Entropy **2015**, *17*, 3989-4027; doi:10.3390/e17063989

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Entropy, Information Theory, Information Geometry and Bayesian Inference in Data, Signal and Image Processing and Inverse Problems †

Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes, UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD, SUPELEC, Plateau de Moulon, 3 rue Juliot-Curie, 91192 Gif-sur-Yvette, France; E-Mail: djafari@lss.supelec.fr; Tel.: +33-169851741

† This paper is an extended version of the paper published in Proceedings of the 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21–26 September 2014.

External Editor: Kevin H. Knuth

Received: 20 November 2014 / Accepted: 5 May 2015 / Published: 12 June 2015

Abstract: The main content of this review article is first to review the main inference tools using Bayes rule, the maximum entropy principle (MEP), information theory, relative entropy and the Kullback–Leibler (KL) divergence, Fisher information and its corresponding geometries. For each of these tools, the precise context of their use is described. The second part of the paper is focused on the ways these tools have been used in data, signal and image processing and in the inverse problems, which arise in different physical sciences and engineering applications. A few examples of the applications are described: entropy in independent components analysis (ICA) and in blind source separation, Fisher information in data model selection, different maximum entropy-based methods in time series spectral estimation and in linear inverse problems and, finally, the Bayesian inference for general inverse problems. Some original materials concerning the approximate Bayesian computation (ABC) and, in particular, the variational Bayesian approximation (VBA) methods are also presented. VBA is used for proposing an alternative Bayesian computational tool to the classical Markov chain Monte Carlo (MCMC) methods. We will also see that VBA englobes joint maximum *a posteriori* (MAP), as well as the different expectation-maximization (EM) algorithms as particular cases.

Keywords: Bayes; Laplace; entropy; Bayesian inference; maximum entropy principle; information theory; Kullback–Leibler divergence; Fisher information; geometrical science of information; inverse problems

1. Introduction

As this paper is an overview and an extension of my tutorial paper in MaxEnt 2014 workshop [1], this Introduction gives a summary of the content of this paper.

The qualification Bayesian refers to the influence of Thomas Bayes [2], who introduced what is now known as Bayes' rule, even if the idea had been developed independently by Pierre-Simon de Laplace [3]. For this reason, I am asking a question of the community if we shall change Bayes to Laplace and Bayesian to Laplacian or at least mention them both. Whatever the answer, we assume that the reader knows what probability means in a Bayesian or Laplacian framework. The main idea is that a probability law $P(X)$ assigned to a quantity X represents our state of knowledge that we have about it. If X is a discrete valued variable, $\{P(X = x_n) = p_n, n = 1, \cdots, N\}$ with mutually exclusive values x_n is its probability distribution. When X is a continuous valued variable, $p(x)$ is its probability density function from which we can compute $P(a \leq X < b) = \int_a^b p(x)\,\mathrm{d}x$ or any other probabilistic quantity, such as its mode, mean, median, region of high probabilities, *etc*.

In science, it happens very often that a quantity cannot be directly observed or, even when it can be observed, the observations are uncertain (commonly said to be noisy), by uncertain or noisy, here, I mean that, if we repeat the experiences with the same practical conditions, we obtain different data. However, in the Bayesian approach, for a given experiment, we have to use the data as they are, and we want to infer it from those observations. Before starting the observation and gathering new data, we have very incomplete knowledge about it. However, this incomplete knowledge can be translated in probability theory via an *a priori* probability law. We will discuss this point later on regarding how to do this. For now, we assume that this can be done. When a new observation (data D) on X becomes available (direct or indirect), we gain some knowledge via the likelihood $P(D|X)$. Then, our state of knowledge is updated combining $P(D|X)$ and $P(X)$ to obtain an *a posteriori* law $P(X|D)$, which represents the new state of knowledge on X. This is the main esprit of the Bayes rule, which can be summarized as:

$$P(X|D) = P(D|X)P(X)/P(D). \tag{1}$$

As $P(X|D)$ has to be a probability law, we have:

$$P(D) = \sum_X P(D|X)P(X). \tag{2}$$

This relation can be extended to the continuous case. Some more details will be given in Section 2.

Associated with a probability law is the quantity of information it contains. Shannon [4] introduced the notion of quantity of information I_n associated with one of the possible values of x_n of X with

probabilities $P(X = x_n) = p_n$ to be $I_n = \ln \frac{1}{p_n} = -\ln p_n$ and the entropy H as the expected value of I_n:

$$H = -\sum_{n=1}^{N} p_n \ln p_n. \tag{3}$$

The word entropy has also its roots in thermodynamics and physics. However, this notion of entropy has no direct link with entropy in physics, even if for a particular physical system, we may attribute a probability law to a quantity of interest of that system and then define its entropy. This information definition of Shannon entropy became the main basis of information theory in many data analyses and the science of communication. More details and extensions about this subject will be given in Section 3.

As we can see up to now, we did not yet discuss how to assign a probability law to a quantity. For the discrete value variable, when X can take one of the N values $\{x_1, \cdots, x_N\}$ and when we do not know anything else about it, Laplace proposed the "*Principe d'indifférence*", where $P(X = x_n) = p_n = \frac{1}{N}, \forall n = 1, \cdots, N$, a uniform distribution. However, what if we know more, but not enough to be able to assign the probability law $\{p_1, \cdots, p_N\}$ completely?

For example, if we know that the expected value is $\sum_n x_n p_n = d$, this problem can be handled by considering this equation as a constraint on the probability distribution $\{p_1, \cdots, p_N\}$. If we have a sufficient number of constraints (at least N), then we may obtain a unique solution. However, very often, this is not the case. The question now is how to assign a probability distribution $\{p_1, \cdots, p_N\}$ that satisfies the available constraints. This question is an ill-posed problem in the mathematical sense of Hadamard [5] in the sense that the solution is not unique. We can propose many probability distributions that satisfy the constraint imposed by this problem. To answer this question, Jaynes [6–8] introduced the maximum entropy principle (MEP) as a tool for assigning a probability law to a quantity on which we have some incomplete or macroscopic (expected values) information. Some more details about this MEP, the mathematical optimization problem, the expression of the solution and the algorithm to compute it will be given in Sections 3 and 4.

Kullback [9] was interested in comparing two probability laws and introduced a tool to measure the increase of information content of a new probability law with respect to a reference one. This tool is called either the Kullback–Leibler (KL) divergence, cross entropy or relative entropy. It has also been used to update a prior law when new pieces of information in the form of expected values are given. As we will see later, this tool can also be used as an extension to the MEP of Jaynes. Furthermore, as we will see later, this criterion of comparison of two probability laws is not symmetric: one of the probability laws has to be chosen to be the reference, and then, the second is compared to this reference. Some more details and extensions will be given in Section 5.

Fisher [10] wanted to measure the amount of information that a random variable X carries about an unknown parameter θ upon which its probability law $p(x|\theta)$ depends. The partial derivative with respect to θ of the logarithm of this probability law, called the log-likelihood function for θ, is called the score. He showed that the first order moment of the score is zero, but its second order moment is positive and is also equivalent to the expected values of the second derivative of log-likelihood function with respect to θ. This quantity is called the Fisher information. It is also been shown that for the small variations of θ, the Fisher information induces locally a distance in the space of parameters Θ, if we had to compare two very close values of θ. In this way, the notion of the geometry of information is introduced [11,12].

We must be careful here that this geometrical property is related to the space of the parameters Θ for small changes of the parameter for a given family of parametric probability law $p(x|\theta)$ and not in the space of probabilities. However, for two probability laws $p_1(x) = p(x|\theta_1)$ and $p_2(x) = p(x|\theta_2)$ in the same exponential family, the Kullback–Leibler divergence $\mathrm{KL}[p_1 : p_2]$ induces a Bregman divergence $\mathrm{B}[\theta_1 : \theta_2]$ between the two parameters [13,14]. More details will be given in Section 8.

At this stage, we have almost introduced all of the necessary tools that we can use for different levels of data, signal and image processing. In the following, we give some more details for each of these tools and their inter-relations. Then, we review a few examples of their use in different applications. As examples, we demonstrate how these tools can be used in independent components analysis (ICA) and source separation, data model selection, in spectral analysis of the signals and in inverse problems, which arise in many sciences and engineering applications. At the end, we focus more on the Bayesian approach for inverse problems. We present some details concerning unsupervised methods, where the hyper parameters of the problem have to be estimated jointly with the unknown quantities (hidden variables). Here, we will see how the Kullback–Leibler divergence can help approximate Bayesian computation (ABC). In particular, some original materials concerning variational Bayesian approximation (VBA) methods are presented.

2. Bayes Rule

Let us introduce things very simply. If we have two discrete valued related variables X and Y, for which we have assigned probability laws $P(X)$ and $P(Y)$, respectively, and their joint probability law $P(X, Y)$, then from the sum and product rule, we have:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{4}$$

where $P(X, Y)$ is the joint probability law, $P(X) = \sum_Y P(X, Y)$ and $P(Y) = \sum_X P(X, Y)$ are the marginals and $P(X|Y) = P(X, Y)/P(Y)$ and $P(Y|X) = P(X, Y)/P(X)$ are the conditionals. Now, consider the situation where Y can be observed, but not X. Because these two quantities are related, we may want to infer X from the observations on Y. Then, we can use:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{5}$$

which is called the Bayes rule.

This relation is extended to the continuous valued variables using the measure theory [15,16]:

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)} \tag{6}$$

with:

$$p(y) = \int p(y|x)\, p(x)\, \mathrm{d}x. \tag{7}$$

More simply, the Bayes rule is often written as:

$$p(x|y) \propto p(y|x)\, p(x). \tag{8}$$

This writing can be used when we want to use $p(x|y)$ to compute quantities that are only dependent on the shape of $p(x|y)$, such as the mode, the median or quantiles. However, we must be careful that the denominator is of importance in many other cases, for example when we want to compute expected values. There is no need for more sophisticated mathematics here if we want to use this approach.

As we mentioned, the main use of this rule is in particular when X can not be observed (unknown quantity), but Y is observed and we want to infer X. In this case, the term $p(y|x)$ is called the likelihood (of unknown quantity X in the observed data y), $p(x)$ is called *a priori* and $p(x|y)$ *a posteriori*. The likelihood is assigned using the link between the observed Y and the unknown X, and $p(x)$ is assigned using the prior knowledge about it. The Bayes rule then is a way to do state of knowledge fusion. Before taking into account any observation, our state of knowledge is represented by $p(x)$, and after the observation of Y, it becomes $p(x|y)$.

However, in this approach, two steps are very important. The first step is the assigning of $p(x)$ and $p(y|x)$ before being able to use the Bayes rule. As noted in the Introduction and as we will see later, we need other tools for this step. The second important step is after: how to use $p(x|y)$ to summarize it. When X is just a scalar value variable, we can do this computation easily. For example, we can compute the probability that X is in the interval $[a, b]$ via:

$$P(a \leq X < b|y) = \int_a^b p(x|y)\,\mathrm{d}x. \tag{9}$$

However, when the unknown becomes a high dimensional vectorial variable $\boldsymbol{X}$, as is the case in many signal and image processing applications, this computation becomes very costly [17]. We may then want to summarize $p(x|y)$ by a few interesting or significant point estimates. For example, compute the maximum *a posteriori* (MAP) solution:

$$\widehat{x}_{\mathrm{MAP}} = \arg\max_x \left\{p(x|y)\right\}, \tag{10}$$

the expected *a posteriori* (EAP) solution:

$$\widehat{x}_{\mathrm{EAP}} = \int x\, p(x|y)\,\mathrm{d}x, \tag{11}$$

the domains of X which include an integrated probability mass of more than some desired value (0.95 for example):

$$[x_1, x_2] : \int_{x_1}^{x_2} p(x|y)\,\mathrm{d}x = .95, \tag{12}$$

or any other questions, such as median or any α-quantiles:

$$x_q : \int_{-\infty}^{x_q} p(x|y)\,\mathrm{d}x = (1-\alpha). \tag{13}$$

As we see, computation of MAP needs an optimization algorithm, while these last three cases need integration, which may become very complicated for high dimensional cases [17].

We can also just explore numerically the whole space of the distribution using the Markov chain Monte Carlo (MCMC) [18–26] or any other sampling techniques [17]. In the scalar case (one dimension), all of these computations can be done numerically very easily. For the vectorial case, when

the dimensions become large, we need to develop specialized approximation methods, such as VBA and algorithms to do these computations. We give some more details about these when using this approach for inverse problems in real applications.

Remarks on notation used for the expected value in this paper: For a variable X with the probability density function (pdf) $p(x)$ and any regular function $h(X)$, we use indifferently:

$$\mathrm{E}\{X\} = \mathrm{E}_p\{X\} =< X >=< X >_p= \int x\, p(x)\, \mathrm{d}x$$

and:

$$\mathrm{E}\{h(X)\} = \mathrm{E}_p\{h(X)\} =< h(X) >=< h(X) >_p= \int h(x)\, p(x)\, \mathrm{d}x.$$

As an example, as we will say later, the entropy of $p(x)$ is noted indifferently:

$$H[p] = \mathrm{E}\{-\ln(p(X))\} = \mathrm{E}_p\{-\ln p(X)\} =< -\ln p(X) >=< -\ln p(X) >_p= -\int p(x)\, \ln p(x)\, \mathrm{d}x.$$

For any conditional probability density function (pdf) $p(x|y)$ and any regular function $h(X)$, we use indifferently:

$$\mathrm{E}\{X|y\} = \mathrm{E}_{p(x|y)}\{X\} =< X|y >=< X >_{p(x|y)}= \int x\, p(x|y)\, \mathrm{d}x$$

and:

$$\mathrm{E}\{h(X)|y\} = \mathrm{E}_{p(x|y)}\{h(X)\} =< h(X)|y >=< h(X) >_{p(x|y)}= \int h(x)\, p(x|)\, \mathrm{d}x.$$

As another example, as we will see later, the relative entropy of $p(x)$ over $q(x)$ is noted indifferently:

$$D[p|q] = \mathrm{E}_p\left\{-\ln\frac{p(X)}{q(X)}\right\} =< -\ln\frac{p(X)}{q(X)} >_{p(x)}= -\int p(x)\, \ln\frac{p(x)}{q(x)}\, \mathrm{d}x$$

and when there is not any ambiguity in the integration variable, we may omit it. For example, we may note:

$$D[p|q] = \mathrm{E}_p\left\{-\ln\frac{p}{q}\right\} =< -\ln\frac{p}{q} >_p= -\int p\, \ln\frac{p}{q}.$$

Finally, when we have two variables X and Y with their joint pdf $p(x, y)$, their marginals $p(x)$ and $p(y)$ and their conditionals $p(x|y)$ and $p(y|x)$, we may use the following notations:

$$\mathrm{E}\{h(X)|y\} = \mathrm{E}_{p(x|y)}\{h(X)\} = \mathrm{E}_{X|Y}\{h(X)\} =< h(X)|y >=< h(X) >_{p(x|y)}= \int h(x)\, p(x|y)\, \mathrm{d}x.$$

3. Quantity of Information and Entropy

3.1. Shannon Entropy

To introduce the quantity of information and the entropy, Shannon first considered a discrete valued variable X taking values $\{x_1, \cdots, x_N\}$ with probabilities $\{p_1, \cdots, p_N\}$ and defined the quantities of information associated with each of them as $I_n = \ln\frac{1}{p_n} = -\ln p_n$ and its expected value as the entropy:

$$\mathrm{H}[X] = -\sum_{i=1}^{N} p_i\, \ln p_i. \tag{14}$$

Later, this definition is extended to the continuous case by:

$$\mathrm{H}\left[X\right] = -\int p(x)\ \ln p(x)\,\mathrm{d}x. \tag{15}$$

By extension, if we consider two related variables (X, Y) with the probability laws, joint $p(x, y)$, marginals, $p(x)$, $p(y)$, and conditionals, $p(y|x)$, $p(x|y)$, we can define, respectively, the joint entropy:

$$\mathrm{H}\left[X, Y\right] = -\iint p(x, y)\ \ln p(x, y)\,\mathrm{d}x\,\mathrm{d}y, \tag{16}$$

as well as $\mathrm{H}\left[X\right]$, $\mathrm{H}\left[Y\right]$, $\mathrm{H}\left[Y|X\right]$ and $\mathrm{H}\left[X|Y\right]$.

Therefore, for any well-defined probability law, we can have an expression for its entropy. $\mathrm{H}\left[X\right]$, $\mathrm{H}\left[Y\right]$, $\mathrm{H}\left[Y|X\right]$, $\mathrm{H}\left[X|Y\right]$ and $\mathrm{H}\left[X, Y\right]$, which should better be noted as $\mathrm{H}\left[p(x)\right]$, $\mathrm{H}\left[p(y)\right]$, $\mathrm{H}\left[p(y|x)\right]$, $\mathrm{H}\left[p(x|y)\right]$ and $\mathrm{H}\left[p(x, y)\right]$.

3.2. Thermodynamical Entropy

Entropy is also a property of thermodynamical systems introduced by Clausius [27]. For a closed homogeneous system with reversible transformation, the differential entropy δS is related to δQ the incremental reversible transfer of heat energy into that system by $\delta S = \delta Q/T$ with T being the uniform temperature of the closed system.

It is very hard to establish a direct link between these two entropies. However, in statistical mechanics, thanks to Boltzmann, Gibbs and many others, we can establish some link if we consider the microstates (for example, the number, positions and speeds of the particles) and the macrostates (for example, the temperature T, pressure P, volume V and energy E) of the system and if we assign a probability law to microstates and consider the macrostates as the average (expected values) of some functions of those microstates. Let us give a very brief summary of some of those interpretations.

3.3. Statistical Mechanics Entropy

The interpretation of entropy in statistical mechanics is the measure of uncertainty that remains about the state of a system after its observable macroscopic properties, such as temperature (T), pressure (P) and volume (V), have been taken into account. For a given set of macroscopic variables T, P and V, the entropy measures the degree to which the probability of the system is spread out over different possible microstates. In contrast to the macrostate, which characterizes plainly observable average quantities, a microstate specifies all atomic details about the system, including the position and velocity of every atom. Entropy in statistical mechanics is a measure of the number of ways in which the microstates of the system may be arranged, often taken to be a measure of "disorder" (the higher the entropy, the higher the disorder). This definition describes the entropy as being proportional to the natural logarithm of the number of possible microscopic configurations of the system (microstates), which could give rise to the observed macroscopic state (macrostate) of the system. The proportionality constant is the Boltzmann constant.

3.4. Boltzmann Entropy

Boltzmann described the entropy as a measure of the number of possible microscopic configurations Ω of the individual atoms and molecules of the system (microstates) that comply with the macroscopic state (macrostate) of the system. Boltzmann then went on to show that $k \ln \Omega$ was equal to the thermodynamic entropy. The factor k has since been known as Boltzmann's constant.

In particular, Boltzmann showed that the entropy S of an ideal gas is related to the number of states of the molecules (microstates Ω) with a given temperature (macrostate):

$$S = k \ln \Omega \tag{17}$$

3.5. Gibbs Entropy

The macroscopic state of the system is defined by a distribution on the microstates that are accessible to a system in the course of its thermal fluctuations. Therefore, the entropy is defined over two different levels of description of the given system. The entropy is given by the Gibbs entropy formula, named after J. Willard Gibbs. For a classical system (*i.e.*, a collection of classical particles) with a discrete set of microstates, if E_i is the energy of microstate i and p_i is its probability that it occurs during the system's fluctuations, then the entropy of the system is:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i. \tag{18}$$

where k is again the physical constant of Boltzmann, which, like the entropy, has units of heat capacity. The logarithm is dimensionless. It is interesting to note that Relation (17) can be obtained from Relation (18) when the probability distribution is uniform over the volume Ω [28–30].

4. Relative Entropy or Kullback–Leibler Divergence

Kullback wanted to compare the relative quantity of information between two probability laws p_1 and p_2 on the same variable X. Two related notions have been defined:

- Relative Entropy of p_1 with respect to p_2:

$$D\left[p_1 : p_2\right] = -\int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{19}$$

- Kullback–Leibler divergence of p_1 with respect to p_2:

$$\mathrm{KL}\left[p_1 : p_2\right] = -D\left[p_1 : p_2\right] = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{20}$$

We may note that:

- $\mathrm{KL}\left[q : p\right] \geq 0$,
- $\mathrm{KL}\left[q : p\right] = 0$, if $q = p$ and

- $\mathrm{KL}\left[q:p_0\right] \geq \mathrm{KL}\left[q:p_1\right] + \mathrm{KL}\left[p_1:p_0\right]$.
- $\mathrm{KL}\left[q:p\right]$ is invariant with respect to a scale change, but is not symmetric.
- A symmetric quantity can be defined as:

$$\mathrm{J}\left[q,p\right] = \frac{1}{2}\left(\mathrm{KL}\left[q:p\right] + \mathrm{KL}\left[p:q\right]\right). \tag{21}$$

5. Mutual Information

The purpose of mutual information is to compare two related variables Y and X. It can be defined as the expected amount of information that one gains about Xif we observe the value of Y, and *vice versa*. Mathematically, the mutual information between X and Y is defined as:

$$\mathrm{I}\left[Y,X\right] = \mathrm{H}\left[X\right] - \mathrm{H}\left[X|Y\right] = \mathrm{H}\left[Y\right] - \mathrm{H}\left[Y|X\right] \tag{22}$$

or equivalently as:

$$\mathrm{I}\left[Y,X\right] = D\left[p(X,Y):p(X)p(Y)\right]. \tag{23}$$

With this definition, we have the following properties:

$$\mathrm{H}\left[X,Y\right] = \mathrm{H}\left[X\right] + \mathrm{H}\left[Y|X\right] = \mathrm{H}\left[Y\right] + \mathrm{H}\left[X|Y\right] = \mathrm{H}\left[X\right] + \mathrm{H}\left[Y\right] - \mathrm{I}\left[Y,X\right] \tag{24}$$

and:

$$\begin{aligned} \mathrm{I}\left[Y,X\right] &= \mathrm{E}_X\left\{D\left[p(Y|X):p(Y)\right]\right\} \stackrel{\triangle}{=} \int D\left[p(y|x):p(y)\right]p(x)\,\mathrm{d}x \\ &= \mathrm{E}_Y\left\{D\left[p(X|Y):p(X)\right]\right\} \stackrel{\triangle}{=} \int D\left[p(x|y):p(x)\right]p(y)\,\mathrm{d}y. \end{aligned} \tag{25}$$

We may also remark on the following property:

- $\mathrm{I}\left[Y,X\right]$ is a concave function of $p(y)$ when $p(x|y)$ is fixed and a convex function of $p(x|y)$ when $p(y)$ is fixed.
- $\mathrm{I}\left[Y,X\right] \geq 0$ with equality only if X and Y are independent.

6. Maximum Entropy Principle

The first step before applying any probability rules for inference is to assign a probability law to a quantity. Very often, the available knowledge on that quantity can be described mathematically as the constraints on the desired probability law. However, in general, those constraints are not enough to determine in a unique way that probability law. There may exist many solutions that satisfy those constraints. We need then a tool to select one.

Jaynes introduced the MEP [8], which can be summarized as follows: When we do not have enough constraints to determine a probability law that satisfies those constraints, we may select between them the one with maximum entropy.

Let us be now more precise. Let us assume that the available information on that quantity X is the form of:

$$\mathrm{E}\left\{\phi_k(X)\right\} = d_k, \quad k = 1,\ldots,K. \tag{26}$$

where ϕ_k are any known functions. First, we assume that such probability laws exist by defining:

$$\mathcal{P} = \left\{ p(x) : \int \phi_k(x) p(x)\, \mathrm{d}x = d_k, \quad k = 0, \ldots, K \right\}$$

with $\phi_0 = 1$ and $d_0 = 1$ for the normalization purpose. Then, the MEP is written as an optimization problem:

$$p_{ME}(x) = \arg\max_{p \in \mathcal{P}} \left\{ \mathrm{H}[p] = -\int p(x)\, \ln p(x)\, \mathrm{d}x \right\} \tag{27}$$

whose solution is given by:

$$p_{ME}(x) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x) \right] \tag{28}$$

where $Z(\boldsymbol{\lambda})$, called the partition function, is given by: $Z(\boldsymbol{\lambda}) = \int \exp[-\sum_{k=1}^{K} \lambda_k \phi_k(x)]\, \mathrm{d}x$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]'$ have to satisfy:

$$-\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} = d_k, \quad k = 1, \ldots, K \tag{29}$$

which can also be written as $-\nabla_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{\lambda}) = \boldsymbol{d}$. Different algorithms have been proposed to compute numerically the ME distributions. See, for example, [31–37]

The maximum value of entropy reached is given by:

$$H_{\max} = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}' \boldsymbol{d}. \tag{30}$$

This optimization can easily be extended to the use of relative entropy by replacing $H(p)$ by $D[p : q]$ where $q(x)$ is a given reference of *a priori* law. See [9,38,39] and [34,40–42] for more details.

7. Link between Entropy and Likelihood

Consider the problem of the parameter estimation $\boldsymbol{\theta}$ of a probability law $p(x|\boldsymbol{\theta})$ from an n-element sample of data $\boldsymbol{x} = \{x_1, \cdots, x_n\}$.

The log-likelihood of $\boldsymbol{\theta}$ is defined as:

$$L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta). \tag{31}$$

Maximizing $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives what is called the maximum likelihood (ML) estimate of θ.

Noting that $L(\boldsymbol{\theta})$ depends on n, we may consider $\frac{1}{n} L(\boldsymbol{\theta})$ and define:

$$\bar{L}(\boldsymbol{\theta}) = \lim_{n \mapsto \infty} \frac{1}{n} L(\boldsymbol{\theta}) = \mathrm{E}\{\ln p(x|\boldsymbol{\theta})\} = \int p(x|\boldsymbol{\theta}^*)\, \ln p(x|\boldsymbol{\theta})\, \mathrm{d}x, \tag{32}$$

where $\boldsymbol{\theta}^*$ is the right answer and $p(x|\boldsymbol{\theta}^*)$ its corresponding probability law. We may then remark that:

$$D[p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta})] = -\int p(x|\boldsymbol{\theta}^*)\, \ln \frac{p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta}^*)}\, \mathrm{d}x = -\int p(x|\boldsymbol{\theta}^*)\, \ln p(x|\boldsymbol{\theta}^*)\, \mathrm{d}x + \bar{L}(\boldsymbol{\theta}). \tag{33}$$

The first term in the right-hand side being a constant, we derive that:

$$\arg\max_{\boldsymbol{\theta}} \left\{ D\left[p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}) \right] \right\} = \arg\max_{\boldsymbol{\theta}} \left\{ \bar{L}(\boldsymbol{\theta}) \right\}.$$

In this way, there is a link between the maximum likelihood and maximum relative entropy solutions [24].

There is also a link between the maximum relative entropy and the Bayes rule. See, for example, [43,44] and their corresponding references.

8. Fisher Information, Bregman and Other Divergences

Fisher [10] was interested in measuring the amount of information that samples of a variable X carries about an unknown parameter θ upon which its probability law $p(x|\theta)$ depends. For a given sample of observation x and its probability law $p(x|\theta)$, the function $\mathcal{L}(\theta) = p(x|\theta)$ is called the likelihood of θ in the sample x. He called the score of x over θ the partial derivative with respect to θ of the logarithm of this function:

$$S(x|\theta) = \frac{\partial \ln p(x|\theta)}{\partial \theta} \tag{34}$$

He also showed that the first order moment of the score is zero:

$$\mathrm{E}\left\{ S(X|\theta) \right\} = \mathrm{E}\left\{ \frac{\partial \ln p(x|\theta)}{\partial \theta} \right\} = 0 \tag{35}$$

but its second order moment is positive and is also equivalent to the expected values of the second derivative of the log-likelihood function with respect to θ.

$$\mathrm{E}\left\{ S^2(X|\theta) \right\} = \mathrm{E}\left\{ \left| \frac{\partial \ln p(x|\theta)}{\partial \theta} \right|^2 \right\} = \mathrm{E}\left\{ \frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2} \right\} = F \tag{36}$$

This quantity is called the Fisher information [14].

It is also shown that for the small variations of θ, the Fisher information induces locally a distance in the space of parameters Θ, if we had to compare two very close values of θ. In this way, the notion of the geometry of information is introduced. The main steps for introducing this notion are the following: Consider $D\left[p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \right]$ and assume that $\ln p(x|\boldsymbol{\theta})$ can be developed in a Taylor series. Then, keeping the terms up to the second order, we obtain:

$$D\left[p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \right] \simeq \frac{1}{2} \Delta\boldsymbol{\theta}' \boldsymbol{F}(\theta^*) \Delta\boldsymbol{\theta}. \tag{37}$$

where $\boldsymbol{F}$ is the Fisher information:

$$\boldsymbol{F}(\theta^*) = \mathrm{E}\left\{ \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\}. \tag{38}$$

We must be careful here that this geometry property is related to the space of the parameters Θ for a given family of parametric probability law $p(x|\theta)$ and not in the space of probabilities. However, for two probability laws $p_1(x) = p(x|\theta_1)$ and $p_2(x) = p(x|\theta_2)$ in the same exponential family, the Kullback–Leibler divergence $\mathrm{KL}\left[p_1 : p_2 \right]$ induces a Bregman divergence $\mathrm{B}[\theta_1|\theta_2]$ between the two parameters [14,45–48].

To go further into detail, let us extend the discussion about the link between Fisher information and KL divergence, as well as other divergences, such as f-divergences, Rényi's divergences and Bregman divergences.

- f-divergences:

  The f-divergences, which are a general class of divergences, indexed by convex functions f, that include the KL divergence as a special case. Let $f : (0, \infty) \mapsto \mathbf{R}$ be a convex function for which $f(1) = 0$. The f-divergence between two probability measures P and Q is defined by:

  $$D_f[P : Q] = \int q\, f\left(\frac{p}{q}\right) \tag{39}$$

  Every f-divergence can be viewed as a measure of distance between probability measures with different properties. Some important special cases are:

  - $f(x) = x \ln x$ gives KL divergence: $\mathrm{KL}\,[P : Q] = \int p\, \ln(\frac{p}{q})$.
  - $f(x) = |x - 1|/2$ gives total variation distance: $\mathrm{TV}[P, Q] = \int |p - q|/2$.
  - $f(x) = (\sqrt{x} - 1)^2$ gives the square of the Hellinger distance: $H^2[P, Q] = \int (\sqrt{p} - \sqrt{q})^2$.
  - $f(x) = (x - 1)^2$ gives the chi-squared divergence: $\chi^2[P : Q] = \int \frac{(p-q)^2}{q}$.

- Rényi divergences:

  These are another generalization of the KL divergence. The Rényi divergence between two probability distributions P and Q is:

  $$D_\alpha[P : Q] = \frac{1}{\alpha - 1} \ln \int p^\alpha\, q^{1-\alpha}. \tag{40}$$

  When $\alpha = 1$, by a continuity argument, $D_\alpha[P : Q]$ converges to $\mathrm{KL}\,[P : Q]$.

  $D_{1/2}[P, Q] = -2 \ln \int \sqrt{p\, q}$ is called Bhattacharyya divergence (closely related to Hellinger distance). Interestingly, this quantity is always smaller than KL:

  $$D_{1/2}[P : Q] \leq \mathrm{KL}\,[P : Q]\,. \tag{41}$$

  As a result, it is sometimes easier to derive risk bounds with $D_{1/2}$ as the loss function as opposed to KL.

- Bregman divergences:

  The Bregman divergences provide another class of divergences that are indexed by convex functions and include both the Euclidean distance and the KL divergence as special cases. Let ϕ be a differentiable strictly convex function. The Bregman divergence B_ϕ is defined by:

  $$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) \rangle \tag{42}$$

  where $< \boldsymbol{x}, \boldsymbol{y} >= \sum_j x_i\, y_j$ here means the scalar product of $\boldsymbol{x}$ and $\boldsymbol{y}$ and where the domain of ϕ is a space where convexity and differentiability make sense (e.g., whole or a subset of $\mathbf{R}^d$ or an L_p space). For example, $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ on $\mathbf{R}^d$ gives the Euclidean distance:

  $$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) \rangle = \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 - \langle \boldsymbol{x} - \boldsymbol{y}, 2\boldsymbol{y} \rangle = \|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{43}$$

and $\phi(\boldsymbol{x}) = \sum_j x_j \ln x_j$ on the simplex in $\mathbf{R}^d$ gives the KL divergence:

$$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \sum_j x_j \ln x_j - \sum_j y_j \ln y_j - \sum_j (x_j - y_j)(1 + \ln y_j) = \sum_j x_j \ln \frac{x_j}{y_j} = \mathrm{KL}\left[\boldsymbol{x} : \boldsymbol{y}\right] \quad (44)$$

where it is assumed $\sum_j x_j = \sum_j y_j = 1$.

Let X be a quantity taking values in the domain of ϕ with a probability distribution function $p(x)$. Then, $\mathrm{E}_{p(x)}\left\{B_\phi(X, m)\right\}$ is minimized over m in the domain of ϕ at $m = \mathrm{E}\left\{X\right\}$:

$$\widehat{m} = \arg\min_m \left\{B_\phi(X, m)\right\} = \mathrm{E}\left\{X\right\}.$$

Moreover, this property characterizes Bregman divergence. When applied to the Bayesian approach, this means that, using the Bregman divergence as the loss function, the Bayes estimator is the posterior mean. This point is detailed in the following.

Links between all of these through an example:
Let us consider the Bayesian parameter estimation where we have some data $\boldsymbol{y}$, a set of parameters $\boldsymbol{x}$, a likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ and a prior $\pi(\boldsymbol{x})$, which gives the posterior $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})\, \pi(\boldsymbol{x})$. Let us also consider a cost function $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$ in the parameter space $\boldsymbol{x} \in \mathcal{X}$. The classical Bayesian point estimation of $\boldsymbol{x}$ is expressed as the minimizer of an expected risk:

$$\widehat{\boldsymbol{x}} = \arg\min_{\widetilde{\boldsymbol{x}}} \left\{\bar{C}(\widetilde{\boldsymbol{x}})\right\} \quad (45)$$

where:

$$\bar{C}(\widetilde{\boldsymbol{x}}) = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})}\left\{C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\right\} = \int C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$$

It is very well known that the mean squared error estimator, which corresponds to $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}] = \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|^2$, is the posterior mean. It is now interesting to know that choosing $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$ to be any Bregman divergence $B_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$, we obtain also the posterior mean:

$$\widehat{\boldsymbol{x}} = \arg\min_{\widetilde{\boldsymbol{x}}} \left\{\bar{B}_\phi(\widetilde{\boldsymbol{x}})\right\} = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})}\left\{\int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\right\} \quad (46)$$

where:

$$\bar{B}_\phi(\widetilde{\boldsymbol{x}}) = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})}\left\{D_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\right\} = \int B_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$$

Consider now that we have two prior probability laws $\pi_1(\boldsymbol{x})$ and $\pi_2(\boldsymbol{x})$, which give rise to two posterior probability laws $p_1(\boldsymbol{x}|\boldsymbol{y})$ and $p_2(\boldsymbol{x}|\boldsymbol{y})$. If the prior laws and the likelihood are in the exponential families, then the posterior laws are also in the exponential family. Let us note them as $p_1(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters of those posterior laws. We then have the following properties:

- $\mathrm{KL}\left[p_1 : p_2\right]$ is expressed as a Bregman divergence $B[\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2]$.
- A Bregman divergence $B[\boldsymbol{x}_1 : \boldsymbol{x}_2]$ is induced when $\mathrm{KL}\left[p_1 : p_2\right]$ is used to compare the two posteriors.

9. Vectorial Variables and Time Indexed Process

The extension of the scalar variable to the finite dimensional vectorial case is almost immediate. In particular, for the Gaussian case $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{R})$, the mean becomes a vector $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{X}\}$, and the variances are replaced by a covariance matrix: $\boldsymbol{R} = \mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\}$; and almost all of the quantities can be defined immediately. For example, for a Gaussian vector $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{R})$, the entropy is given by [49]:

$$H = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(|\det(\boldsymbol{R})|) \tag{47}$$

and the relative entropy of $\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{R})$ with respect to $\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{S})$ is given by:

$$D = -\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{R}\boldsymbol{S}^{-1}\right) - \log\frac{|\det(\boldsymbol{R})|}{|\det(\boldsymbol{S})|} - n\right). \tag{48}$$

The notion of time series or processes needs extra definitions. For example, for a random time series $X(t)$, we can define $p(X(t)), \forall t$, the expected value time series $\bar{x}(t) = \mathrm{E}\{X(t)\}$ and what is called the autocorrelation function $\Gamma(t, \tau) = \mathrm{E}\{X(t)\,X(t+\tau)\}$. A time series is called stationary when these quantities does not depend on t, *i.e.*, $\bar{x}(t) = m$ and $\Gamma(t, \tau) = \Gamma(\tau)$ [50]. Another quantity of interest for a stationary time series is its power spectral density (PSD) function:

$$S(\omega) = \mathrm{FT}\{\Gamma(\tau)\} = \int \Gamma(\tau)\exp[-j\omega\tau]\,\mathrm{d}\tau. \tag{49}$$

When $X(t)$ is observed on times $t = n\Delta T$ with $\Delta T = 1$, we have $X(n)$, and for a sample $\{X(1), \cdots, X(N)\}$, we may define the mean $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{X}\}$ and the covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\}$.

With these definitions, it can easily been shown that the covariance matrix of a stationary Gaussian process is Toeplitz [49]. It is also possible to show that the entropy of such a process can be expressed as a function of its PSD function:

$$\lim_{n\longrightarrow\infty}\frac{1}{n}H(p) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln S(\omega)\,\mathrm{d}\omega. \tag{50}$$

For two stationary Gaussian processes with two spectral density functions $S_1(\omega)$ and $S_2(\omega)$, we have:

$$\lim_{n\longrightarrow\infty}\frac{1}{n}D[p_1 : p_2] = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{S_1(\omega)}{S_2(\omega)} - \ln\frac{S_1(\omega)}{S_2(\omega)} - 1\right)\mathrm{d}\omega \tag{51}$$

where we find the Itakura–Saito distance in the spectral analysis literature [50–53].

These definitions and expressions have often been used in time series analysis. In what follows, we give a few examples of the different ways these notions and quantities have been used in different applications of data, signal and image processing.

10. Entropy in Independent Component Analysis and Source Separation

Given a vector of time series $\boldsymbol{x}(t)$, the independent component analysis (ICA) consists of finding a separating matrix $\boldsymbol{B}$, such that the components $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ are as independent as possible. The notion of entropy is used here as a measure of independence. For example, to find $\boldsymbol{B}$, we may choose

$D\left[p(\boldsymbol{y}):\prod_j p_j(y_j)\right]$ as a criterion of independence of the components y_j. The next step is to choose a probability law $p(\boldsymbol{x})$ from which we can find an expression for $p(\boldsymbol{y})$ from which we can find an expression for $D\left[p(\boldsymbol{y}):\prod_j p_j(y_j)\right]$ as a function of the matrix $\boldsymbol{B}$, which can be optimized to obtain it.

The ICA problem has a tight link with the source separation problem, where it is assumed that the measured time series $\boldsymbol{x}(t)$ is a linear combination of the sources $\boldsymbol{s}(t)$, *i.e.*, $\boldsymbol{x}(t)=\boldsymbol{A}\boldsymbol{s}(t)$, with $\boldsymbol{A}$ being the mixing matrix. The objective of source separation is then to find the separating matrix $\boldsymbol{B}=\boldsymbol{A}^{-1}$.

To see how the entropy is used here, let us note $\boldsymbol{y}=\boldsymbol{B}\boldsymbol{x}$. Then,

$$p_Y(\boldsymbol{y})=\frac{1}{|\partial\boldsymbol{y}/\partial\boldsymbol{x}|}p_X(\boldsymbol{x})\longrightarrow H(\boldsymbol{y})=-\mathrm{E}\left\{\ln p_Y(\boldsymbol{y})\right\}=\mathrm{E}\left\{\ln|\partial\boldsymbol{y}/\partial\boldsymbol{x}|\right\}-H(\boldsymbol{x}). \tag{52}$$

$H(\boldsymbol{y})$ is used as a criterion for ICA or source separation. As the objective in ICA is to obtain $\boldsymbol{y}$ in such a way that its components become as independent as possible, the separating matrix $\boldsymbol{B}$ has to maximize $H(\boldsymbol{y})$. Many ICA algorithms are based on this optimization [54–65]

11. Entropy in Parametric Modeling and Model Selection

Determining the order of a model, *i.e.*, the dimension of the vector parameter $\boldsymbol{\theta}$ in a probabilistic model $p(\boldsymbol{x}|\boldsymbol{\theta})$, is an important subject in many data and signal processing problems. As an example, in autoregressive (AR) modeling:

$$x(n)=\sum_{k=1}^{K}\theta_k x(n-k)+\epsilon(n) \tag{53}$$

where $\boldsymbol{\theta}=\{\theta_1,\cdots,\theta_K\}$, we may want to compare two models with two different values of K.

When the order K is fixed, the estimation of the parameters $\boldsymbol{\theta}$ is a very well-known problem, and there are likelihood based [66] or Bayesian approaches for that [67]. The determination of the order is however more difficult [68]. Between the tools, we may mention here the Bayesian methods [69–74], but also the use of relative entropy $D\left[p(\boldsymbol{x}|\boldsymbol{\theta}^*):p(\boldsymbol{x}|\boldsymbol{\theta})\right]$, where $\boldsymbol{\theta}^*$ represents the vector of the parameters of dimension K^* and $\boldsymbol{\theta}$ and the vector $\boldsymbol{\theta}$ with dimension $K\leq K^*$. In such cases, even if the two probability laws to be compared have parameters with different dimensions, we can always use the $\mathrm{KL}\left[p(\boldsymbol{x}|\boldsymbol{\theta}^*):p(\boldsymbol{x}|\boldsymbol{\theta})\right]$ to compare them. The famous criterion of Akaike [75–78] uses this quantity to determine the optimal order. For a linear parameter model with Gaussian probability laws and likelihood-based methods, there are analytic solutions for it [68].

12. Entropy in Spectral Analysis

Entropy and MEP have been used in different ways in the spectral analysis problem. It has been an important subject of signal processing for the decades. Here, we are presenting, in a brief way, these different approaches.

12.1. Burg's Entropy-Based Method

A classical one is Burg's entropy method [79], which can be summarized as follows: Let $X(n)$ be a stationary, centered process, and assume we have as data a finite number of samples (lags) of its autocorrelation function:

$$r(k) = \mathrm{E}\left\{X(n)X(n+k)\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)\exp\left[\mathrm{j}k\omega\right]\,\mathrm{d}\omega, \quad k = 0,\ldots,K. \tag{54}$$

The task is then to estimate its power spectral density function:

$$S(\omega) = \sum_{k=-\infty}^{\infty} r(k)\exp\left[-\mathrm{j}k\omega\right]$$

As we can see, due to the fact that we have only the elements of the right-hand for $k = -K,\cdots,+K$, the problem is ill posed. To obtain a probabilistic solution, we may start by assigning a probability law $p(\boldsymbol{x})$ to the vector $\underline{X} = [X(0),\ldots,X(N-1)]'$. For this, we can use the principle of maximum entropy (PME) with the data as constraints (54). As these constraints are the second order moments, the PME solution is a Gaussian probability law: $\mathcal{N}(\boldsymbol{x}|\mathbf{0},\boldsymbol{R})$. For a stationary Gaussian process, when the number of samples $N \longrightarrow \infty$, the expression of the entropy becomes:

$$H = \int_{-\pi}^{\pi} \ln S(\omega)\,\mathrm{d}\omega. \tag{55}$$

This expression is called Burg's entropy [79]. Thus, Burg's method consists of maximizing H subject to the constraints (54). The solution is:

$$S(\omega) = \frac{1}{\left|\sum_{k=-K}^{K}\lambda_k\exp\left[\mathrm{j}k\omega\right]\right|^2}, \tag{56}$$

where $\boldsymbol{\lambda} = [\lambda_0,\cdots,\lambda_K]'$, the Lagrange multipliers associated with the constraints (54), are here equivalent to the AR modeling of the Gaussian process $X(n)$.

We may note that, in this particular case, we have an analytical expression for $\boldsymbol{\lambda}$, which provides the possibility to give an analytical expression for $S(\omega)$ as a function of the data $\{r(k), k = 0,\cdots,K\}$:

$$S(\omega) = \frac{\boldsymbol{\delta}\,\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}}{\boldsymbol{e}\,\boldsymbol{\Gamma}^{-1}\boldsymbol{e}}, \tag{57}$$

where $\boldsymbol{\Gamma} = \text{Toeplitz}(r(0),\cdots,r(K))$ is the correlation matrix and $\boldsymbol{\delta}$ and $\boldsymbol{e}$ are two vectors defined by $\boldsymbol{\delta} = [1,0,\cdots,0]'$ and $\boldsymbol{e} = [1,e^{-\mathrm{j}\omega},e^{-\mathrm{j}2\omega},\cdots,e^{-\mathrm{j}K\omega}]'$.

We may note that we first used MEP to choose a probability law for $X(n)$. With the prior knowledge that we have second order moments, the MEP results in a Gaussian probability density function. Then, as for a stationary Gaussian process, the expression of the entropy is related to the power spectral density $S(\omega)$, and as this is related to the correlation data by a Fourier transform, an ME solution could be computed easily.

12.2. Extensions to Burg's Method

The second approach consists of maximizing the relative entropy $D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ or minimizing $\mathrm{KL}\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ where $p_0(\boldsymbol{x})$ is an *a priori* law. The choice of the prior is important. Choosing a uniform $p_0(\boldsymbol{x})$, we retrieve the previous case [77].

However, choosing a Gaussian law for $p_0(\boldsymbol{x})$, the expression to maximize becomes:

$$D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right] = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{S(\omega)}{S_0(\omega)} - \ln \frac{S(\omega)}{S_0(\omega)} - 1 \right) \mathrm{d}\omega \tag{58}$$

when $N \mapsto \infty$ and where $S_0(\omega)$ corresponds to the power spectral density of the reference process $p_0(\boldsymbol{x})$. Now, the problem becomes: minimize $D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ subject to the constraints (54).

12.3. Shore and Johnson Approach

Another approach is to decompose first the process $X(n)$ on the Fourier basis $\{\cos k\omega t,\ \sin k\omega t\}$, to consider ω to be the variable of interest and $S(\omega)$, normalized properly, to be considered as its probability distribution function. Then, the problem can be reformulated as the determination of the $S(\omega)$, which maximizes the entropy:

$$-\int_{-\pi}^{\pi} S(\omega)\ \ln S(\omega)\, \mathrm{d}\omega \tag{59}$$

subject to the linear constraints (54). The solution is in the form of:

$$S(\omega) = \exp\left[\sum_{k=-K}^{K} \lambda_k \exp\left[\mathrm{j}k\omega\right]\right]. \tag{60}$$

which can be considered as the most uniform power spectral density that satisfies those constraints.

12.4. ME in the Mean Approach

In this approach, we consider $S(\omega)$ as the expected value $Z(\omega)$ for which we have a prior law $\mu(z)$, and we are looking to assign $p(z)$, which maximizes the relative entropy $D\left[p(z) : \mu(z)\right]$ subject to the constraints (54).

When $p(z)$ is determined, the solution is given by:

$$S(\omega) = \mathrm{E}\left\{Z(\omega)\right\} = \int Z(\omega) p(z)\, \mathrm{d}z. \tag{61}$$

The expression of $S(\omega)$ depends on $\mu(z)$. When $\mu(z)$ is Gaussian, we obtain the Rényi entropy:

$$H = \int_{-\pi}^{\pi} S^2(\omega)\, \mathrm{d}\omega. \tag{62}$$

If we choose a Poisson measure for $\mu(z)$, we obtain the Shannon entropy:

$$H = -\int_{-\pi}^{\pi} S(\omega) \ln S(\omega)\, \mathrm{d}\omega, \tag{63}$$

and if we choose a Lebesgue measure over $[0, \infty]$, we obtain Burg's entropy:

$$H = \int_{-\pi}^{\pi} \ln S(\omega)\, \mathrm{d}\omega. \tag{64}$$

When this step is done, the next step becomes maximizing these entropies subject to the constraints of the correlations. The obtained solutions are very different. For more details, see [39,79–85].

13. Entropy-Based Methods for Linear Inverse Problems

13.1. Linear Inverse Problems

A general way to introduce inverse problems is the following: Infer an unknown signal $f(t)$, image $f(x,y)$ or any multi-variable function $f(\boldsymbol{r})$ through an observed signal $g(t)$, image $g(x,y)$ or any multi-variable observable function $g(\boldsymbol{s})$, which are related through an operator $\mathcal{H}: f \mapsto g$. This operator can be linear or nonlinear. Here, we consider only linear operators $g = \mathcal{H}f$:

$$g(\boldsymbol{s}) = \int h(\boldsymbol{r}, \boldsymbol{s}) f(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \tag{65}$$

where $h(\boldsymbol{r}, \boldsymbol{s})$ is the response of the measurement system. Such linear operators are very common in many applications of signal and image processing. We may mention a few examples of them:

- Convolution operations $g = h * f$ in 1D (signal):

$$g(t) = \int h(t - t') f(t') \, \mathrm{d}t' \tag{66}$$

  or in 2D (image):

$$g(x,y) = \iint h(x - x', y - y') f(x', y') \, \mathrm{d}x' \, \mathrm{d}y' \tag{67}$$

- Radon transform (RT) in computed tomography (CT) in the 2D case [86]:

$$g(r, \phi) = \int \int \delta(r - x\cos\phi - y\sin\phi) \, f(x,y) \, \mathrm{d}x \, \mathrm{d}y \tag{68}$$

- Fourier transform (FT) in the 2D case:

$$g(u,v) = \int \int \exp\left[-j(ux + vy)\right] \, f(x,y) \, \mathrm{d}x \, \mathrm{d}y \tag{69}$$

  which arise in magnetic resonance imaging (MRI), in synthetic aperture radar (SAR) imaging or in microwave and diffraction optical tomography (DOT) [86–90].

No matter the category of the linear transforms, when the problem is discretized, we arrive at the relation:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}, \tag{70}$$

where $\boldsymbol{f} = [f_1, \cdots, f_n]'$ represents the unknowns, $\boldsymbol{g} = [g_1, \cdots, g_m]'$ the observed data, $\boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_m]'$ the errors of modeling and measurement and $\boldsymbol{H}$ the matrix of the system response.

13.2. Entropy-Based Methods

Let us consider first the simple no noise case:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f}, \tag{71}$$

where $\boldsymbol{H}$ is a matrix of dimensions $(M \times N)$, which is in general singular or very ill conditioned. Even if the cases $M > N$ or $M = N$ may appear easier, they have the same difficulties as those of the underdetermined case $M < N$ that we consider here. In this case, evidently the problem has an infinite number of solutions, and we need to choose one.

Between the numerous methods, we may mention the minimum norm solution, which consists of choosing between all of the possible solutions:

$$\mathcal{F} = \{\boldsymbol{f} : \boldsymbol{H}\boldsymbol{f} = \boldsymbol{g}\} \tag{72}$$

the one that has the minimum norm:

$$\Omega(\boldsymbol{f}) = \|\boldsymbol{f}\|_2^2 = \sum_j f_j^2. \tag{73}$$

This optimization problem can be solved easily in this case, and we obtain:

$$\widehat{\boldsymbol{f}}_{NM} = \underset{\boldsymbol{f}\in\mathcal{F}}{\arg\min} \left\{\Omega(\boldsymbol{f}) = \|\boldsymbol{f}\|_2^2\right\} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{g}. \tag{74}$$

In fact, we may choose any other convex criterion $\Omega(\boldsymbol{f})$ and satisfy the uniqueness of the solution. For example:

$$\Omega(\boldsymbol{f}) = -\sum_j f_j \ln f_j \tag{75}$$

which can be interpreted as the entropy when $f_j > 0$ and $\sum f_j = 1$, thus considering f_j as a probability distribution $f_j = P(U = u_j)$. The variable U can correspond (or not) to a physical quantity. $\Omega(\boldsymbol{f})$ is the entropy associated with this variable.

If we consider $f_j > 0$ to represent the power spectral density of a physical quantity, then the entropy becomes:

$$\Omega(\boldsymbol{f}) = \sum_j \ln f_j \tag{76}$$

and we can use it as criterion to select a solution to the problem (71).

As we can see, any convex criterion $\Omega(\boldsymbol{f})$ can be used. Here, we mentioned four of them with different interpretations.

- L_2 or quadratic:

$$\Omega(\boldsymbol{f}) = \sum_j f_j^2 \tag{77}$$

  which can be interpreted as the Rényi's entropy with $q = 2$.

- L_β :

$$\Omega(\boldsymbol{f}) = \sum_j |f_j|^\beta \tag{78}$$

  When $\beta < 1$ the criterion is not bounded at zero. When $\beta \geq 1$ the criterion is convex.

- Shannon entropy:

$$\Omega(\boldsymbol{f}) = -\sum_j f_j \ln f_j \tag{79}$$

  which has a valid interpretation if $0 < f_j < 1$,

- The Burg entropy:

$$\Omega(\boldsymbol{f}) = \sum_j \ln f_j \tag{80}$$

  which needs $f_j > 0$.

Unfortunately, only for the first case, there is an analytical solution for the problem, which is $\widehat{\boldsymbol{f}} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')\boldsymbol{g}$. For all of the other cases, we may need an optimization algorithm to obtain a numerical solution [91–95].

13.3. Maximum Entropy in the Mean Approach

A second approach consists of considering $f_j = \mathrm{E}\{U_j\}$ or $\boldsymbol{f} = \mathrm{E}\{\boldsymbol{U}\}$ [41,41,42]. Again, here, U_j or $\boldsymbol{U}$ can, but need not, correspond to some physical quantities. In any case, we now want to assign a probability law $\widehat{p}(\boldsymbol{u})$ to it. Noting that the data $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} = \boldsymbol{H}\mathrm{E}\{\boldsymbol{U}\} = \mathrm{E}\{\boldsymbol{H}\boldsymbol{U}\}$ can be considered as the constraints on it, we may need again a criterion to determine $\widehat{p}(\boldsymbol{u})$. Assuming then having some prior $\mu(\boldsymbol{u})$, we may maximize the relative entropy as that criterion. The mathematical problem then becomes:

$$\text{minimize } D[p(\boldsymbol{u}) : \mu(\boldsymbol{u})] \text{ subject to } \int \boldsymbol{H}\boldsymbol{u}\, p(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = \boldsymbol{g} \tag{81}$$

The solution is:

$$\widehat{p}(\boldsymbol{u}) = \frac{1}{Z(\boldsymbol{\lambda})}\mu(\boldsymbol{u}) \exp\left[-\boldsymbol{\lambda}'\boldsymbol{H}\boldsymbol{u}\right] \tag{82}$$

where:

$$Z(\boldsymbol{\lambda}) = \int \mu(\boldsymbol{u}) \exp\left[-\boldsymbol{\lambda}'\boldsymbol{H}\boldsymbol{u}\right]\, \mathrm{d}\boldsymbol{u}. \tag{83}$$

When $\widehat{p}(\boldsymbol{u})$ is obtained, we may be interested in computing:

$$\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\} = \int \boldsymbol{u}\widehat{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{84}$$

which is the required solution.

Interestingly, if we focus on $\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\}$, we will see that its expression depends on the choice of the prior $\mu(\boldsymbol{u})$. When $\mu(\boldsymbol{u})$ is separable: $\mu(\boldsymbol{u}) = \prod_j \mu_j(u_j)$, the expression of $\widehat{p}(\boldsymbol{u})$ will also be separable.

To go a little more into the details, let us introduce $\boldsymbol{s} = \boldsymbol{H}'\boldsymbol{\lambda}$ and define:

$$G(\boldsymbol{s}) = \ln \int \mu(\boldsymbol{u}) \exp\left[-\boldsymbol{s}'\boldsymbol{u}\right]\, \mathrm{d}\boldsymbol{u} \tag{85}$$

and its conjugate convex:

$$F(\boldsymbol{f}) = \sup_{\boldsymbol{s}} \left\{\boldsymbol{f}'\boldsymbol{s} - G(\boldsymbol{s})\right\}. \tag{86}$$

It can be shown easily that $\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\}$ can be obtained either via the dual $\widehat{\boldsymbol{\lambda}}$ variables:

$$\widehat{\boldsymbol{f}} = G'(\boldsymbol{H}'\widehat{\boldsymbol{\lambda}}) \tag{87}$$

where $\widehat{\boldsymbol{\lambda}}$ is obtained by:

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \{D(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}' \boldsymbol{g}\}, \tag{88}$$

or directly:

$$\widehat{\boldsymbol{f}} = \arg\min_{\{\boldsymbol{f} : \boldsymbol{H}\boldsymbol{f} = \boldsymbol{g}\}} \{F(\boldsymbol{f})\}. \tag{89}$$

$D(\boldsymbol{\lambda})$ is called the dual criterion and $F(\boldsymbol{f})$ primal. However, it is not always easy to obtain an analytical expression for $G(\boldsymbol{s})$ and its gradient $G'(\boldsymbol{s})$. The functions $F(\boldsymbol{f})$ and $G(\boldsymbol{s})$ are conjugate convex.

For the computational aspect, unfortunately, the cases where we may have analytical expressions for $Z(\boldsymbol{\lambda})$ or $G(\boldsymbol{s}) = \ln Z$ or $F(\boldsymbol{f})$ are very limited. However, when there is analytical expressions for them, the computations can be done very easily. In Table 1, we summarizes some of those solutions:

Table 1. Analytical solutions for different measures $\mu(\boldsymbol{u})$

$\mu(\boldsymbol{u}) \propto \exp[-\frac{1}{2}\sum_j u_j^2]$	$\widehat{\boldsymbol{f}} = \boldsymbol{H}'\boldsymbol{\lambda}$	$\widehat{\boldsymbol{f}} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{g}$
$\mu(\boldsymbol{u}) \propto \exp[-\sum_j \lvert u_j \rvert]$	$\widehat{\boldsymbol{f}} = \mathbf{1}./(\boldsymbol{H}'\boldsymbol{\lambda} \pm \mathbf{1})$	$\boldsymbol{H}\widehat{\boldsymbol{f}} = \boldsymbol{g}$
$\mu(\boldsymbol{u}) \propto \exp[-\sum_j u_j^{\alpha-1} \exp[-\beta u_j]], \quad u_j > 0$	$\widehat{\boldsymbol{f}} = \alpha\mathbf{1}./(\boldsymbol{H}'\boldsymbol{\lambda} + \beta\mathbf{1})$	$\boldsymbol{H}\widehat{\boldsymbol{f}} = \boldsymbol{g}$

14. Bayesian Approach for Inverse Problems

In this section, we present in a brief way the Bayesian approach for the inverse problems in signal and image processing.

14.1. Simple Bayesian Approach

The different steps to find a solution to an inverse problem using the Bayesian approach can be summarized as follows:

- Assign a prior probability law $p(\boldsymbol{\epsilon})$ to the modeling and observation errors, here $\boldsymbol{\epsilon}$. From this, find the expression of the likelihood $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)$. As an example, consider the Gaussian case:

$$p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, v_\epsilon \boldsymbol{I}) \longrightarrow p(\boldsymbol{g}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I}). \tag{90}$$

  θ_1 in this case is the noise variance v_ϵ.

- Assign a prior probability law $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$ to the unknown $\boldsymbol{f}$ to translate your prior knowledge on it. Again, as an example, consider the Gaussian case:

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\mathbf{0}, v_f \boldsymbol{I}) \tag{91}$$

  θ_2 in this case is the variance v_f.

- Apply the Bayes rule to obtain the expression of the posterior law:

$$p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)}{p(\boldsymbol{g}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2), \tag{92}$$

where the sign $\propto$ stands for "proportionality to", $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)$ is the likelihood, $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$ the prior model, $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]'$ their corresponding parameters (often called the hyper-parameters of the problem) and $p(\boldsymbol{g}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is called the evidence of the model.

- Use $p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ to infer any quantity dependent of $\boldsymbol{f}$.

For the expressions of likelihood in (90) and the prior in (91), we obtain very easily the expression of the posterior:

$$p(\boldsymbol{f}|\boldsymbol{g}, v_\epsilon, v_f) = \mathcal{N}(\boldsymbol{f}|\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{V}}) \text{ with } \widehat{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \frac{v_\epsilon}{v_f}\boldsymbol{I})^{-1} \text{ and } \widehat{\boldsymbol{f}} = \widehat{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g} \tag{93}$$

When the hyper-parameters $\boldsymbol{\theta}$ can be fixed *a priori*, the problem is easy. In practice, we may use some summaries, such as:

- MAP:
$$\widehat{\boldsymbol{f}}_{\text{MAP}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta})\} \tag{94}$$
- EAP or posterior mean (PM):
$$\widehat{\boldsymbol{f}}_{EAP} = \int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{f} \tag{95}$$

For the Gaussian case of (91), the MAP and EAP are the same and can be obtained by noting that:

$$\widehat{\boldsymbol{f}}_{\text{MAP}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} \text{ with } J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{f}\|_2^2, \text{ where } \lambda = v_\epsilon/v_f. \tag{96}$$

However, in real applications, the computation of even these simple point estimators may need efficient algorithm:

- For MAP, we need optimization algorithms, which can handle the huge dimensional criterion $J(\boldsymbol{f}) = -\ln p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta})$. Very often, we may be limited to using gradient-based algorithms.
- For EAP, we need integration algorithms, which can handle huge dimensional integrals. The most common tool here is the MCMC methods [24]. However, for real applications, very often, the computational costs are huge. Recently, different methods, called approximate Bayesian computation (ABC) [96–100] or VBA, have been proposed [74,96,98,101–107].

14.2. Full Bayesian: Hyperparameter Estimation

When the hyperparameters $\boldsymbol{\theta}$ have also to be estimated, a prior $p(\boldsymbol{\theta})$ is assigned to them, and the expression of the joint posterior:

$$p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})}{p(\boldsymbol{g})} \tag{97}$$

is obtained, which can then be used to infer them jointly. Very often, the expression of this joint posterior law is complex, and any computation may become very costly. The VBA methods try to approximate $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by a simpler distribution, which can be handled more easily. Two particular and extreme cases are:

- Bloc separable, such as $q(\boldsymbol{f},\boldsymbol{\theta}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{\theta})$ or
- Completely separable, such as $q(\boldsymbol{f},\boldsymbol{\theta}) = \prod_j q_{1j}(f_j) \prod_k q_{2k}(\theta_k)$.

Any mixed solution is also valid. For example, the one we have chosen is:

$$q(\boldsymbol{f},\boldsymbol{\theta}) = q_1(\boldsymbol{f}) \prod_k q_{2k}(\theta_k) \tag{98}$$

Obtaining the expressions of these approximated separable probability laws has to be done via a criterion. The natural criterion with some geometrical interpretation for the probability law manifolds is the Kullback–Leibler (KL) criterion:

$$\mathrm{KL}\left[q:p\right] = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q . \tag{99}$$

For hierarchical prior models with hidden variables $\boldsymbol{z}$, the problem becomes more complex, because we have to give the expression of the joint posterior law:

$$p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta}) \tag{100}$$

and then approximate it by separable ones:

$$q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\; q_2(\boldsymbol{z})\; q_3(\boldsymbol{\theta}) \text{ or } q(\boldsymbol{f},\boldsymbol{\theta}) = \prod_j q_{1j}(f_j|z_{f_j}) \prod_j q_{2j}(z_{f_j}) \prod_k q_{3k}(\theta_k) \tag{101}$$

and then use them for estimation. See more discussions in [9,31,38,108–110]

In the following, first the general VBA method is detailed for the inference problems with hierarchical prior models. Then, a particular class of prior model (Student *t*) is considered, and the details of VBA algorithms for that are given.

15. Basic Algorithms of the Variational Bayesian Approximation

To illustrate the basic ideas and tools, let us consider a vector $\boldsymbol{X}$ and its probability density function $p(\boldsymbol{x})$, which we want to approximate by $q(\boldsymbol{x}) = \prod_j q_j(x_j)$. Using the KL criterion:

$$\begin{aligned}
\mathrm{KL}\left[q:p\right] &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} = \int q(\boldsymbol{x}) \ln q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int q(\boldsymbol{x}) \ln p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&= \sum_j \int q_j(x_j) \ln q_j(x_j)\, \mathrm{d}x_j - \left\langle \ln p(\boldsymbol{x}) \right\rangle_q \\
&= \sum_j \int q_j(x_j) \ln q_j(x_j)\, \mathrm{d}x_j - \int q_j(x_j) < \ln p(\boldsymbol{x}) >_{q_{-j}}\, \mathrm{d}x_j
\end{aligned} \tag{102}$$

where we used the notation: $\left\langle \ln p(\boldsymbol{x}) \right\rangle_q = \int q(\boldsymbol{x}) \ln p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$ and $q_{-j}(\boldsymbol{x}) = \prod_{i \neq j} q_i(x_i)$.

From here, trying to find the solution q_i, the basic method is an alternate optimization algorithm:

$$q_j(x_j) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_{-j}} \right] . \tag{103}$$

As we can see, the expression of $q_j(x_j)$ depends on $q_i(x_i), i \neq j$. It is not always possible to obtain analytical expressions for $q_j(x_j)$. It is however possible to show that, if $p(\boldsymbol{x})$ is a member of exponential

families, then $q_j(x_j)$ are also members of exponential families. These iterations then become much simpler, because at each iteration, we need to update the parameters of the exponential families. To go a little more into the details, let us consider some particular simple cases.

15.1. Case of Two Gaussian Variables

In the case of two variables $\boldsymbol{x} = [x_1, x_2]'$, we have:

$$\begin{cases} q_1(x_1) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_2(x_2)}\right] \\ q_2(x_2) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_1(x_1)}\right] \end{cases} \tag{104}$$

As an illustrative example, consider the case where we want to approximate $p(x_1, x_2)$ by $q(x_1, x_2) = q_1(x_1)\, q_2(x_2)$ to be able to compute the expected values:

$$\begin{cases} m_1 = \mathrm{E}\{x_1\} = \int\int x_1\, p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ m_2 = \mathrm{E}\{x_2\} = \int\int x_2\, p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \end{cases} \tag{105}$$

which need double integrations when $p(x_1, x_2)$ is not separable in its two variables. If we can do that separable approximation, then, we can compute:

$$\begin{cases} \widetilde{\mu}_1 = \mathrm{E}\{x_1\} = \int x_1\, q_1(x_1)\, \mathrm{d}x_1 \\ \widetilde{\mu}_2 = \mathrm{E}\{x_2\} = \int x_2\, q_2(x_2)\, \mathrm{d}x_2 \end{cases} \tag{106}$$

which needs only 1D integrals. Let us see if $(\widetilde{\mu}_1, \widetilde{\mu}_2)$ will converge to (m_1, m_2). To illustrate this, let us consider the very simple case of the Gaussian:

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} v_1 & \rho\sqrt{v_1 v_2} \\ \rho\sqrt{v_1 v_2} & v_2 \end{bmatrix} \right). \tag{107}$$

It is then easy to see that $q_1(x_1) = \mathcal{N}(x_1|\widetilde{\mu}_1, \widetilde{v}_1)$ and $q_2(x_2) = \mathcal{N}(x_2|\widetilde{\mu}_2, \widetilde{v}_2)$ and that:

$$\begin{cases} q_1^{(k+1)}(x_1) = p(x_1|x_2 = \widetilde{\mu}_2^{(k)}) = \mathcal{N}\left(x_1|\widetilde{\mu}_1^{(k)}, \widetilde{v}_1^{(k)}\right) \\ q_2^{(k+1)}(x_2) = p(x_2|x_1 = \widetilde{\mu}_1^{(k)}) = \mathcal{N}\left(x_1|\widetilde{\mu}_2^{(k)}, \widetilde{v}_2^{(k)}\right) \end{cases} \tag{108}$$

with:

$$\begin{cases} \widetilde{\mu}_1^{(k+1)} = m_1 + \rho\sqrt{v_1/v_2}(\widetilde{\mu}_2^{(k)} - m_2) \\ \widetilde{v}_1^{(k+1)} = (1-\rho^2)v_1 \\ \widetilde{\mu}_2^{(k+1)} = m_2 + \rho\sqrt{v_2/v_1}(\widetilde{\mu}_1^{(k)} - m_1) \\ \widetilde{v}_2^{(k+1)} = (1-\rho^2)v_2 \end{cases}. \tag{109}$$

See [111] for details and where we showed that, initializing the algorithm with $\widetilde{\mu}_1^{(0)} = 0$ and $\widetilde{\mu}_2^{(0)} = 0$, the means converges to the right values m_1 and m_2, However, we may be careful about the convergence of the variances.

15.2. Case of Exponential Families

As we could see, to be able to use such an algorithm in practical cases, we need to be able to compute $< \ln p(\boldsymbol{x}) >_{q_2(x_2)}$ and $< \ln p(\boldsymbol{x}) >_{q_1(x_1)}$. Only for a few cases can we can do this analytically. Different algorithms can be obtained depending on the choice of a particular family for $q_j(x_j)$ [103,112–120].

To show this, let us consider the exponential family:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \, \exp\left[\boldsymbol{\theta}' \boldsymbol{u}(\boldsymbol{x})\right] \tag{110}$$

where $\boldsymbol{\theta}$ is a vector of parameter and $g(\boldsymbol{\theta})$ and $\boldsymbol{u}(\boldsymbol{x})$ are known functions.

This parametric exponential family has the following conjugacy property: For a given prior $p(\boldsymbol{\theta})$ in the family:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) \, g(\boldsymbol{\theta})^{\eta} \, \exp\left[\boldsymbol{\nu}' \boldsymbol{\theta}\right] \tag{111}$$

the corresponding posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\boldsymbol{\theta}) \; p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) \\ &\propto g(\boldsymbol{\theta})^{\eta+1} \exp\left[[\boldsymbol{\nu} + \boldsymbol{u}(\boldsymbol{x})]' \boldsymbol{\theta}\right] \\ &\propto p(\boldsymbol{\theta}|\eta + 1, \boldsymbol{\nu} + \boldsymbol{u}(\boldsymbol{x})) \end{aligned} \tag{112}$$

is in the same family.

For this family, we have:

$$\left\langle \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \right\rangle_q = \ln g(\boldsymbol{\theta}) + \boldsymbol{\theta}' \left\langle \boldsymbol{u}(\boldsymbol{x}) \right\rangle_q . \tag{113}$$

It is then easy to show that:

$$q_j(x_j) \propto g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\theta}' \left\langle \boldsymbol{u}(\boldsymbol{x}) \right\rangle_{q_{-j}}\right] \tag{114}$$

which are in the same exponential family. This simplifies greatly the computations, thanks to the fact that, in each iteration, we only need to compute $\widetilde{\boldsymbol{u}}(\boldsymbol{x}) = \left\langle \boldsymbol{u}(\boldsymbol{x}) \right\rangle_{q_{-j}}$ and update the parameters.

Now, if we consider:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \, \exp\left[\boldsymbol{\theta}' \boldsymbol{u}(\boldsymbol{x})\right] \tag{115}$$

with a prior on $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) \, g(\boldsymbol{\theta})^{\eta} \, \exp\left[\boldsymbol{\nu}' \boldsymbol{\theta}\right] \tag{116}$$

and the joint $p(\boldsymbol{x}, \boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \, p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu})$, which is not separable in $\boldsymbol{x}$ and $\boldsymbol{\theta}$, and we want to approximate it with the separable $q(\boldsymbol{x}, \boldsymbol{\theta}) = q_1(\boldsymbol{x}) \, q_2(\boldsymbol{\theta})$, then we will have:

$$\left\{ \begin{array}{l} q(\boldsymbol{\theta}) = h(\tilde{\eta}, \widetilde{\boldsymbol{\nu}}) \, g(\boldsymbol{\theta})^{\tilde{\eta}} \, \exp\left[\widetilde{\boldsymbol{\nu}}' \boldsymbol{\theta}\right] \\ q(\boldsymbol{x}) = g(\widetilde{\boldsymbol{\theta}}) \exp\left[\widetilde{\boldsymbol{\theta}}' \boldsymbol{u}(\boldsymbol{x})\right] \end{array} \right. \quad \text{with} \quad \left\{ \begin{array}{ll} \tilde{\eta} & = \eta + 1 \\ \widetilde{\boldsymbol{\nu}} & = \boldsymbol{\nu} + \widetilde{\boldsymbol{u}}(\boldsymbol{x}) \\ \widetilde{\boldsymbol{\theta}} & = \widetilde{\boldsymbol{\nu}} \end{array} \right. \tag{117}$$

where $\widetilde{\boldsymbol{u}} = \left\langle \boldsymbol{u}(\boldsymbol{x}) \right\rangle_{q_1(\boldsymbol{x})}$.

16. VBA for the Unsupervised Bayesian Approach to Inverse Problems

Before going into the details and for similarity with the notations in the next sections, we replace $\boldsymbol{x}$ by $\boldsymbol{f}$, such that now we are trying to approximate $p(\boldsymbol{f},\boldsymbol{\theta}) = p(\boldsymbol{f}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ by a separable $q(\boldsymbol{f},\boldsymbol{\theta}) = q_1(\boldsymbol{f})\,q_2(\boldsymbol{\theta})$. Interestingly, depending on the choice of the family laws for q_1 and q_2, we obtain different algorithms:

- $q_1(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$ and $q_2(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$. In this case, we have:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[< \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_2}\right] \propto \exp\left[\ln p(\boldsymbol{f},\widetilde{\boldsymbol{\theta}})\right] \propto p(\boldsymbol{f},\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}) \propto p(\boldsymbol{f}|\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}) \\ q_2(\boldsymbol{\theta}) \propto \exp\left[< \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_1}\right] \propto \exp\left[\ln p(\widetilde{\boldsymbol{f}},\boldsymbol{\theta})\right] \propto p(\boldsymbol{f}=\widetilde{\boldsymbol{f}},\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{f}=\widetilde{\boldsymbol{f}}) \end{array}\right. \tag{118}$$

  and so:

$$\left\{\begin{array}{l} \widetilde{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}}\left\{p(\boldsymbol{f},\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}})\right\} \\ \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\left\{p(\boldsymbol{f}=\widetilde{\boldsymbol{f}},\boldsymbol{\theta})\right\} \end{array}\right. \tag{119}$$

  which can be interpreted as an alternate optimization algorithm for obtaining the JMAPestimates:

$$(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{f},\boldsymbol{\theta})}\left\{p(\boldsymbol{f},\boldsymbol{\theta})\right\}. \tag{120}$$

  The main drawback here is that the uncertainties of the $\boldsymbol{f}$ are not used for the estimation of $\boldsymbol{\theta}$ and the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- $q_1(\boldsymbol{f})$ is free form and $q_2(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$. In the same way, this time we obtain:

$$\left\{\begin{array}{l} < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_2(\boldsymbol{\theta})} = \ln p(\boldsymbol{f},\widetilde{\boldsymbol{\theta}}) \\ < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_1(\boldsymbol{f})} = < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\theta}})} = Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}) \end{array}\right. \tag{121}$$

  which leads to:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[\ln p(\boldsymbol{f},\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}})\right] \propto p(\boldsymbol{f},\widetilde{\boldsymbol{\theta}}) \\ q_2(\boldsymbol{\theta}) \propto \exp\left[Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}})\right] \longrightarrow \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\left\{Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}})\right\} \end{array}\right. \tag{122}$$

  which can be compared with the Bayesian expectation maximization (BEM) algorithm. The E-step is the computation of the expectation $Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}})$ in (121), and the M-step is the maximization in (122). Here, the uncertainties of the $\boldsymbol{f}$ are used for the estimation of $\boldsymbol{\theta}$, but the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- $q_1(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$ and $q_2(\boldsymbol{\theta})$ is free form. In the same way, this time we obtain:

$$\left\{\begin{array}{l} < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_1(\boldsymbol{f})} = \ln p(\boldsymbol{f}=\widetilde{\boldsymbol{f}},\boldsymbol{\theta}) \\ < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{q_2(\boldsymbol{\theta})} = < \ln p(\boldsymbol{f},\boldsymbol{\theta}) >_{p(\boldsymbol{\theta}|\boldsymbol{f}=\widetilde{\boldsymbol{f}})} = Q(\widetilde{\boldsymbol{f}},\boldsymbol{f}) \end{array}\right. \tag{123}$$

$$\left\{\begin{array}{l} q_2(\theta) \propto \ln p(\boldsymbol{f}=\widetilde{\boldsymbol{f}},\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{f}=\widetilde{\boldsymbol{f}}) \\ q_1(\boldsymbol{f}) \propto \exp\left[Q(\widetilde{\boldsymbol{f}},\boldsymbol{\theta})\right] \longrightarrow \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\left\{Q(\boldsymbol{f}=\widetilde{\boldsymbol{f}},\boldsymbol{\theta})\right\} \end{array}\right. \tag{124}$$

  which can be compared with the classical EM algorithm. Here, the uncertainties of the $\boldsymbol{f}$ are used for the estimation of $\boldsymbol{\theta}$, but the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- Both $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{\theta})$ have free form. The main difficulty here is that, at each iteration, the expression of q_1 and q_2 may change. However, if $p(\boldsymbol{f},\boldsymbol{\theta})$ is in the generalized exponential family, the expressions of $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{\theta})$ will also be in the same family, and we have only to update the parameters at each iteration.

17. VBA for a Linear Inverse Problem with Simple Gaussian Priors

As a simple example, consider the Gaussian case where $p(\boldsymbol{g}|\boldsymbol{f},\theta_1)=\mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f},(1/\theta_1)\boldsymbol{I})$, $p(\boldsymbol{f}|\theta_2)=\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},(1/\theta_2)\boldsymbol{I})$ and $p(\theta_1)=\mathcal{G}(\theta_1|\alpha_{10},\beta_{10})$ $p(\theta_2)=\mathcal{G}(\theta_2|\alpha_{20},\beta_{20})$, and so, we have:

$$\begin{aligned}\ln p(\boldsymbol{f},\theta_1,\theta_2|\boldsymbol{g}) = \frac{M}{2}\ln\theta_1 - \frac{\theta_1}{2}\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2 + \frac{N}{2}\ln\theta_2 - \frac{\theta_2}{2}\|\boldsymbol{f}\|_2^2 \\ + (\alpha_{10}-1)\ln\theta_1 - \beta_{10}\theta_1 + (\alpha_{20}-1)\ln\theta_2 - \beta_{20}\theta_2.\end{aligned} \tag{125}$$

From this expression $J(\boldsymbol{f},\theta_1,\theta_2)=\ln p(\boldsymbol{f},\theta_1,\theta_2|\boldsymbol{g})$, it is easy to obtain the equations of an alternate JMAP algorithm by computing the derivatives of it with respective to its arguments and equating them to zero:

$$\begin{aligned}&\frac{\partial J}{\partial \boldsymbol{f}}=0 \longrightarrow \boldsymbol{f}=(\boldsymbol{H}'\boldsymbol{H}+\lambda\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g} \text{ with } \lambda=\frac{\theta_2}{\theta_1}\\ &\frac{\partial J}{\partial \theta_1}=0 \longrightarrow \theta_1=\frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1} \text{ with } \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2} \text{ and } \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2\\ &\frac{\partial J}{\partial \theta_2}=0 \longrightarrow \theta_1=\frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2} \text{ with } \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{M}{2} \text{ and } \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}\|\boldsymbol{f}\|_2^2\end{aligned} \tag{126}$$

From the expression of the joint probability law $p(\boldsymbol{f},\theta_1,\theta_2|\boldsymbol{g})$, we can also obtain the expressions of the conditionals:

$$\left\{\begin{array}{l} p(\boldsymbol{f}|\boldsymbol{g},\theta_1,\theta_2)=\mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{V}})\\ \quad\text{with}\quad \widetilde{\boldsymbol{V}}=(\boldsymbol{H}'\boldsymbol{H}+\lambda\boldsymbol{I})^{-1},\quad \widetilde{\boldsymbol{f}}=\widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g},\quad \lambda=\frac{\theta_2}{\theta_1}\\ p(\theta_1|\boldsymbol{g},\boldsymbol{f},\theta_2)=\mathcal{G}(\theta_1|\widetilde{\alpha}_1,\widetilde{\beta}_1)\\ \quad\text{with}\quad \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2},\quad \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2\\ p(\theta_2|\boldsymbol{g},\boldsymbol{f},\theta_1)=\mathcal{G}(\theta_2|\widetilde{\alpha}_2,\widetilde{\beta}_2)\\ \quad\text{with}\quad \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{M}{2},\quad \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}\|\boldsymbol{f}\|_2^2\end{array}\right. \tag{127}$$

However, obtaining analytical expressions of the marginals $p(\boldsymbol{f}|\boldsymbol{g})$, $p(\theta_1|\boldsymbol{g})$ and $p(\theta_2|\boldsymbol{g})$ is not easy. We can then obtain approximate expressions $q_1(\boldsymbol{f}|\boldsymbol{g})$, $q_2(\theta_1|\boldsymbol{g})$ and $q_3(\theta_2|\boldsymbol{g})$ using the VBA method. For this case, thanks to the conjugacy property, we have:

$$\left\{\begin{array}{l} q(\boldsymbol{f})=\mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{V}})\\ \quad\text{with}\quad \widetilde{\boldsymbol{V}}=(\boldsymbol{H}'\boldsymbol{H}+\widetilde{\lambda}\boldsymbol{I})^{-1},\quad \widetilde{\boldsymbol{f}}=\widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g},\quad \widetilde{\lambda}=\frac{<\theta_2>}{<\theta_1>};\\ q(\theta_1)=\mathcal{G}(\theta_1|\widetilde{\alpha}_1,\widetilde{\beta}_1)\\ \quad\text{with}\quad \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2},\quad \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}<\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2>\\ p(\theta_2|\boldsymbol{g},\boldsymbol{f})=\mathcal{G}(\theta_2|\widetilde{\alpha}_2,\widetilde{\beta}_2)\\ \quad\text{with}\quad \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{N}{2},\quad \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}<\|\boldsymbol{f}\|_2^2>\end{array}\right. \tag{128}$$

We can then compare the three algorithms in Table 2:

Table 2. Comparision of three algorithms: JMAP, BEM and VBA

JMAP	BEM	VBA
$q(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$	$q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}\|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{V}})$	$q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}\|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{V}})$
$\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$	$\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$	$\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$
$\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$	$\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$	$\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$
$q(\theta_1) = \delta(\theta_1 - \widetilde{\theta}_1)$	$q(\theta_1) = \delta(\theta_1 - \widetilde{\theta}_1)$	$q(\theta_1) = \mathcal{G}(\theta_1\|\widetilde{\alpha}_1, \widetilde{\beta}_1)$
$\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$	$\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$	$\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$
$\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2}\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2$	$\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2} < \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 >$	$\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2} < \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 >$
$\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$	$\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$	$\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$
$q(\theta_2) = \delta(\theta_2 - \widetilde{\theta}_2)$	$q(\theta_2) = \delta(\theta_2 - \widetilde{\theta}_2)$	$q(\theta_2) = \mathcal{G}(\theta_2\|\widetilde{\alpha}_2, \widetilde{\beta}_2)$
$\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{M}{2}$	$\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{M}{2}$	$\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{N}{2}$
$\widetilde{\beta}_2 = \beta_{10} + \frac{1}{2}\|\boldsymbol{f}\|_2^2$	$\widetilde{\beta}_2 = \beta_{20} + \frac{1}{2} < \|\boldsymbol{f}\|_2^2 >$	$\widetilde{\beta}_2 = \beta_{20} + \frac{1}{2} < \|\boldsymbol{f}\|_2^2 >$
$\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$	$\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$	$\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$
$\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$	$\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$	$\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$

It is important to remark that, in JMAP, the computation of $\boldsymbol{f}$ can be done via the optimization of the criterion $J(\boldsymbol{f}, \theta_1, \theta_2) = \ln p(\boldsymbol{f}, \theta_1, \theta_2|\boldsymbol{g})$, which does not need explicitly the matrix inversion of $\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$. However, in BEM and VBA, we need to compute it due to the following requirements:

$$\begin{aligned}
< \boldsymbol{f} >_q &= \widetilde{\boldsymbol{f}}, \\
< \|\boldsymbol{f}\|^2 >_q &= \text{tr}\left(< \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}' >_q \right) = \text{tr}\left(\widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}' + \widetilde{\boldsymbol{V}} \right) = \|\widetilde{\boldsymbol{f}}\|^2 + \text{tr}\left(\widetilde{\boldsymbol{V}} \right), \\
< f_j^2 >_q &= [\widetilde{\boldsymbol{V}}]_{jj} + \widetilde{f}_j^2, \\
< \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 >_q &= [\boldsymbol{g}'\boldsymbol{g} - 2\left\langle \boldsymbol{f} \right\rangle_q' \boldsymbol{H}'\boldsymbol{g} + \boldsymbol{H}' \left\langle \boldsymbol{f}'\boldsymbol{f} \right\rangle_q \boldsymbol{H}] \\
&= [\boldsymbol{g}'\boldsymbol{g} - 2\widetilde{\boldsymbol{f}}'\boldsymbol{H}'\boldsymbol{g} + \boldsymbol{H}'(\widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}')\boldsymbol{H}] \\
&= \|\boldsymbol{g} - \boldsymbol{H}\widetilde{\boldsymbol{f}}\|^2 + \text{tr}\left(\boldsymbol{H}'\widetilde{\boldsymbol{V}}\boldsymbol{H} \right)]
\end{aligned} \tag{129}$$

For some extensions and more details, see [111].

18. Bayesian Variational Approximation with Hierarchical Prior Models

For a linear inverse problem:

$$\mathcal{M} : \boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon} \tag{130}$$

with an assigned likelihood $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)$, when a hierarchical prior model $p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)$ is used and when the estimation of the hyper-parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3]'$ has to be considered, the joint posterior law of all the unknowns becomes:

$$p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) = \frac{p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})}{p(\boldsymbol{g})} = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta})}{p(\boldsymbol{g})}. \tag{131}$$

The main idea behind the VBA is to approximate this joint posterior by a separable one, for example: $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{z})\, q_3(\boldsymbol{\theta})$ and where the expressions of $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ are obtained by minimizing the Kullback–Leibler divergence (99), as explained in previous section. This approach can also be used for model selection based on the evidence of the model $\ln p(\boldsymbol{g})$ [121] where:

$$p(\boldsymbol{g}) = \int\int\int p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{\theta}. \tag{132}$$

Interestingly, it is easy to show that:

$$\ln p(\boldsymbol{g}) = \mathrm{KL}\left[q : p\right] + \mathcal{F}(q) \tag{133}$$

where $\mathcal{F}(q)$ is the free energy associated with q defined as:

$$\mathcal{F}(q) = \left\langle \ln \frac{p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g})}{q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta})} \right\rangle_q \tag{134}$$

Therefore, for a given model $\mathcal{M}$, minimizing $\mathrm{KL}\left[q : p\right]$ is equivalent to maximizing $\mathcal{F}(q)$ and when optimized, $\mathcal{F}(q^*)$ gives a lower bound for $\ln p(\boldsymbol{g})$. Indeed, the name variational approximation is due to the fact that $\ln p(\boldsymbol{g}) \geq \mathcal{F}(q)$, and so, $\mathcal{F}(q)$ is a lower bound to the evidence $\ln p(\boldsymbol{g})$.

Without any other constraint than the normalization of q, an alternate optimization of $\mathcal{F}(q)$ with respect to q_1, q_2 and q_3 results in:

$$\left\{ \begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[- \left\langle \ln p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}) \right\rangle_{q(\boldsymbol{z})q(\boldsymbol{\theta})} \right], \\ q_2(\boldsymbol{z}) \propto \exp\left[- \left\langle \ln p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}) \right\rangle_{q(\boldsymbol{f})q(\boldsymbol{\theta})} \right], \\ q_3(\boldsymbol{\theta}) \propto \exp\left[- \left\langle \ln p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}) \right\rangle_{q(\boldsymbol{f})q(\boldsymbol{z})} \right]. \end{array} \right. \tag{135}$$

Note that these relations represent an implicit solution for $q_1(\boldsymbol{f})$, $q_2(\boldsymbol{z})$ and $q_3(\boldsymbol{\theta})$, which need, at each iteration, the expression of the expectations in the right hand of exponentials. If $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}_1)$ is a member of an exponential family and if all of the priors $p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}_2)$, $p(\boldsymbol{z}|\boldsymbol{\theta}_3)$, $p(\boldsymbol{\theta}_1)$, $p(\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_3)$ are conjugate priors, then it is easy to see that these expressions lead to standard distributions for which the required expectations are easily evaluated. In that case, we may note:

$$q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) = q_1(\boldsymbol{f}|\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}})\, q_2(\boldsymbol{z}|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}})\, q_3(\boldsymbol{\theta}|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{z}}) \tag{136}$$

where the tilded quantities $\widetilde{\boldsymbol{z}}$, $\widetilde{\boldsymbol{f}}$ and $\widetilde{\boldsymbol{\theta}}$ are, respectively, functions of $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})$, $(\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})$ and $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}})$. This means that the expression of $q_1(\boldsymbol{f}|\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}})$ depends on $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})$, the expression of $q_2(\boldsymbol{z}|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}})$ depends on $(\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})$ and the expression of $q_3(\boldsymbol{\theta}|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{z}})$ depends on $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}})$. With this notation, the alternate optimization results in alternate updating of the parameters $(\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}})$ of q_1, the parameters $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}})$ of q_2 and the parameters $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{z}})$ of q_3. Finally, we may note that, to monitor the convergence of the algorithm, we may evaluate the free energy:

$$\begin{aligned} \mathcal{F}(q) &= \left\langle \ln p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}) \right\rangle_q - \left\langle \ln q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q \\ &= \left\langle \ln p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{z}|\boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{\theta}) \right\rangle_q \\ &\quad - \left\langle \ln q(\boldsymbol{f}) \right\rangle_q - \left\langle \ln q(\boldsymbol{z}) \right\rangle_q - \left\langle \ln q(\boldsymbol{\theta}) \right\rangle_q. \end{aligned} \tag{137}$$

Other decompositions for $q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta})$ are also possible. For example: $q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}) = q_1(\boldsymbol{f}|\boldsymbol{z})\, q_2(\boldsymbol{z})\, q_3(\boldsymbol{\theta})$ or even: $q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}) = \prod_j q_{1j}(f_j)\ \prod_j q_{2j}(z_{f_j})\ \prod_l q_{3l}(\theta_l)$. Here, we consider the first case and give some more details on it.

19. Bayesian Variational Approximation with Student *t* Priors

The Student *t* model is:

$$p(\boldsymbol{f}|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \text{ with } \mathcal{S}t(f_j|\nu) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1+f_j^2/\nu\right)^{-(\nu+1)/2} \tag{138}$$

The Cauchy model is obtained when $\nu = 1$. Knowing that:

$$\mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0,1/z_{f_j})\,\mathcal{G}(z_{f_j}|\nu/2,\nu/2)\,\mathrm{d}z_{f_j} \tag{139}$$

we can write this model via the positive hidden variables z_{f_j}:

$$\left\{\begin{array}{ll} p(f_j|z_{f_j}) & = \mathcal{N}(f_j|0,1/z_{f_j}) \propto \exp\left[-\frac{1}{2}z_{f_j}f_j^2\right] \\ p(z_{f_j}|\alpha,\beta) & = \mathcal{G}(z_{f_j}|\alpha,\beta) \propto z_{f_j}^{(\alpha-1)}\exp\left[-\beta z_{f_j}\right] \end{array}\right. \tag{140}$$

Now, let us consider the forward model $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$ and assign a Gaussian law with unknown variance v_{ϵ_i} to the noise ϵ_i, which results in $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{0},\boldsymbol{V}_\epsilon)$ with $\boldsymbol{V}_\epsilon = \text{diag}\,[\boldsymbol{v}_\epsilon]$ with $\boldsymbol{v}_\epsilon = [v_{\epsilon_1},\cdots,v_{\epsilon_M}]$, and so:

$$p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{v}_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f},\boldsymbol{V}_\epsilon) \propto \exp\left[-\frac{1}{2}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\boldsymbol{V}_\epsilon^{-1}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\right]. \tag{141}$$

Let us also note by $z_{\epsilon_i} = 1/v_{\epsilon_i}$, $\boldsymbol{z}_\epsilon = [z_{\epsilon_1},\cdots,z_{\epsilon_M}]$ and $\boldsymbol{Z}_\epsilon = \text{diag}\,[\boldsymbol{z}_\epsilon] = \boldsymbol{V}_\epsilon^{-1}$ and assign a prior on it $p(v_{\epsilon_i}|\alpha_{\epsilon_0},\beta_{\epsilon_0}) = \mathcal{IG}(v_{\epsilon_i}|\alpha_{\epsilon_0},\beta_{\epsilon_0})$ or equivalently:

$$p(z_{\epsilon_i}|\alpha_{\epsilon_0},\beta_{\epsilon_0}) = \mathcal{G}(z_{\epsilon_i}|\alpha_{\epsilon_0},\beta_{\epsilon_0}) \quad \text{and} \quad p(\boldsymbol{z}_\epsilon|\alpha_{\epsilon_0},\beta_{\epsilon_0}) = \prod_i \mathcal{G}(z_{\epsilon_i}|\alpha_{\epsilon_0},\beta_{\epsilon_0}). \tag{142}$$

Let us also note $\boldsymbol{v}_f = [v_{f_1},\cdots,v_{f_N}]$, $\boldsymbol{V}_f = \text{diag}\,[\boldsymbol{v}_f]$, $z_{f_j} = 1/v_{f_j}$, $\boldsymbol{Z}_f = \text{diag}\,[\boldsymbol{z}_f] = \boldsymbol{V}_f^{-1}$ and note:

$$p(\boldsymbol{f}|\boldsymbol{v}_f) = \prod_j p(f_j|v_{f_j}) = \prod_j \mathcal{N}\left(f_j|0,v_{f_j}\right) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{V}_f) \tag{143}$$

and finally,

$$p(\boldsymbol{v}_f|\alpha_{f_0},\beta_{f_0}) = \prod_j \mathcal{G}(v_{f_j}|\alpha_{f_0},\beta_{f_0}). \tag{144}$$

Then, we obtain the following expressions for the VBA:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{V}}_f) = \mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{V}}) \text{ with } \widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g},\ \widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\widetilde{\boldsymbol{V}}_\epsilon^{-1}\boldsymbol{H} + \widetilde{\boldsymbol{Z}}_f)^{-1}; \\ q_{2j}(z_{f_j}) = \mathcal{G}(z_{f_j}|\widetilde{\alpha}_j,\widetilde{\beta}_j) \text{ with } \widetilde{\alpha}_j = \alpha_{00}+1/2,\ \widetilde{\beta}_j = \beta_{00} + <f_j^2>/2; \\ q_3(z_{\epsilon_i}) = \mathcal{G}(z_{\epsilon_i}|\widetilde{\alpha}_{\epsilon_i},\widetilde{\beta}_{\epsilon_i}) \text{ with } \widetilde{\alpha}_{\epsilon_i} = \alpha_{\epsilon_0}+(N+1)/2,\ \widetilde{\beta}_{\epsilon_i} = \beta_{\epsilon 0} + \frac{1}{2} < |g_i - [\boldsymbol{H}\boldsymbol{f}]_i|^2 >; \end{array}\right. \tag{145}$$

where:

$$< |g_i - [\boldsymbol{H}\boldsymbol{f}]_i|^2 >= |g_i - \boldsymbol{H} < \boldsymbol{f} >]_i|^2 + [\boldsymbol{H}'\widetilde{\boldsymbol{V}}\boldsymbol{H}]_{ii},$$
$$< \boldsymbol{f} >= \widetilde{\boldsymbol{\mu}}, \quad < \boldsymbol{f}\boldsymbol{f}' >= \widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}',$$
$$< f_j^2 >= [\widetilde{\boldsymbol{V}}]_{jj} + \widetilde{\mu}_j^2$$

We have implemented these algorithms for many linear inverse problems [102], such as periodic components estimation in time series [122] or computed tomography [123], blind deconvolution [124], blind image separation [125,126] and blind image restoration [89].

20. Conclusions

The main conclusions of this paper can be summarized as follows:

- A probability law is a tool for representing our state of knowledge about a quantity.
- The Bayes or Laplace rule is an inference tool for updating our state of knowledge about an inaccessible quantity when another accessible, related quantity is observed.
- Entropy is a measure of information content in a variable with a given probability law.
- The maximum entropy principle can be used to assign a probability law to a quantity when the available information about it is in the form of a limited number of constraints on that probability law.
- Relative entropy and Kullback–Leibler divergence are tools for updating probability laws in the same context.
- When a parametric probability law is assigned to a quantity and we want to measure the amount of information gain about the parameters when some direct observations of that quantity is available, we can use the Fisher information. The structure of the Fisher information geometry in the space of parameters is derived from the relative entropy by a second order Taylor series approximation.
- All of these rules and tools are used currently in different ways in data and signal processing. In this paper, a few examples of the ways these tools are used in data and signal processing problems are presented. One main conclusion is that each of these tools has to be used in appropriate contexts. The example in spectral estimation shows that it is very important to define the problems very clearly at the beginning and to use appropriate tools and interpret the results appropriately.
- The Laplacian or Bayesian inference is the appropriate tool for proposing satisfactory solutions to inverse problems. Indeed, the expression of the posterior probability law represents the combination of the state of the knowledge in the forward model and the data and the state of the knowledge before using the data.
- The Bayesian approach can also easily be used to propose unsupervised methods for the practical application of these methods.

- One of the main limitation of those sophisticated methods is the computational cost. For this, we proposed to use VBA as an alternative to MCMC methods to propose realistic algorithms in huge dimensional inverse problems where we want to estimate an unknown signal (1D), image (2D), volume (3D) or even more (3D + time or 3D + wavelength), *etc*.

Acknowledgments

The author would like to thank the reviewers who, by their true review work and their extensive comments and remarks, helped to improve this review paper greatly.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Mohammad-Djafari, A. Bayesian or Laplacian inference, entropy and information theory and information geometry in data and signal processing. *AIP Conf. Proc.* **2014**, *1641*, 43–58.
2. Bayes, T. An Essay toward Solving a Problem in the Doctrine of Chances. *Philos. Trans.* **1763**, *53*, 370–418. By the late Rev. Mr. Bayes communicated by Mr. Price, in a Letter to John Canton.
3. De Laplace, P. S. Mémoire sur la probabilité des causes par les évènements. *Mémoires de l'Academie Royale des Sciences Presentés par Divers Savan* **1774**, *6*, 621–656.
4. Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
5. Hadamard, J. *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*; Mémoires présentés par divers savants à l'Académie des sciences de l'Institut de France: Imprimerie nationale, 1908.
6. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
7. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *108*, 171–190.
8. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241.
9. Kullback, S. *Information Theory and Statistics*; Wiley: New York, NY, USA, 1959.
10. Fisher, R. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Stat. Soc. A* **1922**, *222*, 309–368.
11. Rao, C. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Culcutta Math. Soc.* **1945**, *37*, 81–91.
12. Sindhwani, V.; Belkin, M.; Niyogi, P. The Geometric basis for Semi-supervised Learning. In *Semi-supervised Learning*; Chapelle, O., Schölkopf, B., Zien, A., Eds.; MIT press: Cambridge, MA, USA, 2006; pp. 209–226.
13. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
14. Johnson, O.; Barron, A.R. Fisher Information Inequalities and the Central Limit Theorem. *Probab. Theory Relat. Fields* **2004**, *129*, 391–409.
15. Berger, J. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer-Verlag: New York, NY, USA, 1985.

16. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed.; Chapman & Hall/CRC Texts in Statistical Science; Chapman and Hall/CRC: Boca Raton, FL, USA, 2003.
17. Skilling, J. Nested Sampling. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Proceedings of 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Garching, Germany, 25–30 July 2004; Fischer, R., Preuss, R., Toussaint, U.V., Eds.; pp. 395–405.
18. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
19. Hastings, W.K. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika* **1970**, *57*, 97–109.
20. Gelfand, A.E.; Smith, A.F.M. Sampling-Based Approaches to Calculating Marginal Densities. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409.
21. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. Introducing Markov Chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 1–19.
22. Gilks, W.R. Strategies for Improving MCMC. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 89–114.
23. Roberts, G.O. Markov Chain Concepts Related to Sampling Algorithms. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 45–57.
24. Tanner, M.A. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*; Springer series in Statistics; Springer: New York, NY, USA, 1996.
25. Djurić, P.M., Godsill, S.J., Eds. *Special Issue on Monte Carlo Methods for Statistical Signal Processing*; IEEE: New York, NY, USA, 2002.
26. Andrieu, C.; de Freitas, N.; Doucet, A.; Jordan, M.I. An Introduction to MCMC for Machine Learning. *Mach. Learn.* **2003**, *50*, 5–43.
27. Clausius, R. *On the Motive Power of Heat, and on the Laws Which Can be Deduced From it for the Theory of Heat*; Poggendorff's Annalen der Physick,LXXIX, Dover Reprint: New York, NY, USA, 1850; ISBN 0-486-59065-8.
28. Caticha, A. Maximum Entropy, fluctuations and priors. Presented at MaxEnt 2000, the 20th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Gif-sur-Yvette, Paris, France, 8–13 July 2000.
29. Giffin, A.; Caticha, A. Updating Probabilities with Data and Moments. In Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, The Saratoga Hotel Saratoga Springs, New York, NY, USA, 8–13 July 2007.
30. Caticha, A.; Preuss, R. Maximum Entropy and Bayesian Data Analysis: Entropic Priors Distributions. *Phys. Rev. E* **2004**, *70*, 046127.
31. Akaike, H. On Entropy Maximization Principle. In *Applications of Statistics*; Krishnaiah, P.R., Ed.; North-Holland: Amsterdam, The Netherlands, 1977; pp. 27–41.

32. Agmon, N.; Alhassid, Y.; Levine, D. An Algorithm for Finding the Distribution of Maximal Entropy. *J. Comput. Phys.* **1979**, *30*, 250–258.
33. Jaynes, E.T. Where do we go from here? In *Maximum-Entropy and Bayesian Methods in Inverse Problems*; Smith, C.R., Grandy, W.T., Jr., Eds.; Springer: Dordrecht, The Netherlands, 1985; pp. 21–58.
34. Borwein, J.M.; Lewis, A.S. Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.* **1991**, *29*, 325–338.
35. Elfwing, T. On some Methods for Entropy Maximization and Matrix Scaling. *Linear Algebra Appl.* **1980**, *34*, 321–339.
36. Eriksson, J. A note on Solution of Large Sparse Maximum Entropy Problems with Linear Equality Constraints. *Math. Program.* **1980**, *18*, 146–154.
37. Erlander, S. Entropy in linear programs. *Math. Program.* **1981**, *21*, 137–151.
38. Jaynes, E.T. On the Rationale of Maximum-Entropy Methods. *Proc. IEEE* **1982**, *70*, 939–952.
39. Shore, J.E.; Johnson, R.W. Properties of Cross-Entropy Minimization. *IEEE Trans. Inf. Theory* **1981**, *27*, 472–482.
40. Mohammad-Djafari, A. Maximum d'entropie et problèmes inverses en imagerie. *Traitement Signal* **1994**, *11*, 87–116.
41. Bercher, J. Développement de critères de nature entropique pour la résolution des problèmes inverses linéaires. Ph.D. Thesis, Université de Paris–Sud, Orsay, France, 1995.
42. Le Besnerais, G. Méthode du maximum d'entropie sur la moyenne, critère de reconstruction d'image et synthèse d'ouverture en radio astronomie. Ph.D. Thesis, Université de Paris-Sud, Orsay, France, 1993.
43. Caticha, A.; Giffin, A. Updating Probabilities. Presented at MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods, Paris, France, 8–13 July 2006; doi:10.1063/1.2423258.
44. Caticha, A. Entropic Inference. Presented at MaxEnt 2010, the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Chamonix, France, 4–9 July 2010.
45. Costa, S.I.R.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. **2012**, arXiv:1210.2354.
46. Rissanen, J. Fisher Information and Stochastic Complexity. *IEEE Trans. Inf. Theory* **1996**, *42*, 40–47.
47. Shimizu, R. On Fisher's amount of information for location family. In *A Modern Course on Statistical Distributions in Scientific Work*; D. Reidel: Dordrecht, The Netherlands, 1975; Volume 3, pp. 305–312.
48. Nielsen, F.; Nock, R. Sided and Symmetrized Bregman Centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2048–2059.
49. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
50. Schroeder, M.R. Linear prediction, entropy and signal analysis. *IEEE ASSP Mag.* **1984**, *1*, 3–11.

51. Itakura, F.; Saito, S. A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies. *Electron. Commun. Jpn.* **1970**, *53-A*, 36–43.
52. Kitagawa, G.; Gersch, W. *Smoothness Priors Analysis of Time Series*; Lecture Notes in Statistics, Volume 116; Springer: New York, NY, USA, 1996.
53. Rue, H.; Held, L. *Gaussian Markov Random Fields: Theory and Applications*; CRC Press: New York, NY, USA, 2005.
54. Amari, S.; Cichocki, A.; Yang, H.H. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, Proceedings of the Conference on Neural Information Processing Systems 1995 (NIPS 1995), Denver, CO, USA, 27–30 November 1995; pp. 757–763.
55. Amari, S. Neural learning in structured parameter spaces—Natural Riemannian gradient. In *Advances in Neural Information Processing Systems 9*, Proceedings of the Conference on Neural Information Processing Systems 1995 (NIPS 1996), Denver, CO, USA, 2–5 December 1996; pp. 127–133.
56. Amari, S. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
57. Knuth, K.H. Bayesian source separation and localization. *SPIE Proc.* **1998**, *3459*, doi:10.1117/12.323794.
58. Knuth, K.H. A Bayesian approach to source separation. In Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA'99), Aussios, France, 11–15 January 1999; Cardoso, J.-F., Loubaton, P., Eds.; pp. 283–288.
59. Attias, H. Independent Factor Analysis. *Neural Comput.* **1999**, *11*, 803–851.
60. Mohammad-Djafari, A. A Bayesian approach to source separation. Presented at MaxEnt 99, the 19th International Workshop on Bayesian Inference and Maximum Entropy Methods, Boise State University, Boise, ID, USA, 2–6 August 1999; pp. 221–244.
61. Choudrey, R.A.; Roberts, S. Variational Bayesian Mixture of Independent Component Analysers for Finding Self-Similar Areas in Images. In Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, 1–4 April 2003; pp. 107–112.
62. Lopes, H.F.; West, M. Bayesian Model Assessment in Factor Analysis. *Statsinica* **2004**, *14*, 41–67.
63. Ichir, M.; Mohammad-Djafari, A.; Bayesian Blind Source Separation of Positive Non Stationary Sources. In Proceedings of MaxEnt 2004, 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Max-Planck Institiute, Garching, Germany, 25–30 July 2004; pp. 493–500
64. Mohammad-Djafari, A. Bayesian Source Separation: Beyond PCA and ICA. In Proceedings of 14th European Symposium on Artificial Neural Networks (ESANN 2006), Bruges, Belgium, 26–28 April 2006.
65. Comon, P., Jutten, C., Eds. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press: Burlington, MA, USA, 2010.
66. Yuan, M.; Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* **2007**, *94*, 19–35.

67. Fitzgerald, W. Markov Chain Monte Carlo methods with Applications to Signal Processing. *Signal Process.* **2001**, *81*, 3–18.
68. Matsuoka, T.; Ulrych, T. Information theory measures with application to model identification. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 511–517.
69. Bretthorst, G.L. Bayesian Model Selection: Examples Relevant to NMR. In *Maximum Entropy and Bayesian Methods*; Springer: Dordrecht, The Netherlands, 1989; pp. 377–388.
70. Gelfand, A.E.; Dey, D.K. Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 501–514.
71. Mohammad-Djafari, A. Model selection for inverse problems: Best choice of basis function and model order selection. Presented at MaxEnt 1999, the 19th International Workshop on Bayesian Inference and Maximum Entropy Methods, Boise, Idaho, USA, 2–6 August 1999.
72. Clyde, M.A.; Berger, J.O.; Bullard, F.; Ford, E.B.; Jefferys, W.H.; Luo, R.; Paulo, R.; Loredo, T. Current Challenges in Bayesian Model Choice. In *Statistical Challenges in Modern Astronomy IV*, Proceedings of Conference on Statistical Challenges in Modern Astronomy, Penn State University, PA, USA, 12–15 June 2006; Babu, G.J., Feigelson, E.D., Eds.; Volume 71, pp. 224–240.
73. Wyse, J.; Friel, N. Block clustering with collapsed latent block models. *Stat. Comput.* **2012**, *22*, 415–428.
74. Giovannelli, J.F.; Giremus, A. Bayesian noise model selection and system identification based on approximation of the evidence. In Proceedings of 2014 IEEE Statistical Signal Processing Workshop (SSP), Jupiters TBD, Gold Coast, Australia, 29 June–2 July 2014; pp. 125–128.
75. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Control* **1974**, *AC-19*, 716–723.
76. Akaike, H. Power spectrum estimation through autoregressive model fitting. *Ann. Inst. Stat. Math.* **1969**, *21*, 407–419.
77. Farrier, D. Jaynes' principle and maximum entropy spectral estimation. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1176–1183.
78. Wax, M. Detection and Estimation of Superimposed Signals. Ph.D. Thesis, Standford University, CA, USA, March, 1985.
79. Burg, J.P. Maximum Entropy Spectral Analysis. In Proceedings of the 37th Annual International Meeting of Society of Exploration Geophysicists, Oklahoma City, OK, USA, 31 October 1967.
80. McClellan, J.H. Multidimensional spectral estimation. *Proc. IEEE* **1982**, *70*, 1029–1039.
81. Lang, S.; McClellan, J.H. Multidimensional MEM spectral estimation. *IEEE Trans. Acoust. Speech Signal Process.* **1982**, *30*, 880–887.
82. Johnson, R.; Shore, J. Which is Better Entropy Expression for Speech Processing:-SlogS or logS? *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *ASSP-32*, 129–137.
83. Wester, R.; Tummala, M.; Therrien, C. Multidimensional Autoregressive Spectral Estimation Using Iterative Methods. In Proceedings of 1990 Conference Record Twenty-Fourth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 5–7 November 1990; Volume 1, doi:10.1109/ACSSC.1990.523376.

84. Picinbono, B.; Barret, M. Nouvelle présentation de la méthode du maximum d'entropie. *Traitement Signal* **1990**, *7*, 153–158.
85. Borwein, J.M.; Lewis, A.S. Convergence of best entropy estimates. *SIAM J. Optim.* **1991**, *1*, 191–205.
86. Mohammad-Djafari, A., Ed. *Inverse Problems in Vision and 3D Tomography*; digital signal and image processing series; ISTE: London, UK and Wiley: Hoboken, NJ, USA, 2010.
87. Mohammad-Djafari, A.; Demoment, G. Tomographie de diffraction and synthèse de Fourier à maximum d'entropie. *Rev. Phys. Appl. (Paris)* **1987**, *22*, 153–167.
88. Féron, O.; Chama, Z.; Mohammad-Djafari, A. Reconstruction of piecewise homogeneous images from partial knowledge of their Fourier transform. In Proceedings of MaxEnt 2004, 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Max-Planck Institiute, Garching, Germany, 25–30 July 2004; pp.68–75.
89. Ayasso, H.; Mohammad-Djafari, A. Joint NDT Image Restoration and Segmentation Using Gauss–Markov–Potts Prior Models and Variational Bayesian Computation. *IEEE Trans. Image Process.* **2010**, *19*, 2265–2277.
90. Ayasso, H.; DuchÃłne, B.; Mohammad-Djafari, A. Bayesian inversion for optical diffraction tomography. *J. Mod. Opt.* **2010**, *57*, 765–776.
91. Burch, S.; Gull, S.F.; Skilling, J. Image Restoration by a Powerful Maximum Entropy Method. *Comput. Vis. Graph. Image Process.* **1983**, *23*, 113–128.
92. Gull, S.F.; Skilling, J. Maximum entropy method in image processing. *IEE Proc. F* **1984**, *131*, 646–659.
93. Gull, S.F. Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Springer: Dordrecht, The Netherlands, 1989; pp. 53–71.
94. Jones, L.K.; Byrne, C.L. General entropy criteria for inverse problems with application to data compression, pattern classification and cluster analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 23–30.
95. Macaulay, V.A.; Buck, B. Linear inversion by the method of maximum entropy. *Inverse Probl.* **1989**, *5*, doi:10.1088/0266-5611/5/5/013.
96. Rue, H.; Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J. Stat. Plan. Inference* **2007**, *137*, 3177–3192.
97. Wilkinson, R. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. **2009**, arXiv:0811.3355.
98. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 319–392.
99. Fearnhead, P.; Prangle, D. Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC. **2011**, arxiv:1004.1112v2.
100. Turner, B.M.; van Zandt, T. A tutorial on approximate Bayesian computation. *J. Math. Psych.* **2012**, *56*, 69–85.
101. MacKay, D.J.C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Comput.* **1992**, *4*, 448–472.

102. Mohammad-Djafari, A. Variational Bayesian Approximation for Linear Inverse Problems with a hierarchical prior models. In *Geometric Science of Information*, Proceedings of First International Conference on Geometric Science of Information (GSI 2013), Paris, France, 28–30 August 2013; Lecture Notes in Computer Science, Volume 8085; pp. 669–676.
103. Likas, C.L.; Galatsanos, N.P. A Variational Approach For Bayesian Blind Image Deconvolution. *IEEE Trans. Signal Process.* **2004**, *52*, 2222–2233.
104. Beal, M.; Ghahramani, Z. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Stat.* **2006**, *1*, 793–832.
105. Kim, H.; Ghahramani, Z. Bayesian Gaussian Process Classification with the EM-EP Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1948–1959.
106. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **2006**, *37*, 183–233.
107. Forbes, F.; Fort, G. Combining Monte Carlo and Mean-Field-Like Methods for Inference in Hidden Markov Random Fields. *IEEE Trans. Image Process.* **2007**, *16*, 824–837.
108. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. (B)* **1977**, *39*, 1–38.
109. Miller, M.I.; Snyder, D.L. The Role of Likelihood and Entropy in Incomplete-Data Problems: Applications to Estimating Point-Process Intensities and Toeplitz Constrained Covariances. *Proc. IEEE* **1987**, *75*, 892–907.
110. Snoussi, H.; Mohammad-Djafari, A. Information geometry of Prior Selection. In *Bayesian Inference and Maximum Entropy Methods*, Proceedings of 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, University of Idaho, Moscow, Idaho, ID, USA, 3–7 August 2002; Williams, C., Ed.; AIP Conference Proceedings 570.
111. Mohammad-Djafari, A. Approche variationnelle pour le calcul bayésien dans les problémes inverses en imagerie. **2009**, arXiv:0904.4148.
112. Beal, M. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London, UK, 2003.
113. Winn, J.; Bishop, C.M.; Jaakkola, T. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
114. Chatzis, S.; Varvarigou, T. Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-DistributionsClassification of binary random Patterns. *IEEE Trans. Fuzzy Syst.* **2009**, *17*, 505–517.
115. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686.
116. Mohammad-Djafari, A. A variational Bayesian algorithm for inverse problem of computed tomography. In *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*; Censor, Y., Jiang, M., Louis, A.K., Eds.; Publications of the Scuola Normale Superiore/CRM Series; Edizioni della Normale: Rome, Italy, 2008; pp. 231–252.
117. Mohammad-Djafari, A.; Ayasso, H. Variational Bayes and mean field approximations for Markov field unsupervised estimation. In Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Grenoble, France, 2–4 September 2009; pp. 1–6.

118. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
119. He, L.; Chen, H.; Carin, L. Tree-Structured Compressive Sensing With Variational Bayesian Analysis. *IEEE Signal Process. Lett.* **2010**, *17*, 233–236.
120. Fraysse, A.; Rodet, T. A gradient-like variational Bayesian algorithm. In Proceedings of 2011 IEEE Conference on Statistical Signal Processing Workshop (SSP), Nice, France, 28–30 June 2011; pp. 605–608.
121. Johnson, V.E. On Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings. *Bayesian Anal.* **2013**, *8*, 741–758.
122. Dumitru, M.; Mohammad-Djafari, A. Estimating the periodic components of a biomedical signal through inverse problem modeling and Bayesian inference with sparsity enforcing prior. *AIP Conf. Proc.* **2015**, *1641*, 548–555.
123. Wang, L.; Gac, N.; Mohammad-Djafari, A. Bayesian 3D X-ray computed tomography image reconstruction with a scaled Gaussian mixture prior model. *AIP Conf. Proc.* **2015**, *1641*, 556–563.
124. Mohammad-Djafari, A. Bayesian Blind Deconvolution of Images Comparing JMAP, EM and VBA with a Student-t a priori Model. In Proceedings of International Workshops on Electrical and Computer Engineering Subfields, Koc University, Istanbul, Turkey, 22–23 August 2014; pp. 98–103.
125. Su, F.; Mohammad-Djafari, A. An Hierarchical Markov Random Field Model for Bayesian Blind Image Separation. In Proceedings of International Congress on Image and Signal Processing (CISP2008), Sanya, China, 27–30 May 2008.
126. Su, F.; Cai, S.; Mohammad-Djafari, A. Bayesian blind separation of mixed text patterns. In Proceedings of IEEE International Conference on Audio, Language and Image Processing (ICALIP 2008), Shanghai, China, 7–9 July 2008; pp. 1373–1378.