Bayesian or Laplacien inference, Entropy and Information theory and Information Geometry in data and signal processing

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

Abstract. The main object of this tutorial article is first to review the main inference tools using Bayesian approach, Entropy, Information theory and their corresponding geometries. This review is focused mainly on the ways these tools have been used in data, signal and image processing. After a short introduction of the different quantities related to the Bayes rule, the entropy and the Maximum Entropy Principle (MEP), relative entropy and the Kullback-Leibler divergence, Fisher information, we will study their use in different fields of data and signal processing such as: entropy in source separation, Fisher information in model order selection, different Maximum Entropy based methods in time series spectral estimation and finally, general linear inverse problems.

Keywords: Entropy, Information theory, Information geometry, Bayesian inference, Laplacien inference, Data processing, Signal processing

INTRODUCTION

Bayesian inference is nowadays one of the dominant approaches to statistical inference. The word *Bayesian* refers to the influence of Thomas Bayes [1], who introduced what is now known as *Bayes' theorem* even if the idea has been developed prior to him by Pierre-Simon de Laplace [2] [1].

Whatever the answer to the footnote question, the main idea is that a probability law $P(X)$ assigned to a quantity X represents our state of knowledge about it [3]. Before starting new observation and gathering new data, we have an *a priori* probability law. When a new observation (data D) on X is 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 message of the Laplace or Bayes rule which can be summarized as: $P(X|D) \propto P(D|X)P(X)$. Some more details will be given in the following sections.

Shannon [4] introduced the notion of *Quantity of Information* I_n associated to 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 it: $H = -\sum_{n=1}^{N} p_n \ln p_n$. This notion of Entropy, which has no direct link with entropy in thermodynamics, became the foundation

[1] A question to the community: Shall we change Bayes to Laplace and Bayesian to Laplacien?

of the information theory in many data analysis and science of communication. More one details and extensions will be given in following sections.

Up to now, we did not yet discuss how to assign a probability law to a quantity. For the discrete values 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 $P(X = x_n) = p_n = \frac{1}{N}, \forall n = 1, \cdots, N$. But, 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 $\sum_n x_n p_n$ is d. This question is an ill-posed problem (in the mathematical sense of Hadamard) in the sense that the solution is not unique. We can propose many probability distributions which satisfies the constraint imposed by this problem. To answer to this question, Jaynes [5, 6, 7] introduced the *Principle of Maximum Entropy* 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 the optimization and expression of the solution and the algorithm to compute it will be given in the following sections.

Kullback [8] was interested in comparing two probability laws and introduced a tool to measure the quantity of information gain of a new probability law with respect to a reference one. This tool is called either the Kullback-Leibler (KL) divergence or the *relative entropy*. This tool 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 be used as an extension to MEP of Jaynes.

Fisher [9] 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 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 Geometry of information is introduced. 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 Bregmann divergence $\mathrm{B}(\theta_1|\theta_2)$ between the two parameters.

At this stage, we have almost introduced all the necessary tools that we can use in different levels of data, signal and image processing. In the following, we give a little 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 see how these tools can be used in data model selection, in Independent Components Analysis (ICA) and sources separation, in spectral analysis of the signals and in inverse problems.

BAYES OR LAPLACE RULE

Let introduce the things very simply. If we have two discrete valued related variables X and Y, then from the sum and product rule, we have

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \longrightarrow P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1}$$

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.

This relation is easily extended to the continuous valued variables

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} \tag{2}$$

with

$$p(y) = \int p(y|x)\,p(x)\,\mathrm{d}x. \tag{3}$$

More simply, the Bayes' rule is often written as:

$$p(x|y) \propto p(y|x)\,p(x). \tag{4}$$

No need for more sophisticated mathematics here if we want to use this approach. 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 on X. In this case, the terms $p(y|x)$ is called 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 or Laplace rule then is a way to do state of knowledge fusion. Before doing 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, a very important preliminary step the assigning of $p(x)$ and $p(y|x)$. As noted in the introduction and as we will see later, we need other tools for this step. Another important step is after: how to use $p(x|y)$ to summarize it?. For example, compute the Maximum A Posteriori (MAP) solution, the Expected A Posteriori (EAP) solution, the domains of X on which the probabilities are higher than other places or any other questions such as median or quantiles. We can also just explore numerically the whole space of the distribution using the Markov Chain Monte Carlo (MCMC) or any other sampling techniques. In the scalar case (one dimension), all 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 Bayesian Variational Approximation (BVA) and algorithms to do these computations.

QUANTITY OF INFORMATION AND 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 to each of them as $I_n = \ln \frac{1}{p_n} = -\ln p_n$ and so its expected value as the Entropy:

$$\mathrm{H}[X] = -\sum_{i=1}^{N} p_i \ln p_i. \tag{5}$$

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

$$\mathrm{H}[X] = -\int p(x) \ln p(x) \, \mathrm{d}x. \tag{6}$$

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:

$$\mathrm{H}[X,Y] = -\iint p(x,y) \ln p(x,y) \, \mathrm{d}x \, \mathrm{d}y, \tag{7}$$

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

So, for any well defined probability law, we can have an expression for its entropy. $\mathrm{H}[X]$ should better be noted $\mathrm{H}[p(x)]$.

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:

$$\mathrm{D}[p_1 : p_2] = -\int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{8}$$

  and

- Kullback-Leibler Divergence of p_1 with respec to to p_2 :

$$\mathrm{K}[p_1 : p_2] = -\mathrm{D}[p_1 : p_2] = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{9}$$

We may note that:

- $\mathrm{D}[p_1 : p_2]$ is invariant with respect to scale change, but is not symmetric.
- A symmetric quantity can be defined as:

$$\mathrm{J}[p_1, p_2] = \mathrm{D}[p_1 : p_2] + \mathrm{D}[p_2 : p_1]. \tag{10}$$

MUTUAL INFORMATION

The notion of Mutual Information is to compare two related variables Y and X which is defined as:

$$\mathrm{I}[Y,X] = \mathrm{H}[X] - \mathrm{H}[X|Y] = \mathrm{H}[Y] - \mathrm{H}[Y|X] \qquad (11)$$

or equivalently as:

$$\mathrm{I}[Y,X] = \mathrm{D}[p(X,Y) : p(X)p(Y)]. \qquad (12)$$

With this definition, we have the following properties:

$$\mathrm{H}[X,Y] = \mathrm{H}[X] + \mathrm{H}[Y|X] = \mathrm{H}[Y] + \mathrm{H}[X|Y] = \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{I}[Y,X] \qquad (13)$$

and

$$\mathrm{I}[Y,X] = \mathrm{E}_X\{\mathrm{D}[p(Y|X) : p(Y)]\} = \mathrm{E}_Y\{\mathrm{D}[p(X|Y) : p(X)]\}. \qquad (14)$$

We may also remark the following property:

- $\mathrm{I}[Y,X]$ 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.
- $\mathrm{I}[Y,X] \geq 0$ with equality only if X and Y are independent.

MAXIMUM ENTROPY PRINCIPLE (MEP)

One step before applying the probability rules is to assign a probability law to a quantity. MEP can be used as a natural tool to do this when the available information on that quantity is the form of:

$$\mathrm{E}\{\phi_k(X)\} = d_k, \quad k = 1,\ldots,K. \qquad (15)$$

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

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

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

$$p_{ME}(x) = \underset{p \in \mathscr{P}}{\arg\max} \left\{ \mathrm{H}[p] = -\int p(x) \ln p(x) \,\mathrm{d}x \right\} \qquad (16)$$

whose solution is given by:

$$p_{ME}(x) = \frac{1}{Z(\lambda)} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x) \right] \qquad (17)$$

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

$$-\frac{\partial \ln Z(\lambda)}{\partial \lambda_k} = d_k, \quad k = 1, \ldots, K \tag{18}$$

which can also be written as: $-\nabla_{\lambda} \ln Z(\lambda) = d$.

The maximum value of entropy reached is given by:

$$H_{\max} = \ln Z(\lambda) + \lambda^t d. \tag{19}$$

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 [16, 8, 17, 18] and [19, 20, 21, 22] for more details.

LINK BETWEEN ENTROPY AND LIKELIHOOD

Consider the problem of the parameter estimation θ of a probability law $p(x|\theta)$ from an n-sample of data $x = \{x_1, \cdots, x_n\}$. The log-likelihood of θ is defined as

$$L(\theta) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta). \tag{20}$$

Maximizing $L(\theta)$ with respect to θ gives what is called *Maximum Likelihood (ML)* estimate of θ.

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

$$\bar{L}(\theta) = \lim_{n \mapsto \infty} \frac{1}{n} L(\theta) = \mathrm{E}\{\ln p(x|\theta)\} = \int p(x|\theta^*) \ln p(x|\theta)\,\mathrm{d}x, \tag{21}$$

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

$$\mathrm{D}[p(x|\theta^*) : p(x|\theta)] = -\int p(x|\theta^*) \ln \frac{p(x|\theta)}{p(x|\theta^*)}\,\mathrm{d}x = \int p(x|\theta^*) \ln p(x|\theta^*)\,\mathrm{d}x + \bar{L}(\theta). \tag{22}$$

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

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

In this way, there is a link between the Maximum Likelihood and Maximum Relative Entropy solutions.

FISHER INFORMATION

Fisher [9] wanted to measure the amount of information that samples $x = \{x_1, \cdots, x_N\}$ of a variable X carry about an unknown parameter θ upon which its probability law $p(x|\theta)$ depends. For a given sample of observation x, the function $\mathscr{L}(\theta) = p(x|\theta)$ 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:

$$S(x|\theta) = \frac{\partial \ln p(x|\theta)}{\partial \theta} \tag{23}$$

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

$$\mathrm{E}\{S(X|\theta)\} = \mathrm{E}\left\{\frac{\partial \ln p(x|\theta)}{\partial \theta}\right\} = 0 \tag{24}$$

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.

$$\mathrm{E}\left\{S^2(X|\theta)\right\} = \mathrm{E}\left\{\left|\frac{\partial \ln p(x|\theta)}{\partial \theta}\right|^2\right\} = \mathrm{E}\left\{\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2}\right\} = F \tag{25}$$

This quantity is called the Fisher information [10, 11, 12].

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 geometry of information is introduced.

Consider $\mathrm{D}\left[p(x|\theta^*) : p(x|\theta^* + \Delta\theta)\right]$ and assume that $\ln p(x|\theta)$ can be developed in Taylor series. Then, keeping the terms up to the second order, we obtain:

$$\mathrm{D}\left[p(x|\theta^*) : p(x|\theta^* + \Delta\theta)\right] \simeq \frac{1}{2}\Delta\theta^t F(\theta^*)\Delta\theta. \tag{26}$$

where F is the Fisher information:

$$F(\theta^*) = \mathrm{E}\left\{\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^t \partial \theta}|_{\theta=\theta^*}\right\}. \tag{27}$$

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 $\mathrm{KL}(p_1|p_2)$ induces a Bregmann divergence $\mathrm{B}(\theta_1|\theta_2)$ between the two parameters.

VECTORIAL VARIABLES AND TIME INDEXED PROCESS

The extension of the scalar variable to finite dimensional vectorial case is almost immediate. In particular for the Gaussian case, we need to replace the variances by a covariance matrix and almost all the quantities can be defined immediately. For example, for a Gaussian vector $p(x) = \mathcal{N}(x|0,R)$, the entropy is given by:

$$H = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(|\det(R)|) \tag{28}$$

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

$$D = -\frac{1}{2}\left(\operatorname{tr}\left(RS^{-1}\right) - \log\frac{|\det(R)|}{|\det(S)|} - n\right). \tag{29}$$

The notion of time series or processes need extra definitions. For example, for a random time series $X(t)$, we can define $p(X(t)), \forall t$ and the expected time series $\bar{x}(t) = \mathrm{E}\{X(t)\}$. For a stationary time series (when $p(X(t))$ does not depend on t), we can define the correlation function $\Gamma(\tau) = \mathrm{E}\{X(t)X(t+\tau)\}$ and the spectral density as the Fourier Transform (FT) of it:

$$S(\omega) = \int \Gamma(\tau)\exp[-j\omega\tau]\,\mathrm{d}\tau. \tag{30}$$

With these definitions, it is easy to show that the covariance matrix of a stationary Gaussian process is Toeplitz and we have:

$$\lim_{n\longrightarrow\infty}\frac{1}{n}H(p) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln S(\omega)\,\mathrm{d}\omega \tag{31}$$

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

$$\lim_{n\longrightarrow\infty}\frac{1}{n}D(p_1:p_2) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{S_1(\omega)}{S_2(\omega)} - \ln\frac{S_1(\omega)}{S_2(\omega)} - 1\right)\mathrm{d}\omega \tag{32}$$

where we find the Itakura-Saito distance in Spectral analysis literature [13, 14, 15].

ENTROPY IN INDEPENDENT COMPONENT ANALYSIS AND SOURCES SEPARATION

Given a vector of time series $x(t)$ the Independent Component Analysis (ICA) consists in finding a Separating matrix B such that the components $y(t) = Bx(t)$ be as independent as possible. The notion of entropy is used here as a measure of independence.

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

To see how the entropy is used here, let note $y = Bx$ or more generally $y_i = g([Bx]_i)$ where g can be any increasing monotonic function. Then,

$$p_Y(y) = \frac{1}{|\partial y/\partial x|}p_X(x) \longrightarrow H(y) = -\mathrm{E}\{\ln p_Y(y)\} = \mathrm{E}\{\ln|\partial y/\partial x|\} - H(x). \tag{33}$$

$H(y)$ is then used as a criterion for ICA or sources separation.

ENTROPY IN PARAMETRIC MODELING AND MODEL SELECTION

Determining the order of a model, i.e. the dimension of the vector parameter θ in a probabilistic model $p(x|\theta)$ in many data and signal processing is an important subject. When the order is fixed, the estimation of the parameters is a very well known problem and there are Likelihood based or Bayesian approaches for that. The determination of the order is however more difficult. Between the tools, we may mention the use of relative entropy $\mathrm{D}\left[p(x|\theta^*) : p(x|\theta)\right]$, where θ^* represents the vector of the parameters of dimension k^* et θ and the vector θ with dimension $k \leq k^*$. The famous criterion of Akaike [23, 24, 25, 26, 27] uses this quantity to determine the optimal order where for linear models with Gaussian models and likelihood based methods, there is analytic solutions for it [28].

ENTROPY IN SPECTRAL ANALYSIS

Entropy and MEP have been used in different ways in spectral analysis problem which has been a great subject of signal processing. Here, we are presenting in a synthetic way, these different approaches.

Burg's method

The first and classical one is the Burg method[29] 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

$$r(k) = \mathrm{E}\left\{X(n)X(n+k)\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)\exp\left[\mathrm{j}k\omega\right]\,\mathrm{d}\omega, \quad k = 0,\ldots,K. \tag{34}$$

The question is then to estimate its 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 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 $p(x)$ to the vector $\underline{X} = [X(0),\ldots,X(N-1)]^t$. For this, we can use PME with the data or constraints (34). The answer is a Gaussian probability law: $\mathcal{N}(x|0,R)$. For a stationary Gaussian process, when the number of samples $N \longrightarrow \infty$, the expression of the entropy becomes:

$$H = \int_{-\pi}^{\pi} \ln S(\omega)\,\mathrm{d}\omega. \tag{35}$$

Now, Burg method consisted in maximizing H subject to the constraints (34). The solution is:

$$S(\omega)=\frac{1}{\left|\sum_{k=-K}^{K}\lambda_k \exp\left[\mathrm{j}k\omega\right]\right|^2}, \tag{36}$$

where $\lambda=[\lambda_0,\cdots,\lambda_K]^t$, the Lagrange multipliers associated to the constraints (34), 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 λ, which gives the possibility to give an analytical expression for $S(\omega)$ as a function of the data $\{r(k),k=0,\cdots,K\}$:

$$S(\omega)=\frac{\delta\Gamma^{-1}\delta}{e\Gamma^{-1}e}, \tag{37}$$

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

We may note that, first we used MEP to choose a probability law for $X(n)$. With the prior knowledge that we have second order moments, the MEP results to 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, a ME solution could be computed easily. This method is called Burg's maximum entropy method [**?**].

Extension to Burg's method

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

But choosing a Gaussian law for $p_0(x)$, the expression to maximize becomes:

$$\mathrm{D}\left[p(x):p_0(x)\right]=\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{S(\omega)}{S_0(\omega)}-\ln\frac{S(\omega)}{S_0(\omega)}-1\right)\mathrm{d}\omega \tag{38}$$

when $N\mapsto\infty$ and where $S_0(\omega)$ corresponds to the power spectral density of the reference process $p_0(x)$.

Shore and Johnson approach

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

$$-\int_{-\pi}^{\pi}S(\omega)\ln S(\omega)\,\mathrm{d}\omega \tag{39}$$

subject to the linear constraints (34). The solution is in the form of:

$$S(\omega)=\exp\left[\sum_{k=-K}^{K}\lambda_k\exp\left[\mathrm{j}k\omega\right]\right]. \tag{40}$$

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

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 for assigning $p(z)$ which maximizes the relative entropy $D(p(z);\mu(z))$ subject to the constraints (34).

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

$$S(\omega)=\mathrm{E}\{Z(\omega)\}=\int Z(\omega)p(z)\,\mathrm{d}z. \tag{41}$$

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

$$H=\int_{-\pi}^{\pi}S^2(\omega)\,\mathrm{d}\omega. \tag{42}$$

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

$$H=-\int_{-\pi}^{\pi}S(\omega)\ln S(\omega)\,\mathrm{d}\omega, \tag{43}$$

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

$$H=\int_{-\pi}^{\pi}\ln S(\omega)\,\mathrm{d}\omega. \tag{44}$$

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 [29, 30, 31, 32, 33, 38, 22].

ENTROPY BASED METHODS FOR LINEAR INVERSE PROBLEMS

Let consider the discretized linear inverse problem

$$y=Ax, \tag{45}$$

where A 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 that the under determined case $M < N$ that we consider here. In this case, evidently the problem has infinite number of solutions and we need to choose one.

Between the numerous methods, we may mention the minimum norm solution:

$$\widehat{x}_{NM} = \underset{\{x:y=Ax\}}{\arg\max} \left\{\Omega(x) = \|x\|^2\right\} = A^t(AA^t)^{-1}y. \tag{46}$$

In fact, we may choose any convex criterion $\Omega(x)$ and satisfy the uniqueness of the solution.

The second solution is then to choose

$$\Omega(x) = -\sum_j x_j \ln x_j \tag{47}$$

which can be interpreted as the entropy when $x_j > 0$ and $\sum x_j = 1$, thus assimilating x_j as a probability distribution $x_j = P(U = u_j)$. The variable U can correspond (or not) to a physical quantity. $\Omega(x)$ is the entropy associated to this variable [34, 35, 36, 37].

A second approach consists in considering $x_j = \mathrm{E}\left\{U_j\right\}$ or $x = \mathrm{E}\{U\}$. Again here,U_j or U can correspond to some physical quantities or not. In any case, we know want to assign a probability law $\widehat{p}(u)$ to it. Noting that the data $y = Ax = A\mathrm{E}\{U\} = \mathrm{E}\{AU\}$ can be considered as the constraints on it, we may need again a criterion to determine $\widehat{p}(u)$. Assuming then to have some prior $\mu(u)$, we may maximize the relative entropy as that criterion. The mathematical problem then becomes:

$$\widehat{p}(u) = \underset{\{x:y=\int Au\,p(u)\,\mathrm{d}u\}}{\arg\max} \{D[p(u) : \mu(u)]\} \tag{48}$$

The solution is :

$$\widehat{p}(u) = \frac{1}{Z(\lambda)}\mu(u)\exp\left[-\lambda^t Au\right] \tag{49}$$

and interestingly, if we focus on $\widehat{x} = \mathrm{E}\{U\}$, we will see that its expression depend greatly on the choice of the prior $\mu(u)$. The following table summarizes those solutions:

$\mu(u) \propto \exp[-\frac{1}{2}\sum_j u_j^2]$	$\widehat{x} = A^t\lambda$	$\widehat{x} = A^t(AA^t)^{-1}y$
$\mu(u) \propto \exp[-\sum_j \lvert u_j\rvert]$	$\widehat{x} = 1./(A^t\lambda \pm 1)$	$A\widehat{x} = y$
$\mu(u) \propto \exp[-\sum_j u_j^{\alpha-1}\exp\left[-\beta u_j\right]], \quad u_j > 0$	$\widehat{x} = \alpha 1./(A^t\lambda + \beta 1)$	$A\widehat{x} = y$

In the general case, replacing (49) in (48) and defining $Z(\lambda) = \displaystyle\int \mu(u)\exp\left[-\lambda^t Au\right]\,\mathrm{d}u$, $G(s) = \ln\displaystyle\int \mu(u)\exp\left[-s^t u\right]\,\mathrm{d}u$ and its conjugate convex $F(x) = \sup_s\{x^t s - G(s)\}$, it can be shown easily that $\widehat{x} = \mathrm{E}\{U\}$ can be obtained either via the dual $\widehat{\lambda}$ variables $\widehat{x} = G'(A^t\widehat{\lambda})$ with $\widehat{\lambda}$ is obtained by:

$$\widehat{\lambda} = \underset{\lambda}{\arg\min}\left\{D(\lambda) = \ln Z(\lambda) + \lambda^t y\right\}, \tag{50}$$

or directly

$$\widehat{x} = \underset{\{x:Ax=y\}}{\arg\min} \; \{F(x)\}. \tag{51}$$

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

KULLBACK-LEIBLER DIVERGENCE AS A TOOL FOR APPROXIMATE BAYESIAN COMPUTATION (ABC)

In this final section, we show how the Kullback-Leibler divergence can be used in the Bayesian approach for the computational purpose when handling high dimensional inverse problems. To present is simply, let consider a linear inverse problem $g = Hf + \varepsilon$ and the Bayesian approach which consists in estimating f given the observations g via the Bayes or Laplace rule:

$$p(f|g,\theta) \propto p(g|f,\theta_1)\, p(f|\theta_2) \tag{52}$$

where $p(g|f,\theta_1)$ is the likelihood, $p(f,\theta_2)$ is the prior and and $p(f|g,\theta)$ is the posterior and where $\theta = (\theta_1,\theta_2)$ are the hyper parameters of the problem. In practical applications, they also have to be inferred and so we have:

$$p(f,\theta|g) \propto p(g|f,\theta_1)\, p(f|\theta_2), p(\theta) \tag{53}$$

Even, in the simplest cases with choosing parametric exponential families for $p(g|f,\theta_1)$ and $p(f|\theta_2)$ and conjugate priors for the hyper parameters $p(\theta)$, hadling the joint posterior $p(f,\theta|g)$ for inferring both unknown quantities f and θ is not easy or even easy very costly. We then need to do approximations. The Bayesian Variational Approximation methods consists in first approximating $p(f,\theta|g)$ by a simpler probability law $q(f,\theta)$ for example a separable one $q(f,\theta) = q_1(f)q_2(\theta)$ by choosing them in an appropriate families and then use them for doing computations. A natural criterion to choose to do this approximation is the KL divergence

$$\begin{aligned} \mathrm{KL}(q:p) &= \int\!\!\int q \ln q/p = \int\!\!\int q_1 q_2 \ln \frac{q_1 q_2}{p} \\ &= \int q_1 \ln q_1 + \int q_2 \ln q_2 - \int\!\!\int q \ln p \\ &= -H(q_1) - H(q_2) - < \ln p >_q \end{aligned} \tag{54}$$

and a simple algorithm is alternate optimization: $q_1 = \arg\min_{q_1} \{\mathrm{KL}(q_1 q_2 : p)\}$ and $q_2 = \arg\min_{q_2} \{\mathrm{KL}(q_1 q_2 : p)\}$ until the convergence. By doing so, we obtain the following iterations:

$$\left\{ \begin{aligned} q_1(f) &\propto \exp\left[\langle \ln p(g,f,\theta)\rangle_{q_2(\theta)}\right] \\ q_2(\theta) &\propto \exp\left[\langle \ln p(g,f,\theta)\rangle_{q_1(f)}\right] \end{aligned} \right. \tag{55}$$

where

$$p(g,f,\theta) = p(g|f,\theta_1)\, p(f|\theta_2), p(\theta) \tag{56}$$

The last step of simplification before obtaining a practical algorithm which can be really implemented is to choose easy to use parametric families for $q_1(f)$ and $q_2(\theta)$. For a few example, I refer the readers to some of my PhD students papers presented in this workshop.

CONCLUSIONS

A probability law is a tool for representing our state of knowledge about a quantity. 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. 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 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 show that it is very important to define the problems very clearly at the beginning and use appropriate tools and interpret the results appropriately.

ACKNOWLEDGEMENTS

I would like to thank John Skilling for reviewing this paper and his comments on it. Even if I agree with almost all of his comments, I did not want to include all. I wanted this tutorial to be a descriptive of known and established methods based on probability theory, information theory and the different expressions of entropy and the way they have been used in data and signal processing. I did not want to be too critical on each of these methods, because each method is based on some hypothesis that we have to satisfy when we use them. The main point is to use them appropriately and to interpret appropriately the results we can obtain by applying them.

REFERENCES

1. Bayes T. (1763), Price R., “An Essay Towards Solving a Problem in the Doctrine of Chances, Philosophical Transactions of the Royal Society of London, 53, 370-418.

2. Laplace P. S., “Essai Philosophique sur les Probabilités,”, English translation in Truscott, F.W. and Emory, F.L. (2007) from (1902) as "A Philosophical Essay on Probabilities". ISBN 1602063281, translated from the French 6th ed. (1840).
3. Cox, R. T. 1964. "Probability, frequency, and reasonable expectation," Am. J. Phys. 14, 1-13. `http://www.cco.caltech.edu/ jimbeck/summerlectures/references/ProbabilityFrequen`
4. Shannon C. and Weaver W., “The mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
5. Jaynes E. T., “Information theory and statistical mechanics i,” *Physical review*, vol. 106, pp. 620–630, 1957.
6. Jaynes E. T., “Information theory and statistical mechanics ii,” *Physical review*, vol. 108, pp. 171–190, 1957.
7. Jaynes E. T., “Prior probabilities,” *IEEE Transactions on Systems Science and Cybernetics*, vol. SSC-4, pp. 227–241, September 1968.
8. Kullback S., *Information Theory and Statistics*. New York: Wiley, 1959.
9. R. Fisher, "On the mathematical foundations of theoretical statistics," Philosophical Transactions of the Royal Society, A, 222: 309âĂŞ368. (1922) doi:10.1098/rsta.1922.0009
10. C.R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," Bulletin of the Calcutta Mathematical Society, 37:81-91, 1945
11. S. Amari, H. Nagaoka, "Methods of information geometry", Translations of mathematical monographs; v. 191, American Mathematical Society, 2000 (ISBN 978-0821805312)
12. B.R. Frieden, "Science from Fisher Information", Cambridge, 2004
13. Itakura and Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” *Electron. and Commun.*, vol. 53-A, pp. 36–43, 1970.
14. Knockaert L., “A class of statistical and spectral distance measures based on Bose-Einstein statistics,” *IEEE Transactions on Signal Processing*, vol. 41, no. 11, pp. 3171–3174, 1963.
15. Schroeder M., “Linear prediction, entropy and signal analysis,” *IEEE ASSP Magazine*, pp. 3–11, juillet 1984.
16. Jaynes E. T., “On the rationale of maximum-entropy methods,” *Proceedings of the IEEE*, vol. 70, pp. 939–952, September 1982.
17. Shore J. and Johnson R., “Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,” *IEEE Transactions on Information Theory*, vol. IT-26, pp. 26–37, January 1980.
18. Shore J. E. and Johnson R. W., “Properties of cross-entropy minimization,” *IEEE Transactions on Information Theory*, vol. IT-27, pp. 472–482, July 1981.
19. Mohammad-Djafari A., “Maximum d’entropie et problèmes inverses en imagerie,” *Traitement du Signal*, pp. 87–116, 1994.
20. Bercher J.-F., *Développement de critères de nature entropique pour la résolution des problèmes inverses linéaires*. Thèse de Doctorat, Université de Paris-Sud, Orsay, février 1995.
21. Le Besnerais G., *Méthode du maximum d’entropie sur la moyenne, critères de reconstruction d’image et synthèse d’ouverture en radio-astronomie*. Thèse de Doctorat, Université de Paris-Sud, Orsay, décembre 1993.
22. Borwein J. and Lewis A., “Duality relationships for entropy-like minimization problems,” *SIAM Journal of Control*, vol. 29, pp. 325–338, March 1991.
23. Akaike H., “Power spectrum estimation through autoregressive model fitting,” *Annals of Institute of Statistical Mathematics*, vol. 21, pp. 407–419, 1969. JFG.
24. Akaike H., “A new look at the statistical model identification,” *IEEE Transactions on Automatic and Control*, vol. AC-19, pp. 716–723, December 1974. JFG.
25. Farrier D., “Jaynes’ principle and maximum entropy spectral estimation,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-32, pp. 1176–1183, 1984.
26. Wax M. and Kailath T., “Detection of signals by information theoretic criteria,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 33, pp. 387–392, avril 1985.
27. Wax M., “Detection and localization of multiple sources via the stochastic signals model,” *IEEE Transactions on Signal Processing*, vol. SP-39, pp. 2450–2456, novembre 1991. FD.
28. Matsuoka T., “Information theory measures with application to model identification,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 34, pp. 511–517, juin 1986.

29. Burg J. P., “Maximum entropy spectral analysis,” in *Proc. of the 37th Meeting of the Society of Exploration Geophysicists*, (Oklahoma City), pp. 34–41, October 1967.
30. Shore J., “Minimum cross-entropy spectral analysis,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-29, pp. 230–237, April 1981.
31. Johnson R. and Shore J., “Minimum cross-entropy spectral analysis,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-29, pp. 230–237, avril 1981.
32. McClellan J., “Multidimensional spectral estimation,” *Proceedings of the IEEE*, vol. 70, pp. 1029–1039, septembre 1982.
33. Johnson R. and Shore J., “Which is better entropy expression for speech processing : -slogs or logs?,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-32, pp. 129–137, 1984.
34. S. Burch, S. Gull, and J. Skilling, *Comput. Vis. Graph. Im. Process.* **23**, 113–128 (1983).
35. S. Gull, and J. Skilling, *IEE Proceedings* **131, Pt. F** (1984).
36. J. Skilling, and S. Gull, *IEE Proceedings* **131**, 646–659 (1984).
37. J. Skilling, and S. Gull, *SIAM-AMS Proceedings* **14** (1984).
38. Picinbono B. and Barret M., “Nouvelle présentation de la méthode du maximum d’entropie,” *Traitement du Signal*, vol. 7, no. 2, pp. 153–158, 1990.
39. Bregman, L. M. (1967). "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming". USSR Computational Mathematics and Mathematical Physics 7 (3) 200âĂŞ217. doi:10.1016/0041-5553(67)90040-7.