ELSEVIER

International Journal of Mass Spectrometry 12013 (2002) 1–19

International Journal of Mass Spectrometry

www.elsevier.com/locate/ijms

Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems

A. Mohammad-Djafari*, J.-F. Giovannelli, G. Demoment, J. Idier

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

Received 16 July 2001; accepted 5 November 2001

Abstract

This paper is a synthetic overview of regularization, maximum entropy and probabilistic methods for some inverse problems such as deconvolution and Fourier synthesis problems which arise in mass spectrometry. First we present a unified description of such problems and discuss the reasons why simple naïve methods cannot give satisfactory results. Then we briefly present the main classical deterministic regularization methods, maximum entropy-based methods and the probabilistic Bayesian estimation framework for such problems. The main idea is to show how all these different frameworks converge to the optimization of a compound criterion with a data adequation part and an a priori part. We will however see that the Bayesian inference framework gives naturally more tools for inferring the uncertainty of the computed solutions, for the estimation of the hyperparameters or for handling the myopic or blind inversion problems. Finally, based on Bayesian inference, we present a few advanced methods particularly designed for some mass spectrometry data processing problems. Some simulation results illustrate mainly the effect of the prior laws or equivalently the regularization functionals on the results one can obtain in typical deconvolution or Fourier synthesis problems arising in different mass spectrometry technique. (Int J Mass Spectrom, in press) © 2002 Published by Elsevier Science B.V.

Keywords: Regularization; Maximum entropy; Bayesian inference; Deconvolution; Fourier synthesis

1. Introduction

1.1. Data processing problems in mass spectrometry

In mass spectrometry, the data acquisition and processing is an essential part of the final measurement process. Even if, in some cases, only some pre-processing is done during the acquisition process, the post-acquisition data processing is a vital part of many new mass spectrometry instruments. The main reason is that the raw data do not, in general, directly represent the parameters of interest. These raw data are, in general, transformed and distorted version of the ideal physical quantity of interest which is the mass distribution of the object under the test.

Some distortions are related directly to the measurement system, for example the blurring effect of the time-of-flight (TOF) [1] mass spectrometry data can be written as a simple one-dimensional convolution equation:

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

where $h(t)$ is the point spread function (psf) of blurring effect, $f(t)$ the desired mass distribution and $g(t)$ the data. Fig. 1 shows an example where in place of

* Corresponding author. E-mail: djafari@lss.supelec.fr

1387-3806/02/$20.00
PII S1387-3806(01)00562-0

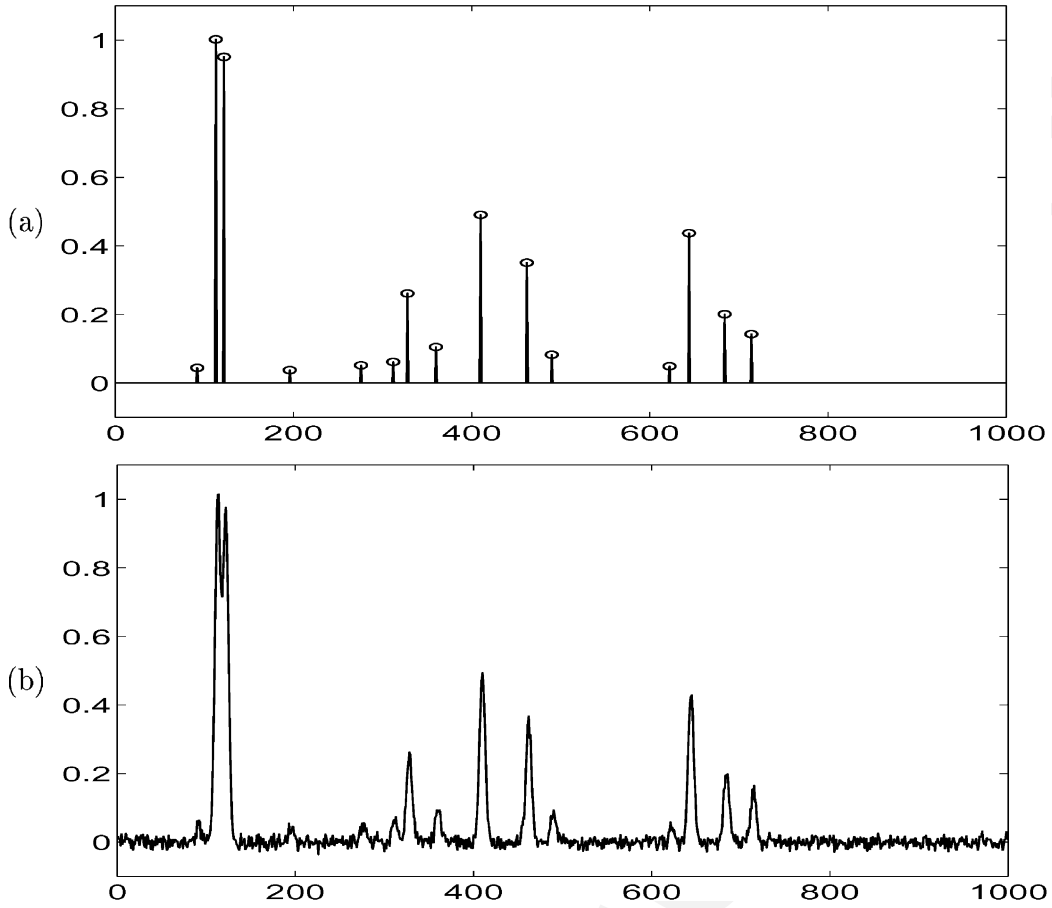

Fig. 1. Blurring effect in TOF mass spectrometry data: (a) desired spectra; (b) observed data.

observing the signal $f(t)$ in (a) the signal $g(t)$ in (b) has been observed.

Some others are due to the output parts of the instrument, for example the interaction and coupling effect of focal plane detectors (FPD) [2] or non-uniformity of ion conversion devices (electron multipliers) in general and in matrix-assisted laser desorption ionization (MALDI) techniques in particular. These distortions can be written as a two-dimensional convolution equation:

$$g(x', y') = \iint f(x, y)h(x' - x, y' - y)\,\mathrm{d}x\,\mathrm{d}y. \quad (2)$$

In some other mass spectrometry techniques such as Fourier transform ion cyclotron resonance (FT-ICR), the observed data are related to the Fourier transform (FT) or Laplace transform (LT) of the mass distribution:

$$g(\tau) = \int f(s)\exp\{-s\tau\}\,\mathrm{d}\omega, \quad \text{with } s = j\omega \text{ or } s = j\omega + \alpha, \quad (3)$$

where α is an attenuation factor. Fig. 2 shows an example of the theoretical spectrum $f(s)$ in (a) and the corresponding observed data $g(\tau)$ in (b). We may observe that, due to the attenuation and the noise in the data, a simple inversion by inverse FT (c) may not give satisfactory result.

In this paper we try to give a unified approach to deal with all these problems. For this purpose, first we

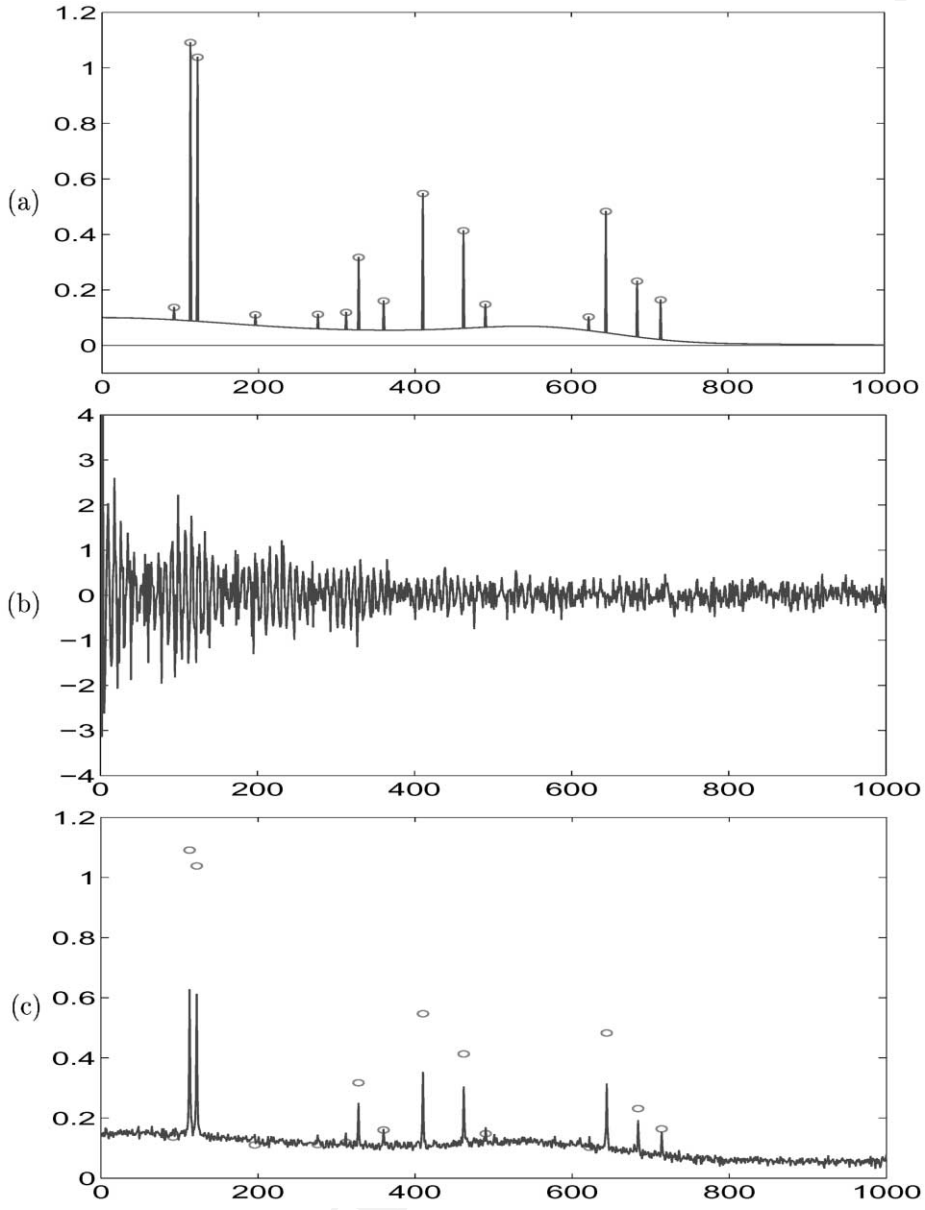

Fig. 2. The reference spectrum (a), its corresponding simulated data in FT-ICR (b) and the inverse FT of the data (c).

note that all these problems are special cases of

$$g(\boldsymbol{s}) = \int f(\boldsymbol{r}) h(\boldsymbol{r}, \boldsymbol{s})\, \mathrm{d}\boldsymbol{r}. \quad (4)$$

Then, we assume that the unknown function $f(\boldsymbol{r})$ can be described by a finite number of parameters $\boldsymbol{x} = [x_1, \ldots, x_n]$:

$$f(\boldsymbol{r}) = \sum_{j=1}^{n} x_j b_j(\boldsymbol{r}), \quad (5)$$

where $b_j(\boldsymbol{r})$ are known basis functions. With this assumption the raw data $y(i) = g(\boldsymbol{s}_i)$, $i = 1, \ldots, m$ are

related to the unknown parameters $\boldsymbol{x}$ by

$$y(i) = g(\boldsymbol{s}_i) = \sum_{j=1}^{n} A_{i,j} x_j,$$

$$\text{with } A_{i,j} = \int b_j(\boldsymbol{r}) h(\boldsymbol{r}, \boldsymbol{s}_i)\, \mathrm{d}\boldsymbol{r}, \tag{6}$$

which can be written in the simple matrix form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. The inversion problem can then be simplified to the estimation of $\boldsymbol{x}$ given $\boldsymbol{A}$ and $\boldsymbol{y}$. Two approaches are then in competition: (a) 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; (b) the more general regularization approach where a classical sampling basis for $b_j(\boldsymbol{r})$ with desired resolution is choose 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 hard to give a very fine mathematical model to take account for all the different steps of the measurement process. However, very often, we can find a rough linear model for the relation between the data and the unknowns (one- or two-dimension convolution or FT or any other linear transformation). But this model may depend on some unknown parameters $\boldsymbol{\theta}$, for example the amplitude and the width of the Gaussian shape psf. It is then usual to write

$$\boldsymbol{y} = \boldsymbol{A}_{\boldsymbol{\theta}}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\epsilon}$ is a random vector accounting for the remaining uncertainties of the model and the measurement noise process.

When the direct model is perfectly known, the main objective of the data processing step is to obtain an estimate $\hat{\boldsymbol{x}}$ of the $\boldsymbol{x}$ such that $\hat{\boldsymbol{x}}$ optimizes some optimality criteria. We will see that, very often, a data matching criterion such as a least square (LS) criterion $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$ does not give satisfactory results. This is, in general due to *ill-posedness* of the problem which, in the case of linear problems, results in *ill-conditioned* linear systems of equations [3]. To obtain a satisfactory result, we need to introduce some *prior information* about the errors and about the unknowns $\boldsymbol{x}$. This can be done through the general *regularization theory* or in a more general way through the *probabilistic inference and statistical estimation*. In probabilistic methods, the rough prior informations about the errors $\boldsymbol{\epsilon}$ and the unknowns $\boldsymbol{x}$ are used to assign the prior probability distribution $p(\boldsymbol{\epsilon}|\boldsymbol{\phi}_1)$ and $p(\boldsymbol{x}|\boldsymbol{\phi}_2)$ where $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are their respective parameters.

Thus, the first steps of solving the problem are to clearly identify $\boldsymbol{x}$, $\boldsymbol{A}$, $\boldsymbol{\theta}$ and $\boldsymbol{y}$ and to define an optimality criterion for $\hat{\boldsymbol{x}}$ which may also depends on the hyperparameters $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2]$. The next step is to find an efficient algorithm to optimize it, and finally, the third step is to characterize the obtained solution. We will however see that these steps are forcibly dependent to each other.

In this paper we focus on this general problem. We first consider the case where the model is assumed to be perfectly known ($\boldsymbol{A}$ and $\boldsymbol{\theta}$ known). This is the simple *inversion problem*. Then we consider the more general case where we have also to infer about $\boldsymbol{\theta}$. This is the *myopic* or *blind inversion* problem. We may also want to infer on the hyperparameters $\boldsymbol{\phi}$ from the data (unsupervised inversion). In some cases, we may have two sets of data, one with known input (for calibration or point spread function estimation purposes) and one with unknown input. Finding an optimal solution for the psf and the unknown input from the two sets of data can be considered as *multi-channel blind deconvolution*.

1.2. Why simple naïve methods do not give satisfaction?

When the degradation model is assumed to be perfectly known, we are face to a simple inversion problem. However, even in this case

- the operator $\boldsymbol{A}$ may not be invertible ($\boldsymbol{A}^{-1}$ does not exists);
- it may admit more than one inverse ($\exists \boldsymbol{B}_1$ and $\boldsymbol{B}_2|\boldsymbol{B}_1(\boldsymbol{A}) = \boldsymbol{B}_2(\boldsymbol{A}) = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity operator);

- it may be ill-posed or ill-conditioned meaning that there exists $\boldsymbol{x}$ and $\boldsymbol{x} + \alpha\delta\boldsymbol{x}$ for which $\|\boldsymbol{A}^{-1}(\boldsymbol{x}) - \boldsymbol{A}^{-1}(\boldsymbol{x} + \alpha\delta\boldsymbol{x})\|$ 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 [4–6]. This explains the reason for which, in general, even in this simple case, many naïve methods based on generalized inversion or on least squares may not give satisfactory results. Fig. 3 shows, in a simple way, the ill-posedness of a deconvolution problem. In this figure, we see that three different input signals can result three outputs which are practically indistinguishable from each other. This means that, data adequation alone cannot distinguish between any of these inputs.

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, we will see that, to do this, there are, at least, three approaches: (i) classical determinist regularization approach; (ii) information theory and entropy-based approach; and (iii) probabilistic and more specifically the Bayesian estimation approach.

The main idea of this paper is to show how all these different frameworks converge to the optimization of a compound criterion: a data adequation part (likelihood) and an a priori part (or penalization). We will see however that the Bayesian framework gives more tools, for example, for inferring the uncertainty of the computed solutions, for accounting for more specific knowledge of the errors and noise and for the estimation of the hyperparameters and for handling myopic and blind inversion problems.

2. Regularization methods

Conceptually, regularization means finding a unique and stable solution to an ill-posed inverse problem. A review of the regularization theory and its different presentations is out of the scope of this paper. Here, we adopt a practical discrete approach, i.e., when the problem is discretized and we are faced to a linear system of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ which may be either under or over-determined.

In the first case the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has more than one solution and one way to obtain a unique solution is to define a criterion, for example $\Delta(\boldsymbol{x}, \boldsymbol{m})$ to choose that unique solution by

$$\hat{\boldsymbol{x}} = \arg\min_{\{\boldsymbol{x};\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}\}} \Delta(\boldsymbol{x}, \boldsymbol{m}), \tag{8}$$

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

The solution to this constrained optimization can be obtained via Lagrangian techniques [7] which consists of defining the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \Delta(\boldsymbol{x}, \boldsymbol{m}) + \boldsymbol{\lambda}^{\mathrm{t}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$ and searching for $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{x}})$ through

$$\begin{cases} \hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\{\mathcal{D}(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\}, \\ \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}\{\mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}})\}. \end{cases} \tag{9}$$

As an example, when $\Delta(\boldsymbol{x}, \boldsymbol{m}) = 1/2\|\boldsymbol{x} - \boldsymbol{m}\|^2$ then the solution is given by

$$\hat{\boldsymbol{x}} = \boldsymbol{m} + \boldsymbol{A}^{\mathrm{t}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{t}})^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{m}). \tag{10}$$

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

Another example is the case where $\Delta(\boldsymbol{x}, \boldsymbol{m}) = \sum_j x_j \ln(x_j/m_j)$ which is more detailed in Section 3.1.

The main issue here is that, this approach provides a unique solution to the inverse problem, but in general, this solution remains sensitive to error on the data.

In the second case the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ may not even has a solution. One then can try to define a solution by

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}\{\Delta(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x})\}, \tag{11}$$

where $\Delta(\boldsymbol{y}, \boldsymbol{z})$ is a distance measure between $\boldsymbol{y}$ and $\boldsymbol{z}$.

The case where $\Delta(\boldsymbol{y}, \boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{z}\|^2$ is the well-known least squares (LS) method. In this case, it is easy to see that any $\hat{\boldsymbol{x}}$ which satisfies the normal equation $\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^{\mathrm{t}}\boldsymbol{y}$ is a LS solution. If $\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A}$ is invertible and well-conditioned then $\hat{\boldsymbol{x}} = (\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{t}}\boldsymbol{y}$ is again the unique generalized inverse solution. But,

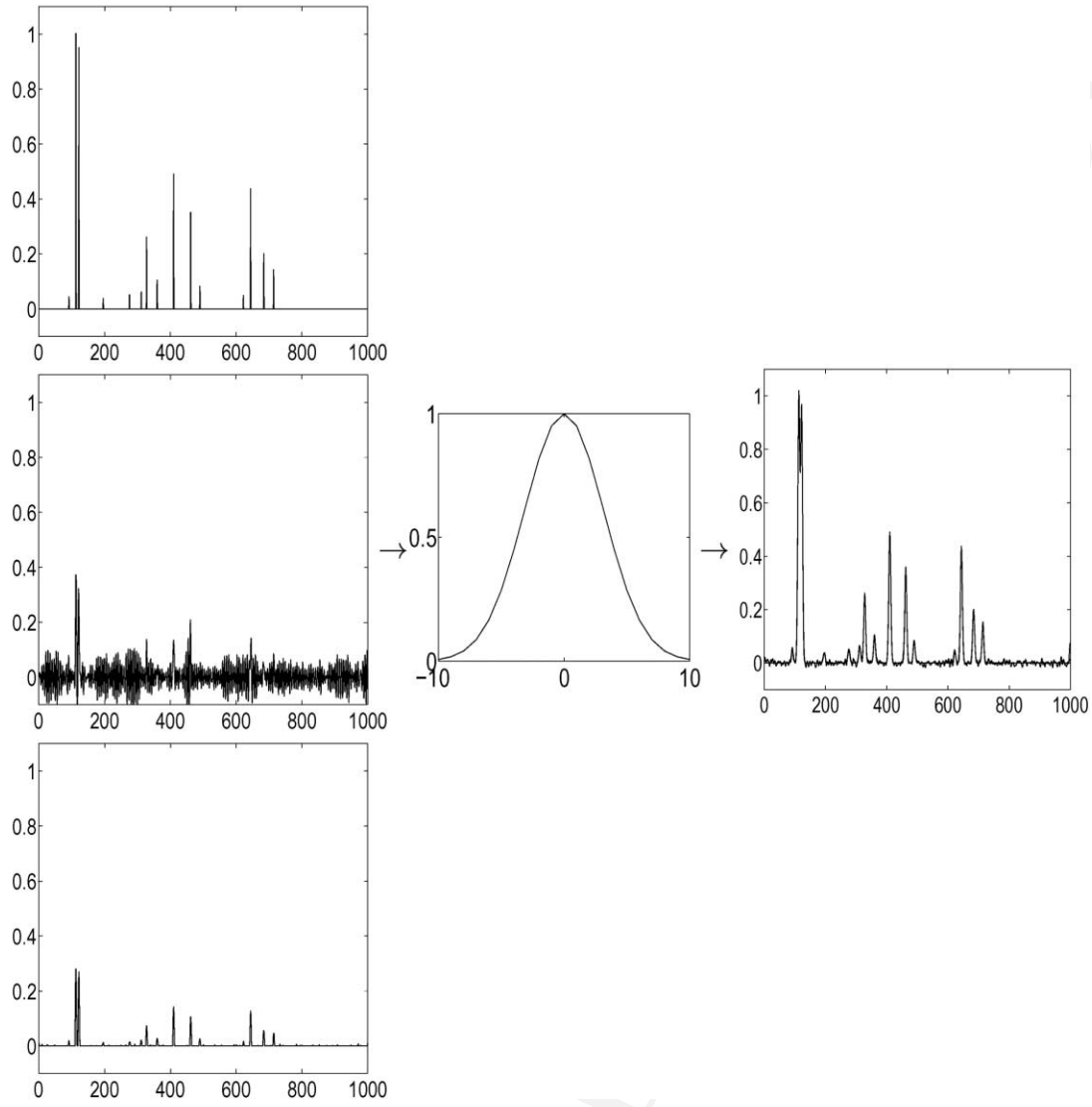

Fig. 3. Ill-posedness of a deconvolution problem: inputs on the left give practically indistinguishable outputs.

in general, this is not the case: $\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A}$ is rank deficient or ill-conditioned and we need to constrain the space of the admissible solutions. The constraint LS is then defined as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}}\{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2\}, \tag{12}$$

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 $\mathcal{C} = \{\boldsymbol{x} : \|\boldsymbol{x}\|^2 \leq \alpha\}$ where the solution can be computed via the optimization of

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x}\|^2. \tag{13}$$

The main technical difficulty is the relation between α and λ. The minimum norm LS solution can also be computed using the singular values decomposition, where there is a link between the choice of the

threshold for truncation of the singular values and α or λ.

In the general case, it is always possible to define a unique solution as the optimizer of a compound criterion $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\mathcal{F}(\boldsymbol{x})$ or the more general criterion

$$J(\boldsymbol{x}) = \Delta_1(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) + \lambda\Delta_2(\boldsymbol{x}, \boldsymbol{m}), \tag{14}$$

where Δ_1 and Δ_2 are two distances or discrepancy measures, λ a regularization parameter and $\boldsymbol{m}$ is 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 approach is determination of the regularization parameter λ which is discussed at the end of this paper, but its description is out of the scope of this paper.

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

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^t\boldsymbol{W}\boldsymbol{A} + \lambda\boldsymbol{Q})^{-1}(\boldsymbol{A}^t\boldsymbol{W}\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{m}), \tag{15}$$

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

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

3. Maximum entropy methods

3.1. Classical ME methods

The notion of entropy has been used in different ways in inversion problems. The classical approach is considering $\boldsymbol{x}$ as a distribution and the data $\boldsymbol{y}$ as linear constraints on them. Then, assuming that the data constraints are satisfied by a non-empty set of solutions, a unique solution is chosen by maximizing the entropy

$$S(\boldsymbol{x}) = -\sum_j x_j \ln x_j, \tag{16}$$

or by minimizing the cross-entropy or the Kullback–Leibler distance between $\boldsymbol{x}$ and a default solution $\boldsymbol{m}$

$$\mathrm{KL}(\boldsymbol{x}, \boldsymbol{m}) = \sum_j x_j \ln\frac{x_j}{m_j} - (x_j - m_j), \tag{17}$$

subject to the linear constraints $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. This method can be considered as a special case of the regularization technique described in previous section for the under-determined case. Here, we have $\Delta(\boldsymbol{x}, \boldsymbol{m}) = \mathrm{KL}(\boldsymbol{x}, \boldsymbol{m})$ and the solution is given by

$$\hat{x}_j = m_j \exp[-[\boldsymbol{A}^t\hat{\boldsymbol{\lambda}}]_j],$$

$$\text{with } \hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\{\mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} - \mathcal{G}(\boldsymbol{A}^t\boldsymbol{\lambda}, \boldsymbol{m})\}, \tag{18}$$

where $\mathcal{G}(\boldsymbol{s}, \boldsymbol{m}) = \sum_j m_j(1 - \exp[-s_j])$. Unfortunately here $\mathcal{D}(\boldsymbol{\lambda})$ is not a quadratic function of $\boldsymbol{\lambda}$ and thus there is not an analytic expression for $\hat{\boldsymbol{\lambda}}$. However, it can be computed numerically and many algorithms have been proposed for its efficient computation. See for example [8,9] and the cited references for more discussions on the computational issues and algorithm implementation.

For other choices of entropy expressions and the presentation of the optimization problem in continuous case (functions and operators in place of vectors and matrices) see [10].

However, even if in these methods, thanks to convex analysis and Lagrangian techniques, the constrained optimization of (16) or (17) can be replaced by an equivalent unconstrained optimization, the ob-

tained solutions satisfy the uniqueness condition of well-posedness but not always the stability one [5,6].

3.2. Entropy as a regularization functional

Entropy (16) or cross-entropy (17) has also been used as a regularization functional $\Delta_2(\boldsymbol{x}, \boldsymbol{m})$ in (14). The main difficulty in this approach is the determination and proper signification of the regularization parameter λ. Note that the criterion

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\,\mathrm{KL}(\boldsymbol{x}, \boldsymbol{m}), \tag{19}$$

is convex on $\mathbb{R}^n_+$ and the solution, when exists, is unique and can be obtained either by any simple gradient-based algorithm or by using the same Lagrangian technique giving:

$$\hat{x}_j = m_j \exp[-[\boldsymbol{A}^{\mathrm{t}}\hat{\boldsymbol{\lambda}}]_j],$$

with

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{ \mathcal{D}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - \mathcal{G}(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda}, \boldsymbol{m}) + \frac{1}{\lambda}\|\boldsymbol{\lambda}\|^2 \right\}. \tag{20}$$

Note that the only difference between (18) and (20) is the extra term $1/\lambda\|\boldsymbol{\lambda}\|^2$ in $\mathcal{D}(\boldsymbol{\lambda})$. Note also that the solution is not a linear function of the data $\boldsymbol{y}$, but a linear approximation to it can be obtained by replacing $\mathrm{KL}(\boldsymbol{x}, \boldsymbol{m})$ by its Taylor series expansion up to the second order which writes

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda(\boldsymbol{x} - \boldsymbol{m})^{\mathrm{t}}\mathrm{diag}[\boldsymbol{m}]^{-1}(\boldsymbol{x} - \boldsymbol{m}),$$

which gives

$$\hat{\boldsymbol{x}} = \boldsymbol{m} + \mathrm{diag}[\boldsymbol{m}](\boldsymbol{A}\,\mathrm{diag}[\boldsymbol{m}]\boldsymbol{A}^{\mathrm{t}} + \lambda^{-1}\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{m}).$$

3.3. Maximum entropy in the mean

The following summarizes the different steps of the approach:

- Consider $\boldsymbol{x}$ as the mean value of a quantity $\boldsymbol{X} \in \mathcal{C}$, where $\mathcal{C}$ is a compact set on which a probability law P is defined: $\boldsymbol{x} = E_P\{\boldsymbol{X}\}$, and the data $\boldsymbol{y}$ as exact equality constraints on it: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}E_P\{\boldsymbol{X}\}$.
- Determine P by minimizing $\mathrm{KL}(P; \mu)$ subject to the data constraints. Here $\mu(\boldsymbol{x})$ is a reference measure corresponding to the prior information on the solution. The solution is obtained via the Lagrangian and is given by

$$\mathrm{d}P(\boldsymbol{x}, \boldsymbol{\lambda}) = \exp[\boldsymbol{\lambda}^{\mathrm{t}}[\boldsymbol{A}\boldsymbol{x}] - \ln Z(\boldsymbol{\lambda})]\,\mathrm{d}\mu(\boldsymbol{x}),$$

$$\text{where } Z(\boldsymbol{\lambda}) = \int_{\mathcal{C}} \exp[\boldsymbol{\lambda}^{\mathrm{t}}[\boldsymbol{A}\boldsymbol{x}]]\,\mathrm{d}\mu(\boldsymbol{x}).$$

  The Lagrange parameters are obtained by searching the unique solution of $\partial \ln Z(\boldsymbol{\lambda})/\partial\lambda_i = y_i$, $i = 1, \cdots, M$.
- The solution to the inverse problem is then defined as the expected value of this distribution: $\hat{\boldsymbol{x}}(\boldsymbol{\lambda}) = E_P\{\boldsymbol{X}\} = \int \boldsymbol{x}\,\mathrm{d}P(\boldsymbol{x}, \boldsymbol{\lambda})$.

  These steps are very formal. In fact, it is possible to show that the solution $\hat{\boldsymbol{x}}(\hat{\boldsymbol{\lambda}})$ can be computed in two ways:
- Via optimization of a dual criterion: the solution $\hat{\boldsymbol{x}}$ is expressed as a function of the dual variable $\hat{\boldsymbol{s}} = \boldsymbol{A}^{\mathrm{t}}\hat{\boldsymbol{\lambda}}$ by $\hat{x}(\hat{\boldsymbol{s}}) = \nabla_{\boldsymbol{s}} G(\hat{\boldsymbol{s}}, \boldsymbol{m})$ where

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

$$\boldsymbol{m} = E_\mu\{\boldsymbol{X}\} = \int_{\mathcal{C}} \boldsymbol{x}\,\mathrm{d}\mu(\boldsymbol{x}) \text{ and } \hat{\boldsymbol{\lambda}}$$

$$= \arg\max_{\boldsymbol{\lambda}}\{D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{t}}\boldsymbol{y} - G(\boldsymbol{A}^{\mathrm{t}}\boldsymbol{\lambda})\}.$$

- Via optimization of a primal or direct criterion:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}}\{H(\boldsymbol{x}, \boldsymbol{m})\}$$

$$\text{s.t., } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \text{ where } H(\boldsymbol{x}, \boldsymbol{m})$$

$$= \sup_{\boldsymbol{s}}\{\boldsymbol{s}^{\mathrm{t}}\boldsymbol{x} - G(\boldsymbol{s}, \boldsymbol{m})\}.$$

What is interesting here is the link between these two options. Note that

- Functions G and H depend on the reference measure $\mu(\boldsymbol{x})$.
- The dual criterion $D(\boldsymbol{\lambda})$ depends on the data and the function G.
- 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 $\boldsymbol{x} \notin \mathcal{C}$.

- 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

$$G(\boldsymbol{s}, \boldsymbol{m}) = \sum_j g_j(s_j, m_j),$$

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

where g_j is the logarithmic Laplace 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	$\exp[-(1/2)(x-m)^2]$	$1/2(s-m)^2$	$1/2(x-m)^2$
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 Burg and Shannon entropies are obtained as special cases. For more details see [11–21].

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 where ever 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. As a final remark, we see that even if this information theory approach gives some more insights for the choice of criteria to optimize, it is more difficult to account for the errors on the data and there is no tools for the determination of the hyperparameters.

4. Bayesian estimation approach

In Bayesian approach, the main idea is to translate our prior knowledge about the errors and about the unknowns to prior probability laws. Then, using the Bayes rule the posterior law of the unknowns is obtained from which we deduce an estimate for them.

To illustrate this, let consider the case of linear inverse problems $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$. 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 correlation for the errors) and assuming that we may only have some idea about its energy $\sigma_\epsilon^2 = 1/(2\phi)$, and using either the intuition or the maximum entropy principle (MEP) lead to a Gaussian prior law: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, 1/(2\phi)\boldsymbol{I})$. Then, using the direct model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$ with this assumption leads to

$$p(\boldsymbol{y}|\boldsymbol{x}, \phi) \propto \exp[-\phi\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2]. \tag{21}$$

The next step is to assign a prior law to the unknowns $\boldsymbol{x}$. This step is more difficult and needs more caution.

Again here, let illustrate it through a few examples. 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

$$p(\boldsymbol{x}) = \prod_j p_j(x_j). \tag{22}$$

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 around these mean values. This again leads us to Gaussian laws $\mathcal{N}(m_j, \sigma_{x_j}^2)$, and if we assume $\sigma_{x_j}^2 = 1/(2\theta), \forall j$, we obtain

$$p(\boldsymbol{x}) \propto \exp[-\theta \sum_j |x_j - m_j|^2] = \exp[-\theta\|\boldsymbol{x} - \boldsymbol{m}\|^2]. \tag{23}$$

With these assumptions, using the Bayes rule, we obtain

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp[-\phi\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 - \theta\|\boldsymbol{x} - \boldsymbol{m}\|^2]. \tag{24}$$

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. For this reason, in general, we may choose a *point estimator* to summarize it (*best representing value*). For example, we can choose the value $\hat{\boldsymbol{x}}$ which corresponds to the maximum of $p(\boldsymbol{x}|\boldsymbol{y})$—the *maximum a posteriori* (MAP) estimate, or the value $\hat{\boldsymbol{x}}$ which corresponds to the mean of this posterior—the *posterior mean* (PM) estimate. We can also generate samples (using any Monte Carlo method) 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 from which we can infer on the uncertainty of the proposed solutions.

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

$$J(\boldsymbol{x}) = -\ln p(\boldsymbol{x}|\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x} - \boldsymbol{m}\|^2, \quad \text{with } \lambda = \frac{\theta}{\phi} = \frac{\sigma_\epsilon^2}{\sigma_x^2}. \tag{25}$$

We may note here the analogy with the quadratic regularization criterion (14) with the emphasis that the choice $\Delta_1(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$ 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 and the unknowns $\boldsymbol{x}$.

The Gaussian choice for $p_j(x_j)$ may not always be 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 [22]. This knowledge can be translated by choosing a generalized Gaussian law

$$p(x_j) \propto \exp\left[-\frac{1}{2\sigma_x^2}|x_j - m_j|^p\right], \quad 1 \le p \le 2. \tag{26}$$

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

$$p(x_j) = \mathcal{G}(\alpha, m_j) \propto (x_j/m_j)^{-\alpha} \exp[-x_j/m_j], \tag{27}$$

would be a better choice.

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

$$p(x_j) \propto \frac{m_j^{x_j}}{x_j!} \exp[-m_j] \tag{28}$$

is a better choice.

In all these cases, the MAP estimates are always obtained by minimizing the criterion $J(\boldsymbol{x}) = -\ln p(\boldsymbol{x}|\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\mathcal{F}(\boldsymbol{x})$ where $\mathcal{F}(\boldsymbol{x}) = -\ln p(\boldsymbol{x})$. It is interesting to note the different expressions we obtain for $\mathcal{F}(\boldsymbol{x})$ for these choices contain also different entropy expressions for the $\boldsymbol{x}$.

When, a priori we know that x_j are not independent, for example when they represents the pixels of an image, we may use a Markovian modeling

$$p(x_j|x_k, k \in \mathcal{S}) = p(x_j|x_k, k \in \mathcal{N}(j)), \tag{29}$$

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

With some assumptions about the border limits [23], such models again result to the optimization of the same criterion with

$$\mathcal{F}(\boldsymbol{x}) = \Delta_2(\boldsymbol{x}, \boldsymbol{z}) = \sum_j \phi(x_j, z_j) \quad \text{where } z_j = \psi(x_k, k \in \mathcal{N}(j)), \tag{30}$$

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 [24–26] for some more discussion and properties of these potential functions.

5. Main conclusion and unifying viewpoint

As one of the main conclusions here, we see that, a common tool between the three previous approaches is defining the solution as the optimizer of a compound criterion: a data dependent part $\Delta_1(\boldsymbol{y}, \boldsymbol{Ax})$ and an a priori part $\Delta_2(\boldsymbol{x}, \boldsymbol{m})$. In all cases, the expression of $\Delta_1(\boldsymbol{y}, \boldsymbol{Ax})$ depends on the direct model and the hypothesis on the noise and the expression of $\Delta_2(\boldsymbol{x}, \boldsymbol{m})$ depends on our prior knowledge of $\boldsymbol{x}$. The only difference between the three approaches is the arguments leading to these choices. In classical regularization, the arguments are based on notion of energy, in maximum entropy approach they are based on information theory, and in Bayesian approach, they are based on the choice of the prior probability laws.

However, the Bayesian approach has some more extra features: it gives naturally the tools to account for uncertainties and errors of modeling and data through the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. It also gives natural tools to account for any prior information about the unknown signal through the prior probability law $p(\boldsymbol{x})$. We also 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 uncertainty. For example, in the Gaussian case, we can use the diagonal elements of posterior covariance matrix to put error bars on the computed solution. 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 last section, we have finer tools for hyperparameters estimation and for handling myopic or blind deconvolution problems. In the following we keep this approach and present methods with finer prior modeling more appropriate for mass spectrometry signal processing applications.

6. Advanced methods

6.1. Bernoulli–Gamma and generalized Gaussian modeling

In mass spectrometry, the unknown quantity $\boldsymbol{x}$ is mainly composed of positive pulses. One way to model this prior knowledge is to 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

$$p(x_j|z_j) = z_j p_0(x_j), \tag{31}$$

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

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}). \tag{32}$$

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 approach is the detection step which needs the computation of

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{z}) \int p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{x} \tag{33}$$

and then the 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 [27]. To avoid complex and costly algorithms of detection–estimation and still be able to catch the mass spectrometry pulse shape prior information, there is a simpler modeling: *generalized Gaussian modeling* which consist of assuming $p(\boldsymbol{x}) \propto \exp[-\theta \sum_j |x_j|^\alpha]$, $1 \leq \alpha \leq 2$ or $p(\boldsymbol{x}) \propto \exp[-\theta \sum_j |x_j - x_{j-1}|^\alpha]$ or still a combination of them

$$p(\boldsymbol{x}) \propto \exp[-\theta_0 \sum_j |x_j|^{\alpha_0} - \theta_1 \sum_j |x_j - x_{j-1}|^{\alpha_1}]. \tag{34}$$

The first one translates the fact that, if we plot the histogram of a typical spectrum, we see that great number of samples are near to zero, but there are

samples which can go very far from this axis. The second expression translates the same fact but on the differences between two consecutive samples and the third choice combines the two facts. The more interesting fact of such a choice as a prior law for $\boldsymbol{x}$ is that the corresponding MAP criterion is convex and the computation of the solutions can be done easily by any gradient-based type algorithm.

6.2. A mixed background and impulsive signal modeling

In some techniques of mass spectrometry, a better model for $\boldsymbol{x}$ is to assume it as the sum of two components $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$: a smooth background $\boldsymbol{x}_1$ and pulse shape $\boldsymbol{x}_2$. To catch the smoothness of $\boldsymbol{x}_1$ we can assign a Gaussian distribution $p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_{1_0}, \boldsymbol{R}_{x_1})$ and to catch the pulse shape of $\boldsymbol{x}_2$ we can again either use the Bernoulli–Gamma or Bernoulli–Gaussian models of the previous section or use a generalized Gaussian prior

$$p(\boldsymbol{x}_2) \propto \exp[-\theta \sum_j |x_{2_j}|^\alpha]. \tag{35}$$

The inference can then be done through the joint posterior $p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x}_1)\, p(\boldsymbol{x}_2)$ which writes

$$\begin{aligned} \ln p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{y}) = {} & \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x}_1 + \boldsymbol{x}_2)\|^2 \\ & + (\boldsymbol{x}_1 - \boldsymbol{x}_{1_0})^{\mathrm{t}} \boldsymbol{R}_{x_1}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_{1_0}) \\ & - \theta \sum_j |x_{2_j}|^\alpha. \end{aligned} \tag{36}$$

One possible way to estimate $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is the joint optimization of this posterior through the following relaxation iterations:

$$\begin{cases} \hat{\boldsymbol{x}}_1 = (\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A} + \lambda_1 \boldsymbol{R}_{x_1}^{-1})^{-1} (\boldsymbol{A}^{\mathrm{t}}\boldsymbol{y}_1 + \lambda_1 \boldsymbol{m}_1), \\ \hat{\boldsymbol{x}}_2 = \arg\max_{\boldsymbol{x}_2} \{\ln p(\hat{\boldsymbol{x}}_1, \boldsymbol{x}_2|\boldsymbol{y})\}. \end{cases}$$

6.3. Hierarchical modeling

Another approach is a hierarchical modeling. As an appropriate example, we propose $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}, \sigma_z^2 \boldsymbol{I})$ and $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_z)$ with $\boldsymbol{R}_z = \sigma_z^2 (\boldsymbol{D}^{\mathrm{t}}\boldsymbol{D})^{-1}$ which leads to

$$-\ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x} - \boldsymbol{z}\|^2 + \mu \|\boldsymbol{D}\boldsymbol{z}\|^2. \tag{37}$$

Its joint optimization can be obtained through the following relaxation iterations:

$$\begin{cases} \hat{\boldsymbol{x}} = (\boldsymbol{A}^{\mathrm{t}}\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{A}^{\mathrm{t}}\boldsymbol{y} + \lambda \hat{\boldsymbol{z}}), \\ \hat{\boldsymbol{z}} = \lambda \left(\boldsymbol{D}^{\mathrm{t}}\boldsymbol{D} + \dfrac{\lambda}{\mu} \boldsymbol{I} \right)^{-1} \hat{\boldsymbol{x}}. \end{cases} \tag{38}$$

A better choice for $p(\boldsymbol{x}|\boldsymbol{z})$ is $p(\boldsymbol{x}|\boldsymbol{z}) \propto \exp[-\theta \sum_j |x_j - z_j|^\alpha]$ which leads to

$$\begin{aligned} -\ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = {} & \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \mu \sum_j |x_j - z_j|^\alpha \\ & + \lambda \|\boldsymbol{D}\boldsymbol{z}\|^2. \end{aligned} \tag{39}$$

The main drawback of this model is that $-\ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y})$ is neither quadratic in $\boldsymbol{z}$ nor in $\boldsymbol{x}$. However, the solution can be obtained via an iterative gradient-based algorithm.

7. Numerical experiment

The main objective of this section is to illustrate some of the points we discussed in previous sections. As we discussed, one of the main critical points in inverse problems is the choice of appropriate prior laws. In this paper, we only focus on this point and we give a very brief comparison of results obtained with some of the aforementioned prior law choices. We have limited ourselves to the prior laws which result to concave MAP criteria to avoid the difficult task of global optimization problems.

We also limit ourselves to two inverse problems: deconvolution and Fourier synthesis. This comparison can be done objectively on simulated data. However, we must generate data representing some real and difficult situations to be able to see the differences between different methods. For this reason, we simulated two spectra:

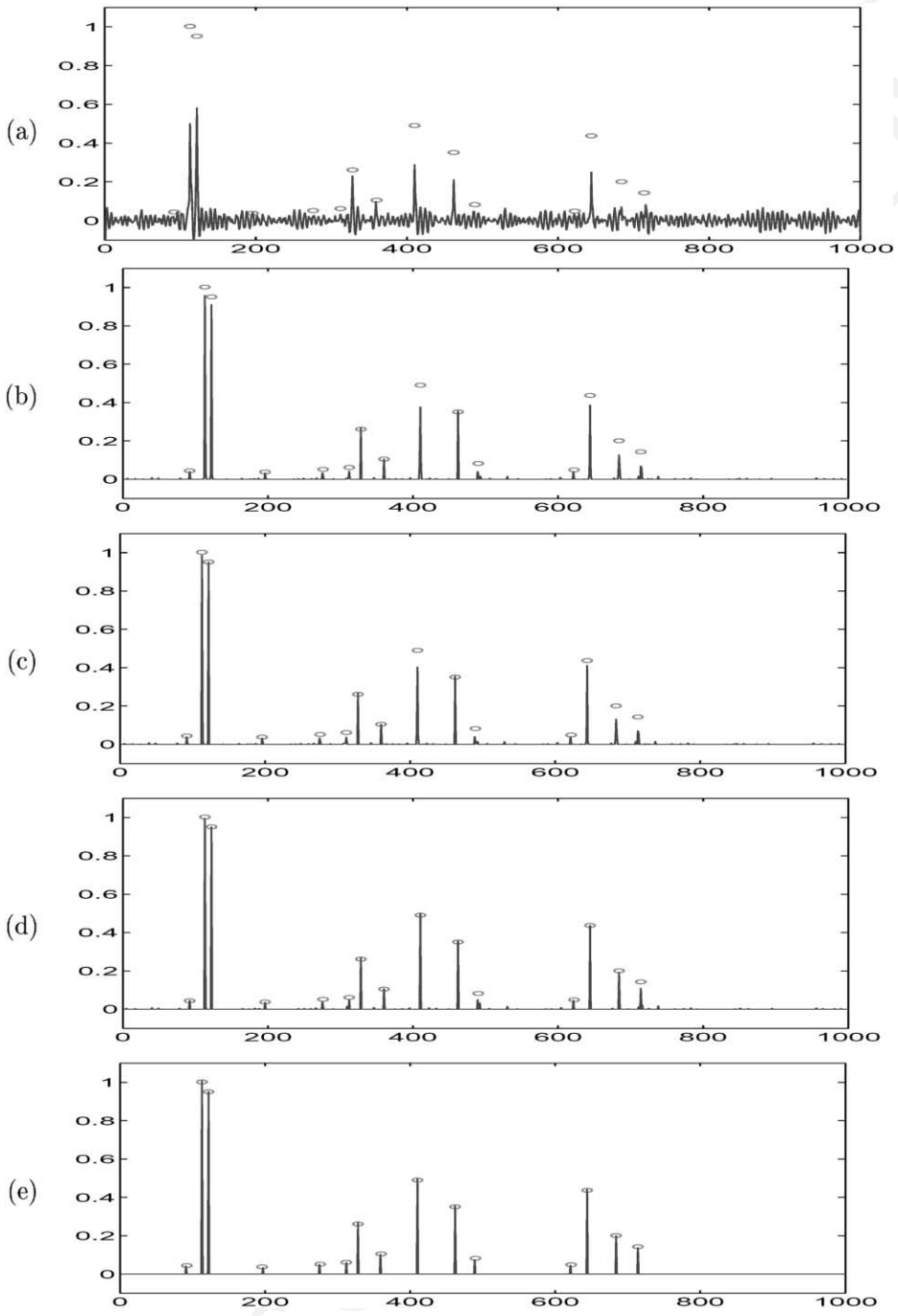

Fig. 4. Simple deconvolution results for the first reference spectrum. The original spectrum and data are those of Fig. 1. (a) Quadratic regularization (QR); (b) QR with positivity constraint; (c) MAP estimation with generalized Gaussian prior; (d) MAP estimation with $-x \ln x$ prior; (e) MAP estimation with $\ln x$ prior.

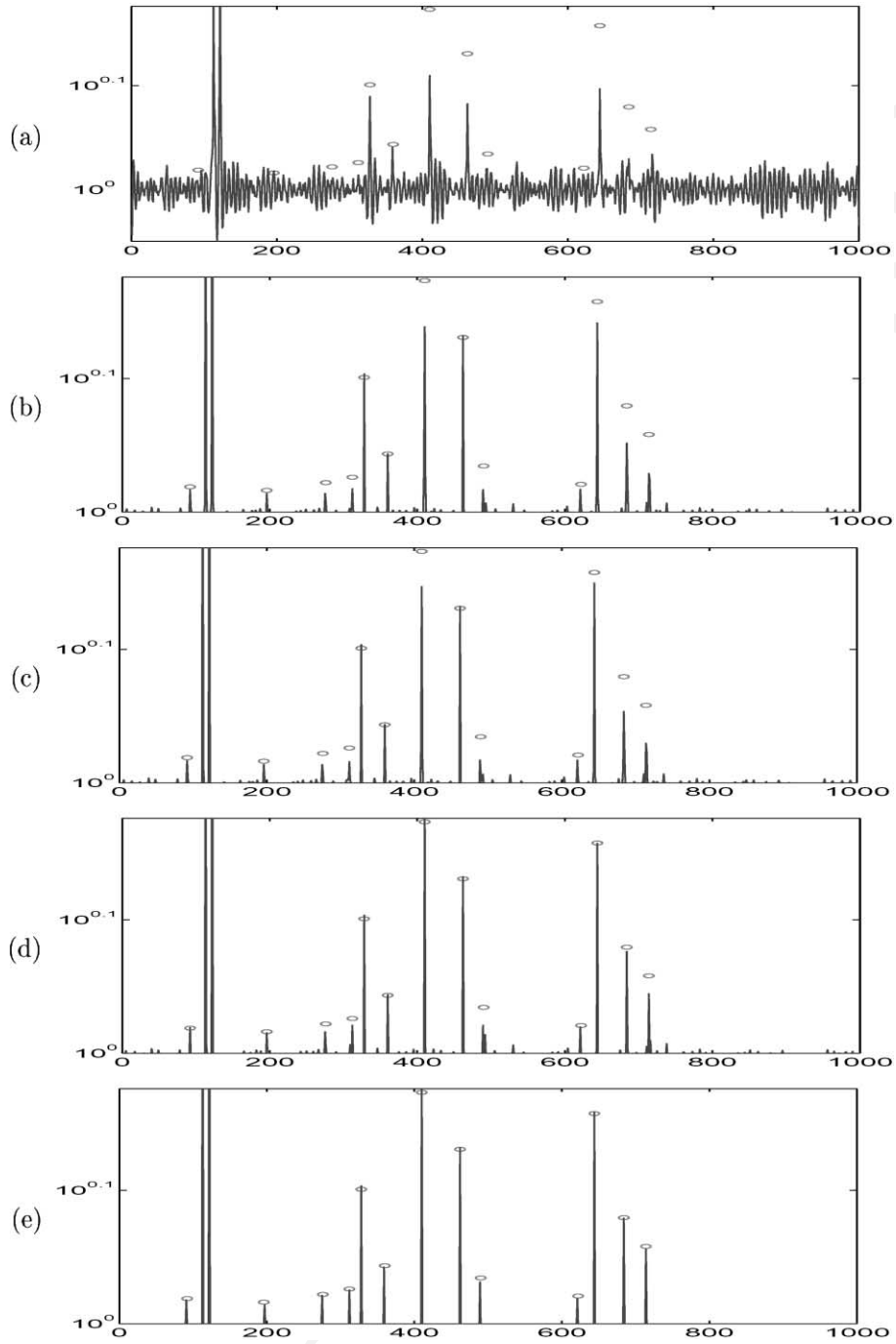

Fig. 5. Deconvolution results of Fig. 4 showed in logarithmic scale: (a) Gaussian prior; (b) truncated Gaussian prior; (c) truncated generalized Gaussian prior; (d) entropic $x \ln x - x$ prior; (e) entropic $\ln x + x$ prior.

(i) a simple case where the background is flat (Fig. 1a) and
(ii) a more complicated case where the background is not flat (Fig. 2a).

We used these spectra as references for measuring the performances of the proposed data processing methods.

7.1. Simple deconvolution

For this case, first we used the first spectrum as the reference. Then using it, we simulated data by convoluting it with a Gaussian shape psf and added some noise (white Gaussian such that SNR = 20 dB). Fig. 1 shows this original spectrum and the associated simulated data. Then, using these data, we applied some of the different methods previously explained.

Fig. 4 shows these results. All these results are obtained by optimizing the MAP criterion

$$J(\boldsymbol{x}) = -\ln p(\boldsymbol{x}|\boldsymbol{y}) \propto \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\phi(\boldsymbol{x}),$$

with different prior laws $p(\boldsymbol{x}) \propto \exp[-\lambda\,\phi(\boldsymbol{x})]$. The main objective of these experiments is to show the effects of the prior law $p(\boldsymbol{x})$ or equivalently the choice of the regularization functional $\phi(\boldsymbol{x})$ on the results. We limited ourselves here to the following choices:

(a) Gaussian or equivalently quadratic regularization $\phi(\boldsymbol{x}) = \alpha \sum x_j^2, \alpha > 0$;
(b) Gaussian truncated on positive axis or equivalently quadratic regularization with positivity constraint $\phi(\boldsymbol{x}) = \alpha \sum x_j^2, x_j > 0, \alpha > 0$;
(c) Generalized Gaussian or equivalently L_p regularization with $\phi(\boldsymbol{x}) = \alpha \sum |x_j|^p$, $p = 1.1, x_j > 0, \alpha > 0$;
(d) Shannon $(x \ln x)$ entropy $\phi(\boldsymbol{x}) = \alpha(\sum x_j \ln x_j - x_j), x_j > 0, \alpha > 0$;
(e) Burg $(\ln x)$ entropy or equivalently Gamma prior $\phi(\boldsymbol{x}) = \alpha(\sum \ln x_j + x_j), x_j > 0, \alpha > 0$.

Fig. 5 shows the same result on a logarithmic scale for the amplitudes to show in more detail the low amplitude pulses. We used $\log(1 + y)$ scale in place of y scale which has the advantage of being equal to zero for $y = 0$.

As it can be seen from these results, Gaussian prior or equivalently quadratic regularization does not give satisfactory result, but in almost all the other cases the results are satisfactory, because the corresponding priors are more in agreement with the nature of the unknown input signal. The Gaussian prior (a) is not at all appropriate, Gaussian truncated to positive axis (b) is a better choice. The generalized Gaussian (c) and the $-x \ln x$ entropic priors (d) give also practically 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 Fig. 6. The Gamma prior is sharper near to zero and has longer tail than other priors. It thus enforces 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. Description and discussion of the methods to estimate λ from the data is out of

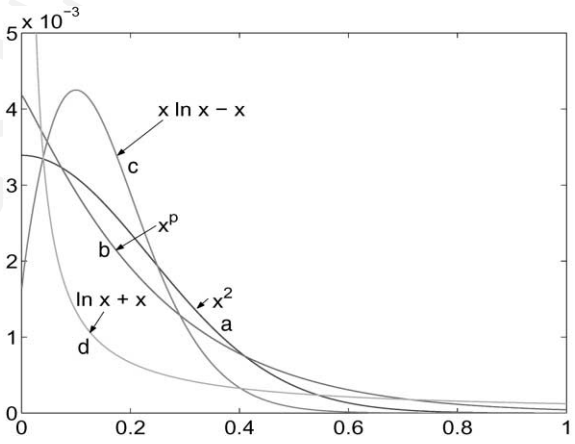

Fig. 6. Plots of the different prior laws $p(x) \propto \exp[-\alpha\phi(x)]$: (a) truncated Gaussian $\phi(x) = x^2$, $\alpha = 3$; (b) truncated generalized Gaussian $\phi(x) = x^p$, $p = 1.1$, $\alpha = 4$; (c) entropic $\phi(x) = x \ln x - x$, $\alpha = 10$; (d) entropic $\phi(x) = \ln x + x$, $\alpha = 0.1$.

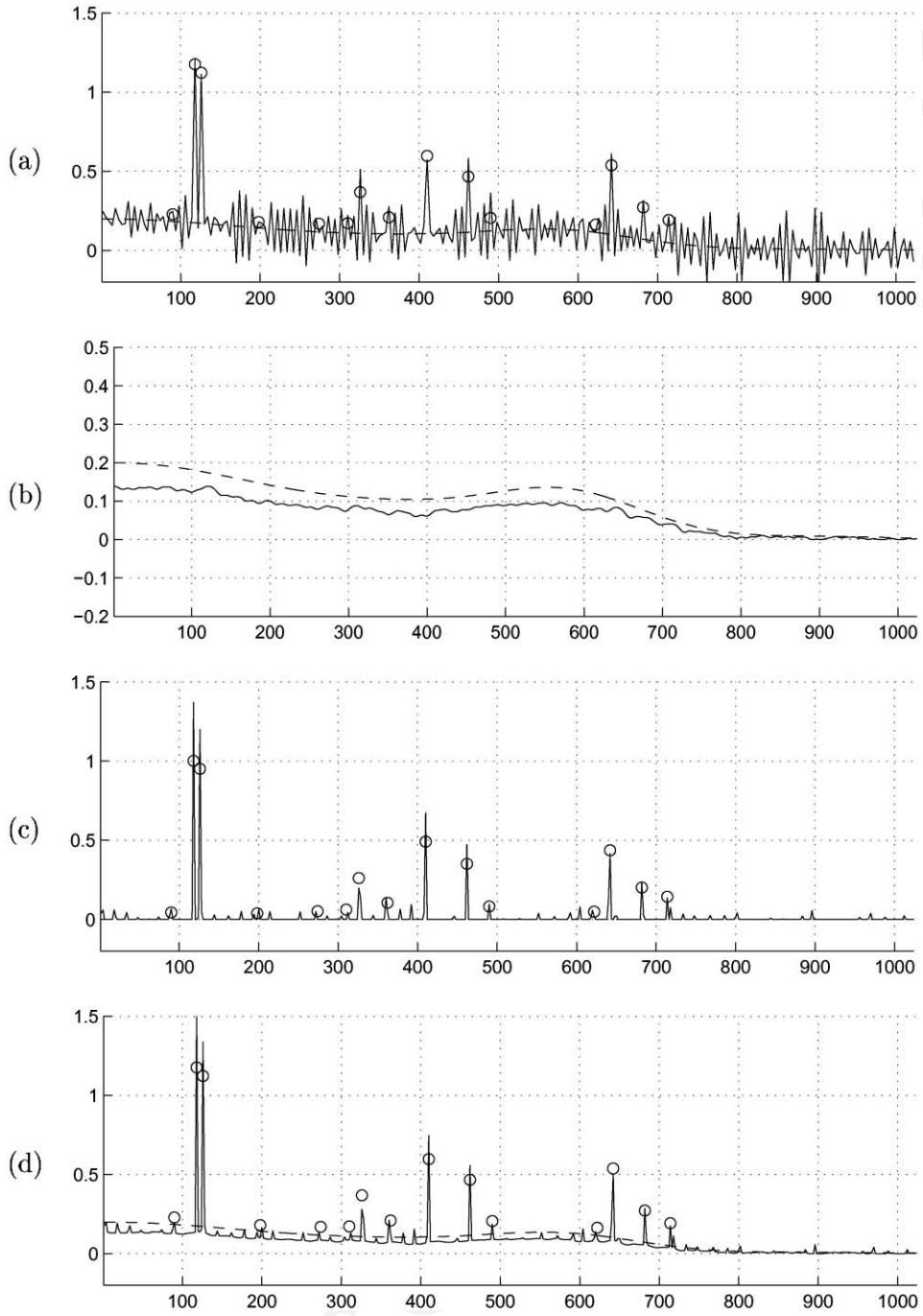

Fig. 7. Reconstructed spectra in FT-NMR data: (a) shows the weighted FFT solution; (b), (c) and (d), respectively, gives $\hat{x}_1$, $\hat{x}_2$ and $\hat{x} = \hat{x}_1 + \hat{x}_2$. The true peaks are given by circles and the true background is given by dashed lines.

focus of this paper. We can, however, mention that, in general, the results are not too sensitive to this value when it is fixed to the right scale.

7.2. Fourier synthesis inversion in NMR mass spectrometry

As a second example, we used the second spectrum as the reference. But here, we simulated the FID data that one could observe using a relaxation of $\tau = 0.2$. Here also, we added some noise on the data and then, using them, we applied the mixed background *and* pulse shape signal model previously explained in this paper. Fig. 7 shows the result which is obtained more precisely by optimizing the following criterion:

$$J(\boldsymbol{x}_1, \boldsymbol{x}_2) = -\ln p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x}_1 + \boldsymbol{x}_2)\|^2 + \lambda_1 \sum_j (x_1(j+1) - x_1(j))^2 + \lambda_0 \sum_j |x_2(j)|,$$

which involves a usual data-based term and two regularization terms: the first one addresses the smooth background $\boldsymbol{x}_1$ and the second one addresses the impulsive component $\boldsymbol{x}_2$. The chosen heavy-tailed $L_2 - L_1$ potential function is a hyperbolic cost [28,29]. So that, J is strictly convex and the estimated object is defined as the minimizer of J over $\mathbb{R}^n_+$. The optimization is achieved by an iterative coordinate descent algorithm [7]. The minimizers $\hat{\boldsymbol{x}}_1$, $\hat{\boldsymbol{x}}_2$ and $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_1 + \hat{\boldsymbol{x}}_2$ are given in Fig. 7(b)–(d). It is to be compared to the "weighted FFT" solution of Fig. 7(a). The proposed solution accounts for positivity and clearly separates background and peaks. Moreover, the peaks are more accurately identified.

8. Conclusions

In this paper we presented a synthetic overview of regularization, maximum entropy and probabilistic methods for linear inversion problems arising in mass spectrometry. We discussed the reasons why simple naïve methods cannot give satisfactory results and the need for some prior knowledge about the unknowns to obtain satisfactory results. We then presented briefly the main classical regularization, maximum entropy based and the Bayesian estimation-based methods. We showed how all these different frameworks converge to the optimization of a compound criterion. We discussed the superiority of the Bayesian framework which gives more tools for the estimation of the hyperparameters or for inferring the uncertainty of the computed solutions or for handling the myopic or blind inversion problems. Finally, we presented some advanced methods based on Bayesian inference and particularly designed for some mass spectrometry data processing problems. We illustrated some numerical results simulating deconvolution and Fourier synthesis problems to illustrate the results we can obtain using some of the presented methods. The main objective of these numerical experiments was to show the effect of different choices for prior laws or equivalently the regularization functional on the result.

However, 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 $\boldsymbol{x}$. In practical applications, we can only formalize these hypothesis either through prior probabilities or through regularization functionals depending on some hyperparameters (regularization parameter for example). Determination of these hyperparameters from the data becomes then a crucial part of the problem. Description of the methods to handle this problem is out of focus of this paper. Interested readers can refer to [30] for deterministic methods such as cross-validation technics or to [31–42] for Bayesian inference-based methods.

Another point we did not discussed is the validity of linear model with additive noise $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\epsilon}$ and all the hypothesis needed to write down the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. For example, we assumed $\boldsymbol{\epsilon}$ to be additive and independent of the input $\boldsymbol{x}$. This may not be true, but it simplifies the derivation of $p(\boldsymbol{y}|\boldsymbol{x})$ from $p_\epsilon(\boldsymbol{\epsilon})$. If this hypothesis is correct, then $p(\boldsymbol{y}|\boldsymbol{x}) = p_\epsilon(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})$. If this is not the case, we have to account for it in the expression of $p(\boldsymbol{y}|\boldsymbol{x})$. Then, all the other steps of the Bayesian inference do not change. However, if

$\ln p(\boldsymbol{y}|\boldsymbol{x})$ is not a quadratic function of $\boldsymbol{x}$, the consequent computations of the posterior law summaries or its sampling may be more difficult. This is also true for the hypothesis that $\boldsymbol{\epsilon}$ is white. This assumption is also used to simplify the expression of $p(\boldsymbol{y}|\boldsymbol{x})$, but this can be handled more easily than the previous hypothesis if it is not true. For example, if we can assume it to Gaussian and model its covariance matrix $\boldsymbol{R}_\epsilon$, we can use it easily in the expression of the likelihood which becomes $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}, \boldsymbol{R}_\epsilon)$. Also, as mentioned by one of reviewers of this paper, in some techniques of mass spectrometry, the Gaussian assumption for $\boldsymbol{\epsilon}$ may not be valid, because what is measured is proportional to the number of ions. Then, a Poisson distribution for $p(\boldsymbol{y}|\boldsymbol{x})$ will be a better choice.

Other problems we did not consider in this paper are myopic or blind inverse problems. As a typical example, consider deconvolution problems (1) or (2) where the psfs $h(t)$ or $h(x, y)$ are partially known. For example, we know that they have a 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 $\boldsymbol{y} = \boldsymbol{A}_{\boldsymbol{\theta}}\boldsymbol{x} + \boldsymbol{\epsilon}$. The case where we know exactly the shape but not the gain a is called *auto-calibration* and the case where we only know the support of the psf but not its shape is called *blind deconvolution*. In the first case $\boldsymbol{\theta} = a$ and in the second case $\boldsymbol{\theta} = [h(0), \ldots, h(p)]$. We must note however that, in general, the blind inverse 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 $h \times x$. This means that, to find satisfactory methods for these problems need much more precise 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. Interested readers can refer to the following papers [27,43] for a few examples. We are still working on these points. We have also to mention that we have not yet applied these methods to real data in spectrometry and we are interested and prospective to evaluate them on real data.

References

[1] R.J.E. Cotter, in: Proceedings of the Oxford ACS Symposium Series on Time-of-Flight Mass Spectrometry, Vol. 549, Oxford, UK, 1994.
[2] K. Birkinshaw, Fundamentals of focal plane detectors, J. Mass Spectrom. 32 (1997) 795–806.
[3] G. Demoment, Image reconstruction and restoration: overview of common estimation structure and problems, IEEE Trans. Acoustics, Speech Signal Proceedings ASSP 37 (12) (1989) 2024–2036.
[4] J. Hadamard, Sur les Problems aux Drives Partielles et Leur Signification Physique, Princeton University Bulletein, 13.
[5] M.Z. Nashed, G. Wahba, Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operators equations, SIAM J. Math. Anal. 5 (1974) 974–987.
[6] M.Z. Nashed, Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory, IEEE Trans. Antennas Propagat. 29 (1981) 220–231.
[7] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1995.
[8] J. Skilling, Theory of maximum entropy image reconstruction, in: J.H. Justice (Ed.), Proceedings of the Fourth Max. Ent. Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics, Cambridge University Press, Calgary, 1984.
[9] J. Skilling, in: J. Skilling (Ed.), Maximum Entropy and Bayesian Methods: Classical Maximum Entropy, Kluwer Academic Publishers, Dordrecht, 1989, pp. 45–52.
[10] J.M. Borwein, A.S. Lewis, Duality relationships for entropy-like minimization problems, SIAM J. Control 29 (2) (1991) 325–338.
[11] D. Dacunha-Castelle, F. Gamboa, Maximum d'entropie et problème des moments, Annal. Institut Henri Poincaré 26 (4) (1990) 567–596.
[12] F. Gamboa, Méthode du Maximum d'Entropie sur la Moyenne et Applications, Thèse de Doctorat, Université de Paris-Sud, Orsay, Décembre 1989.
[13] J. Navaza, On the maximum entropy estimate of the electron density function, Acta Crystallogr. A 41 (1985) 232–244.
[14] G. Le Besnerais, 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.
[15] J.-F. Bercher, G. Le Besnerais, G. Demoment, The maximum entropy on the mean method, noise and sensitivity, Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, Cambridge, UK, 1994, pp. 223–232.

[16] J.-F. Bercher, G. Le Besnerais, G. Demoment, Building convex criteria for solving linear inverse problems, in: Proceedings of the International Workshop on Inverse Problems, Ho-Chi-Minh City, Vietnam, 1995, pp. 33–44.
[17] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.
[18] R.T. Rockafellar, Lagrange multipliers and optimality, SIAM Rev. 35 (2) (1993) 183–238.
[19] J.M. Borwein, A.S. Lewis, Partially finite convex programming. Part I. Quasi relative interiors and duality theory, Math. Program. 57 (1992) 15–48.
[20] J.M. Borwein, A.S. Lewis, Partially finite convex programming. Part II. Explicit lattice models, Math. Program. 57 (1992) 49–83.
[21] A. Decarreau, D. Hilhorst, C. Lemaréchal, J. Navaza, Dual methods in entropy maximization: application to some problems in crystallography, SIAM J. Optimiz. 2 (2) (1992) 173–197.
[22] C.A. Bouman, K.D. Sauer, A unified approach to statistical tomography using coordinate descent optimization, IEEE Trans. Image Processing 5 (3) (1996) 480–492.
[23] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Machine Intell. PAMI 6 (6) (1984) 721–741.
[24] A. Mohammad-Djafari, J. Idier, Scale invariant Bayesian estimators for linear inverse problems, in: Proceedings of the First ISBA Meeting, San Francisco, CA, 1993.
[25] S. Brette, J. Idier, A. Mohammad-Djafari, Scale invariant Markov models for Bayesian inversion of linear inverse problems, in: J. Skilling, S. Sibusi (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, Cambridge, UK, 1994, pp. 199–212.
[26] S. Brette, J. Idier, A. Mohammad-Djafari, Scale invariant Markov models for linear inverse problems, in: Proceedings of the Section on Bayesian Statistical Sciences, American Statistical Association, Alicante, 1994, pp. 266–270.
[27] F. Champagnat, Y. Goussard, J. Idier, Unsupervised deconvolution of sparse spike trains using stochastic approximation, IEEE Trans. Signal Processing 44 (12) (1996) 2988–2998.
[28] P. Ciuciu, J. Idier, J.-F. Giovannelli, Regularized estimation of mixed spectra using a circular Gibbs–Markov model, IEEE Trans. Signal Processing 49 (10) (2001) 2201–2213.
[29] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[30] G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (2) (1979) 215–223.
[31] P. Hall, D.M. Titterington, Common structure of techniques for choosing smoothing parameter in regression problems, J. R. Stat. Soc. B 49 (2) (1987) 184–198.
[32] T.J. Hebert, R. Leahy, Statistic-based MAP image reconstruction from Poisson data using Gibbs prior, IEEE Trans. Signal Processing 40 (9) (1992) 2290–2303.
[33] V. Johnson, W. Wong, X. Hu, C.-T. Chen, Image restoration using Gibbs priors: boundary modeling, treatement of blurring, and selection of hyperparameter, IEEE Trans. Pattern Anal. Machine Intell. PAMI 13 (5) (1984) 413–425.
[34] D.M. Titterington, Common structure of smoothing techniques in statistics, Int. Stat. Rev. 53 (2) (1985) 141–170.
[35] L. Younès, Estimation and annealing for Gibbsian fields, Annal. Institut Henri Poincaré 24 (2) (1988) 269–294.
[36] L. Younes, Parametric inference for imperfectly observed Gibbsian fields, Prob. Th. Rel. Fields 82 (1989) 625–645.
[37] C.A. Bouman, K.D. Sauer, Maximum likelihood scale estimation for a class of Markov random fields penalty for image regularization, in: Proceedings of the International Conference on Acoustic, Speech and Signal Processing, Vol. V, 1994, pp. 537–540.
[38] J.A. Fessler, A.O. Hero, Complete data spaces and generalized EM algorithms, in: Proceedings of the International Conference on Acoustic, Speech and Signal Processing, Minneapolis, MN, 1993, pp. 1–4.
[39] K.-Y. Liang, D. Tsou, Empirical Bayes and conditional inference with many nuisance parameters, Biometrika 79 (2) (1992) 261–270.
[40] N. Fortier, G. Demoment, Y. Goussard, GCV and ML methods of determining parameters in image restoration by regularization: fast computation in the spatial domain and experimental comparison, J. Vis. Commun. Image Represent. 4 (2) (1993) 157–170.
[41] A. Mohammad-Djafari, On the estimation of hyperparameters in Bayesian approach of solving inverse problems, in: Proceedings of the International Conference on Acoustic, Speech and Signal Processing, IEEE, Minneapolis, MN, 1993, pp. 567–571.
[42] A. Mohammad-Djafari, A full Bayesian approach for inverse problems, in: K. Hanson, R.N. Silver (Eds.), Kluwer Academic Publishers, Santa Fe, NM, 1996, pp. 135–143.
[43] A. Mohammad-Djafari, N. Qaddoumi, R. Zoughi, A blind deconvolution approach for resolution enhancement of near-field microwave images, in: F. Prêteux, A. Mohammad-Djafari, E. Dougherty (Eds.), Mathematical Modeling, Bayesian Estimation and Inverse Problems, Vol. 3816, SPIE 99, Denver, CO, USA, 1999, pp. 274–281.