Bayesian inference for inverse problems

Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes,
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette, France

Abstract. Traditionally, the MaxEnt workshops start by a tutorial day. This paper summarizes my talk during 2001'th workshop at John Hopkins University. The main idea in this talk is to show how the Bayesian inference can naturally give us all the necessary tools we need to solve real inverse problems: starting by simple inversion where we assume to know exactly the forward model and all the input model parameters up to more realistic advanced problems of myopic or blind inversion where we may be uncertain about the forward model and we may have noisy data.

Starting by an introduction to inverse problems through a few examples and explaining their ill posedness nature, I briefly presented the main classical deterministic methods such as data matching and classical regularization methods to show their limitations. I then presented the main classical probabilistic methods based on likelihood, information theory and maximum entropy and the Bayesian inference framework for such problems. I show that the Bayesian framework, not only generalizes all these methods, but also gives us natural tools, for example, for inferring the uncertainty of the computed solutions, for the estimation of the hyperparameters or for handling myopic or blind inversion problems. Finally, through a deconvolution problem example, I presented a few state of the art methods based on Bayesian inference particularly designed for some of the mass spectrometry data processing problems.

INTRODUCTION

Forward and inverse problems

In experimental science, it is hard to find an example where we can measure directly a desired quantity. Describing *mathematical models* to relate the measured quantities to the unknown quantity of interest is called *forward modeling problem*. The main object of a forward modeling is to be able to generate data which are as likely as possible to the observed data if the unknown quantity was known. But, almost always, we want to use this model and the observed data to make inference on the unknown quantity of interest: This is the *inversion problem*. To be more explicit, let take an example that we will use all along this paper to illustrate the different aspects of inverse problems. The example is taken from the mass spectrometry where the ideal physical quantity of interest is the components mass distribution of the material under the test. There are many techniques used in mass spectrometry. The Time-of-Flight (TOF) technique is one of them. In this technique, one measures the electrical current generated on the surface of a detector by the charged ions generated by the material under the test. Finding a very fine physical model to relate the time variation of this current to the distribution of the arrival times of the charged ions, which is itself related to the components mass distribution of the material under the test, is not an easy task. However, in a first approximation, assuming

that the instrument is linear and its characteristics do not change during the acquisition time of the experiment, a very simple convolution model relates the raw data $g(t)$ to the unknown quantity of interest $f(t)$:

$$g(\tau) = \int f(t)\, h(\tau - t)\, \mathrm{d}t, \tag{1}$$

where $h(t)$ is the point spread function (psf) of the instrument. Figure 1 shows an example of data observed (signal in b) for a theoretical mass distribution (signal in a).

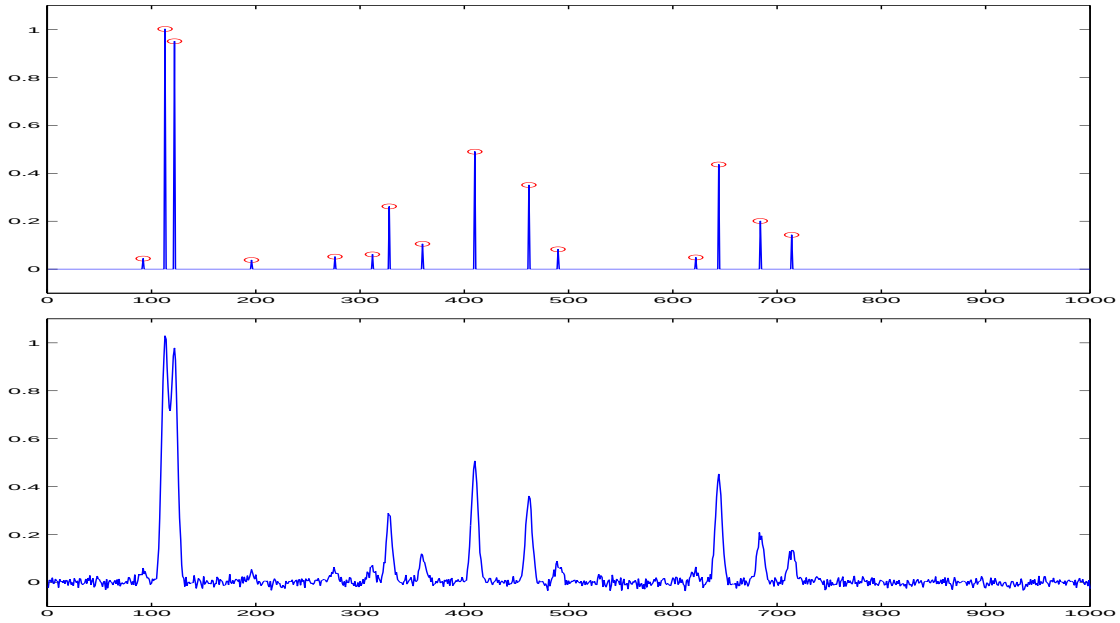

FIGURE 1. Blurring effect in TOF mass spectrometry data: a) desired or theoretical spectrum, b) observed data.

In this example, the forward problem consists in computing g given f and h which is given by a simple *convolution* operation. The inverse problem of inferring f given g and h is called *deconvolution*, the inverse problem of inferring h given g and f is called *psf identification* and the inverse problem of inferring h and f given only g is called *blind deconvolution.*

In my talk, I have given many more examples such as image restoration

$$g(x', y') = \int\int f(x, y)\, h(x' - x, y' - y)\, \mathrm{d}x\, \mathrm{d}y, \tag{2}$$

or Fourier synthesis inversion

$$g(\tau) = \int f(\omega) \exp\{-j\omega\tau\}\, \mathrm{d}\omega \tag{3}$$

as well as a few non linear inverse problems. I am not going to detail them here, but I try to give a unified method to deal with all these problems. For this purpose, first we note that, in all these problems, we have always limited the number of data, for example

$y_i = g(\tau_i),\ i = 1, \ldots, m$. We also note that, to be able to do numerical computation, we need to model the unknown function f by a finite number of parameters $\boldsymbol{x} = [x_1, \ldots, x_n]$. As an example, we may assume that

$$f(t) = \sum_{j=1}^{n} x_j b_j(t) \tag{4}$$

where $b_j(t)$ are known basis functions. With this assumption the raw data $\boldsymbol{y} = [y_1, \ldots, y_m]$ are related to the unknown parameters $\boldsymbol{x}$ by

$$y_i = g(\tau_i) = \sum_{j=1}^{n} H_{i,j}\, x_j \quad \text{with} \quad H_{i,j} = \int b_j(t) h(t - \tau_i)\, \mathrm{d}t \tag{5}$$

which can be written in the simple matrix form $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$. The inversion problem can then be simplified to the estimation of $\boldsymbol{x}$ given $\boldsymbol{H}$ and $\boldsymbol{y}$. Two approaches are then in competition:
i) the dimensional control approach which consists in an appropriate choice of the basis functions $b_j(\boldsymbol{r})$ and $n \leq m$ in such a way that the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ be well conditioned;
ii) the more general regularization approach where a classical sampling basis for $b_j(\boldsymbol{r})$ with desired resolution is chosen no matter if $n > m$ or if $\boldsymbol{A}$ is ill conditioned. In the following, we follow the second approach which is more flexible for adding more general prior information on $\boldsymbol{x}$.

We must also remark that, in general, it is very difficult to give a very fine mathematical model to take account for all the different quantities affecting the measurement process. However, we can almost always come up with a more general relation such as

$$y_i = \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \epsilon_i, \quad i = 1, \ldots, m \tag{6}$$

where $\boldsymbol{\theta}$ represents the unknown parameters of the forward model (for example the amplitude and the width of a Gaussian shape psf in a deconvolution problem) and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_m]$ represents all the errors (measurement noise, discretization errors and all the other uncertainties of the model). For the case of linear models we have

$$\boldsymbol{y} = \boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{x} + \boldsymbol{\epsilon}. \tag{7}$$

In this paper we focus on this general problem. We first consider the case where the model is assumed to be perfectly known. This is the simple *inversion problem*. Then we consider the more general case where we have also to infer on $\boldsymbol{\theta}$. This is the *myopic* or *blind inversion* problem.

Even in the simplest case of perfectly known linear system and exact data:
i) the operator $\boldsymbol{H}$ may not be invertible ($\boldsymbol{H}^{-1}$ does not exist);
ii) it may admit more than one inverse ($\exists \boldsymbol{G}_1$ and $\boldsymbol{G}_2 | \boldsymbol{G}_1(\boldsymbol{H}) = \boldsymbol{G}_2(\boldsymbol{H}) = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity operator); or
iii) it may be very ill-posed or ill-conditioned (meaning that there exists $\boldsymbol{x}$ and $\boldsymbol{x} + \alpha\delta\boldsymbol{x}$ for which $\left\| \boldsymbol{H}^{-1}(\boldsymbol{x}) - \boldsymbol{H}^{-1}(\boldsymbol{x} + \alpha\delta\boldsymbol{x}) \right\|$ never vanishes even if $\alpha \mapsto 0$ [**?** , **?**].

These are the three necessary conditions of *existence*, *uniqueness* and *stability* of Hadamard for the well-posedness of an inversion problem. This explains the fact that,

in general, even in this simple case, many naïve methods based on generalized inversion or on least squares may not give satisfactory results. The following figure shows, in a simple way, the ill-posedness of a deconvolution problem. On this figure, we see that three different input signals can result three outputs which are practically indistinguishable from each other. This means that, data matching alone can not distinguish between any of these inputs.

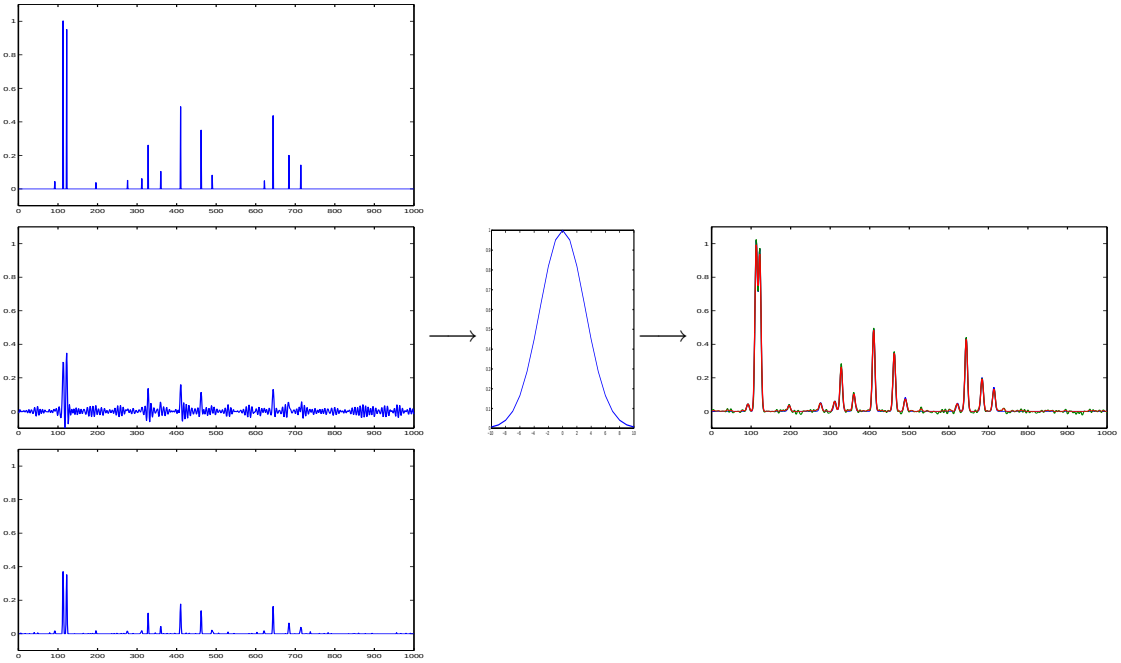

FIGURE 2. Ill-posedness of a deconvolution problem: Inputs on the left give practically indistinguishable outputs.

As a conclusion, we see that, apart from the data, we need extra information. The art of *inversion* in a particular inverse problem is how to include *just enough prior information* to obtain a satisfactory result. In the following, first we summarize the classical deterministic approaches of data matching and regularization. Then, we focus on probabilistic approaches where errors and uncertainties are taken into account through the probability laws. Here, we distinguish, three classes of methods: those which only account for the data errors (error probability distribution matching and likelihood based methods), those which only account for uncertainties of unknown parameters (entropy based methods) and those which account for both of them (Bayesian inference approach).

DATA MATCHING AND REGULARIZATION METHODS

Exact data matching

Let consider the discretized equation $y_i = h_i(\boldsymbol{x}) + \epsilon_i,\ i = 1, \ldots, m$; and assume first that the model and data are exact ($\epsilon_i = 0$). We can then write $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$.

Assume now the system of equations is under determined, *i.e.*, there is more than one solution satisfying it (for example when the number of data is less than the number of unknowns). Then, one way to obtain a unique solution is to define an *a priori* criterion, for example $\Delta(\boldsymbol{x},\boldsymbol{m})$ to choose that unique solution by

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{y}}{\arg\min} \{\Delta(\boldsymbol{x},\boldsymbol{m})\} \tag{8}$$

where $\boldsymbol{m}$ is an *a priori* solution and Δ a distance measure.

In the linear inverse problems case, the solution to this constrained optimization can be obtained via Lagrangian techniques which consists in defining the Lagrangian $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = \Delta(\boldsymbol{x},\boldsymbol{m}) + \boldsymbol{\lambda}^t(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})$ and searching for $(\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{x}})$ through

$$\left\{\begin{array}{rcl} \widehat{\boldsymbol{\lambda}} &=& \arg\min_{\boldsymbol{\lambda}} \{\mathcal{D}(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})\} \\ \widehat{\boldsymbol{x}} &=& \arg\min_{\boldsymbol{x}} \left\{\mathcal{L}(\boldsymbol{x},\widehat{\boldsymbol{\lambda}})\right\} \end{array}\right. \tag{9}$$

Noting that $\nabla_{\boldsymbol{x}}\mathcal{L} = \nabla_{\boldsymbol{x}}\Delta(\boldsymbol{x},\boldsymbol{m}) - \boldsymbol{H}^t\boldsymbol{\lambda}$ and $\nabla_{\boldsymbol{\lambda}}\mathcal{L} = \boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}$ and defining $\mathcal{G}(\boldsymbol{s},\boldsymbol{m}) = \sup_{\boldsymbol{x}}\{\boldsymbol{x}^t\boldsymbol{s} - \Delta(\boldsymbol{x},\boldsymbol{m})\}$ the algorithm to find the solution $\widehat{\boldsymbol{x}}$ becomes:
– Determine $\mathcal{G}(\boldsymbol{s},\boldsymbol{m}) = \sup_{\boldsymbol{x}}\{\boldsymbol{x}^t\boldsymbol{s} - \Delta(\boldsymbol{x},\boldsymbol{m})\}$;
– Find $\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - \mathcal{G}(\boldsymbol{H}^t\boldsymbol{\lambda},\boldsymbol{m})\right\}$;
– Determine $\widehat{\boldsymbol{x}} = \nabla_{\boldsymbol{s}}\mathcal{G}(\boldsymbol{H}^t\widehat{\boldsymbol{\lambda}})$.

As an example, when $\Delta(\boldsymbol{x},\boldsymbol{m}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{m}\|^2$ then $\mathcal{G}(\boldsymbol{s},\boldsymbol{m}) = \boldsymbol{m}^t\boldsymbol{s} + \frac{1}{2}\|\boldsymbol{s}\|^2$, $\nabla_{\boldsymbol{s}}\mathcal{G} = \boldsymbol{m}+\boldsymbol{s}$ and $\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - \boldsymbol{m}^t\boldsymbol{H}^t\boldsymbol{\lambda} + \frac{1}{2}\left\|\boldsymbol{H}^t\boldsymbol{\lambda}\right\|^2$ which results to $\widehat{\boldsymbol{\lambda}} = (\boldsymbol{H}\boldsymbol{H}^t)^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{m})$ and the solution is given by

$$\widehat{\boldsymbol{x}} = \boldsymbol{m} + \boldsymbol{H}^t(\boldsymbol{H}\boldsymbol{H}^t)^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{m}). \tag{10}$$

One can remark that, when $\boldsymbol{m} = \boldsymbol{0}$ we have $\widehat{\boldsymbol{x}} = \boldsymbol{H}^t(\boldsymbol{H}\boldsymbol{H}^t)^{-1}\boldsymbol{y}$ and this is the classical minimum norm generalized inverse solution.

Another example is the classical *Maximum Entropy* method case where $\Delta(\boldsymbol{x},\boldsymbol{m}) = \mathrm{KL}(\boldsymbol{x},\boldsymbol{m})$ is the Kullback-Leibler distance or cross entropy between $\boldsymbol{x}$ and the *a priori* solution $\boldsymbol{m}$:

$$\mathrm{KL}(\boldsymbol{x},\boldsymbol{m}) = \sum_j x_j \ln\frac{x_j}{m_j} - (x_j - m_j) \tag{11}$$

Here, the solution is given by

$$\widehat{x}_j = m_j \exp\left[-[\boldsymbol{A}^t\widehat{\boldsymbol{\lambda}}]_j\right] \text{ with } \widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\arg\min} \left\{\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - \mathcal{G}(\boldsymbol{A}^t\boldsymbol{\lambda},\boldsymbol{m})\right\} \tag{12}$$

where $\mathcal{G}(\boldsymbol{s},\boldsymbol{m}) = \sum_j m_j\left(1-\exp\left[-s_j\right]\right)$. But, unfortunately here $\mathcal{D}(\boldsymbol{\lambda})$ is not a quadratic function of $\boldsymbol{\lambda}$ and thus there is not an analytic expression for $\widehat{\boldsymbol{\lambda}}$. However, it can be computed numerically and many algorithms have been proposed for its efficient computation. See for example [**?**] and the cited references for more discussions on the computational issues and algorithm implementation.

The main issue here is that, this approach gives a satisfactory solution to the uniqueness of the inverse problem, but in general, the performances obtained by the resulting algorithms stay sensitive to error on the data.

Least squares data matching and regularization

When the discretized equation $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ is over-determined, *i.e.*, there is no solution satisfying it exactly (for example when the number of data is greater than the number of unknowns or when the data are not exact), one can try to estimate them by:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \{\Delta(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))\}, \tag{13}$$

where $\Delta(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))$ is a distance measure in the data space. The case where $\Delta(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \|\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{x})\|^2$ is the classical Least Squares (LS) criterion.

For a linear inversion problem $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$, it is easy to see that any $\widehat{\boldsymbol{x}}$ which satisfies the normal equation $\boldsymbol{H}^t\boldsymbol{H}\widehat{\boldsymbol{x}} = \boldsymbol{H}^t\boldsymbol{y}$ is a LS solution. If $\boldsymbol{H}^t\boldsymbol{H}$ is invertible and well-conditioned then $\widehat{\boldsymbol{x}} = (\boldsymbol{H}^t\boldsymbol{H})^{-1}\boldsymbol{H}^t\boldsymbol{y}$ is again the unique generalized inverse solution. But, in general, this is not the case: $\boldsymbol{H}^t\boldsymbol{H}$ is rank deficient and we need to constrain the space of the admissible solutions. The constraint LS is then defined as

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2\right\}. \tag{14}$$

where $\mathcal{C}$ is a convex set. The choice of the set $\mathcal{C}$ is primordial to satisfy the three conditions of a well-posed solution. An example is the positivity constraint: $\mathcal{C} = \{\boldsymbol{x} : \forall j,\ x_j > 0\}$. Another example is $\mathcal{C} = \{\boldsymbol{x} :\ \|\boldsymbol{x}\| < \alpha\}$ where the solution can be computed via the optimization of

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x}\|. \tag{15}$$

The main technical difficulty is the relation between α and λ. The minimum norm LS solution can also be computed using the singular value decomposition [**?**]. The main issue here is that, even if this approach has been well understood and commonly used, it assumes implicitly that the noise and the $\boldsymbol{x}$ are Gaussian. This may not be suitable in some applications, and more specifically in mass spectrometry data processing where the unknowns are spiky spectra.

A more general regularization procedure is to define the solution to the inversion problem $\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x}) + \boldsymbol{\epsilon}$ as the optimizer of a compound criterion $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \lambda\phi(\boldsymbol{x})$ or the more general criterion

$$J(\boldsymbol{x}) = \Delta_1(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x}) + \lambda\Delta_2(\boldsymbol{x}, \boldsymbol{m}). \tag{16}$$

where Δ_1 and Δ_2 are two distances or discrepancy measures, λ a regularization parameter and $\boldsymbol{m}$ an *a priori* solution[**?**]. The main questions here are: i) how to choose Δ_1 and Δ_2 and ii) how to determine λ and $\boldsymbol{m}$.

For the first question, many choices exist:
– Quadratic or L_2 distance: $\Delta(\boldsymbol{x},\boldsymbol{z}) = \|\boldsymbol{x}-\boldsymbol{z}\|^2 = \sum_j (x_j - z_j)^2$;
– L_p distance: $\Delta(\boldsymbol{x},\boldsymbol{z}) = \|\boldsymbol{x}-\boldsymbol{z}\|^p = \sum_j |x_j - z_j|^p$;
– Kullback distance: $\Delta(\boldsymbol{x},\boldsymbol{z}) = \sum_j x_j \ln(x_j/z_j) - (x_j - z_j)$;
– Roughness distance: $\Delta(\boldsymbol{x},\boldsymbol{z})$ any of the previous distances with $z_j = x_{j-1}$ or $z_j = (x_{j-1} + x_{j+1})/2$ or any linear function $z_j = \psi(x_k, k \in \mathcal{N}(j))$ where $\mathcal{N}(j)$ stands for the neighborhood of j. (One can see the link between this last case and the Gibbsian energies in the Markovian modeling of signals and images.)

The second difficulty in this deterministic approach is the determination of the regularization parameter λ. Even if there are some techniques based on cross validation [**?** , **?** , **?**], there is not natural tools for their extension to other hyperparameters in a natural way.

As a simple example, we consider the case where both Δ_1 and Δ_2 are quadratic: $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2_{\boldsymbol{W}} + \lambda \|\boldsymbol{x} - \boldsymbol{m}\|^2_{\boldsymbol{Q}}$. The optimization problem, in this case, has an analytic solution:

$$\widehat{\boldsymbol{x}} = \left(\boldsymbol{H}^t\boldsymbol{W}\boldsymbol{H} + \lambda\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{H}^t\boldsymbol{W}\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{m})\right) \tag{17}$$

which can also be written

$$\widehat{\boldsymbol{x}} = \boldsymbol{m} + \boldsymbol{Q}^{-1}\boldsymbol{H}^t\left(\boldsymbol{H}\boldsymbol{Q}^{-1}\boldsymbol{H}^t + \lambda^{-1}\boldsymbol{W}^{-1}\right)^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{m}) \tag{18}$$

which is a linear function of the *a priori* solution $\boldsymbol{m}$ and the data $\boldsymbol{y}$. Note also that when $\boldsymbol{m} = 0$, $\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{W} = \boldsymbol{I}$ we have $\widehat{\boldsymbol{x}} = \left(\boldsymbol{H}^t\boldsymbol{H} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{H}^t\boldsymbol{y}$ or $\widehat{\boldsymbol{x}} = \boldsymbol{H}^t\left(\boldsymbol{H}\boldsymbol{H}^t + \lambda^{-1}\boldsymbol{I}\right)^{-1}\boldsymbol{y}$ and when $\lambda = 0$ we obtain the generalized inverse solutions $\widehat{\boldsymbol{x}} = \left(\boldsymbol{H}^t\boldsymbol{H}\right)^{-1}\boldsymbol{H}^t\boldsymbol{y}$ or $\widehat{\boldsymbol{x}} = \boldsymbol{H}^t\left(\boldsymbol{H}\boldsymbol{H}^t\right)^{-1}\boldsymbol{y}$.

As we mentioned before, the main practical difficulties in this approach are the choice of Δ_1 and Δ_2 and the determination of the hyperparameters λ and the inverse covariance matrices $\boldsymbol{W}$ and $\boldsymbol{Q}$.

As a main conclusion on these deterministic inversion methods, we can say that, even if, in practice, they are used and give satisfaction, they lack tools to handle with uncertainties and to account for more precise *a priori* knowledge of statistical properties of errors and unknown parameters. The probabilistic methods can exactly handle more easily these problems as we will see in the following.

PROBABILISTIC METHODS

Probability distribution matching and maximum likelihood

The main idea here is to account for data and model uncertainty through the assignment of a theoretical distribution $p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$ to the data. In probability distribution matching method, the main idea is to determine the unknown parameters $\boldsymbol{x}$ by minimizing a distance measure $\Delta(\rho, p)$ between the empirical histogram ρ of the data defined

as

$$\rho(\boldsymbol{z}) \stackrel{\triangle}{=} \frac{1}{N}\sum_i \delta(z_i - y_i) \tag{19}$$

and the theoretical distribution of the data $p_{Y|X}(\boldsymbol{z}|\boldsymbol{x})$.

When $\Delta(p,\rho)$ is choosed to be the Kullback-Leibler mismatch measure

$$\begin{aligned} KL[\rho,p] &\stackrel{\triangle}{=} \int \rho(\boldsymbol{z}) \ln \frac{\rho(\boldsymbol{z})}{p_{Y|X}(\boldsymbol{z}|\boldsymbol{x})}\, \mathrm{d}\boldsymbol{z} \\ &= -\int \rho(\boldsymbol{z}) \ln p_{Y|X}(\boldsymbol{z}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{z} + \int \rho(\boldsymbol{z}) \ln \rho(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \end{aligned} \tag{20}$$

we have

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \{KL[\rho,p]\} = \arg\min_{\boldsymbol{x}} \left\{ -\int \rho(\boldsymbol{z}) \ln p_{Y|X}(\boldsymbol{z}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{z} \right\}. \tag{21}$$

It is then easy to see that, for the i.i.d. data, this estimate becomes equivalent to the maximum likelihood (ML) estimate

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\{ -\ln p_{Y|X}(\boldsymbol{z}|\boldsymbol{x})|_{\boldsymbol{z}=\boldsymbol{y}} \right\} = \arg\max_{\boldsymbol{x}} \left\{ p_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) \right\}. \tag{22}$$

In the case of a linear model and Gaussian noise, it is easy to show that the ML estimate becomes equivalent to the LS one, which in general, does not give satisfactory results as we have discussed it in the previous section.

The important point to note here is that, in this approach, only the data uncertainty is considered and modeled through the probabilty law $p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$. We will see in the following that, in contrary to this approach, in information theory and maximum entropy methods, the data and model are assumed to be exact and only the uncertainty of x is modeled through an *a priori* reference measure $\mu(\boldsymbol{x})$ which is updated to an *a posteriori* probabilty law $p(\boldsymbol{x})$ by optimizing the KL mismatch $\mathrm{KL}(p,\mu)$ subject to the data constraints.

Maximum entropy in the mean

The main idea in this approach is to consider $\boldsymbol{x}$ as the mean value of a quantity $\boldsymbol{X} \in \mathcal{C}$, where $\mathcal{C}$ is a compact set on which we want to define a probability law P: $\boldsymbol{x} = \mathrm{E}_P\{\boldsymbol{X}\}$ and the data $\boldsymbol{y}$ as exact equality constraints on it:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}\mathrm{E}_P\{\boldsymbol{X}\} = \int_{\mathcal{C}} \boldsymbol{H}\boldsymbol{x}\, \mathrm{d}P(\boldsymbol{x}). \tag{23}$$

Then, assuming that we can translate our prior information on the unknowns through a prior law (a reference measure) $\mathrm{d}\mu(\boldsymbol{x})$, we can determine the distribution P by:

$$\text{maximize} \quad -\int_{\mathcal{C}} \ln \frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}\mu(\boldsymbol{x})}\, \mathrm{d}P(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}\mathrm{E}_P\{\boldsymbol{X}\}. \tag{24}$$

The solution is obtained via the Lagrangian:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = \int_{\mathcal{C}} \left[\ln \frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}\mu(\boldsymbol{x})} - \boldsymbol{\lambda}^t(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}) \right] \mathrm{d}P(\boldsymbol{x})$$

and is given by: $\mathrm{d}P(\boldsymbol{x},\boldsymbol{\lambda}) = \exp\left[\boldsymbol{\lambda}^t[\boldsymbol{H}\boldsymbol{x}] - \ln Z(\boldsymbol{\lambda})\right] \mathrm{d}\mu(\boldsymbol{x})$, where $Z(\boldsymbol{\lambda}) = \displaystyle\int_{\mathcal{C}} \exp\left[\boldsymbol{\lambda}^t[\boldsymbol{H}\boldsymbol{x}]\right] \mathrm{d}\mu(\boldsymbol{x})$. The Lagrange parameters are obtained by searching the unique solution (if exists) of the following system of non linear equations:

$$\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_i} = y_i, \quad i = 1, \cdots, M. \tag{25}$$

Then, naturally, the solution to the inverse problem is defined as the expected value of this distribution: $\widehat{\boldsymbol{x}}(\boldsymbol{\lambda}) = \mathrm{E}_P\{\boldsymbol{X}\} = \int \boldsymbol{x}\ \mathrm{d}P(\boldsymbol{x},\boldsymbol{\lambda})$. The interesting point here is that, the solution $\widehat{\boldsymbol{x}}(\widehat{\boldsymbol{\lambda}})$ can be computed without actually computing P in two ways:
– Via optimization of a dual criterion: The solution $\widehat{\boldsymbol{x}}$ is expressed as a function of the dual variable $\widehat{\boldsymbol{s}} = \boldsymbol{H}^t\widehat{\boldsymbol{\lambda}}$ by $\widehat{\boldsymbol{x}}(\widehat{\boldsymbol{s}}) = \nabla_{\boldsymbol{s}} G(\widehat{\boldsymbol{s}},\boldsymbol{m})$ where

$$G(\boldsymbol{s},\boldsymbol{m}) = \ln Z(\boldsymbol{s},\boldsymbol{m}) = \ln \int_{\mathcal{C}} \exp\left[\boldsymbol{s}^t\boldsymbol{x}\right] \mathrm{d}\mu(\boldsymbol{x}), \quad \boldsymbol{m} = \mathrm{E}_\mu\{\boldsymbol{X}\} = \int_{\mathcal{C}} \boldsymbol{x}\ \mathrm{d}\mu(\boldsymbol{x})$$

and $\widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \left\{ D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - G(\boldsymbol{H}^t\boldsymbol{\lambda}) \right\}$.
– Via optimization of a primal or direct criterion:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \{H(\boldsymbol{x},\boldsymbol{m})\} \ \text{ s.t. } \ \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} \ \text{ where } \ H(\boldsymbol{x},\boldsymbol{m}) = \sup_{\boldsymbol{s}} \{\boldsymbol{s}^t\boldsymbol{x} - G(\boldsymbol{s},\boldsymbol{m})\}.$$

Another interesting point is the link between these two options:
i) Functions G and H depend on the reference measure $\mu(\boldsymbol{x})$;
ii) The dual criterion $D(\boldsymbol{\lambda})$ depends on the data and the function G;
iii) The primal criterion $H(\boldsymbol{x},\boldsymbol{m})$ is a distance measure between $\boldsymbol{x}$ and $\boldsymbol{m}$ which means: $H(\boldsymbol{x},\boldsymbol{m}) \geq 0$ and $H(\boldsymbol{x},\boldsymbol{m}) = 0$ iff $\boldsymbol{x} = \boldsymbol{m}$; $H(\boldsymbol{x},\boldsymbol{m})$ is differentiable and convex on $\mathcal{C}$ and $H(\boldsymbol{x},\boldsymbol{m}) = \infty$ if $\boldsymbol{x} \notin \mathcal{C}$;
iv) If the reference measure is separable: $\mu(\boldsymbol{x}) = \prod_{j=1}^N \mu_j(x_j)$ then P is too:
$\mathrm{d}P(\boldsymbol{x},\boldsymbol{\lambda}) = \prod_{j=1}^N \mathrm{d}P_j(x_j,\boldsymbol{\lambda})$ and we have

$$G(\boldsymbol{s},\boldsymbol{m}) = \sum_j g_j(s_j,m_j), \quad H(\boldsymbol{x},\boldsymbol{m}) = \sum_j h_j(x_j,m_j), \quad \widehat{x}_j = g_j'(s_j,m_j).$$

where g_j is the log Laplace transform (Cramer transform) of μ_j:

$$g_j(s) = \ln \int \exp[sx]\ \mathrm{d}\mu_j(x);$$

and h_j is the convex conjugate of g_j: $h_j(x) = \max_{\boldsymbol{s}} \{sx - g_j(s)\}$.

The following table gives three examples of choices for μ_j and the resulting expressions for g_j and h_j:

	$\mu_j(x)$	$g_j(s)$	$h_j(x,m)$
Gaussian:	$\exp\left[-(1/2)(x-m)^2\right]$	$(1/2)(s-m)^2$	$(1/2)(x-m)^2$
Poisson:	$(m^x/x!)\exp[-m]$	$\exp[m-s]$	$-x\ln(x/m)+m-x$
Gamma:	$x^{\alpha-1}\exp[-(x/m)]$	$\ln(s-m)$	$-\ln(x/m)+(x/m)-1$

We may remark that the two famous expressions of the Burg $\ln x$ and Shannon $-x\ln x$ entropies are obtained as special cases.

As a conclusion, we see that the Maximum entropy in mean extends in some way the classical ME approach by giving other expressions for the criterion to optimize. Indeed, it can be shown that when we optimize a convex criterion subject to the data constraints we are optimizing the entropy of some quantity related to the unknowns and *vise versa*. However, as we have mentioned, basically, in this approach the data and the model are assumed to be exact even if some extensions to the approach gives the possibility to account for the errors [**?**]. In the next section, we see how the Bayesian approach can naturally account for both uncertainties on the data and on the unknown parameters $\boldsymbol{x}$.

BAYESIAN INFERENCE APPROACH

In Bayesian approach, the main idea is to translate our prior knowledge on the errors $\boldsymbol{\epsilon}$ and on the unknowns $\boldsymbol{x}$ to prior probability laws $p(\boldsymbol{\epsilon})$ and $p(\boldsymbol{x})$. The next step is to use the forward model and $p(\boldsymbol{\epsilon})$ to deduce $p(\boldsymbol{y}|\boldsymbol{x})$. The Bayes rule can then be used to determine the posterior law of the unknowns $p(\boldsymbol{x}|\boldsymbol{y})$ from which we can deduce any information about the unknowns $\boldsymbol{x}$. The posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is thus the final product of the Bayesian approach. However, very often, we need a last step which is to take out the necessary information about $\boldsymbol{x}$ from this posterior. The tools for this last step are the decision and estimation theories.

To illustrate this, let consider the case of linear inverse problems $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\epsilon}$. The first step is to write down explicitly our hypothesis: starting by the hypothesis that $\boldsymbol{\epsilon}$ is zero-mean (no systematic error), white (no time correlation for the errors) and assuming that we may only have some idea about its energy $\sigma_\epsilon^2 = 1/(2\phi_1)$, and using either the intuition or the Maximum Entropy Principle (MEP) lead to a Gaussian prior law: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, 1/(2\phi_1)\boldsymbol{I})$. Then, using the forward model with this assumption leads to:

$$p(\boldsymbol{y}|\boldsymbol{x},\phi_1) \propto \exp\left[-\phi_1\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2\right]. \tag{26}$$

The next step is to assign a prior law to the unknowns $\boldsymbol{x}$. This step is more difficult and needs more caution. In inverse problems, as we presented, $\boldsymbol{x}$ represents the samples of a signal or the pixel values of an aerian image. Very often then we have ensemblist prior knowledge about the signals or images concerned by the application and we can model them. The art of the engineer is then to choose the appropriate model and to translate this information to a probability law to reflect it.

Again here, let illustrate this step, first through a few general examples and then more specifically the case of mass spectrometry deconvolution problem.

In the first example, we assume that, *a priori* we do not have (or we do not want or we are not able to account for) any knowledge about the correlation between the components of $\boldsymbol{x}$. This leads us to

$$p(\boldsymbol{x}) = \prod_j p_j(x_j). \tag{27}$$

Now, we have to assign $p_j(x_j)$. For this, we may assume to know the mean values m_j and some idea about the dispersions about these mean values. This again leads us to Gaussian laws $\mathcal{N}(m_j, \sigma_{x_j}^2)$, and if we assume the same dispersions $\sigma_{x_j}^2 = 1/(2\phi_2), \forall j$ we obtain

$$p(\boldsymbol{x}) \propto \exp\left[-\phi_2 \sum_j |x_j - m_j|^2\right] = \exp\left[-\phi_2 \|\boldsymbol{x} - \boldsymbol{m}\|^2\right] \tag{28}$$

With these assumptions, using the Bayes rule, we obtain

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\phi_1 \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 - \phi_2 \|\boldsymbol{x} - \boldsymbol{m}\|^2\right]. \tag{29}$$

This posterior law contains all the information we can have on $\boldsymbol{x}$ (combination of our prior knowledge and data). If $\boldsymbol{x}$ was a scalar or a vector of only two components, we could plot the probability distribution and look at it. But, in practical applications, $\boldsymbol{x}$ may be a vector with huge number of components. Then, even if we can obtain an expression for this posterior, we may need to summarize its information content. In general then, we may choose, equivalently, between summarizing it by its mode, mean, marginal modes, *etc* ... , or use the decision and estimation theory to define *point estimators* to be used to compute (*best representing values*). For example, we can choose the value $\widehat{\boldsymbol{x}}$ which corresponds to the mode of $p(\boldsymbol{x}|\boldsymbol{y})$– the *Maximum a posteriori (MAP)* estimate, or the value $\widehat{\boldsymbol{x}}$ which corresponds to the mean of this posterior– the *Posterior mean (PM)* estimate, or when interested to the component x_j, to choose $\widehat{x}_j$ corresponding to the mode of the posterior marginal $p(x_j|\boldsymbol{y})$.

We can also generate samples from this posterior and just look at them as a movie or use them to compute the PM estimate. We can also use it to compute the posterior covariance matrix ($\boldsymbol{P} = \mathrm{E}\{(\boldsymbol{x} - \widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}})^t\}$ where $\widehat{\boldsymbol{x}}$ is the posterior mean), from which we can infer on the uncertainty of the proposed solutions.

In the Gaussian priors case we just presented, it is easy to see that, the posterior law is also Gaussian and all these estimates are the same and can be computed by minimizing

$$J(\boldsymbol{x}) = -\ln p(\boldsymbol{x}|\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x} - \boldsymbol{m}\|^2 \quad \text{with } \lambda = \frac{\phi_2}{\phi_1} = \frac{\sigma_\epsilon^2}{\sigma_x^2}. \tag{30}$$

We may note here the analogy with the quadratic regularization criterion (16) with the emphasis that the choice $\Delta_1(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2$ and $\Delta_2(\boldsymbol{x}, \boldsymbol{m}) = \|\boldsymbol{x} - \boldsymbol{m}\|^2$ are the direct consequences of Gaussian choices for prior laws of the noise $p(\boldsymbol{\epsilon})$ and the unknowns $p(\boldsymbol{x})$.

The Gaussian choice for $p_j(x_j)$ is not always a pertinent one. For example, we may *a priori* know that the distribution of x_j around their means m_j are more concentrated

but great deviations from them are also more likely than a Gaussian distribution. This knowledge can be translated by choosing a Generalized Gaussian law:

$$p(x_j) \propto \exp\left[-\frac{1}{2\sigma_x^2}\left|x_j - m_j\right|^p\right], \quad 1 \leq p \leq 2. \tag{31}$$

In some cases we may know more, for example we may know that x_j are positive values. Then a Gamma prior law

$$p(x_j) = \mathcal{G}(\alpha, m_j) \propto (x_j/m_j)^{-\alpha} \exp\left[-x_j/m_j\right] \tag{32}$$

would be a better choice.

In some other cases we may know that x_j are discrete positive values. Then a Poisson prior law

$$p(x_j) \propto \frac{m_j^{x_j}}{x_j!} \exp\left[-m_j\right] \tag{33}$$

is a better choice.

In all these cases, the expression of the posterior is $p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-J(\boldsymbol{x})\right]$ with $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \lambda\phi(\boldsymbol{x})$ where $\phi(\boldsymbol{x}) = -\ln p(\boldsymbol{x})$. It is interesting to note the different expressions of $\phi(\boldsymbol{x})$ for the prior laws discussed and remark that they contain different *entropy expressions* for the $\boldsymbol{x}$.

The last general example is the case where *a priori* we know that x_j are not independent, for example when they represents the pixels of an aerian image. We may then use a Markovian modeling

$$p(x_j|x_k, k \in \mathcal{S}) = p(x_j|x_k, k \in \mathcal{N}(j)), \tag{34}$$

where $\mathcal{S} = \{1, \ldots, N\}$ stands for the whole set of pixels and $\mathcal{N}(j) = \{k : |k - j| \leq r\}$ stands for r-th order neighborhood of j.

With some assumptions on the border limits, such models again result to the optimization of the same criterion with

$$\phi(\boldsymbol{x}) = \Delta_2(\boldsymbol{x}, \boldsymbol{z}) = \sum_j \phi(x_j, z_j) \text{ where } z_j = \psi(x_k, k \in \mathcal{N}(j)) \tag{35}$$

with different potential functions $\phi(x_j, z_j)$.

A simple example is the case where $z_j = x_{j-1}$ and $\phi(x_j, z_j)$ any function in between the following:

$$\left\{|x_j - z_j|^\alpha, \quad \alpha \ln\frac{x_j}{z_j} + \frac{x_j}{z_j}, \quad x_j \ln\frac{x_j}{z_j} + (x_j - z_j)\right\}$$

See ([**?** , **?** , **?** , **?**]) for some more discussion and properties of these potential functions.

As one of the main conclusions here, we see that, as it concerns the MAP estimation, the Bayesian approach is equivalent to the general regularization. However, here the

choice of the distance measure $\Delta_1(\boldsymbol{y}, \boldsymbol{H}\boldsymbol{x})$ depends on the forward model and the hypothesis on the noise and the choice of the distance measure $\Delta_2(\boldsymbol{x}, \boldsymbol{m})$ depends on the prior law chosen for $\boldsymbol{x}$.

One more extra feature here is that, we have access to the whole posterior $p(\boldsymbol{x}|\boldsymbol{y})$ from which, not only we can define an estimate but also, we can quantify its corresponding remained uncertainty. We can also compare posterior and prior laws of the unknowns to measure the amount of information contained in the observed data. Finally, as we will see in the following, we have finer tools to model unknown signals or images and to estimate the hyperparameters.

OPEN PROBLEMS AND ADVANCED METHODS

As we have remarked in previous sections, in general, the solution of an inverse problem depends on our prior hypothesis on errors $\boldsymbol{\epsilon}$ and on $\boldsymbol{x}$. Before applying the Bayes rule, we have to assign the prior laws to them. From the forward model and assumptions on $\boldsymbol{\epsilon}$ we assign $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\phi}_1)$ and from the assumptions on $\boldsymbol{x}$ we assign $p(\boldsymbol{x}|\boldsymbol{\phi}_2)$. This step is one of the most crucial part of the applicability of the Bayesian framework for inverse problems. Modeling a signal and finding the corresponding expression for the prior law $p(\boldsymbol{x}|\boldsymbol{\phi}_2)$ is not an easy task. This choice may have many consequences: the complexity of the computation of the posterior and consequently the computation of any point estimators such as MAP (which needs optimization) or PM (which needs integration either analytically or by Monte Carlo methods). This modeling depends also on the application. We discuss this point through the particular deconvolution problem in mass spectrometry.

Appropriate modeling of input signal

We actually had started this discussion in previous section and we saw that, at least for linear inverse problems with a white Gaussian assumption of the noise, the posterior has for expression: $p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-J(\boldsymbol{x})\right]$ with

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \lambda\phi(\boldsymbol{x}) \tag{36}$$

with $\phi(\boldsymbol{x}) = -\ln p(\boldsymbol{x})$. Thus the expression and properties of $J(\boldsymbol{x})$, and consequently those of the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ depend on the prior $p(\boldsymbol{x})$. For example if $p(\boldsymbol{x})$ is Gaussian then

$$\phi(\boldsymbol{x}) = -\ln p(\boldsymbol{x}) = \sum_j x_j^2$$

is a quadratic function of $\boldsymbol{x}$. Then the MAP or PM estimates have the same values and their computation needs the optimization of a quadratic criterion which can be done either analytically or by using any simple gradient based algorithm. But the Gaussian modeling is not always an appropriate one. Let take our example of deconvolution of

mass spectrometry data. We know *a priori* that the input signal must be positive. Then a truncated Gaussian will be a better choice:

$$\phi(\boldsymbol{x}) = \sum_j x_j^2, \quad \text{if } \; x_j \geq 0; \quad \text{else } \; \phi(\boldsymbol{x}) = \infty.$$

But, we know still more about the input signal: it has pulse shapes, meaning that, if we look at the histogram of the samples of a typical signal, we see that great number of samples are near to zero but great deviations from this background are not rare. Thus, a generalized Gaussian

$$\phi(\boldsymbol{x}) = \sum_j |x_j|^p \; \text{ with } 1 \leq p \leq 2; \quad \text{if } \; x_j \geq 0; \quad \text{else } \; \phi(\boldsymbol{x}) = \infty.$$

or a Gamma prior law

$$\phi(\boldsymbol{x}) = \sum_j \ln x_j + x_j \text{ if } x_j \geq 0; \text{else } \phi(\boldsymbol{x}) = \infty.$$

would be better choices.

We can also go further in details and want to account for the fact that we are looking for atomic pulses. Then we can imagine a binary valued random vector $\boldsymbol{z}$ with $p(z_j = 1) = \alpha$ and $p(z_j = 0) = 1 - \alpha$, and describe the distribution of $\boldsymbol{x}$ hierarchically:

$$p(x_j|z_j) = z_j \, p_0(x_j) \tag{37}$$

with $p_0(x_j)$ being either a Gaussian $p(x_j) = \mathcal{N}(m, \sigma^2)$ or a Gamma law $p(x_j) = \mathcal{G}(a, b)$. The second choice is more appropriate while the first results on simpler estimation algorithms. The inference can then be done through the joint posterior

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}) \, p(\boldsymbol{x}|\boldsymbol{z}) \, p(\boldsymbol{z}) \tag{38}$$

The estimation of $\boldsymbol{z}$ is then called *Detection* and that of $\boldsymbol{x}$ *Estimation*. The case where we assume $p(\boldsymbol{z}) = \prod_j p(z_j) = \alpha^{n_1}(1-\alpha)^{(n-n_1)}$ with n_1 the number of ones and n the length of the vector $\boldsymbol{z}$, is called Bernoulli process and this modelization for $\boldsymbol{x}$ is called *Bernoulli-Gaussian* or *Bernoulli-Gamma* as a function of the choice for $p_0(x_j)$.

The difficult step in this modeling is the detection step which needs the computation of

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{z}) \int p(\boldsymbol{y}|\boldsymbol{x}) \, p(\boldsymbol{x}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{x} \tag{39}$$

and then its optimization over $\{0, 1\}^n$ where n is the length of the vector $\boldsymbol{z}$. The cost of the computation of the exact solution is huge (a combinatorial problem).

Many approximations to this optimization have been proposed which result to different algorithms for this detection-estimation problem [**?**]. Many Monte Carlo techniques have also been proposed for generating samples of $\boldsymbol{z}$ and $\boldsymbol{x}$ from the posterior and thus compute the PM estimates of $\boldsymbol{x}$. Giving more details on this modeling and details of corresponding algorithms is out of the scope of this paper.

The results on the following figure illustrate this discussion. Here, we used the data in figure 1 and computed $\boldsymbol{x}$ by optimizing the MAP criterion (36), with different prior laws $p(\boldsymbol{x}) \propto \exp\left[-\lambda\phi(\boldsymbol{x})\right]$ in between the following choices:
a) Gaussian: $\phi(\boldsymbol{x}) = \sum x_j^2$,
b) Gaussian truncated on positive axis: $\phi(\boldsymbol{x}) = \sum x_j^2,\ x_j > 0$,
c) Generalized Gaussian truncated on positive axis: $\phi(\boldsymbol{x}) = \sum |x_j|^p$ with $p = 1.1,\ x_j > 0$.
d) Entropic prior $\phi(\boldsymbol{x}) = \sum x_j \ln x_j - x_j,\ x_j > 0$,
e) Gamma prior: $\phi(\boldsymbol{x}) = \sum \ln x_j + x_j,\ x_j > 0$.

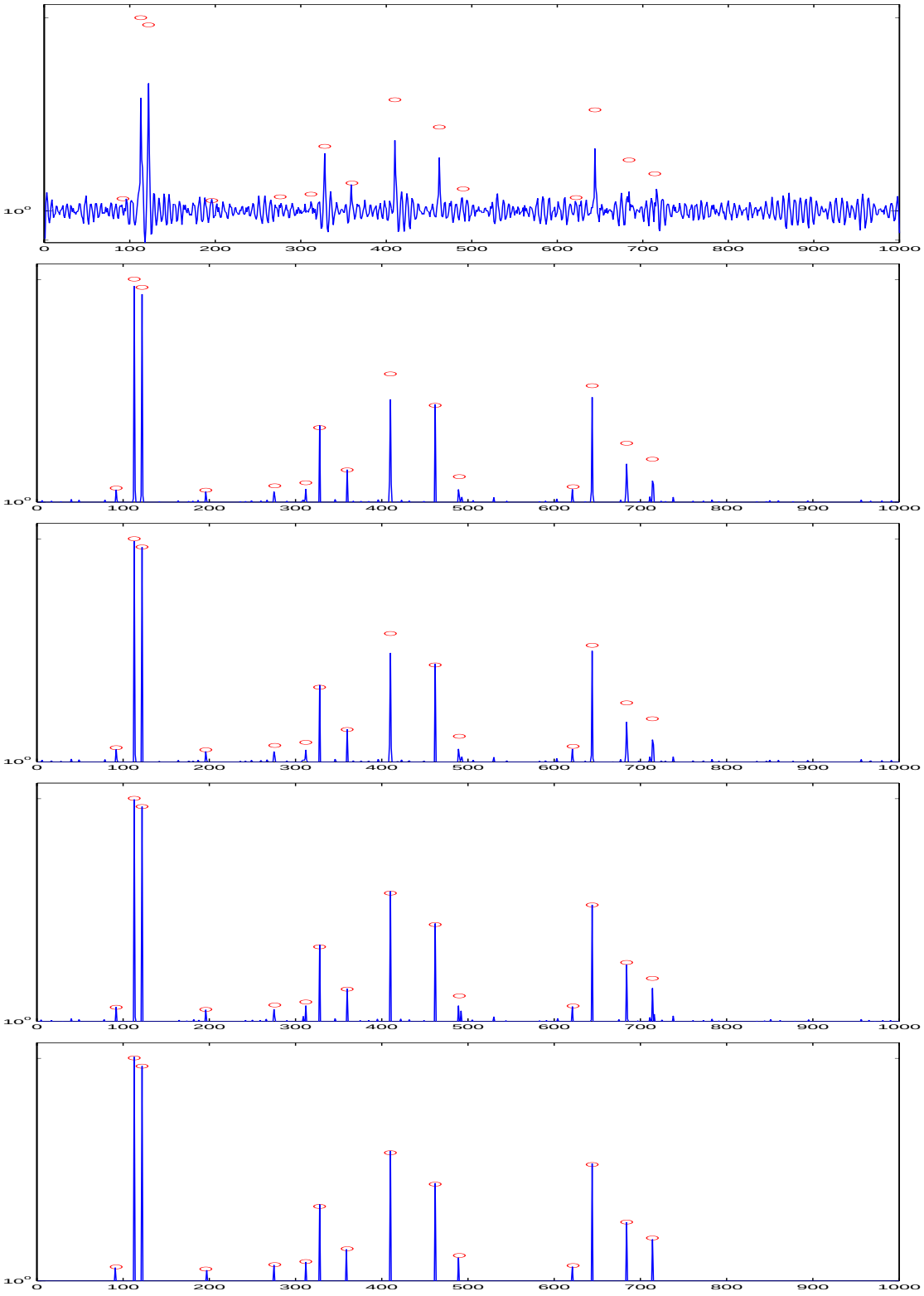

FIGURE 3. Deconvolution results with different priors: a) Gaussian b) Gaussian truncated to positive axis c) Generalized Gaussian. d) $-x \ln x$ entropic prior e) $\ln x$ entropic or Gamma prior.

As it can be seen from these results [1] , for this application, the Gaussian prior does not give satisfactory result, but in almost all the other cases the results are more satisfactory, because the corresponding priors are more in agreement with the nature of the unknown input signal.

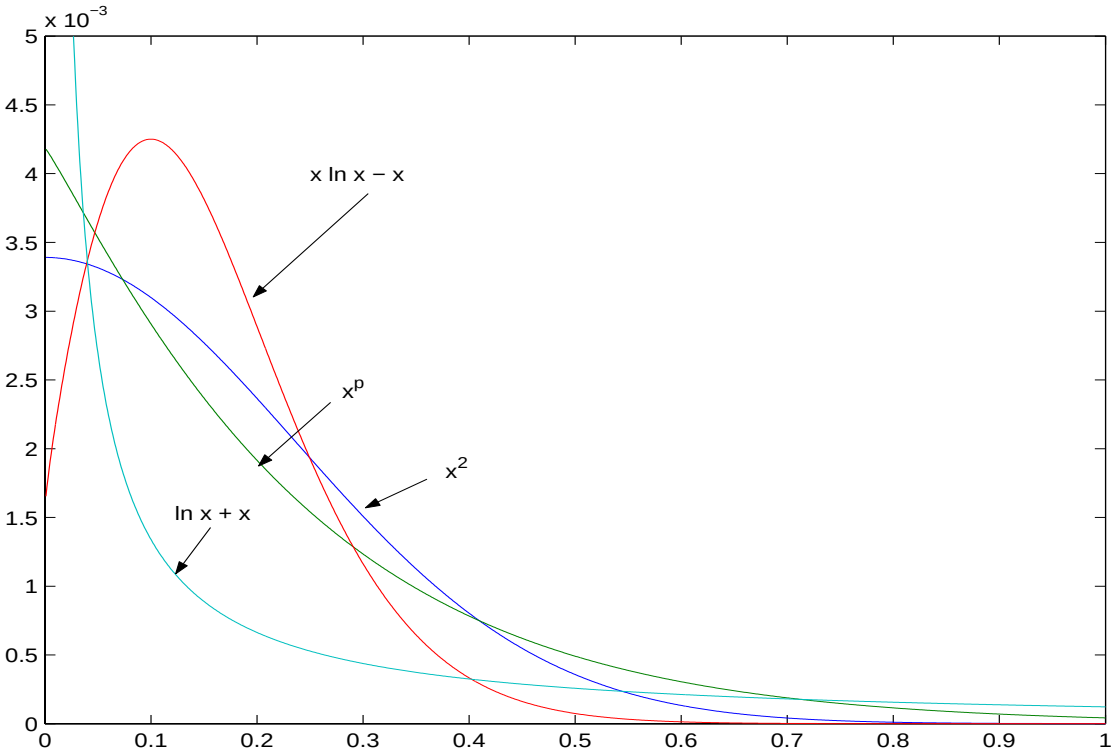

FIGURE 4. Plots of the different prior laws $p(x) \propto \exp[-\lambda\phi(x)]$: a) Truncated Gaussian $\phi(x) = x^2, \lambda = 3$ b) Truncated generalized Gaussian $\phi(x) = x^p,\ p = 1.1,\ \lambda = 4$; c) Entropic $\phi(x) = x\ln x - x, \lambda = 10$ d) Entropic $\phi(x) = \ln x + x, \lambda = 0.1$.

The Gaussian prior (a) is not at all appropriate, Gaussian truncated to positive axis (b) is a better choice. The generalized Gaussian truncated to positive axis (c) and the $-x\ln x$ entropic priors (d) give also almost the same results than the truncated Gaussian case. The Gamma prior (e) seems to give slightly better result (less missing and less artifacts) than all the others. This can be explained if we compare the shape of all these priors shown in figure (4). The Gamma prior is sharper near to zero and has longer tail than other priors. It thus favorites signals with greater number of samples near to zero and still leaves the possibility to have very high amplitude pulses. However, we must be careful on this interpretation, because all these results depend also on the hyperparameter λ whose value may be critical for this conclusion. In these experiments we used the same value for all cases. This brings us to the next open problem which is the determination of the hyperparameters.

---

[1] Remark that the results are presented on a logarithmic scale for the amplitudes to show in more detail the low amplitude pulses. We used $\log(1+a)$ scale in place of y scale which has the advantage of being equal to zero for a=0.

Hyperparameter estimation

The Bayesian approach can be exactly applied when the direct (prior) probability laws $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\phi}_1)$ and $p(\boldsymbol{x}|\boldsymbol{\phi}_2)$ are assigned. Even, when we have chosen appropriate laws, still we have to determine their parameters $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2]$. This problem has been addressed by many authors and the subject is an active area in statistics. See [**?** , **?** , **?** , **?**], [**?** , **?** , **?** , **?** , **?**] and also [**?** , **?** , **?**].

The Bayesian approach gives natural tools to handle this problem by considering $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ as extra unknown parameters to infer on. We may then assign a prior law $p(\boldsymbol{\phi})$ to them too. However, the way to do this is also still an open problem. We do not discuss it more in this paper. The readers are invited to see [**?**] for some extended discussions and references. When this step is done, we can again use the Bayesian approach and compute the joint posterior $p(\boldsymbol{x},\boldsymbol{\phi}|\boldsymbol{y})$ from which we can follow three main directions:

– Joint MAP optimization: In this approach one tries to estimate both the hyperparameters and the unknown variables $\boldsymbol{x}$ directly from the data by defining:

$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\phi}}) = \arg\max_{(\boldsymbol{x},\boldsymbol{\phi})} \{p(\boldsymbol{x},\boldsymbol{\phi}|\boldsymbol{y})\} \text{ where } p(\boldsymbol{x},\boldsymbol{\phi}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\phi})\, p(\boldsymbol{x}|\boldsymbol{\phi})\, p(\boldsymbol{\phi}) \tag{40}$$

and where $p(\boldsymbol{\phi})$ is an appropriate prior law for $\boldsymbol{\phi}$. Many authors used the non informative prior law for them.

– Marginalization: The main idea in this approach is to distinguish between the two sets of unknowns: a high dimensional vector $\boldsymbol{x}$ representing in general a physical quantity and a low dimensional vector $\boldsymbol{\phi}$ representing the parameters of its prior probability laws. This argument leads to estimate first the hyperparameters by marginalizing over the unknown variables $\boldsymbol{x}$:

$$p(\boldsymbol{\phi}|\boldsymbol{y}) \propto p(\boldsymbol{\phi}) \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\phi})\, p(\boldsymbol{x}|\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{x} \tag{41}$$

and then, using them in the estimation of the unknown variables $\boldsymbol{x}$:

$$\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \{p(\boldsymbol{\phi}|\boldsymbol{y})\} \longrightarrow \widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \left\{ p(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\phi}}) \right\}. \tag{42}$$

Note also that when $p(\boldsymbol{\phi})$ is choosed to be uniform, then $p(\boldsymbol{\phi}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\phi})$ which is the likelihood of the hyperparameters $\boldsymbol{\phi}$ and the corresponding maximum likelihood (ML) estimate has all the good asymptotic properties which may not be the case for the joint MAP estimation. However, for practical applications with finite data we may not care too much about the asymptotic properties of these estimates.

– Nuisance parameters: In this approach the hyperparameters are considered as the nuisance parameters, so integrated out of $p(\boldsymbol{x},\boldsymbol{\phi}|\boldsymbol{y})$ to obtain $p(\boldsymbol{x}|\boldsymbol{y})$ and $\boldsymbol{x}$ is estimated by

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y})\} \quad \text{where} \quad p(\boldsymbol{x}|\boldsymbol{y}) = \int p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi} \tag{43}$$

– Joint Posterior Mean: Here, $\boldsymbol{x}$ and $\boldsymbol{\phi}$ are estimated as the posterior means:

$$\widehat{\boldsymbol{x}} = \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y}\} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad \widehat{\boldsymbol{\phi}} = \mathrm{E}\{\boldsymbol{\phi}|\boldsymbol{y}\} = \int \boldsymbol{\phi}\, p(\boldsymbol{\phi}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\phi}. \tag{44}$$

The main issue here is that, excepted the first approach, all the others need integrations for which, in general, there is not analytical expressions and their numerical computation cost may be very high. At the other hand, unfortunately, the estimation by the joint maximization has not the good asymptotic properties (when number of data goes to infinity) of the estimators obtained through the marginalization or expectation. However, in finite number of data, a comparison of their relative properties is still to be done. To see some more discussions and different possible implementations of these approaches see [**?**]. We have also to mention that, we can always use the Markov Chain Monte Carlo (MCMC) techniques to generate samples from the joint posterior $p(\boldsymbol{x}, \boldsymbol{\phi}|\boldsymbol{y})$ and then compute the joint posterior means and corresponding variances. It seems that these techniques are growing up. However, I see two main limitations for their application on real data: their huge computational cost and the need for some discussions on the tools to control their convergences.

Myopic or blind inversion problems

Consider the deconvolution problems (1) or (2) and assume now that the psf $h(t)$ or $h(x, y)$ are partially known. For example, we know they have Gaussian shape, but the amplitude a and the width σ of the Gaussian are unknown. Noting by $\boldsymbol{\theta} = (a, \sigma)$ the problem then becomes the estimation of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$ from $\boldsymbol{y} = \boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{x} + \boldsymbol{\epsilon}$. The case where we know only the support of the psf but not its shape can also be casted in the same way with $\boldsymbol{\theta} = [h(0), \ldots, h(p)]$

Before going more in details, we must note that, in general, the blind inversion problems are much harder than the simple inversion. Taking the deconvolution problem, we have seen in introduction that, the problem even when the psf is given is ill-posed. The blind deconvolution then is still more ill-posed, because here there are more fundamental under-determinations. For example, it is easy to see that, we can find an infinite number of pairs (h, x) which result to the same convolution product $h * x$. This means that, to find satisfactory methods and algorithms for these problems need much more prior knowledge both on x and on h, and in general, the inputs must have more structures (be rich in information content) to be able to obtain satisfactory results.

Conceptually however, the problem is identical to the estimation of hyperparameters in previous section and any of the four approaches presented there can be used. One may wish however to distinguish between these parameters of the system $\boldsymbol{\theta} = (a, \sigma)$ and those hyperparameters of the prior law model descriptions $\boldsymbol{\phi} = (\sigma_\epsilon^2, \sigma_x^2, \ldots)$. In that case, one can try to write down $p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ and use one of the following:

– Joint MAP estimation of $\boldsymbol{x}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$: $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) = \underset{(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})}{\arg\max} \{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})\}$.

– Marginalize over $\boldsymbol{x}$ and estimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using: $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) = \underset{(\boldsymbol{\theta}, \boldsymbol{\phi})}{\arg\max} \{p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})\}$ and

then, estimate $\boldsymbol{x}$ using: $\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \left\{ p(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) \right\}$.

– Marginalize over $\boldsymbol{x}$ and $\boldsymbol{\theta}$ and estimate $\boldsymbol{\phi}$ using: $\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \{ p(\boldsymbol{\phi}|\boldsymbol{y}) \}$,, then estimate $\boldsymbol{\theta}$ using: $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\boldsymbol{\theta}|\boldsymbol{y}, \widehat{\boldsymbol{\phi}}) \right\}$ and finally, estimate $\boldsymbol{x}$ using: $\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \left\{ p(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) \right\}$.

– Joint Posterior Mean: Here, $\boldsymbol{x}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated through their respective posterior means: $\widehat{\boldsymbol{x}} = \mathrm{E}\{\boldsymbol{x}|\boldsymbol{y}\}$, $\widehat{\boldsymbol{\theta}} = \mathrm{E}\{\boldsymbol{\theta}|\boldsymbol{y}\}$ and $\widehat{\boldsymbol{\phi}} = \mathrm{E}\{\boldsymbol{\phi}|\boldsymbol{y}\}$.

Here again, the joint optimization stays the simpler but we must be careful on interpretation of the results. For others, one can either use the Expectation-Maximization (EM) algorithms and/or MCMC sampling tools to approximately compute the necessary integration or expectation computations and overcome the computational cost issues.

CONCLUSIONS

In this paper I presented a synthetic overview of methods for inversion problems starting by deterministic data matching and regularization methods followed by a general presentation of the probabilistic methods such as error probability law matching and likelihood based and the information theory and maximum entropy based methods. Then, I focused on the Bayesian inference. I show that, as it concerns the maximum *a posteriori* estimation method, one can see easily the link with regularization methods. We discussed however the superiority of the Bayesian framework which gives naturally the necessary tools for inferring the uncertainty of the computed solution, for the estimation of the hyperparameters or for handling myopic and blind inversion problems. We saw also that probabilistic modeling of signal and images is more flexible for introduction of practical prior knowledge about them. Finally, we illustrated some of these discussions through a deconvolution example in mass spectrometry data processing.