case and the two following problems and their corresponding solutions:

Data type 1: $d_k = \mathrm{E}\{\phi_k(X)\} = \int p(x)\,\phi_k(x)\,\mathrm{d}x,\ k = 1, \cdots, K$

ME solution: $p(x;\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x)\right]$

λ solution of: $-\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} = d_k, \quad k = 1, \cdots, K$

Data type 2: N direct samples: $\boldsymbol{x} = \{x_1, \cdots, x_N\}$

Choosing a param. family: $p(x;\theta) = \frac{1}{Z(\theta)} \exp\left[-\sum_{k=1}^{K} \theta_k \phi_k(x)\right]$

and assuming x_j iid: $p(\boldsymbol{x};\theta) = \prod_{j=1}^{N} \frac{1}{Z(\theta)} \exp\left[-\sum_{k=1}^{K} \theta_k \phi_k(x_j)\right]$

we can define the Likelihood: $\mathcal{L}(\boldsymbol{x}|\theta) = \frac{1}{Z^n(\theta)} \exp\left[-\sum_{j=1}^{N}\sum_{k=1}^{K} \theta_k \phi_k(x_j)\right]$

and the maximum likelihood (ML) solution $\widehat{\theta} = \arg\max_\theta \{\mathcal{L}(\boldsymbol{x}|\theta)\}$ is given by: $-\frac{\partial \ln Z(\theta)}{\partial \theta_k} = \frac{1}{n}\sum_{j=1}^{N} \phi_k(x_j)$

We can then easily see the link between the two problems. We may emphasize again that this link is one of the properties of the expontential family of probability density functions. See [7, 8, 9] for more details.

UPDATING PROBABILITIES

Updating a prior probability distribution to a posterior probability distribution concerning a quantity X also depends on the nature of the new knowledge. Here too, we consider two cases: i) a set of expected values and ii) a set of direct or indirect observations on X. The main tool for the first is the Minimum Relative Entropy Principle (MREP) and the Bayesian approach for the second. We then will see the link between the two approaches.

Minimum Relative Entropy Principle

The mathematical problem is stated as: Given the prior probabilities $\boldsymbol{q}$ and a set of data type 1: $d_k = \mathrm{E}\{\phi_k(X)\} = \sum_{j=1}^{n} p_j\,\phi_k(\omega_j),\ k = 1, \cdots, K$, update $\boldsymbol{q}$ to $\boldsymbol{p}$.

The Minimum Relative Entropy Principle (MREP) writes:

$$\text{minimize } K(\boldsymbol{p} : \boldsymbol{q}) = \sum_{j=1}^{n} p_j \ln p_j/q_j \quad \text{s.t.} \quad \sum_{j=1}^{n} p_j\,\phi_k(\omega_j) = d_k,\ k = 1, \cdots, K$$

The solution is given by:

$$p_j = \frac{q_j}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\omega_j)\right] \text{ where } Z(\boldsymbol{\lambda}) = \sum_j q_j \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\omega_j)\right]$$

For the continuous case, we have:

$$\text{minimize } K(p : q) = \int p(x) \ln \frac{p(x)}{q(x)}\,\mathrm{d}x \text{ s.t. } \int p(x)\,\phi_k(x)\,\mathrm{d}x = d_k,\ k = 1, \cdots, K$$

and the solution is given by

$$p(x) = \frac{q(x)}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x)\right] \text{ where } Z(\boldsymbol{\lambda}) = \int q(x) \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x)\right] \, \mathrm{d}x$$

More details can be found in the following works [10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Bayesian approach

As in the ML approach, if we have a set of samples $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ of X for which we have choosed a parametric family $p(x|\theta)$ and a likelihood function $p(\boldsymbol{x}|\theta) = \prod_{j=1}^{N} p(x_j|\theta)$ and if we also have some prior knowledge on the unknown parameters θ in the form of a prior probability $\pi(\theta)$, then the Bayesian approach consists in computing the posterior probability

$$p(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\, p(\boldsymbol{x}|\theta)}{p(\boldsymbol{x})} = \frac{\pi(\theta)\, p(\boldsymbol{x}|\theta)}{\int \pi(\theta)\, p(\boldsymbol{x}|\theta)\, \mathrm{d}(\theta)}$$

and then choosing an estimate for θ from this posterior. The general approach is to choose a utility function $u(\theta, \tilde{\theta})$, compute its expected value $\bar{u}(\tilde{\theta}) = \int u(\theta, \tilde{\theta})\, p(\theta|\boldsymbol{x})\, \mathrm{d}\theta$ and choose as a point estimator $\widehat{\theta} = \arg\min_{\tilde{\theta}} \left\{ \bar{u}(\tilde{\theta}) \right\}$.

Of particular interest is the case of exponential families for $p(x|\theta)$ and for $\pi(\theta)$ for which we can try to see some link between MRE and the Bayesian solutions.

Link between MKL and Bayesian approach

Considering the continuous case of X with prior $q(x|\boldsymbol{\lambda}_0)$ and the
Data type 1: $d_k = \mathrm{E}\{\phi_k(X)\} = \int p(x)\, \phi_k(x)\, \mathrm{d}x, \;\; k = 1, \cdots, K,$
the MKL solution is given by

$$p(x|\boldsymbol{\lambda}) = \frac{q(x|\boldsymbol{\lambda}_0)}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x)\right]$$

where $\boldsymbol{\lambda}$ is the solution of: $\quad -\dfrac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} = d_k, \quad k = 1, \cdots, K.$
we note the relation between the prior and the posterior:

$$\begin{array}{rcll} p(x|\boldsymbol{\lambda}) & \propto & q(x|\boldsymbol{\lambda}_0) & \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x)\right] \\ \text{a posteriori} & \propto & \text{a priori} & \text{Data type 1 likelihood} \end{array}$$

Now, considering the
Data type 2: N direct samples: $\quad \boldsymbol{x} = \{x_1, \cdots, x_N\}$

with the following:

$$\begin{array}{ll} \text{Choose a param. family:} & p(x|\theta) = \frac{1}{Z(\theta)} \exp\left[-\sum_{k=1}^{K} \theta_k \phi_k(x)\right] \\ \text{Define the Likelihood:} & \mathcal{L}(\boldsymbol{x}|\theta) = \frac{1}{Z^n(\theta)} \exp\left[-\sum_{j=1}^{N}\sum_{k=1}^{K} \theta_k \phi_k(x_j)\right] \\ \text{Assign a prior on: } \theta & \pi(\theta|\boldsymbol{x}_0) \end{array}$$

and applying Bayes rule, we have:

$$\begin{array}{ccccc} p(\theta|\boldsymbol{x}) & \propto & \pi(\theta|\boldsymbol{x}_0) & \exp\left[-\sum_{j=1}^{N}\sum_{k=1}^{K} \theta_k \phi_k(x_j)\right] \\ \text{a posteriori} & \propto & \text{a priori} & \text{Data type 2 likelihood} \end{array}$$

We can then compare the two approaches. However, we may note that in MKL, we have a posterior law $p(x|\lambda)$ on x which is related to the prior law $q(x|\boldsymbol{\lambda}_0)$ and in the Bayesian approach, we have a posterior law $p(\theta|\boldsymbol{x})$ on θ which is related to the prior $\pi(\theta|\boldsymbol{x}_0)$. Note that we introduced $q(x|\boldsymbol{\lambda}_0)$ and $\pi(\theta|\boldsymbol{x}_0)$ for symmetry and for some more detailed developments. To develop more deeply these relations, consider any point estimators of θ such as:

$$\begin{array}{ll} \text{the mean:} & \widehat{\theta} = \int \theta\, p(\theta|\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{\int \theta\, \mathcal{L}(\boldsymbol{x}|\theta)\pi(\theta)\, \mathrm{d}\boldsymbol{x}}{\int \mathcal{L}(\boldsymbol{x}|\theta)\pi(\theta)\, \mathrm{d}\boldsymbol{x}} \\ \text{or the mode:} & \widehat{\theta} = \arg\max_\theta \{\pi(\theta)\, \mathcal{L}(\boldsymbol{x}|\theta)\} \end{array}$$

then, we can question ourselves on the signification of $p(\boldsymbol{x}|\widehat{\theta})$ and its link with $p(\boldsymbol{x}|\boldsymbol{\lambda})$ and a few more questions:

- How to assign $q(x|\boldsymbol{\lambda}_0)$ or $\pi(\theta|\boldsymbol{x}_0)$?
- How to use $p(x|\boldsymbol{\lambda})$ or $p(\theta|\boldsymbol{x})$?
- How to compute $\mathrm{E}\{X\}$ using $p(x|\boldsymbol{\lambda})$ or $\mathrm{E}\{\theta\}$ using $p(\theta|\boldsymbol{x})$?
- Any link between $q(x|\boldsymbol{\lambda}_0)$ and $\pi(\theta|\boldsymbol{x}_0)$ or between $p(x|\boldsymbol{\lambda})$ and $p(\theta|\boldsymbol{x})$?

MULTIVARIATE EXTENSIONS

Consider $\boldsymbol{X}$ a random vector with pdf $p(\boldsymbol{x})$, the prior $q(\boldsymbol{x}|\boldsymbol{\lambda}_0)$ and the Data type 1:

$$d_k = \mathrm{E}\{\phi_k(\boldsymbol{X})\} = \int p(\boldsymbol{x})\, \phi_k(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},\ k = 1, \cdots, K$$

The ME and MKL relations can easily be extended to this multivariate case and we have:

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) \quad \propto \quad q(\boldsymbol{x}|\boldsymbol{\lambda}_0) \quad \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{x})\right]$$

Then, the following properties can be established:

- Minimizing $K(p:q)$ becomes equivalent to minimizing is a distance measure $D(\boldsymbol{\lambda};\boldsymbol{\lambda}_0)$ between the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_0$ (Primal-Dual optimization), whose expression depends on $q(\boldsymbol{x}|\boldsymbol{\lambda}_0)$;

- If $q(\boldsymbol{x}|\boldsymbol{\lambda}_0)$ is separable then $p(\boldsymbol{x}|\boldsymbol{\lambda})$ is also separable;
- If we note by

$$\mathrm{E}_q\{\boldsymbol{X}\} = \int \boldsymbol{x}\, q(\boldsymbol{x}|\lambda_0)\, \mathrm{d}\boldsymbol{x} = \boldsymbol{x}_q \quad \text{and} \quad \mathrm{E}_p\{\boldsymbol{X}\} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\lambda)\, \mathrm{d}\boldsymbol{x} = \boldsymbol{x}_p$$

  then minimizing $K(p:q) \longrightarrow$ minimizing $\Delta(\boldsymbol{x}_p : \boldsymbol{x}_q)$.

Now, we consider the Data type 3: M indirect samples: $\boldsymbol{y} = \{y_1, \cdots, y_M\}$ where $\boldsymbol{A}$ is a $M \times N$ matrix and $\boldsymbol{y} = \mathrm{E}\{\boldsymbol{A}\boldsymbol{X}\} = \boldsymbol{A}\mathrm{E}\{\boldsymbol{X}\}$ and the prior measure $q(\boldsymbol{x}|\lambda_0)$. Then, again, it is easy to show that

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) \quad \propto \quad q(\boldsymbol{x}|\boldsymbol{\lambda}_0) \quad \exp\left[-\sum_{k=1}^{K} \lambda_k [\boldsymbol{A}\boldsymbol{x}]_k)\right]$$

and we have the following properties:

- Minimizing $K(p:q)$ becomes equivalent to minimizing $D(\boldsymbol{\lambda};\boldsymbol{\lambda}_0)$ and if we are only interested on the mean values $\boldsymbol{x}$, it can be obtained by minimizing a distance measure $\Delta(\boldsymbol{x}:\boldsymbol{x}_0)$ between $\boldsymbol{x}$ and $\boldsymbol{x}_0$ subject to the data constraints $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$. The expression of $\Delta(\boldsymbol{x};\boldsymbol{x}_0)$ depends on the family form of $q(\boldsymbol{x}|\boldsymbol{\lambda}_0)$;
- If $q(\boldsymbol{x}|\boldsymbol{\lambda}_0)$ is separable then $\Delta(\boldsymbol{x};\boldsymbol{x}_0) = \sum_{j=1}^{N} \Delta_j(x_j;x_{0j})$;
- If $q(\boldsymbol{x})$ is a Gaussian, then $D(\boldsymbol{\lambda};\boldsymbol{\lambda}_0) = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|^2$ and $\Delta(\boldsymbol{x};\boldsymbol{x}_0) = \|\boldsymbol{x} - \boldsymbol{x}_0\|^2$;
- If $q(\boldsymbol{x})$ is a Poisson measure, then $\Delta(\boldsymbol{x};\boldsymbol{x}_0) = \sum_j x_j \ln(x_j/x_{0j}) + (x_j - x_{0j})$.

See [15, 16, 18, 19, 20] for more details.

BAYESIAN APPROACH FOR INVERSE PROBLEMS

Finally, we consider the Data type 4: M indirect samples: $\boldsymbol{y} = \{y_1, \cdots, y_M\}$ where $\boldsymbol{A}$ is a $M \times N$ matrix and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$ and the prior probability laws:

$$p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|\theta_1) = \frac{1}{Z(\theta_1)} \exp\left[-\theta_1^t \boldsymbol{Q}(\boldsymbol{\epsilon})\right] \text{ and } p(\boldsymbol{x}|\theta_2) = \frac{1}{Z(\theta_2)} \exp\left[-\theta_2^t \boldsymbol{\phi}(\boldsymbol{x})\right]$$

and we consider the problem of inferring on $\boldsymbol{x}$ and the hyperparameters θ_1 and θ_1. Here, the appropriate tool is the Bayesian one.

The case where θ_1 and θ_2 are known is now classical. We have to write down the expression of the posterior

$$\begin{array}{lrcl} & p(\boldsymbol{x}|\boldsymbol{y},\theta) & = & p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\, p(\boldsymbol{x}|\theta_2)/p(\boldsymbol{y}|\theta), \quad \theta = (\theta_1,\theta_2) \\ \text{where} & p(\boldsymbol{y}|\boldsymbol{x},\theta_1) & = & p_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}|\theta_1) \\ \text{and} & p(\boldsymbol{y}|\theta) & = & \int p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\, p(\boldsymbol{x}|\theta_2)\, \mathrm{d}\boldsymbol{x} \end{array}$$

and then infer $\boldsymbol{x}$ using:

$$\begin{array}{llll} \text{Mode} & \widehat{\boldsymbol{x}}(\theta) & = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y},\theta)\} & \text{which needs optimization;} \\ \text{Mean} & \widehat{\boldsymbol{x}}(\theta) & = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y},\theta)\, \mathrm{d}\boldsymbol{x} = \frac{\int \boldsymbol{x}\, p(\boldsymbol{y}|\boldsymbol{x},\theta) p(\boldsymbol{x}|\theta)\, \mathrm{d}\boldsymbol{x}}{\int p(\boldsymbol{y}|\boldsymbol{x},\theta) p(\boldsymbol{x}|\theta)\, \mathrm{d}\boldsymbol{x}} & \text{which needs integation;} \\ \text{Sampling} & \boldsymbol{x} & \sim p(\boldsymbol{x}|\boldsymbol{y},\theta) & \text{which needs Monte Carlo techniques.} \end{array}$$

When θ_1 and θ_2 are unknown, then we have to write down the joint posterior:

$$p(\boldsymbol{x},\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\, p(\boldsymbol{x}|\theta_2)\, \pi(\theta), \quad \theta = (\theta_1,\theta_2)$$

and then, depending on the final objective, do one of the following:

- inferring $\boldsymbol{x}$:
  $p(\boldsymbol{x}|\boldsymbol{y}) = \int p(\boldsymbol{x},\theta|\boldsymbol{y})\, \mathrm{d}\theta$
  Mode $\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y})\}$
  Mean $\widehat{\boldsymbol{x}} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} = \int\int \boldsymbol{x}\, p(\boldsymbol{x},\theta|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\theta$

- inferring θ :
  $p(\theta|\boldsymbol{y}) = \int p(\boldsymbol{x},\theta|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$
  Mode $\widehat{\theta} = \arg\max_{\theta} \{p(\theta|\boldsymbol{y})\}$
  Mean $\widehat{\theta} = \int \theta\, p(\theta|\boldsymbol{y})\, \mathrm{d}\theta = \int\int \theta\, p(\boldsymbol{x},\theta|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\theta$

- inferring $(\boldsymbol{x},\theta)$: $(\boldsymbol{x},\theta) \sim p(\boldsymbol{x},\theta|\boldsymbol{y})$

  Joint MAP : $(\widehat{\boldsymbol{x}},\widehat{\theta}) = \arg\max_{\boldsymbol{x},\theta} \{p(\boldsymbol{x},\theta|\boldsymbol{y})\}$
  Gibbs sampling : $\theta \sim p(\theta|\boldsymbol{x},\boldsymbol{y}) \longrightarrow \boldsymbol{x} \sim p(\boldsymbol{x}|\theta,\boldsymbol{y})$ iterative
  Joint sampling : $\theta \sim p(\theta|\boldsymbol{y}) \longrightarrow \boldsymbol{x} \sim p(\boldsymbol{x}|\theta,\boldsymbol{y})$

Looking at these relations:

$$\begin{aligned} p(\theta,\boldsymbol{x}|\boldsymbol{y}) &= \frac{p(\boldsymbol{x}|\theta,\boldsymbol{y})\, p(\boldsymbol{y}|\theta)\, \pi(\theta)}{p(\boldsymbol{y})} \\ p(\theta|\boldsymbol{x},\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta,\boldsymbol{x})\, p(\boldsymbol{x}|\theta)}{p(\boldsymbol{y}|\theta)} \\ p(\theta|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta)\, \pi(\theta)}{p(\boldsymbol{y})} \end{aligned}$$

we see that a key term in all these relations is the incomplete likelihood (or evidence) of the parameters $p(\boldsymbol{y}|\theta)$ which is related to the complete likelihood $p(\boldsymbol{y},\boldsymbol{x}|\theta)$ by the following integral equation

$$p(\boldsymbol{y}|\theta) = \int p(\boldsymbol{y},\boldsymbol{x}|\theta)\, \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{y}|\boldsymbol{x},\theta)\, p(\boldsymbol{x}|\theta)\, \mathrm{d}\boldsymbol{x}$$

which, unfortunately, excepted the Gaussian case, has not an analytical solution. Also, noting that

$$\begin{aligned} \ln p(\boldsymbol{y}|\theta) &= \ln \int q(\boldsymbol{x}|\theta')\, \frac{p(\boldsymbol{y},\boldsymbol{x}|\theta)}{q(\boldsymbol{x}|\theta')}\, \mathrm{d}\boldsymbol{x} \\ &\geq \int q(\boldsymbol{x}|\theta') \ln \frac{p(\boldsymbol{y},\boldsymbol{x}|\theta)}{q(\boldsymbol{x}|\theta')}\, \mathrm{d}\boldsymbol{x} = H(q(\boldsymbol{x}|\theta')) + \mathrm{E}_{q(\boldsymbol{x}|\theta')} \{\ln p(\boldsymbol{y},\boldsymbol{X}|\theta)\} \end{aligned}$$

which is valid for any $q(\boldsymbol{x}|\theta')$ leads to the EM algorithm with $q(\boldsymbol{x}|\theta') = p(\boldsymbol{x}|\boldsymbol{y},\theta')$ which is the posterior law for $\boldsymbol{x}$ with the value of the parameters θ' at previous iteration.

In the same way, we have

$$\begin{aligned}\ln p(\boldsymbol{y}) &= \ln \int\int q(\boldsymbol{x},\theta)\frac{p(\boldsymbol{y},\boldsymbol{x},\theta)}{q(\boldsymbol{x},\theta)}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\theta \\ &\geq \int q(\boldsymbol{x},\theta)\ln\frac{p(\boldsymbol{y},\boldsymbol{x},\theta)}{q(\boldsymbol{x},\theta)}\,\mathrm{d}\boldsymbol{x} = H(q(\boldsymbol{x},\theta)) + \langle \ln p(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Theta})\rangle_{q(\boldsymbol{x},\theta)}\end{aligned}$$

where $\langle \ln p(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Theta})\rangle_{q(\boldsymbol{x},\theta)} = \mathrm{E}_{q(\boldsymbol{x},\theta)}\{\ln p(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\Theta})\}$. This inequality relation will lead, as we will see in the next section, to the variational Bayes when $q(\boldsymbol{x},\theta)$ is choosed to be separable, i.e; $q(\boldsymbol{x},\theta) = q_1(\boldsymbol{x}|\boldsymbol{y})\,q_2(\theta|\boldsymbol{y})$. See [13, 21, 22, 14, 15, 16, 17, 18, 19].

COMPUTATIONAL ASPECTS OF THE BAYESIAN APPROACH

Despite of the seemingly ever growing computing power, there are still problems (e.g. in image processing) for which it is difficult to optimize or integrate or sample from the joint posterior $p(\boldsymbol{x},\theta|\boldsymbol{y})$. This constitutes a need for its approximatation by simpler expressions. One of the classical tools is the Laplace approximation which can be a valid one when this joint posterior is unimodal. The second classical one is separable approximation or Variational Bayes which is summarized below.

Variational Bayes

The main idea here is that $p(\boldsymbol{x},\theta|\boldsymbol{y})$ is not, in general, separable in $\boldsymbol{x},\theta$ neither in components of $\boldsymbol{x}$ nor in components of θ. A first step then is to find two distributions $q_1(\boldsymbol{x}|\boldsymbol{y})$ and $q_2(\theta|\boldsymbol{y})$ such that $p(\boldsymbol{x},\theta|\boldsymbol{y})$ can be approximated by $p(\boldsymbol{x},\theta|\boldsymbol{y}) = q_1(\boldsymbol{x}|\boldsymbol{y})\,q_2(\theta|\boldsymbol{y})$. Then all computations are easier using $q(\boldsymbol{x},\theta|\boldsymbol{y})$ in place of $p(\boldsymbol{x},\theta|\boldsymbol{y})$. The two free distributions $q_1(\boldsymbol{x}|\boldsymbol{y})$ and $q_2(\theta|\boldsymbol{y})$ are then to be found such that $K\left(q_1\,q_2 : p\right)$ or $K\left(p : q_1\,q_2\right)$ be minimized. Writing the first one:

$$\begin{aligned}K\left(q_1\,q_2 : p\right) &= \int\int q_1(\boldsymbol{x}|\boldsymbol{y})\,q_2(\theta|\boldsymbol{y})\ln\frac{q_1(\boldsymbol{x}|\boldsymbol{y})q_2(\theta|\boldsymbol{y})}{p(\boldsymbol{x},\theta|\boldsymbol{y})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\theta \\ &= \int q_1(\boldsymbol{x}|\boldsymbol{y})\left(\int q_2(\theta|\boldsymbol{y})\ln\frac{q_1(\boldsymbol{x}|\boldsymbol{y})q_2(\theta|\boldsymbol{y})}{p(\boldsymbol{x},\theta|\boldsymbol{y})}\,\mathrm{d}\theta\right)\mathrm{d}\boldsymbol{x} \\ &= \int q_2(\theta|\boldsymbol{y})\left(\int q_1(\boldsymbol{x}|\boldsymbol{y})\ln\frac{q_1(\boldsymbol{x}|\boldsymbol{y})q_2(\theta|\boldsymbol{y})}{p(\boldsymbol{x},\theta|\boldsymbol{y})}\,\mathrm{d}\boldsymbol{x}\right)\mathrm{d}\theta\end{aligned}$$

and noting that $K\left(q_1\,q_2 : p\right)$ is a convex function of q_1 and q_2, this optimization can be done iteratively

$$\begin{aligned}\widehat{q}_1^{(t+1)}(\boldsymbol{x}|\boldsymbol{y}) &= \arg\min_{q_1}\left\{K\left(q_1\,\widehat{q}_2^{(t)} : p\right)\right\} = \frac{1}{Z_1}\exp\left[< \ln p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\Theta}) >_{q_2^{(t)}(\theta|\boldsymbol{y})}\right] \\ \widehat{q}_2^{(t+1)}(\theta|\boldsymbol{y}) &= \arg\min_{q_2}\left\{K\left(\widehat{q}_1^{(t)}\,q_2 : p\right)\right\} = \frac{\pi(\theta)}{Z_2}\exp\left[< \ln p(\boldsymbol{X},\boldsymbol{y}|\theta) >_{q_1^{(t)}(\boldsymbol{x}|\boldsymbol{y})}\right]\end{aligned}$$

where t notes the iteration number and $< . >_q$ mean the expectation over q. For more details on this approach see [23, 24].

WHERE DO WE HAVE TO GO NOW?

The main idea in this tutorial was first to give a brief review of the main established concepts. Now, the question is what are the directions to follow. Some of the different aspects which will be discussed, I am sure, in this workshop are the following:

- There are still great place to the reserach on finding axioms needed to define a quantity which will represents the *information* or the *entropy*. Depending on different levels of those axioms, we may find different expressions for the *entropy*. Then, it will be interesting to study more in details those expressions and solutions we may obtain for assigning or updating probability distributions.
- As we could see in this paper, depending on the nature of the data we may have, the tools for assigning or updating probability distributions are different. More insights and studies are still needed to establish and to interpret the possible links between them.

I can also give here some directions in relation to the subjects on which my PhD students and myself are interested. These subjects are related to the applied inverse problems. Here, I summarize those directions:

- Forward modeling and assigning a probability law to the errors which leads us to the likelihood expression is one of the crucial steps. In fact, choosing appropriate unknown quantities $\boldsymbol{x}$ and appropriate observable quantities $\boldsymbol{y}$ and finding a simple forward model:

  $$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) + \boldsymbol{\epsilon} \longrightarrow \mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \theta_1) = q_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})|\theta_1) = p(\boldsymbol{y}|\boldsymbol{x}, \theta_1)$$

  relating them in such a way that the errors $\boldsymbol{\epsilon}$ can be approximated to be independent of $\boldsymbol{x}$, centered, white and having an appropriate probability distribution is one of the first and crucial steps for real applications. In engineering sciences, this can be done through good knowledge of the physics of the problem. A few cases of such linear or nonlinear modelings can be found in [18, 19, 25, 26, 27].
- Modeling unknown quantities $\boldsymbol{x}$ and assigning probability laws:
  Simple models: $p(\boldsymbol{x}|\theta_1)$
  Models with hidden variables: $p(\boldsymbol{x}|\boldsymbol{z}, \theta_2), \quad p(\boldsymbol{z}|\theta_3)$
  Here, in general, we use Markovian models directly for $p(\boldsymbol{x}|\theta_1)$ or Hierarchical Markovian models for $p(\boldsymbol{x}|\boldsymbol{z}, \theta_2)$ and/or for $p(\boldsymbol{z}|\theta_3)$. A few cases of such linear or nonlinear modelings can be found in [26, 27].
- Assigning prior laws to the hyperparameters $p(\theta)$:
  For this step, we use often Jeffreys, Entropic [28, 29, 30, 31, 32, 33] or Conjugate priors which are inter-related. For practical applications, the Conjugate priors have been used with success in many applications.

- Obtaining expressions of the posterior laws

$$\begin{array}{lcll} p(\boldsymbol{x},\theta|\boldsymbol{y}) & \propto & p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\, p(\boldsymbol{x}|\theta_2)\, \pi(\theta), & \theta = (\theta_1,\theta_2) \\ p(\boldsymbol{x},\boldsymbol{z},\theta|\boldsymbol{y}) & \propto & p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\, p(\boldsymbol{x}|\boldsymbol{z},\theta_2)\, \pi(\boldsymbol{z}|\theta_3)\, \pi(\theta), & \theta = (\theta_1,\theta_2,\theta_3) \end{array}$$

- Using posterior laws to give practical solutions:
  From this point, the Bayesian interpretation gives us a lot of possibilities. For summarizing the posterior, one can choose between Joint Modes, Means, Marginal modes, or just sampling using the MCMC methods. However, we must be aware that:
  - Computing modes needs huge dimensional multivariate optimization;
  - Computing means needs huge dimensional multivariate integration;
  - Sampling is a good tool for exploring the whole probability density and computing approximate means. However, sampling from a non-separable multivariate probability law is not so easy.
- Finding appropriate approximations to do fast computations:
  Laplace approximation, Separable approximation, Variational and Mean Field approximations are the main tools.
- Evaluating the performances of the obtained algorithms is also one of the main crucial points.
- Evaluating the uncertainties when a solution is given should not to be forgotten.

REFERENCES

1. C. E. Shannon and W. Weaver, "The mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
2. E. T. Jaynes, "Information theory and statistical mechanics I," *Physical review*, vol. 106, pp. 620–630, 1957.
3. E. T. Jaynes, "Information theory and statistical mechanics II," *Physical review*, vol. 108, pp. 171–190, 1957.
4. E. T. Jaynes, "Prior probabilities," *IEEE Trans. Systems Science and Cybern.*, vol. SSC-4, no. 3, pp. 227–241, 1968.
5. E. T. Jaynes, "Where do we stand on maximum entropy ?," in *The Maximum Entropy Formalism* (R. D. Levine and M. Tribus, eds.), Cambridge, MA: M.I.T. Press, 1978.
6. A. Mohammad-Djafari, *A Matlab Program to Calculate the Maximum Entropy Distributions*, pp. 221–233. Laramie, WY: Kluwer Academic Publ., T.W. Grandy ed., 1991.
7. A. Mohammad-Djafari, *Maximum Entropy and Linear Inverse Problems; A Short Review*, pp. 253–264. Paris, France: Kluwer Academic Publ., ali mohammad–djafari and guy demoment ed., 1992.
8. A. Mohammad-Djafari, "On the estimation of hyperparameters in Bayesian approach of solving inverse problems," in *Proc. IEEE ICASSP*, (Minneapolis, MN), pp. 567–571, IEEE, Apr. 1993.
9. A. Mohammad-Djafari, "Maximum d'entropie et problèmes inverses en imagerie," *Traitement du Signal*, pp. 87–116, 1994.
10. G. Le Besnerais, J.-F. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1565–1578, July 1999.
11. J.-F. Bercher, G. Le Besnerais, and G. Demoment, *The maximum entropy on the mean method, noise and sensitivity*, pp. 223–232. Maximum Entropy and Bayesian Methods, Cambridge, UK: Kluwer Academic Publ., 1994.
12. C. Heinrich, J.-F. Bercher, and G. Demoment, "The maximum entropy on the mean method, correlations and implementation issues," in *Maximum Entropy and Bayesian Methods*, MaxEnt Workshops, 1996.

13. A. Mohammad-Djafari and J. Idier, "A scale invariant Bayesian method to solve linear inverse problems," in *Maximum Entropy and Bayesian Methods*, pp. 121–134, Kluwer Academic Publ., G. Heidbreder ed., 1996.
14. A. Mohammad-Djafari, "A comparison of two approaches: Maximum entropy on the mean (MEM) and Bayesian estimation (BAYES) for inverse problems," in *Maximum Entropy and Bayesian Methods*, (Berg–en–Dal, South Africa), Kluwer Academic Publ., Aug. 1996.
15. A. Mohammad-Djafari, "Entropie en traitement du signal," *Traitement du signal*, vol. Num. spécial, volume 15, no. 6, pp. 545–551, 1999.
16. A. Mohammad-Djafari, "Model selection for inverse problems: Best choice of basis function and model order selection," in *Bayesian Inference and Maximum Entropy Methods* (J. R. G. Erikson and C. Smith, eds.), p. to appear on may 2001, to appear in Amer. Inst. Physics, July 1999.
17. C. Heinrich, *Distances entropiques et informationnelles en traitement de données.* Phd thesis, Université de Paris-Sud, Orsay, France, July 1997.
18. A. Mohammad-Djafari, "Model selection for inverse problems: Best choice of basis function and model order selection.," in *Maximum Entropy and Bayesian Methods, Boise, Idaho, USA, 2-5 August. 1999* (J. Rychert, G. Erikson, and C. Smith, eds.), pp. 71–88, AIP Conference Proceedings 567, May 2001.
19. A. Mohammad-Djafari, J.-F. Giovannelli, G. Demoment, and J. Idier, "Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems," *Int. Journal of Mass Spectrometry*, vol. 215, pp. 175–193, Apr. 2002.
20. G. Demoment, J. Idier, J.-F. Giovannelli, and A. Mohammad-Djafari, "Problèmes inverses en traitement du signal et de l'image," vol. TE 5 235 of *Traité Télécoms*, pp. 1–25, Paris, France: Techniques de l'Ingénieur, 2001.
21. A. Mohammad-Djafari, *A full Bayesian approach for inverse problems*, pp. 135–143. Santa Fe, NM: Kluwer Academic Publ., K. Hanson and R.N. Silver ed., 1996.
22. A. Mohammad-Djafari, "Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems," in *Proc. IEEE ICIP*, vol. II, (Lausanne, Switzerland), pp. 473–477, 1996.
23. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.
24. Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, no. 29, pp. 245–273, 1997.
25. J. Idier, ed., *Approche bayésienne pour les problèmes inverses.* Paris: Traité IC2, Série traitement du signal et de l'image, Hermès, 2001.
26. O. Féron, B. Duchêne, and A. Mohammad-Djafari, "Microwave imaging of inhomogeneous objects made of a finite number of dielectric and conductive materials from experimental data," *accepted in Journal of Inverse Problems*, october 2005.
27. O. Féron, B. Duchêne, and A. Mohammad-Djafari, "Microwave imaging: characterization of unknown dielectric or conductive materials," in *MaxEnt05*, august 2005.
28. R. E. Kass, "The geometry of asymptotic inference," *Statistical Science*, vol. 4, no. 3, pp. 188–234, 1989.
29. R. E. Kass and L. Wasserman, "The selection of prior distributions by formal rules," *J. Amer. Statist. Assoc.*, vol. 91, pp. 1343–1370, 1996.
30. C. Rodríguez, "Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models," in *Bayesian Inference and Maximum Entropy Methods* (R. L. FRY, ed.), pp. 410–432, MaxEnt Workshops, Amer. Inst. Physics, Aug. 2001.
31. H. Snoussi and A. Mohammad-Djafari, "Penalized maximum likelihood for multivariate gaussian mixture," in *Bayesian Inference and Maximum Entropy Methods* (R. L. Fry, ed.), pp. 36–46, MaxEnt Workshops, Amer. Inst. Physics, Aug. 2001.
32. H. Snoussi and A. Mohammad-Djafari, "Information Geometry and Prior Selection.," in *Bayesian Inference and Maximum Entropy Methods* (C. Williams, ed.), pp. 307–327, MaxEnt Workshops, Amer. Inst. Physics, Aug. 2002.
33. H. Snoussi and A. Mohammad-Djafari, "Sélection d'a priori et géométrie de l'information," in *IEEE International Conference on Electronic Sciences,Information Technology and Telecommunication*, (Mahdia, Tunisia), SETIT, Mar. 2003.