Mohammad-Djafari *EURASIP Journal on Advances in Signal Processing* 2012, **2012**:52
http://asp.eurasipjournals.com/content/2012/1/52

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

REVIEW Open Access

Bayesian approach with prior models which enforce sparsity in signal and image processing

Ali Mohammad-Djafari

Abstract

In this review article, we propose to use the Bayesian inference approach for inverse problems in signal and image processing, where we want to infer on sparse signals or images. The sparsity may be directly on the original space or in a transformed space. Here, we consider it directly on the original space (impulsive signals). To enforce the sparsity, we consider the probabilistic models and try to give an exhaustive list of such prior models and try to classify them. These models are either heavy tailed (generalized Gaussian, symmetric Weibull, Student-t or Cauchy, elastic net, generalized hyperbolic and Dirichlet) or mixture models (mixture of Gaussians, Bernoulli-Gaussian, Bernoulli-Gamma, mixture of translated Gaussians, mixture of multinomial, etc.). Depending on the prior model selected, the Bayesian computations (optimization for the joint maximum a posteriori (MAP) estimate or MCMC or variational Bayes approximations (VBA) for posterior means (PM) or complete density estimation) may become more complex. We propose these models, discuss on different possible Bayesian estimators, drive the corresponding appropriate algorithms, and discuss on their corresponding relative complexities and performances.

Keywords: sparsity, Bayesian approach, sparse priors, inverse problems

1 Introduction

In many generic inverse problems in signal and image processing we want to infer on an unknown signal $f(t)$ or an unknown image $f(\boldsymbol{r})$ with $\boldsymbol{r} = (x, y)$ through an observed signal $g(s)$ or an observed image $g(\boldsymbol{s})$ related between them through an operator $\mathcal{H}$ such as convolution $g = h * f$ or any other linear or non linear transformation $g = \mathcal{H}f$. When this relation is linear and we have discretized the problem, we arrive to the relation:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{f} = [f_1, ..., f_n]'$ represents the unknowns, $\boldsymbol{g} = [g_1, ..., g_m]'$ the observed data, $\epsilon = [\epsilon_1, ..., \epsilon_m]'$ the errors of modeling and measurement and $\boldsymbol{H}$ the matrix of the system response. We may note that, even if the noise could be neglected ($\epsilon = 0$) and the matrix $\boldsymbol{H}$ invertible ($m = n$), in general, the solution $\widehat{\boldsymbol{f}} = \boldsymbol{H}^{-1}\boldsymbol{g}$ is not forcibly the good solution, because this solution may be too sensitive to small changes in the data due to the ill-conditioning of this matrix. for the general case of $m \neq n$, one tries to obtain a regularized solution, for example by defining it as the optimizer of a two parts criterion

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}}\{J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda\|\boldsymbol{f}\|^2\} \tag{2}$$

which is given by $\widehat{\boldsymbol{f}} = [\boldsymbol{H}\boldsymbol{H}' + \lambda\boldsymbol{I}]^{-1}\boldsymbol{H}'\boldsymbol{g}$. When the regularization parameter $\lambda = 0$, one gets a generalized inverse $\widehat{\boldsymbol{f}} = [\boldsymbol{H}\boldsymbol{H}']^{-1}\boldsymbol{H}'\boldsymbol{g}$ and when $\boldsymbol{H}$ invertible, one gets the normal inverse solution $\widehat{\boldsymbol{f}} = \boldsymbol{H}^{-1}\boldsymbol{g}$. The regularization theory has been developed since the pioneer work of Tikhonov [1] and Tikhonov and Arsénine [2] who had introduced a quadratic regularization terms to account for some prior properties of the solution (smoothness). Since that, many different regularization terms have been proposed. In particular, in place of L_2 norm: $L_2(\boldsymbol{f}) = \|\boldsymbol{f}\|_2^2 = \sum_j |f_j|^2$, it has been proposed to use the L_0 norm $L_0(\boldsymbol{f}) = \|\boldsymbol{f}\|_0 = \sum_j \delta(f_j)$ or the L_1 norm $L_1(\boldsymbol{f}) = ||\boldsymbol{f}||_1 = \Sigma_j |f_j|$ to enforce the sparsity of the solution [3-11]. Then, due to the fact that $L_0(\boldsymbol{f})$ is not convex and $L_1(\boldsymbol{f})$ is convex, but not continuous, the optimization of a criterion with these expressions becomes more difficult than the L_2 norm case. For this reason, there was a great number of works who

Correspondence: djafari@lss.supelec.fr
Laboratoire des signaux et systèmes (L2S), UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD, SUPELEC, Plateau de Moulon, 91192 Gif-sur-Yvette, France

Springer

© 2012 Mohammad-Djafari; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

specialized in proposing algorithms for the optimization of such criteria.

Interestingly, defining the solution of the problem (1) as the optimization of a criterion with two parts can be assimilated to a maximum a posteriori (MAP) solution in a Bayesian approach where the first term of the criterion (2) can be related to the likelihood and the second term to a prior model as we will see in the following where the main objective is to show how the Bayesian approach can go farther than the regularization in at least the following aspects:

- A better account for the noise term characteristics;
- A better and easier way for translating the prior knowledge and in particular the sparsity;
- New tools for assessing the regularization parameter, a great subject of discussion for all those work with regularization theory;
- New solutions and new tools for doing computations (optimizations and integrations).

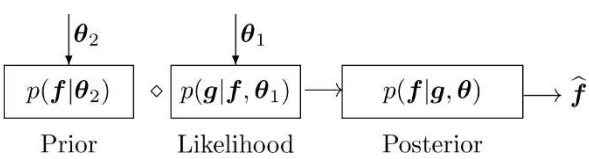
1.1 The Bayesian approach

The Bayesian inference approach is based on the posterior law:

$$p(f\,|g,\theta_1,\theta_2) = \frac{p(g\,|f,\theta_1)\,p(f\,|\theta_2)}{p(g\,|\theta_1,\theta_2)} \propto p(g\,|f,\theta_1)p(f|\theta_2) \tag{3}$$

where the sign $\propto$ stands for "proportional to", $p(g|f, \theta_1)$ is the likelihood, $p(f|\theta_2)$ the prior model, $\theta = (\theta_1, \theta_2)$ are their corresponding parameters (often called the hyper parameters of the problem) and $p(g|\theta_1, \theta_2)$ is called the evidence of the model.

This general Bayesian approach is illustrated as follows:



In this approach, the likelihood $p(g|f, \theta_1)$ summarizes our knowledge about the noise and the model linking the observed data g to the unknowns f and the prior term $p(f|\theta_2)$ summarizes our incomplete prior knowledge about the unknowns and the posterior law $p(f|g, \theta)$ combines these two terms and contains all our state of knowledge about the unknowns f after accounting for the prior and the observed data.

As a very simple example, when the noise is assumed to be Gaussian, then the MAP solution $\widehat{f} = \arg\max_f\{p(f\,|g,\theta)\}$ is obtained as the optimizer of the criterion $J(f) = ||g - Hf||^2 + \lambda\Omega(f)$ where the expression of $\Omega(f)$ depends on the prior law. When the prior knowledge is translated as a Gaussian probability law, then $\Omega(f) = \|f\|_2^2$ and when it is translated as a Laplace probability law, then $\Omega(f) = ||f||_1$ [12-14].

The first interest of using the Bayesian approach to the regularization approach is to have new tools for handling the hyper parameters [15].

1.2 Full Bayesian approach

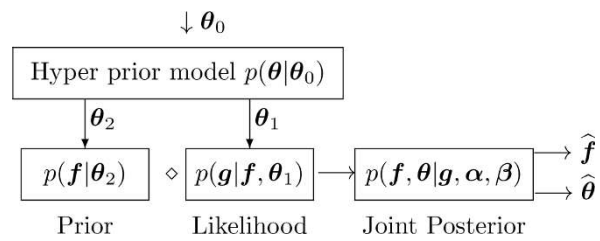
When the parameters θ have to be estimated too, we can assign them a prior $p(\theta|\theta_0)$ with fixed values for θ_0 (often called hyper-hyper-parameters) and express the joint posterior

$$p(f,\theta\,|g,\theta_0) = \frac{p(g\,|f,\theta_1)p(f\,|\theta_2)p(\theta\,|\theta_0)}{p(g\,|\theta_0)} \tag{4}$$

and then try to estimate them jointly, for example joint MAP [16]:

$$(\widehat{f},\widehat{\theta}) = \arg\max_{(f,\theta)}\{p(f,\theta\,|g,\theta_0)\} \tag{5}$$

This Full Bayesian approach is illustrated as follows:



One may also first integrate out one of them, for example f to obtain

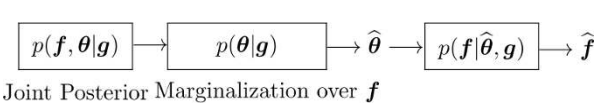
$$p(\theta\,|g,\theta_0) = \int p(f,\theta\,|g,\theta_0)\,\mathrm{d}f, \tag{6}$$

estimate θ, for example by

$$\widehat{\theta} = \arg\max_{\theta}\{p(\theta\,|g,\theta_0)\} \tag{7}$$

and then use it for the estimation of the other one using $p(f\,|g,\widehat{\theta})$.

This approach (called sometimes type II maximum likelihood) is illustrated as follows:



However, very often this marginalization cannot be done analytically and so the optimization for the estimation of θ cannot be achieved. In such cases, the expectation-maximization (EM) algorithms can be helpful [17]. Considering g as *incomplete data*, f as *hidden variable*, (g, f) as *complete data* and noting $\ln p(g|\theta)$ as

incomplete data log-likelihood and $\ln p(\boldsymbol{g}, \boldsymbol{f}|\boldsymbol{\theta})$ *complete data log-likelihood*, the classical EM algorithm writes:

$$\begin{cases} \text{E - step} : q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}_{p(\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)})}\{\ln p(\boldsymbol{g}, \boldsymbol{f}|\boldsymbol{\theta})\} \\ \text{M - step}: \widehat{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}} \left\{ q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k-1)}) \right\} \end{cases} \tag{8}$$

The Bayesian version (Bayesian EM) is not very far and differs only by the introduction of $p(\boldsymbol{\theta})$:

$$\begin{cases} \text{E - step} : q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}_{p(\boldsymbol{f}|\boldsymbol{g}, \widehat{\boldsymbol{\theta}}^{(k)})}\{\ln p(\boldsymbol{g}, \boldsymbol{f}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\} \\ \text{M - step} : \widehat{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}} \left\{ q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k-1)}) \right\} \end{cases} \tag{9}$$

This is illustrated as follows:

$$\boxed{p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\text{EM, Bayes EM}} \longrightarrow \widehat{\boldsymbol{\theta}} \longrightarrow \boxed{p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}}, \boldsymbol{g})} \longrightarrow \widehat{\boldsymbol{f}}$$

As we mentioned before, one of the main steps in the Bayesian approach is the prior modeling which has the role of translating our prior knowledge on the unknown signal or image in a probability law. Sparsity is one of the prior knowledge we may translate. The main objective of this article is to see what are the different possibilities.

1.3 Prior modeling

In this article, we propose different prior modeling for signals and images which can be used in a Bayesian inference approach in many inverse problems in signal and image processing where we want to infer on sparse signals or images. The sparsity may be directly on the original space or in a transformed space (see Figures 1, 2, 3, and 4). In this article, we consider the sparsity directly in the original domain.

The prior models discussed are the following:

- generalized Gaussian (GG) with Gaussian (G) and Laplace or double exponential (DE) as particular cases;
- symmetric Weibull (W) with symmetric Rayleigh (R) and again the DE as particular cases;
- Student-t (St) with Cauchy (C) as particular case;
- Elastic net prior model;
- generalized hyperbolic model;
- Dirichlet and symmetric Dirichlet;
- Mixture of two centered Gaussians (MoG2), one with very small and one with a large variances;
- Bernoulli-Gaussian (BG), also called *Spike and slab*;
- Mixture of two Gammas (MoGamm);
- Bernoulli-Gamma (BGamma);
- Mixture of three Gaussians (MoG3), one centered with very small variance and two symmetrically centered on positive and negative axes and large variances;
- Mixture of one Gaussian and two Gammas (MoG-Gammas), and in a more summary the case of
- Bernoulli-Multinomial (BMult) or mixture of Dirichlet (MoD).

Some of these models are well-known [12-14,18-26], some others less. In general, we can classify them into two categories: (i) simple non Gaussian models with heavy tails and (ii) mixture models with hidden variables which result to hierarchical models.

In the Section 2, we give more details about the sparsity and all these prior models which enforce the sparsity.

1.4 Bayesian computation

The second main step in the Bayesian approach is to do the computations. Depending on the prior model selected, the Bayesian computations needed are:

- For simple prior models:
  - Simple optimization of $p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g})$ for the MAP:

$$\boxed{p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Simple Optimization}\\ \text{Algorithm}\end{array}} \longrightarrow \widehat{\boldsymbol{f}}$$

  - Joint optimization $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ for joint MAP:

$$\boxed{p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Joint Optimization}\\ \text{Algorithm}\end{array}} \begin{array}{l} \longrightarrow \widehat{\boldsymbol{f}} \\ \longrightarrow \widehat{\boldsymbol{\theta}} \end{array}$$

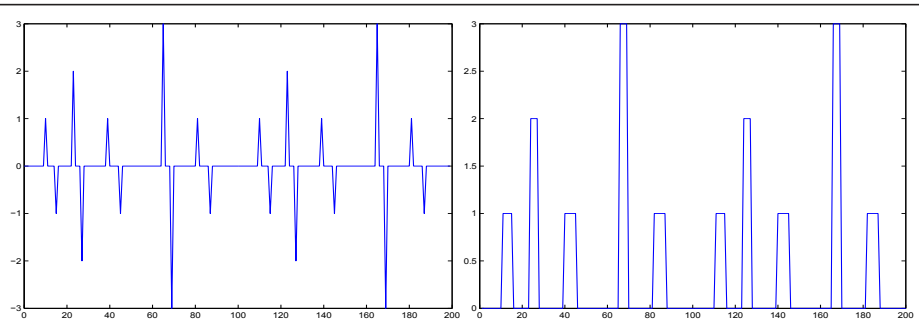


Figure 1 Sparsity: explicite sparse signals. The signal at the right is sparse, but its derivative (signal at the left) is still more sparse.

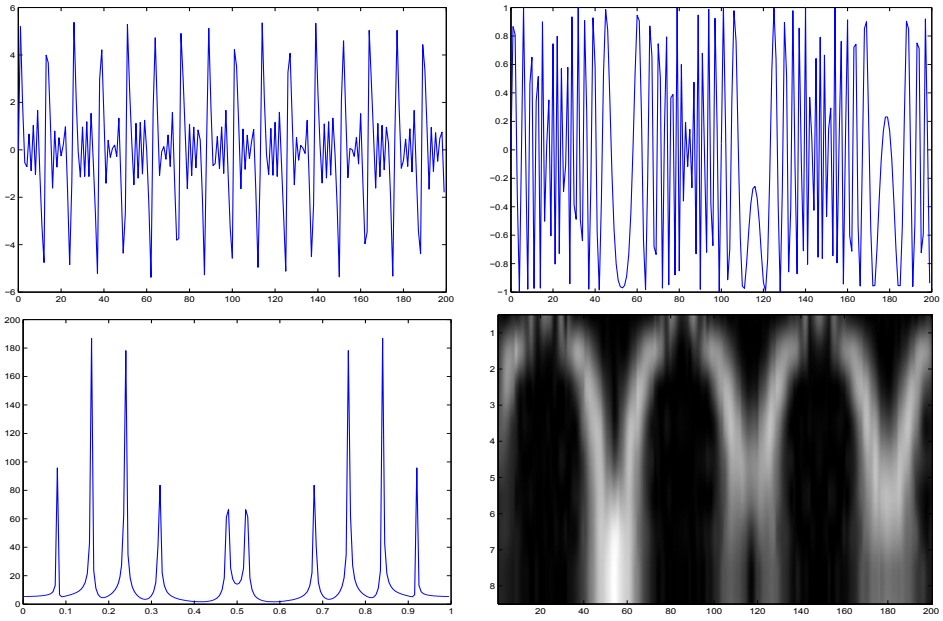


Figure 2 Sparsity: sparse signals in a transformed domaine (Fourier or wavelet). First row: signals, second row: Fourier or wavelet transforms.

- Generation of samples from the conditionals $p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g})$ and $p(\boldsymbol{\theta}|\boldsymbol{f}, \boldsymbol{g})$ for the MCMC Gibbs sampling methods,

$$\boxed{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Gibbs sampling}\\ \boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{g})\\ \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{f},\boldsymbol{g})\end{array}} \begin{array}{l}\longrightarrow \widehat{\boldsymbol{f}}\\ \longrightarrow \widehat{\boldsymbol{\theta}}\end{array}$$

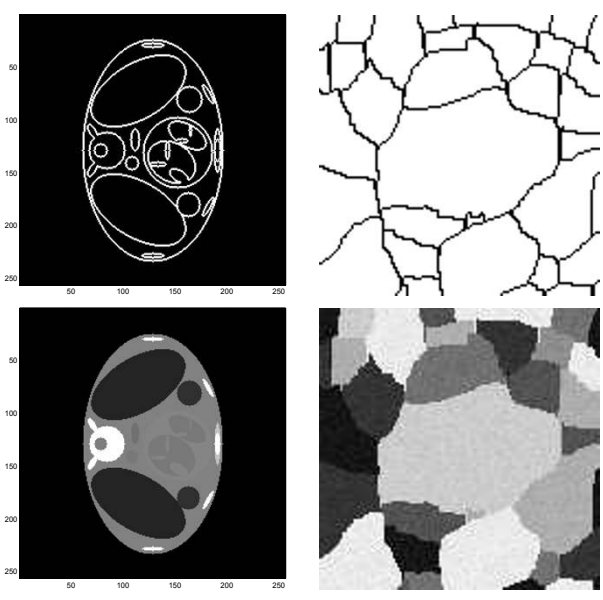


Figure 3 Sparsity: explicite sparse images. The images at the top are sparse. The images at the bottom are not sparse, but their Laplaciens are (images at top).

- Variational approximation (VA) of the joint $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by a separable

$$q(f,\boldsymbol{\theta}\,|\boldsymbol{g}) = q_1(f\,|\widetilde{\boldsymbol{\theta}}\,,\boldsymbol{g})\; q_2(\boldsymbol{\theta}\,|\widetilde{f},\boldsymbol{g})$$

and then using them for estimation

$$\boxed{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Variational}\\ \text{Bayesian}\\ \text{Approximation}\end{array}} \begin{array}{l}\longrightarrow q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\theta}},\boldsymbol{g}) \longrightarrow \widehat{\boldsymbol{f}}\\ \longrightarrow q_2(\boldsymbol{\theta}|\widetilde{\boldsymbol{f}},\boldsymbol{g}) \longrightarrow \widehat{\boldsymbol{\theta}}\end{array}$$

• For hierarchical prior models with hidden variables $\boldsymbol{z}$:

- Joint optimization $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ for joint MAP,

$$\boxed{p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Joint}\\ \text{Optimization}\\ \text{Algorithm}\end{array}} \begin{array}{l}\longrightarrow \widehat{\boldsymbol{f}}\\ \longrightarrow \widehat{\boldsymbol{z}}\\ \longrightarrow \widehat{\boldsymbol{\theta}}\end{array}$$

- Generation of samples from the conditionals $p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g})$, $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{g})$ and $p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{g})$ for the MCMC Gibbs sampling methods:

$$\boxed{p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Gibbs sampling}\\ \boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})\\ \boldsymbol{f} \sim p(\boldsymbol{z}|\boldsymbol{f},\boldsymbol{\theta},\boldsymbol{g})\\ \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{f},\boldsymbol{z},\boldsymbol{g})\end{array}} \begin{array}{l}\longrightarrow \widehat{\boldsymbol{f}}\\ \longrightarrow \widehat{\boldsymbol{z}}\\ \longrightarrow \widehat{\boldsymbol{\theta}}\end{array}$$

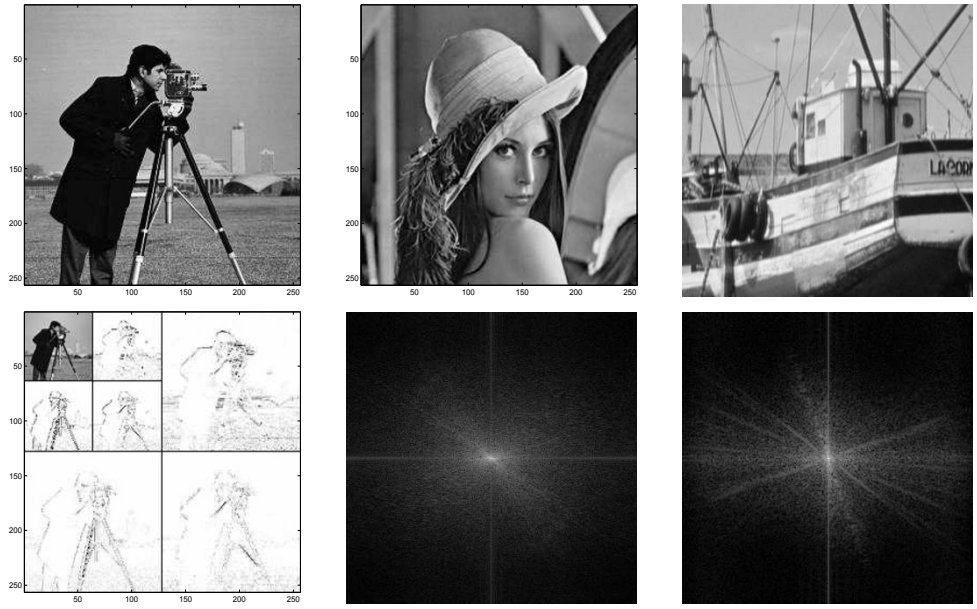


Figure 4 Sparsity: sparse images in a transformed domain (Fourier or wavelet). First row: images, second row: Fourier or wavelet transforms.

- Variational approximation (VA) of the joint $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ by a separable

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

and then using them for estimation

$$\boxed{p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Variational}\\ \text{Bayesian}\\ \text{Approximation}\end{array}} \begin{array}{l} \longrightarrow q_1(\boldsymbol{f}|\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{g}) \longrightarrow \widehat{\boldsymbol{f}} \\ \longrightarrow q_2(\boldsymbol{z}|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{g}) \longrightarrow \widehat{\boldsymbol{z}} \\ \longrightarrow q_3(\boldsymbol{\theta}|\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{f}}, \boldsymbol{g}) \longrightarrow \widehat{\boldsymbol{\theta}} \end{array}$$

The second main objective of this article is to discuss on the relative complexities and performances of the algorithms obtained with the proposed prior law.

The rest of the article is organized as follows:

In Section 2, we present in details the proposed prior models and discuss their properties. For example, we will see that the Student-t model can be interpreted as an infinite mixture with a variance hidden variable or that the BG model can be considered as the degenerate case of a MoG2 where one of the variances go to zero. Also, we will examine the less known models of MoG3 and MoGGammas where the heavy tails are obtained by combining a centered Gaussian and two large variance non-centered Gaussians or Gammas.

In Section 3, we examine the expression of the posterior laws that we obtain using these priors and discuss then on complexity of the Bayesian computation of the algorithms. In particular for the mixture models, we give details of the joint estimation of the signal and the hidden variable as well as the hyper parameters (parameters of the mixtures and the noise) for unsupervised cases.

In Section 4, we give more details on the variational Bayesian approximation method, first for the general case and then for the case of mixture laws and more specifically the case of the Student-t considered as a continuous mixture.

Finally, we present the main conclusions of this article in Section 5.

2 Prior models enforcing sparsity

First, as we mentioned, the sparsity is a property which can be described either directly for the signal itself or after some transformation, for example on the derivative of the signal, or in more general on the coefficients of the projection of the signal on any basis or any set of functions.

Different prior models have been used to enforce sparsity.

2.1 Generalized Gaussian (GG), Gaussian (G) and double exponentials (DE) models

This is the simplest and the most used model (see for example, [27]). Its expression is:

$$p(\boldsymbol{f} \,|\gamma, \beta) = \prod_j \mathcal{GG}(f_j \,|\gamma, \beta) \propto \exp\left\{-\gamma \sum_j |f_j|^\beta\right\} \quad (10)$$

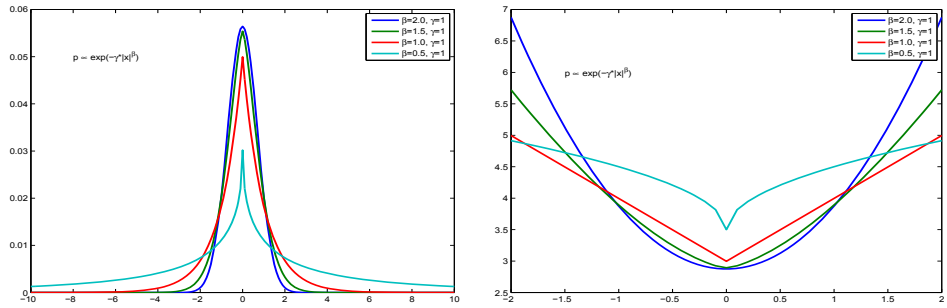


Figure 5 Generalized Gaussian family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

where

$$\mathcal{GG}(f_j\,|\gamma,\beta) = \frac{\beta\gamma}{2\Gamma(1/\beta)}\exp\{-\gamma|f_j|^{\beta}\}. \tag{11}$$

Two particular cases are of importance:

- $\beta = 2$ (Gaussian):

$$p(f\,|\gamma) = \prod_j \mathcal{N}(f_j\,|0, 1/(2\gamma)) \propto \exp\left\{-\gamma\sum_j |f_j|^2\right\} \propto \exp\{-\gamma\,\|f\|_2^2\} \tag{12}$$

- $\beta = 1$ (double exponential or Laplace):

$$p(f\,|\gamma) = \prod_j \mathcal{DE}(f_j\,|\gamma) \propto \exp\left\{-\gamma\sum_j |f_j|\right\} \propto \exp\{-\gamma\,\|f\|_1\} \tag{13}$$

The general shape of these priors are shown in Figure 5, where the cases $\beta = 1$ and $0 < \beta < 1$, which are of great interest for sparsity enforcing are compared to the Gaussian case $\beta = 2$.

2.2 Symmetric Weibull (W) and symmetric Rayleigh (R) models

The second model we consider is the symmetric Weibull probability density function (pdf):

$$p(f\,|\gamma,\beta) = \prod_j \mathcal{W}(f_j\,|\gamma,\beta) \propto \exp\left\{-\gamma\sum_j |f_j|^{\beta} + (\beta-1)\log|f_j|\right\} \tag{14}$$

where

$$\mathcal{W}(f_j\,|\gamma,\beta) = c|f_j|^{(\beta-1)}\exp\{-\gamma|f_j|^{\beta}\} \tag{15}$$

and where $\gamma > 0$ and $\beta > 0$, and the particular cases of $\beta = 1$ is the double exponential and $\beta = 2$ is the symmetric Rayleigh distribution:

$$p(f\,|\gamma,\beta) = \prod_j \mathcal{R}(f_j\,|\gamma) \propto \exp\left\{-\gamma\sum_j |f_j|^2 + \log|f_j|\right\} \tag{16}$$

the cases where $0 < \beta < 1$ are of great interest for sparsity enforcing. This family of models are illustrated on Figure 6.

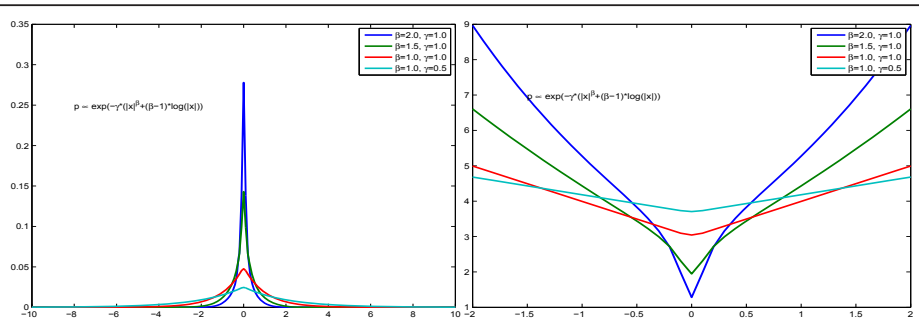


Figure 6 Symmetric Weibull family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

2.3 Student-t (St) and Cauchy (C) models

The second simplest model is the Student-t model:

$$p(f\,|\nu) = \prod_j \mathcal{S}t(f_j\,|\nu) \propto \exp\left\{-\frac{\nu+1}{2}\sum_j \log(1+f_j^2/\nu)\right\} \quad (17)$$

where

$$\mathcal{S}t(f_j\,|\nu) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}(1+f_j^2/\nu)^{-(\nu+1)/2} \quad (18)$$

Knowing that

$$\mathcal{S}t(f_j\,|\nu) = \int_0^\infty \mathcal{N}(f_j\,|0,1/\tau_j)\mathcal{G}(\tau_j\,|\nu/2,\nu/2)\,\mathrm{d}\tau_j \quad (19)$$

we can write this model via the positive hidden variables τ_j :

$$\begin{aligned} p(f,\boldsymbol{\tau}) &= \prod_j p(f_j\,|\tau_j) = \prod_j \mathcal{N}(f_j\,|0,1/\tau_j) \\ &\propto \exp\left\{-\frac{1}{2}\sum_j \tau_j f_j^2\right\} \\ p(\tau_j\,|a,b) &= \mathcal{G}(\tau_j\,|a,b) \propto \tau_j^{(a-1)}\exp\{-b\tau_j\} \\ &\text{with } a=b=\nu/2 \end{aligned} \quad (20)$$

Cauchy model is obtained when $\nu = 1$:

$$p(f) = \prod_j \mathcal{C}(f_j) \propto \exp\left\{-\sum_j \log(1+f_j^2)\right\} \quad (21)$$

This family of models are illustrated on Figure 7.

2.4 Elastic Net (EN) prior model

A prior model inspired from elastic net regression literature [28] is:

$$p(f\,|\nu) = \prod_j \mathcal{EN}(f_j\,|\nu) \propto \exp\left\{-\sum_j(\gamma_1|f_j|+\gamma_2 f_j^2)\right\} \quad (22)$$

where

$$\mathcal{EN}(f_j\,|\nu) = \mathcal{N}(0,1/\gamma_1)\mathcal{DE}(\gamma_1) \propto \exp\left\{-\gamma_1|f_j| - \gamma_2 f_j^2)\right\} \quad (23)$$

which is a product of a Gaussian and a double exponential pdfs. This family of models are illustrated on Figure 8.

2.5 Generalized hyperbolic (GH) prior model

Another general prior model which can be used is:

$$\begin{aligned} p(f\,|\delta,\nu,\beta) = \prod_j (\delta^2+f_j^2)^{(\nu-1/2)/2}\exp\{\beta x\} \\ K_{\nu-1/2}(\alpha\sqrt{\delta^2+f_j^2}) \end{aligned} \quad (24)$$

where $K_{\nu\text{-}1/2}$ is the second kind Bessel function of order (ν - 1/2). This family of models are illustrated on Figure 9.

2.6 Dirichlet (D) and symmetric Dirichlet (SD) models

When f_j are positive and sums to one, we can use the Dirichlet model

$$\mathcal{D}(f\,|\boldsymbol{\alpha}) \propto \prod_j f_j^{\alpha_j-1} \quad \text{with } f_j>0, \quad \sum_j f_j = 1 \quad (25)$$

where $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_N\}$ with $\alpha_j > 0$. The proportionality constant is

$$B(\boldsymbol{\alpha}) = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma\left(\sum_j \Gamma(\alpha_j)\right)} \quad (26)$$

It is noted that the support of this distribution is $[0,1]^N$ and $||\boldsymbol{f}||_1 = \Sigma_j f_j = 1$.

It is also interesting to note that the domain of the Dirichlet distribution is itself a probability distribution, specifically a N-dimensional discrete distribution and the set of points in the support of a N-dimensional Dirichlet distribution is the open standard N - 1-simplex, which is a generalization of a triangle, embedded in the next-higher dimension.

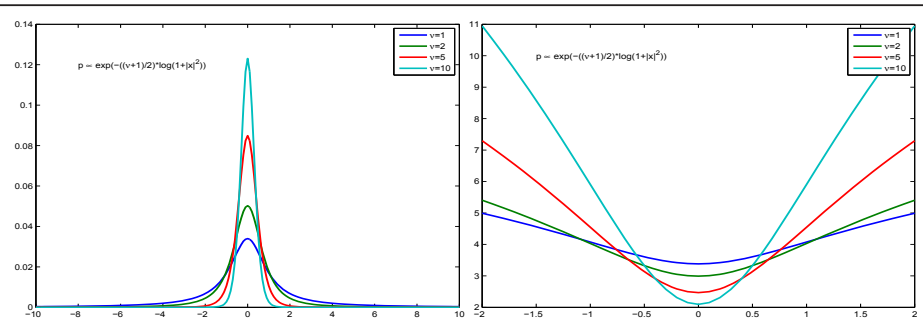


Figure 7 Student-t and Cauchy family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

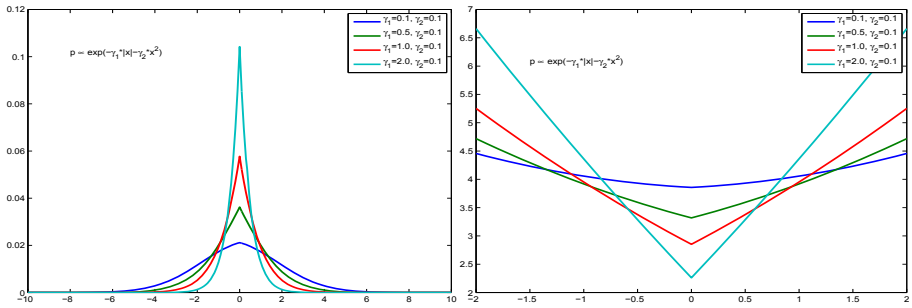


Figure 8 Elastic net family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

A very common special case is the *symmetric Dirichlet* (SD) distribution, where all of the elements making up the parameter vector $\boldsymbol{\alpha}$ have the same value α called the concentration parameter:

$$\mathcal{D}(f\,|\alpha\,) \propto \prod_j f_j^{\alpha-1} \text{ with } f_j > 0, \quad \sum_j f_j = 1 \tag{27}$$

When $\alpha > 1$, the symmetric Dirichlet distribution is equivalent to a uniform distribution over the open standard standard N - 1-simplex, i.e., it is uniform over all points in its support. $\alpha > 1$ prefer variants that are dense, evenly-distributed distributions, i.e., all probabilities f_j returned are similar to each other. $\alpha < 1$ prefer sparse distributions, i.e., most of the probabilities f_j returned will be close to 0, and the vast majority of the mass will be concentrated in a few of them. This is the case on which we are interested. An illustration of this family of models are illustrated on Figure 10.

2.7 Mixture of two Gaussians (MoG2) model

The mixture models are also very commonly used as prior models. In particular the mixture of two Gaussians (MoG2) model:

$$p(f\,|\lambda, \nu_1, \nu_0\,) = \prod_j (\lambda\mathcal{N}(f_j\,|0, \nu_1\,) + (1-\lambda)\mathcal{N}(f_j\,|0, \nu_0\,)) \tag{28}$$

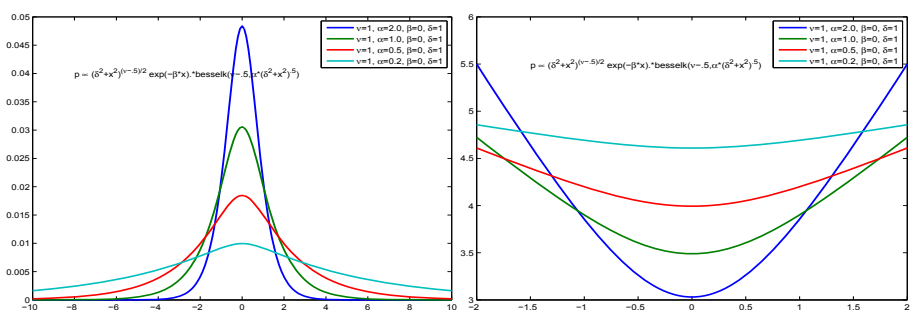


Figure 9 Generalized hyperbolic family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

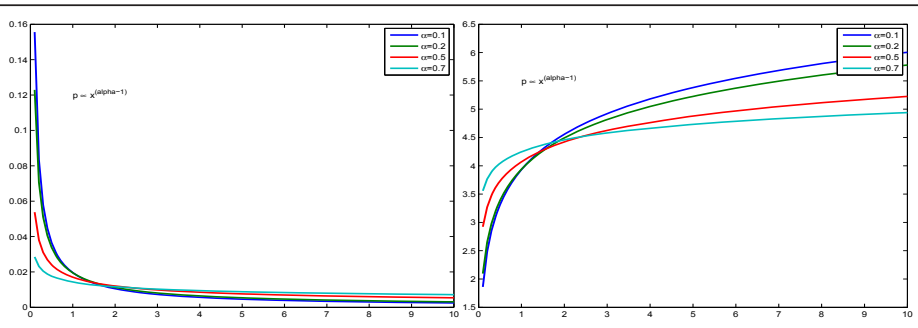


Figure 10 Dirichlet family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

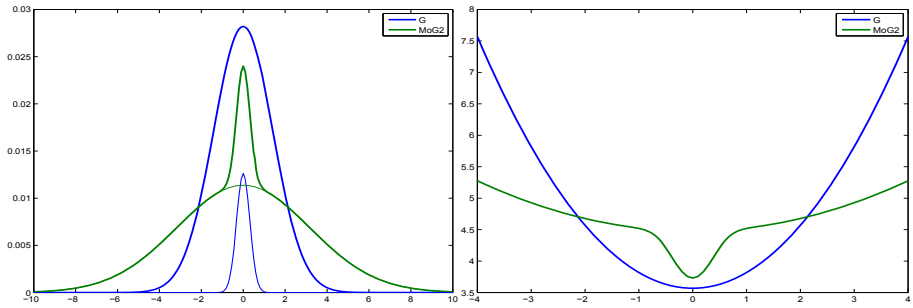


Figure 11 Mixture of two Gaussians family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

which can also be expressed through the binary valued hidden variables $z_j \in \{0,1\}$

$$\begin{cases} p(f\,|z) = \prod_j p(f_j\,|z_j) = \prod_j \mathcal{N}(f_j\,|0, v_{z_j}) \\ \qquad\qquad \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_{z_j}}\right\} \\ P(z_j = 1) = \lambda, \qquad P(z_j = 0) = 1 - \lambda \end{cases} \tag{29}$$

In general $v_1 >> v_0$ and λ measures the sparsity ($0 < \lambda << 1$). This family of models are illustrated on Figure 11.

2.8 Bernoulli-Gaussian (BG) model

The Bernoulli-Gaussian model can be considered as the particular case of the MoG2 with the particular degenerate case of $v_0 = 0$:

$$p(f\,|\lambda, v) = \prod_j p(f_j) = \prod_j (\lambda\mathcal{N}(f_j\,|0, v) + (1-\lambda)\delta(f_j)) \tag{30}$$

which can also be written as

$$\begin{cases} p(f\,|z) = \prod_j p(f_j\,|z_j) = \prod_j [\mathcal{N}(f_j\,|0, v)]^{\delta(z_j)} \\ \qquad\qquad \prod_j [\delta(f_j)]^{\delta(1-z_j)} \\ P(z_j = 1) = \lambda, \qquad P(z_j = 0) = 1 - \lambda \end{cases} \tag{31}$$

This model has also been called *spike and slab*. This family of models are illustrated on Figure 12.

2.9 Mixture of three Gaussians (MoG3) model

Another mixture model proposed is using a Mixture of three Gaussians, one centered at zero and two symmetrically placed:

$$\begin{aligned} p(f\,|\lambda, v_0, v_{+1}, v_{-1}, \beta) = \prod_j \big[&(1-\lambda)\mathcal{N}(f_j\,|0, v_0) \\ &+ (\lambda/2)\mathcal{N}(f_j\,|+\beta, v_{+1}) \\ &+ (\lambda/2)\mathcal{N}(f_j\,|-\beta, v_{-1})\big] \end{aligned} \tag{32}$$

which can also be expressed through the ternary valued hidden variables $z_j \in \{-1, 0, +1\}$

$$\begin{cases} p(f\,|z) = \prod_j p(f_j\,|z_j) = \prod_j \mathcal{N}(f_j\,|z_j\beta, v_{z_j}) \\ P(z_j = 1) = \lambda/2, \\ P(z_j = -1) = \lambda/2, \\ P(z_j = 0) = 1 - \lambda. \end{cases} \tag{33}$$

In general $v_{+1} = v_{-1} = v >> v_0$ and λ measures the sparsity ($0 < \lambda << 1$). This family of models are illustrated on Figure 13.

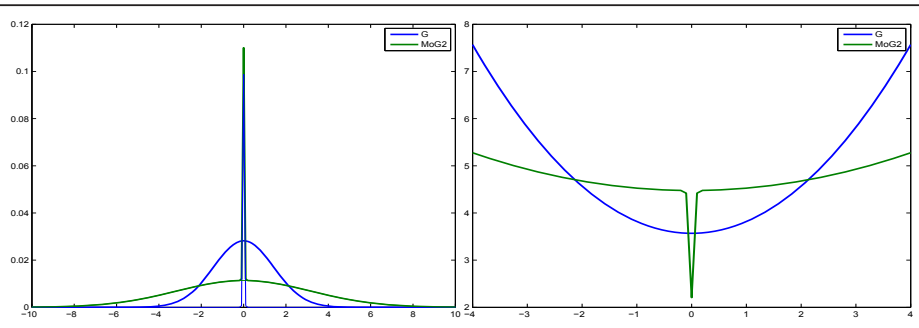


Figure 12 Bernouilli-Gaussian family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

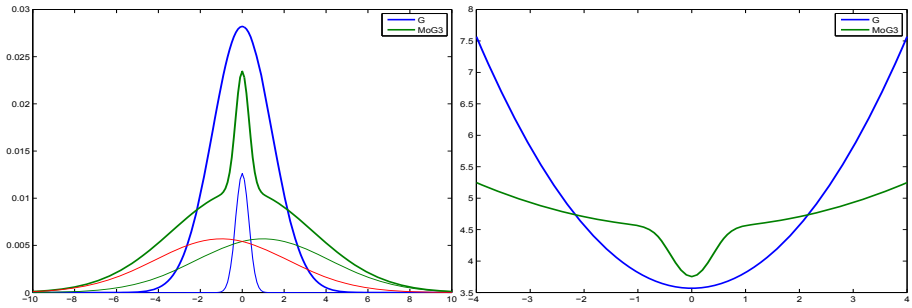


Figure 13 Mixture of three Gaussians family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

2.10 Mixture of one Gaussian and two Gammas (MoGGammas) model

Another mixture model proposed is using a mixture of one central Gaussian and two symmetric Gammas:

$$p(f\,|\lambda, v_0, \alpha, \beta) = \prod_j \big[(1-\lambda)\mathcal{N}(f_j\,|0, v_0) + (\lambda/2)\mathcal{G}(f_j\,|\alpha, \beta) + (\lambda/2)\mathcal{G}(-f_j\,|\alpha, \beta)\big] \tag{34}$$

which can also be expressed through the ternary valued hidden variables $z_j \in \{-1, 0, +1\}$

$$\begin{cases} p(f\,|z) = \prod_j (f_j\,|z_j) = [\mathcal{N}(f_j\,|0, v_0)]^{\sum_j \delta(z_j)} \times [\mathcal{G}(f_j\,|\alpha, \beta)]^{\sum_j \delta(z_j-1)} \times [\mathcal{G}(-f_j\,|\alpha, \beta)]^{\sum_j \delta(z_j+1)} \\ P(z_j = 1) = \lambda/2, \\ P(z_j = -1) = \lambda/2, \\ P(z_j = 0) = 1-\lambda. \end{cases} \tag{35}$$

This family of models are illustrated on Figure 14.

2.11 Bernoulli-Gamma (BGamma) model

As in the BG model, when we want to enforce both sparsity and positivity, we can use the BGamma model:

$$p(f\,|\lambda, \alpha, \beta) = \prod_j [\lambda\delta(f_j) + (1-\lambda)\mathcal{G}(f_j\,|\alpha, \beta)] \tag{36}$$

or

$$\begin{cases} p(f\,|z) = \prod_j p(f_j\,|z_j) = \prod_j [z_j\mathcal{G}(f_j\,|\alpha, \beta)] \prod_j [(1-z_j)\delta(f_j)] \\ P(z_j = 1) = \lambda, \qquad P(z_j = 0) = 1-\lambda \end{cases} \tag{37}$$

A particular case of this model is Bernoulli-exponential (BExponential) which obtained when $\alpha = 1$. These families of models are illustrated on Figure 15 and Figure 16.

2.12 Mixture of Dirichlet (MoD) model

- Mixture of Dirichlet model

$$p(f\,|\lambda, \alpha_1, \alpha_2) = \lambda\mathcal{D}(f\,|\alpha_1) + (1-\lambda)\mathcal{D}(f\,|\alpha_2) \tag{38}$$

where

$$\mathcal{D}(f\,|\alpha) \propto \prod_j f_j^{\alpha-1} \text{ with } f_j > 0, \sum_j f_j = 1 \tag{39}$$

is the symmetric Dirichlet distribution. We need to choose $\alpha_1 > 1$ for dense part and $0 < \alpha_2 < 1$ for the sparse part.

2.13 Bernoulli-multinomial (BMultinomial) model

As in the BG or BGamma model, when we know that the signal is sparse and can only take one of the K

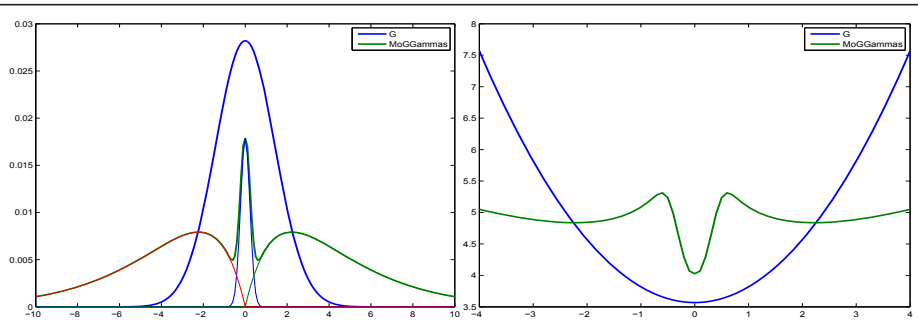


Figure 14 Mixture of one Gaussian and two Gammas family. The probability density function $p(x)$ is shown in the left and - ln$p(x)$ is shown in the right.

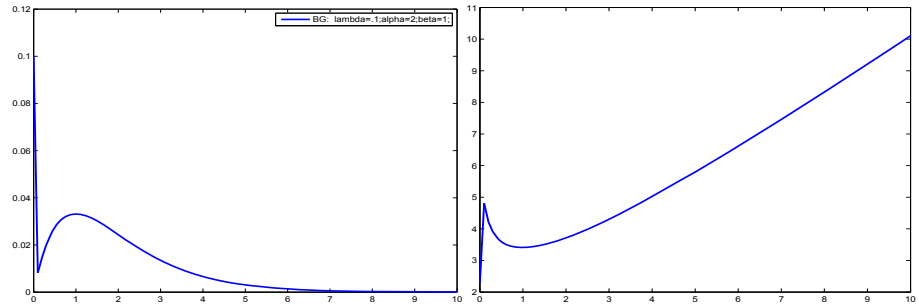


Figure 15 Bernouilli-Gamma family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

discrete values $\{a_1, ..., a_K\}$, we can use the BMultinomial model:

$$p(f\,|\lambda, \boldsymbol{a}, \boldsymbol{\alpha}) = \prod_j \lambda \mathcal{M}ult(f_j\,|\boldsymbol{a}, \boldsymbol{\alpha}) + (1-\lambda)\delta(f_j) \tag{40}$$

where $\boldsymbol{a} = \{a_1, ..., a_K\}$ and $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$ with $\Sigma_k\, \alpha_k = 1$ and

$$\mathcal{M}ult(f_j\,|\boldsymbol{a}, \boldsymbol{\alpha}) = \frac{n!}{a_1! \ldots a_K!} \prod_k \alpha_k^{a_j}$$

or

$$\begin{cases} p(f\,|\boldsymbol{z}) = \prod_j p(f_j\,|z_j) = \prod_j \left[z_j \mathcal{M}ult(f_j\,|\boldsymbol{\alpha})\right] \\ \qquad\qquad\qquad\qquad \prod_j \left[(1-z_j)\delta(f_j)\right] \\ P(z_j = 1) = \lambda, \qquad\qquad P(z_j = 0) = 1-\lambda \end{cases} \tag{41}$$

3 Bayesian inference with sparsity enforcing priors

The priors proposed can be used in a Bayesian approach to infer on $\boldsymbol{f}$ given the observed data $\boldsymbol{g}$ through the posterior law given in Equation (3). First let assume the error ϵ to be centered, Gaussian and white: $\varepsilon \sim \mathcal{N}(\varepsilon\,|0, v_\varepsilon \boldsymbol{I})$. Then, using the forward model (1) we have

$$p(g\,|f) = \mathcal{N}(\boldsymbol{H}f, v_\varepsilon \boldsymbol{I}) \propto \exp\left\{\frac{1}{2v_\varepsilon}\|g - \boldsymbol{H}f\|^2\right\} \tag{42}$$

Now, we consider different priors.

3.1 Simple prior models

Given $p(\boldsymbol{g}|\boldsymbol{f})$ and any simple prior law $p(\boldsymbol{f})$, the posterior law is written:

$$p(f\,|g) \propto p(g\,|f)p(f) \propto \exp\{J(f)\} \tag{43}$$

with

$$J(f) = \frac{1}{2v_\varepsilon}\|g - \boldsymbol{H}f\|^2 + \Omega(f) \tag{44}$$

where $\Omega(\boldsymbol{f})$ = -ln $p(\boldsymbol{f})$ and so the Maximum A Posteriori (MAP) solution is expressed as the minimizer of this criterion which has two parts: the first part is due to the likelihood and the second part is due to the prior:

$$\boxed{p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g})} \longrightarrow \boxed{\begin{array}{c}\text{Optimization of} \\ J(\boldsymbol{f}) = \frac{1}{2v_\varepsilon}\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \Omega(\boldsymbol{f}))\end{array}} \longrightarrow \widehat{\boldsymbol{f}}$$

Thus, depending on the choice of the prior we obtain different expressions for $\Omega(\boldsymbol{f})$. For example for the GG model of (10) we get

$$\Omega(f) = \gamma \sum_j |f_j|^\beta). \tag{45}$$

For the symmetric Weibull model (14) we get

$$\Omega(f) = -\gamma \sum_j |f_j|^\beta + (\beta - 1)\log|f_j|\,. \tag{46}$$

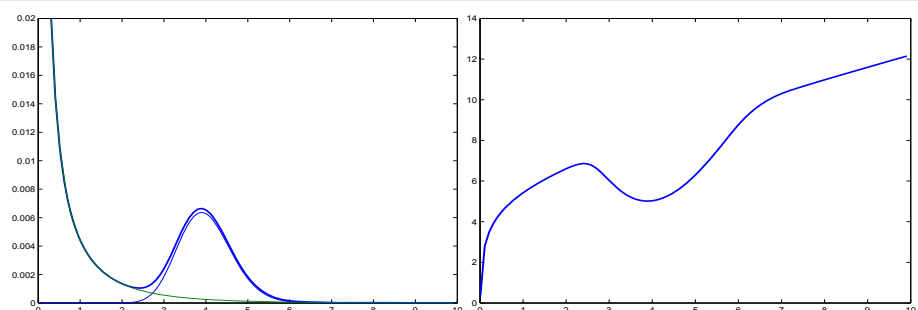


Figure 16 Mixture of 2 Gammas family. The probability density function $p(x)$ is shown in the left and - ln $p(x)$ is shown in the right.

For the Student-t model (17) we get

$$\Omega(f) = \frac{\nu+1}{2}\sum_j \log(1+f_j^2/\nu). \tag{47}$$

For the elastic net model we get

$$\Omega(f) = \sum_j \left[\gamma_1 |f_j| + \gamma_2 f_j^2\right] \tag{48}$$

and for the Dirichlet model we get

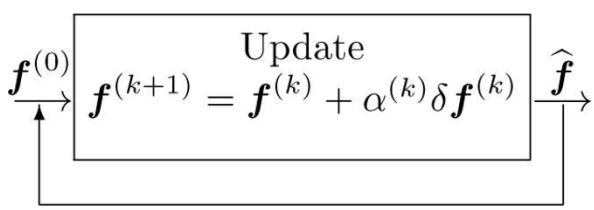
$$\Omega(f) = \sum_j f_j^{\alpha-1}, \quad f_j > 0, \quad \sum_j f_j = 1. \tag{49}$$

For each of these cases, we may discuss on the unimodality and convexity of the criterion $J(\boldsymbol{f})$ which depends mainly on its Hessian

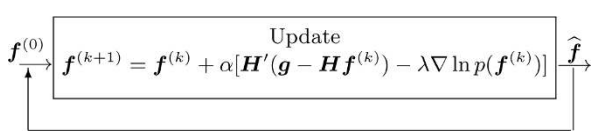
$$\begin{aligned}\boldsymbol{\Delta} J(f) &= \left[\frac{\partial^2 J(f)}{\partial f_j \partial f_i}\right] = \boldsymbol{H}'\boldsymbol{H} + \left[\frac{\partial^2 \Omega(f)}{\partial f_i \partial f_j}\right] \\ &= \boldsymbol{H}'\boldsymbol{H} + \left[\frac{\partial^2 \Omega(f)}{\partial f_j^2}\right]\end{aligned} \tag{50}$$

We may look at each case to examine the range of the parameters for which this Hessian matrix is positive definite.

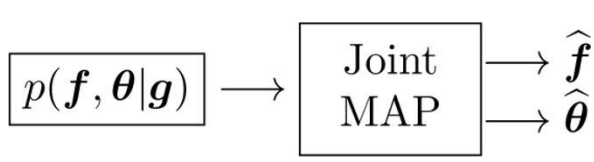
The optimization is done iteratively:



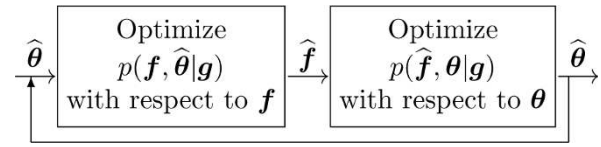
Update operation can be additive, multiplicative or more complex. Updating steps $\alpha^{(k)}$ can be fixed or computed adaptively at each step (steepest descent for example). $\delta\boldsymbol{f}^{(k)}$ can be, for example proportional to the gradient, in which case, we have



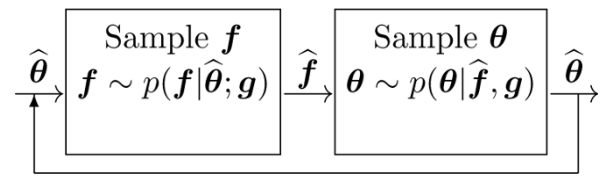
We may also consider to estimate some of these parameters by assigning them appropriate priors and then express the joint $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}, \boldsymbol{\theta}_0)$ as given in Equation (4) and then try to estimate them jointly, for example joint MAP:



or alternate optimization:

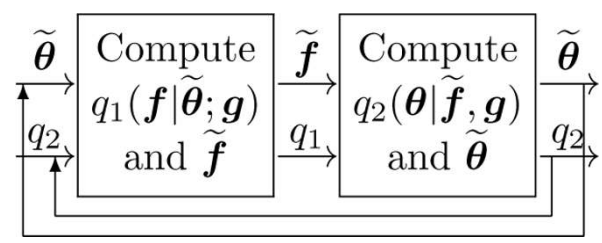


We may also want to explore this joint posterior by generating samples from it. This can be done, for example, through the following Gibbs sampling scheme:



When a great number of samples are thus generated, we may compute their means, variances or any other statistics about them.

Finally, we may try to approximate this joint posterior by a simpler one, for example by a separable $q(\boldsymbol{f}, \boldsymbol{\theta}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{\theta})$ using the variational approximation (VA). The main idea and the main basic steps to achieve this is more detailed in the following section. Here, however, we present the result on the following scheme:



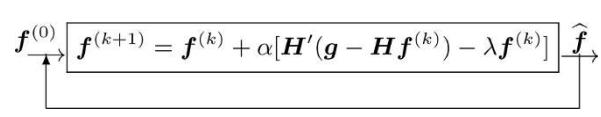
To illustrate the differences, we may consider the simple case of a linear forward model and Gaussian priors:

$$\begin{cases} p(\boldsymbol{g}|f, v_\varepsilon) = \mathcal{N}(\boldsymbol{H}f, v_\varepsilon \boldsymbol{I}) \\ p(f|v_f) = \mathcal{N}(0, v_f \boldsymbol{I}) \end{cases} \tag{51}$$

In this case, if we know $\boldsymbol{\theta} = (v_\epsilon, v_f)$, then

$$\begin{aligned} &p(f|\boldsymbol{g}, v_\varepsilon, v_f) = \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}) \\ &\text{with:} \\ &\widehat{\mu} = [\boldsymbol{H}'\boldsymbol{H} + \lambda I]^{-1}\boldsymbol{H}'\boldsymbol{g} \\ &\widehat{\Sigma} = [\boldsymbol{H}'\boldsymbol{H} + \lambda I]^{-1} \end{aligned} \tag{52}$$

with $\lambda = \frac{v_\varepsilon}{v_f}$. So, we have $\widehat{f} = \widehat{\boldsymbol{\mu}}$ which can be computed by optimizing $J(\boldsymbol{f}) = ||\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}||^2 + \lambda||\boldsymbol{f}||^2$. A gradient based algorithm is shown below:



Now putting inverse Gamma priors on ν_ϵ and ν_f, or equivalently Gamma priors on $\tau_\epsilon = 1/\nu_\epsilon$ and $\tau_f = 1/\nu_f$:

$$\begin{cases} p(\tau_\varepsilon \,|\alpha_{\tau 0}, \beta_{\tau 0}) = \mathcal{G}(\alpha_{\varepsilon 0}, \beta_{\varepsilon 0}) \\ p(\tau_f \,|\alpha_{f0}, \beta_{f0}) = \mathcal{G}(\alpha_{f0}, \beta_{f0}) \end{cases} \tag{53}$$

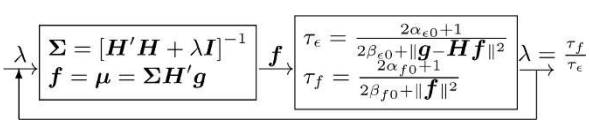
we have

$$\begin{cases} p(\tau_\varepsilon \,|\boldsymbol{f}, \boldsymbol{g}, \alpha_{\tau 0}, \beta_{\tau 0}) = \mathcal{G}(\widehat{\alpha}_\varepsilon, \widehat{\beta}_\varepsilon) \\ p(\tau_f \,|\boldsymbol{f}, \alpha_{f0}, \beta_{f0}) = \mathcal{G}(\widehat{\alpha}_f, \widehat{\beta}_f) \end{cases} \tag{54}$$

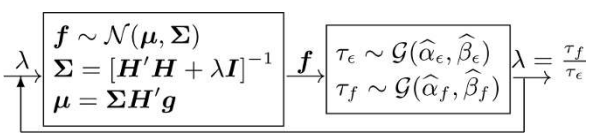
with

$$\begin{cases} \widehat{\alpha}_\varepsilon = \alpha_{\varepsilon 0} + 1/2 \\ \widehat{\beta}_\varepsilon = \beta_{\varepsilon 0} + \left\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\right\|^2/2 \\ \widehat{\tau}_\varepsilon = \widehat{\alpha}_\tau/\widehat{\beta}_\tau = \dfrac{2\alpha_{\varepsilon 0} + 1}{2\beta_{\varepsilon 0} + \left\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\right\|^2} \\ \widehat{\alpha}_f = \alpha_{f0} + 1/2 \\ \widehat{\beta}_f = \beta_{f0} + \left\|\boldsymbol{f}\right\|^2/2 \\ \widehat{\tau}_f = \widehat{\beta}_f/\widehat{\alpha}_f = \dfrac{2\alpha_{f0} + 1}{2\beta_{f0} + \left\|\boldsymbol{f}\right\|^2} \end{cases} \tag{55}$$

and $\widehat{\lambda} = \dfrac{\widehat{\tau}_f}{\widehat{\tau}_\varepsilon}$. Then, the alternate optimization of the JMAP estimate algorithm becomes



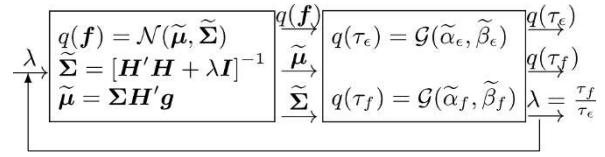
The Gibbs sampling algorithm becomes



The VBA algorithm becomes

$$\begin{cases} q(f) = \mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}) \\ \widetilde{\Sigma} = [\boldsymbol{H}'\boldsymbol{H} + \lambda \boldsymbol{I}]^{-1} \\ \widetilde{\boldsymbol{\mu}} = \widetilde{\Sigma}\boldsymbol{H}'\boldsymbol{g} \\ \\ q(\tau_\varepsilon) = \mathcal{G}(\widetilde{\alpha}_\varepsilon, \widetilde{\beta}_\varepsilon) \\ \widetilde{\alpha}_\varepsilon = \alpha_{\varepsilon 0} + 1/2 \\ \widetilde{\beta}_\varepsilon = \beta_{\varepsilon 0} + \left\|\boldsymbol{g} - \boldsymbol{H} < f >\right\|^2/2 \\ \\ q(\tau_f) = \mathcal{G}(\widetilde{\alpha}_f, \widetilde{\beta}_f) \\ \widetilde{\alpha}_f = \alpha_{f0} + 1/2 \\ \widetilde{\beta}_f = \beta_{f0} + \dfrac{1}{2}\left\|\langle f^2 \rangle\right\| = \beta_{f0} + \dfrac{1}{2}\sum_j \langle f_j^2 \rangle \\ \\ < f >= \widetilde{\boldsymbol{\mu}} \\ \langle f^2 \rangle = \widetilde{\boldsymbol{\mu}}^2 + \mathrm{diag}\left[\widetilde{\Sigma}\right] \end{cases} \tag{56}$$

and $\widehat{\lambda} = \dfrac{\widehat{\tau}_f}{\widehat{\tau}_\varepsilon}$:



We recently implemented these algorithms for different applications such as: synthetic aperture radar (SAR) Imaging [29], ...

3.2 Mixture models

For the mixture models, and in general for the models which can be expressed via the hidden variables, we want to estimate jointly the original unknowns $\boldsymbol{f}$ and the hidden variables: $\boldsymbol{\tau}$ in Cauchy model, $\boldsymbol{z}$ in MoG2, BG or BGam models and $\boldsymbol{z}$ in MoG3 or MoGGammas. Let examine these a little in details.

3.3 Student-t and Cauchy models

In this case the joint prior law can be written as:

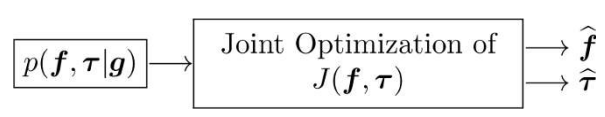
$$\begin{aligned} p(f, \boldsymbol{\tau}) &= \prod_j p(f_j \,|\tau_j) p(\tau_j) = \prod_j \mathcal{N}(f_j \,|0, 1/\tau_j) p(\tau_j) \\ &\propto \exp\left\{-\frac{1}{2}\sum_j \tau_j f_j^2 + a \ln \tau_j - b\tau_j\right\} \text{with } a = b = \nu/2 \end{aligned} \tag{57}$$

such that

$$p(f, \boldsymbol{\tau} \,|\boldsymbol{g}) \propto p(\boldsymbol{g} \,|f) p(f, \boldsymbol{\tau}) \propto \exp\{-J(f, \boldsymbol{\tau})\} \tag{58}$$

where

$$j(f, \boldsymbol{\tau}) = \frac{1}{2\nu_\varepsilon}\left\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\right\|^2 + \sum_j \frac{1}{2}\tau_j f_j^2 - a \ln \tau_j + b\tau_j \tag{59}$$



Joint optimization of this criterion, alternatively with respect to $\boldsymbol{f}$ (with fixed $\boldsymbol{\tau}$)

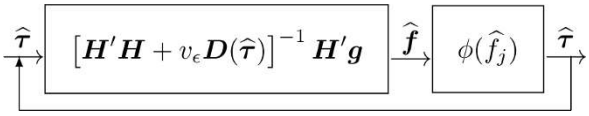
$$\begin{aligned} \widehat{f} &= \arg\min_f \{J(f, \boldsymbol{\tau})\} \\ &= \arg\min_f \left\{\frac{1}{2\nu_\varepsilon}\left\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\right\|^2 + \sum_j \frac{1}{2}\tau_j f_j^2\right\} \end{aligned} \tag{60}$$

and with respect to $\boldsymbol{\tau}$ (with fixed $\boldsymbol{f}$)

$$\begin{aligned} \widehat{\boldsymbol{\tau}} &= \arg\min_{\boldsymbol{\tau}} \{J(f, \boldsymbol{\tau})\} \\ &= \arg\min_{\boldsymbol{\tau}} \left\{\sum_j \frac{1}{2}\tau_j f_j^2 - a \ln \tau_j + b\tau_j\right\} \end{aligned} \tag{61}$$

results in the following iterative algorithm:

$$\begin{cases} \widehat{f} = [H'H + v_\varepsilon D(\widehat{\tau})]^{-1} H'g \\ \widehat{\tau}_j = \phi(\widehat{f}_j) = \dfrac{a}{\widehat{f}_j^2 + b} \\ D(\widehat{\tau}) = \mathrm{diag}[1/\widehat{\tau}_j, j = 1, ..., n] \end{cases} \tag{62}$$



Note that, τ_j is the inverse of a variance and we have $1/\tau_j = \dfrac{f_j^2 + b}{a}$. We can interpret this as an iterative quadratic regularization inversion followed by the estimation of variances τ_j which are used in the next iteration to define the variance matrix $D(\tau)$.

Here too, we may study the conditions on which the joint criterion is uni-modal and its alternate optimization converges to its unique solution.

We may also consider a Gibbs sampling scheme

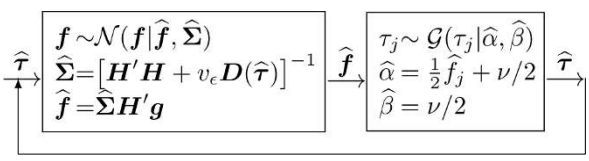
$$\begin{cases} f \sim p(f \,|\, \tau, g) \propto p(g \,|\, f) p(f \,|\, u) = \mathcal{N}(f \,|\, \widehat{f}, \widehat{\Sigma}) \\ \tau \sim p(\tau \,|\, f, g) \propto p(f \,|\, \tau) p(\tau) = \prod_j \mathcal{G}(\tau_j \,|\, \widehat{\alpha}, \widehat{\beta}) \end{cases} \tag{63}$$

where

$$\begin{cases} \widehat{\Sigma} = [H'H + v_\varepsilon D(\tau)]^{-1} \\ \widehat{f} = \widehat{\Sigma} H'g \end{cases} \tag{64}$$

and

$$\begin{cases} \widehat{\alpha} = \dfrac{1}{2}\widehat{f}_j + a = \dfrac{1}{2}\widehat{f}_j + \nu/2 \\ \widehat{\beta} = b \qquad\qquad = \nu/2 \end{cases} \tag{65}$$

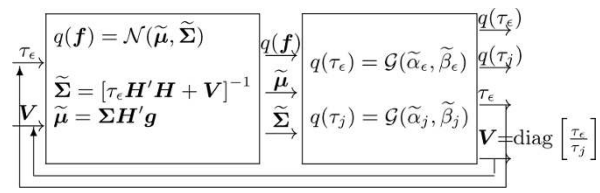


For the VBA, we have

$$\begin{cases} p(g \,|\, f, v_\varepsilon) = \mathcal{N}(g \,|\, Hf, v_\varepsilon I), \quad \tau_\varepsilon = 1/v_\varepsilon \\ p(\tau_\varepsilon) = \mathcal{G}(\tau_\varepsilon \,|\, \alpha_{\varepsilon 0}, \beta_{\varepsilon 0}) \\ p(f \,|\, v) = \prod_j p(f_j \,|\, v_j) = \prod_j \mathcal{N}(f_j \,|\, 0, v_j) = \mathcal{N}(f \,|\, 0, V) \\ V = \mathrm{diag}[v], \quad \tau_j = 1/v_j, \quad \tau = \mathrm{diag}[\tau] = V^{-1} \\ p(\tau) = \prod_j \mathcal{G}(\tau_j \,|\, \alpha_0, \beta_0) \end{cases} \tag{66}$$

$$\begin{cases} \widetilde{q}(f) = \mathcal{N}(f \,|\, \widetilde{\mu}, \widetilde{\Sigma}) \\ \widetilde{\mu} = \widetilde{\Sigma} H'g \\ \widetilde{\Sigma} = (\widetilde{\tau}_\varepsilon H'H + \widetilde{V})^{-1}, \text{with } \widetilde{V} = \mathrm{diag}[\widetilde{v}] \end{cases} \tag{67}$$

$$\begin{cases} \widetilde{q}(\tau_\varepsilon) = \mathcal{G}(\tau_\varepsilon \,|\, \widetilde{\alpha}_\varepsilon, \widetilde{\beta}_\varepsilon), \\ \widetilde{\alpha}_\varepsilon = \alpha_{\varepsilon 0} + (n+1)/2 \\ \widetilde{\beta}_\varepsilon = \beta_{\varepsilon 0} + 1/2 \\ \widetilde{\tau} = \widetilde{\alpha}_\tau / \widetilde{\beta}_\tau \\ \widetilde{q}(\tau_j) = \mathcal{G}(\tau_j \,|\, \widetilde{\alpha}_j, \widetilde{\beta}_j) \\ \widetilde{\alpha}_j = \alpha_{00} + 1/2 \\ \widetilde{\beta}_j = \beta_{00} + < f_j^2 > /2 \\ \widetilde{z}_j = \widetilde{\beta}_j / \widetilde{\alpha}_j \end{cases} \tag{68}$$



3.4 Mixture of two Gaussians (MoG2) model

In this case, following the same arguments, we obtain:

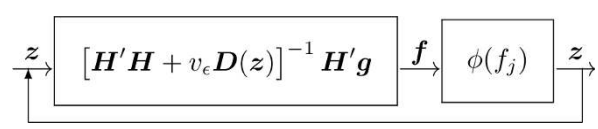
$$p(f, z \,|\, g) \propto p(g \,|\, f) p(f, z) \propto \exp\{-J(f, z)\} \tag{69}$$

where

$$\begin{aligned} J(f, z) = & \frac{1}{2v_\varepsilon} \|g - Hf\|^2 \\ & + \sum_j \frac{f_j^2}{2v_{z_j}} + z_j \ln \lambda + (1 - z_j) \ln(1 - \lambda) \end{aligned} \tag{70}$$

Again, in this case also, the optimization of this criterion, alternatively with respect to f and z results in the following iterative algorithm:

$$\begin{cases} \widehat{f} = [H'H + v_\varepsilon D(z)]^{-1} H'g \\ \widehat{z}_j = \phi(f_j) = \begin{cases} 1, \text{ if } \widehat{f}_j^2 \geq (v_1 - v_0) \ln \dfrac{1-\lambda}{\lambda} \\ 0, \text{ if } \widehat{f}_j^2 < (v_1 - v_0) \ln \dfrac{1-\lambda}{\lambda} \end{cases} \\ D(\widehat{z}) = \mathrm{diag}\left[v_{\widehat{z}_j}, j = 1, .., n\right] \end{cases} \tag{71}$$



Here too, we may also consider a Gibbs sampling scheme

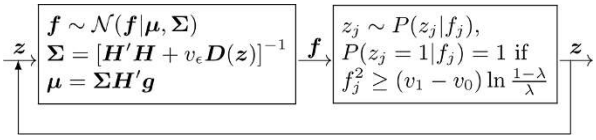
$$\begin{cases} f \sim p(f \,|\, z, g) \propto p(g \,|\, f) p(f \,|\, u) = \mathcal{N}(f \,|\, \widehat{f}, \widehat{\Sigma}) \\ z \sim p(z \,|\, f, g) \propto p(f \,|\, z)\, p(z) \qquad = \prod_j P(z_j = k \,|\, f_j) \end{cases} \tag{72}$$

where

$$\begin{cases} \widehat{\Sigma} = [H'H + v_\varepsilon D(z)]^{-1} \\ \widehat{f} = \widehat{\Sigma} H'g \end{cases} \tag{73}$$

and

$$\begin{cases} P(z_j = 1 \,|f_j) = 1, \text{ if } f_j^2 \geq (v_1 - v_0) \ln \dfrac{1-\lambda}{\lambda} \\ P(z_j = 0 \,|f_j) = 1, \text{if } f_j^2 < (v_1 - v_0) \ln \dfrac{1-\lambda}{\lambda} \end{cases} \tag{74}$$



3.5 BG model

For the case of BG we have to be more careful, because the joint probability laws are degenerated. Two approaches are then possible:

i) Considering them as the particular case of the MoG models where the variance v_0 is fixed to a small value or reduced gradually during the iterations.

ii) Trying first to integrate out $\boldsymbol{f}$ from the expression of $p(\boldsymbol{f}, \boldsymbol{z}|\boldsymbol{g})$ to obtain $p(\boldsymbol{z}|\boldsymbol{g})$ and optimize it with respect to $\boldsymbol{z}$ (detection step) and then use it for the estimation step.

To go further in detail of the second approach, we may remark that for the given $\boldsymbol{z}$, the expression of $p(\boldsymbol{f}, \boldsymbol{z}|\boldsymbol{g})$ as a function of $\boldsymbol{f}$ is Gaussian and so it can be easily integrated out and we obtain:

$$\begin{aligned} p(z\,|g) &\propto p(g\,|z)p(z) \\ &\propto \mathcal{N}(g\,|0\,, \boldsymbol{H}(v\text{diag}[z_j, j = 1,\ldots,n])\boldsymbol{H}' + v_\varepsilon \boldsymbol{I}) \\ &\quad \lambda^{\sum_j z_j}(1-\lambda)^{\sum_j (1-z_j)} \end{aligned} \tag{75}$$

Now writing the expression of $\mathcal{L}(\boldsymbol{z}) = -\ln p(z\,|g)$ and keeping only all terms depending on $\boldsymbol{z}$ we obtain:

$$\mathcal{L}(\boldsymbol{z}) = -\boldsymbol{g}'\boldsymbol{B}^{-1}(\boldsymbol{z})\boldsymbol{g} - \ln|\boldsymbol{B}(\boldsymbol{z})| - 2n \ln \frac{1-\lambda}{\lambda} \tag{76}$$

where $\boldsymbol{B}(\boldsymbol{z}) = \boldsymbol{H}(v\text{diag}\,[z_j, j = 1, \ldots, n])\boldsymbol{H}' + v_\epsilon \boldsymbol{I}$. We see the complexity of this expression which needs the inversion of the matrix $\boldsymbol{B}$ and its optimization which is a combinatorial optimization needing to evaluate this expression 2^n times.

However, we may also remark that when $\boldsymbol{z}$ obtained, the estimation of $\boldsymbol{f}$ is easy. We have:

$$\widehat{f} = \boldsymbol{H}'\boldsymbol{B}^{-1}\boldsymbol{g}. \tag{77}$$

which needs again the inversion of the matrix $\boldsymbol{B}$.

The exact computations of $\widehat{z}$ and $\widehat{f}$ are often too costly, one may try to obtain approximate solutions. Many approximations have been proposed. A good overview of these methods can be found in [30, Chap. 5] and also in [31,32].

3.6 BGamma and MoGGammas model

In these cases, it is no more possible to integrate out $\boldsymbol{f}$ analytically as it was the case with Gaussians. One strategy here is to use the MCMC methods to generate samples from the joint posterior. The second approach is to approximate the joint posterior by a simpler one, for example by a separable one on $\boldsymbol{f}$ and the hidden variables $\boldsymbol{z}$ in the BGamma or the MoGGammas cases. Very often then we can do the computations analytically. However, it may happens that, even after these separable approximations, still we need to use the MCMC methods on some of variables. Detailed explanation of these general methods is out of focus of this article. See [30,33,34]. Here, we just give the details for the case of the Gaussian mixtures (MoG2 or MoG3).

4 Variational Bayesian approximation for the case of mixture laws

To start and to be complete as to propose an unsupervised method, we include also the estimation of the parameters $\boldsymbol{\theta}$ and write the joint posterior law of all the unknowns:

$$p(f, z, \boldsymbol{\theta}\,|g) \propto p(g\,|f, \boldsymbol{\theta})p(f\,|z\,, \boldsymbol{\theta})p(z\,|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{78}$$

which can also be written as

$$q(f, z, \boldsymbol{\theta}\,|g) = p(f\,|z, \boldsymbol{\theta}; g)p(z\,|\boldsymbol{\theta}; g)p(\boldsymbol{\theta}\,|g) \tag{79}$$

where

$$p(f\,|z\,, \boldsymbol{\theta}; g) = p(g\,|f, \boldsymbol{\theta})p(f\,|z\,, \boldsymbol{\theta})/p(g\,|z\,, \boldsymbol{\theta}) \tag{80}$$

with

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

and

$$p(z\,|\boldsymbol{\theta}; g) = p(g\,|z, \boldsymbol{\theta})p(z\,|\boldsymbol{\theta})/p(g\,|\boldsymbol{\theta}) \tag{81}$$

with

$$p(g\,|\boldsymbol{\theta}) = \int p(g\,|z, \boldsymbol{\theta})p(z\,|\boldsymbol{\theta})\,\mathrm{d}z$$

or

$$p(g\,|\boldsymbol{\theta}) = \sum_z p(g\,|z, \boldsymbol{\theta})p(z\,|\boldsymbol{\theta})$$

when $\boldsymbol{z}$ are discrete valued, and finally

$$p(\boldsymbol{\theta}\,|g) = p(g\,|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(g) \tag{82}$$

with

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

One can also write:

$$p(z\,|\boldsymbol{\theta},\boldsymbol{g}) = \int p(f,z\,|\boldsymbol{\theta},\boldsymbol{g})\,\mathrm{d}f \tag{83}$$

and

$$p(\boldsymbol{\theta}\,|\boldsymbol{g}) = \int\int p(f,z,\boldsymbol{\theta}\,|\boldsymbol{g})\,\mathrm{d}f\,\mathrm{d}z = \int p(z\,|\boldsymbol{\theta};\boldsymbol{g})\,\mathrm{d}z \tag{84}$$

or

$$p(\boldsymbol{\theta}\,|\boldsymbol{g}) = \sum_z \int p(f,z,\boldsymbol{\theta}\,|\boldsymbol{g})\,\mathrm{d}f = \sum_z p(z\,|\boldsymbol{\theta};\boldsymbol{g}) \tag{85}$$

when z are discrete valued.

We see that the first term

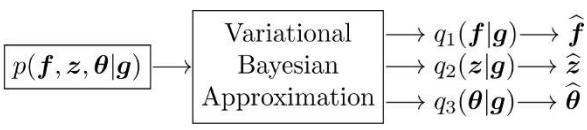
$$p(f\,|z,\boldsymbol{\theta},\boldsymbol{g}) \propto p(\boldsymbol{g}\,|f,\boldsymbol{\theta})p(f\,|z,\boldsymbol{\theta}) \tag{86}$$

will be easy to handle because it is the product of two Gaussians and so it is a multivariate Gaussian. But the two others are not.

The main idea behind the VBA is to approximate the joint posterior $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ by a separable one, for example

$$q(f,z,\boldsymbol{\theta}\,|\boldsymbol{g}) = q_1(f\,|\boldsymbol{g})\,q_2(z\,|\boldsymbol{g})q_3(\boldsymbol{\theta}\,|\boldsymbol{g}) \tag{87}$$

illustrated here:



and where the expressions of $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ is obtained by minimizing the Kullback-Leibler divergence

$$\mathrm{KL}(q:p) = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q \tag{88}$$

It is then easy to show that

$$\mathrm{KL}(q:p) = \ln p(\boldsymbol{g}\,|\mathcal{M}) - \mathcal{F}(q) \tag{89}$$

where $p(\boldsymbol{g}\,|\mathcal{M})$ is the likelihood of the model

$$p(\boldsymbol{g}\,|\mathcal{M}) = \int\int\int p(f,z,\boldsymbol{\theta},\boldsymbol{g}\,|\mathcal{M})\,\mathrm{d}f\,\mathrm{d}z\,\mathrm{d}\boldsymbol{\theta} \tag{90}$$

with

$$p(f,z,\boldsymbol{\theta},\boldsymbol{g}\,|\mathcal{M}) = p(\boldsymbol{g}\,|f,\boldsymbol{\theta})p(f\,|z,\boldsymbol{\theta})p(z\,|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{91}$$

and $\mathcal{F}(q)$ is the free energy associated to q defined as

$$\mathcal{F}(q) = \left\langle \ln \frac{p(f,z,\boldsymbol{\theta},\boldsymbol{g}\,|\mathcal{M})}{q(f,z,\boldsymbol{\theta})} \right\rangle_q \tag{92}$$

So, for a given model $\mathcal{M}$, minimizing $\mathrm{KL}(q:p)$ is equivalent to maximizing $\mathcal{F}(q)$ and when optimized, $\mathcal{F}(q^*)$ gives a lower bound for $\ln p(\boldsymbol{g}\,|\mathcal{M})$.

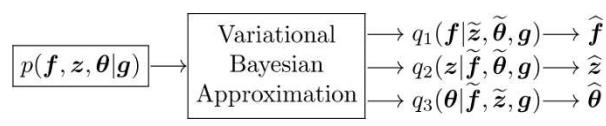
Without any other constraint than the normalization of q, an alternate optimization of $\mathcal{F}(q)$ with respect to q_1, q_2, and q_3 results in

$$\begin{aligned}
q_1(f) &\propto \exp\left\{-\langle \ln p(f,z,\boldsymbol{\theta},\boldsymbol{g})\rangle_{q(z)q(\boldsymbol{\theta})}\right\}\\
q_2(z) &\propto \exp\left\{-\langle \ln p(f,z,\boldsymbol{\theta},\boldsymbol{g})\rangle_{q(f)q(\boldsymbol{\theta})}\right\}\\
q_3(\boldsymbol{\theta}) &\propto \exp\left\{-\langle \ln p(f,z,\boldsymbol{\theta},\boldsymbol{g})\rangle_{q(f)q(z)}\right\}
\end{aligned}$$

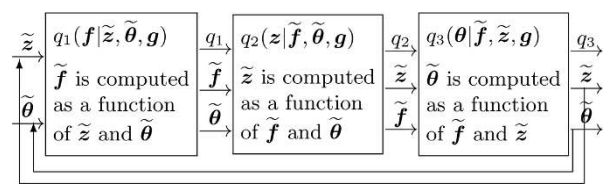
Note that these relations represent an implicit solution for $q_1(\boldsymbol{f})$, $q_2(\boldsymbol{z})$, and $q_3(\theta)$ which need, at each iteration, the expression of the expectations in the right hand of exponentials. If $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}_1)$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

$$q(f,z,\boldsymbol{\theta}\,|\boldsymbol{g}) = q_1(f\,|\widetilde{z},\widetilde{\boldsymbol{\theta}};\boldsymbol{g})q_2(z\,|\widetilde{f},\widetilde{\boldsymbol{\theta}};\boldsymbol{g})q_3(\boldsymbol{\theta}\,|\widetilde{f},\widetilde{z};\boldsymbol{g}) \tag{93}$$

where the tilded quantities $\widetilde{z}, \widetilde{f}$ and $\widetilde{\boldsymbol{\theta}}$ are, respectively functions of $(\widetilde{f},\widetilde{\boldsymbol{\theta}})$, $(\widetilde{z},\widetilde{\boldsymbol{\theta}})$ and $(\widetilde{f},\widetilde{z})$:



and where the alternate optimization results to alternate updating of the parameters $(\widetilde{z},\widetilde{\boldsymbol{\theta}})$ for q_1, the parameters $(\widetilde{f},\widetilde{\boldsymbol{\theta}})$ of q_2 and the parameters $(\widetilde{f},\widetilde{z})$ of q_3.



Finally, we may note that, to monitor the convergence of the algorithm, we may evaluate the free energy

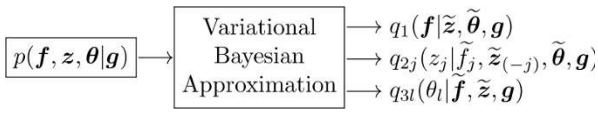
$$\begin{aligned}
\mathcal{F}(q) &= \left\langle \ln \frac{p(f,z,\boldsymbol{\theta},\boldsymbol{g}\,|\mathcal{M})}{q(f,z,\boldsymbol{\theta})} \right\rangle_q\\
&= \langle \ln p(f,z,\boldsymbol{\theta},\boldsymbol{g})\,|\mathcal{M}\rangle_q + \langle -\ln q(f,z,\boldsymbol{\theta})\rangle_q\\
&= \langle \ln p(\boldsymbol{g}\,|f,z,\boldsymbol{\theta})\rangle_q + \langle \ln p(f\,|z,\boldsymbol{\theta})\rangle_q + \langle \ln p(z\,|\boldsymbol{\theta})\rangle_q\\
&\quad + \langle -\ln q(f)\rangle_q + \langle -\ln q(z)\rangle_q + \langle -\ln q(\boldsymbol{\theta})\rangle_q
\end{aligned} \tag{94}$$

where all the expectations are with respect to q.

Other decompositions are also possible:

$$q(f,z,\theta\,|g) = q_1(f\,|\widetilde{z},\widetilde{\theta};g)\prod_j q_{2j}(z_j\,|\widetilde{f},\widetilde{z}_{(-j)},\widetilde{\theta};g)\prod_j q_{3l}(\theta_l\,|\widetilde{f},\widetilde{z},\widetilde{\theta}_{(-l)};g) \tag{95}$$

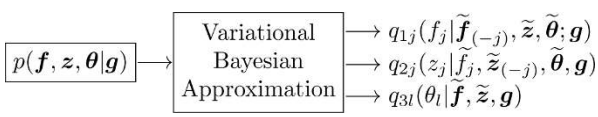
illustrated here:



or even by:

$$q(f,z,\theta\,|g) = \prod_j q_{1j}(f_j\,|\widetilde{f}_{(-j)},\widetilde{z},\widetilde{\theta};g)\prod_j q_{2j}(z_j\,|\widetilde{f},\widetilde{z}(-j),\theta;g)\prod_l q_{3l}(\theta_l\,|\widetilde{f},\widetilde{z},\widetilde{\theta}_{(-l)};g) \tag{96}$$

illustrated here:



Here, we consider the second case (Equation (95)) and give some more details on it. First to simplify the notations, we write it as:

$$q(f,z,\theta) = q_1(f)\prod_j q_{2j}(z_j)\prod_l q_{3l}(\theta_l) \tag{97}$$

where it can be shown that:

$$q_1(f) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q_2(z)q_3(\theta)}\right\}$$
$$q_{2j}(z_j) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q1(f)q_3(\theta)q_2(z_{(-j)})}\right\}$$
$$q_{3l}(\theta_l) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q_1}(f)_{q_2}(z)_{q_3}(\theta_{(-l)})\right\}$$

where $p(f, z, \theta, g) = p(g|f, \theta)p(f|z, \theta)p(z|\theta)p(\theta)$ and where $q_2(z) = \Pi_j\, q_{2j}(z_j)$, $q_3(\theta) = \Pi_l\, q_{3l}(\theta_l)$, $q_2(z_{(-j)}) = \Pi_{i\neq j}\, q_{2j}(z_j)$, $\langle.\rangle_q$ means expected value with respect to q.

In that case, with appropriate models for the priors (exponential families) and hyper parameters (conjugate priors), we see that $q(f)$ is a multivariate Gaussian $g(f) = \mathcal{N}(f\,|\widetilde{\mu},\widetilde{\Sigma})$, $q(\theta_l)$ are either Gaussians (for the means) or Inverse Gammas (for the variances) and $q(z_j)$ are discrete distributions whose expressions can be written easily.

To illustrate this in more detail, we consider the case of the Student-t model.

4.1 Student-t model

In this case, we have the following relations for the forward model and the prior laws:

$$\begin{cases} p(g\,|f,v_\varepsilon) = \mathcal{N}(g\,|Hf,v_\varepsilon I), \quad \tau = 1/v_\varepsilon \\ p(f\,|z) = \prod_j p(f_j\,|z_j) = \prod_j \mathcal{N}(z_j\,|0,z_j) = \mathcal{N}(f\,|0,Z) \\ Z = \mathrm{diag}[z], \quad a_j = 1/z_j, \quad A = \mathrm{diag}[a] = Z^{-1} \\ p(a) = \prod_j \mathcal{G}(a_j\,|\alpha_0,\beta_0) \\ p(\tau) = \mathcal{G}(\tau\,|\alpha_{\tau 0},\beta_{\tau 0}) \end{cases} \tag{98}$$

Then, we obtain the following expressions for the VBA:

$$\begin{cases} \widetilde{q}(f) = \mathcal{N}(f\,|\widetilde{\mu},\widetilde{\Sigma}) \\ \widetilde{\mu} = <\tau> \widetilde{\Sigma}H'g \\ \widetilde{\Sigma} = (<\tau> H'H + \widetilde{Z})^{-1}, \text{ with } \widetilde{Z} = \widetilde{A}^{-1} = \mathrm{diag}[\widetilde{a}] \end{cases} \tag{99}$$

$$\begin{cases} \widetilde{q}(\tau) = \mathcal{G}(\tau\,|\widetilde{\alpha}_\tau,\widetilde{\beta}_\tau), \\ \widetilde{\alpha}_\tau = \alpha_{\tau 0} + (n+1)/2 \\ \widetilde{\beta}_\tau = \beta_{\tau 0} + 1/2\left[\|g\|^2 - 2<f>' H'g + H'<ff'>H\right] \\ \widetilde{q}(a_j) = \mathcal{G}(a_j\,|\widetilde{\alpha}_j,\widetilde{\beta}_j) \\ \widetilde{\alpha}_j = \alpha_{00} + 1/2 \\ \widetilde{\beta}_j = \beta_{00} + <f_j^2>/2 \end{cases} \tag{100}$$

where the expressions of the expectations needed are:

$$\begin{cases} <f> = \widetilde{\mu} \\ <ff'> = \Sigma + \mu\mu' \\ <f_j^2> = [\Sigma]_{jj} + \mu_j^2 \\ <\tau> = \widetilde{\tau} = \widetilde{\alpha}_\tau/\widetilde{\beta}_\tau \\ <a_j> = \widetilde{a}_j = \widetilde{\alpha}_j/\widetilde{\beta}_j \end{cases} \tag{101}$$

We can also express the free energy expression:

$$\begin{aligned} \mathcal{F}(q) &= \left\langle \ln \frac{p(f,a,\tau,g\,|\mathcal{M})}{q(f,a,\tau)} \right\rangle \\ &= \langle \ln p(g\,|f,a,\tau)\rangle + \langle \ln p(f\,|a,\tau)\rangle + \langle \ln p(a\,|\tau)\rangle \\ &\quad + \langle -\ln q(f)\rangle + \langle -\ln q(a)\rangle + \langle -\ln q(\tau)\rangle \end{aligned} \tag{102}$$

where

$$\langle \ln p(g\,|f,\tau)\rangle = \frac{n}{2}\left(<\ln\tau> - \ln(2\pi)\right) - \frac{1}{2}\left\{<\tau> g'g - 2<f>' H'g + H'<ff'>H\right\}$$

$$\langle -\ln p(f\,|a)\rangle = -\frac{n+1}{2}\ln(2\pi) - \frac{1}{2}\left\{\sum_j <\ln\alpha_j><\alpha_j><f_j^2>\right\}$$

$$\langle -\ln p(a)\rangle = -(n+1)\alpha_{\varepsilon_0}\ln(\beta_{\varepsilon_0}) + (\alpha_{\varepsilon_0} - 1)\sum_j [<\ln\alpha_j> - \beta <\alpha_j)] - (n+1)\ln\Gamma(\alpha)$$

$$\langle p(\tau))\rangle = c\ln d + (c-1) <\ln\tau> - d <\tau> - \ln\Gamma(c)$$

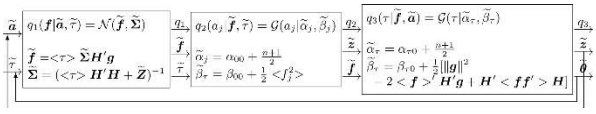
and

$$\langle -\ln q(\boldsymbol{f})\rangle = -\frac{n+1}{2}(1+\ln(2\pi)) - \frac{1}{2}\ln\left|\Sigma_j\right|$$
$$\langle -\ln q(\boldsymbol{a})\rangle = -\sum_j \left[\widetilde{\alpha}_j \ln(\widetilde{\beta}_j) + (\widetilde{\alpha}_j - 1) < \ln\widetilde{\alpha}_j > -\widetilde{\beta}_j < \alpha_j > -\ln\Gamma(\widetilde{\alpha}_j)\right]$$
$$\langle q(\tau))\rangle = \tilde{c}\ln\tilde{d} + (\tilde{c}-1) < \ln\tau) > -\tilde{d} < \tau > -\ln\Gamma(\tilde{c})$$

In these equations,

$$\begin{cases} < \ln a_j >= \psi(\tilde{a}_j) - \ln\tilde{b}_j \\ < \ln\tau >= \psi(\tilde{c}) - \ln\tilde{d} \\ \psi(a) = \dfrac{\partial \ln\Gamma(a)}{\partial a} \end{cases} \tag{103}$$

The resulting algorithm can be summarized as follows



5 Conclusion

The sparsity is a required property in many signal and image processing applications. In this article, first we reviewed the main steps of the Bayesian approach for inverse problems in signal and image processing. Then we presented in a synthetic way the different prior models which can be used to enforce the sparsity. These models have been presented in two categories: simple and hierarchical with hidden variables. For each of these prior models, we discuss their properties and the way to use them in a Bayesian approach resulting to many different inversion algorithms.

We have applied these Bayesian algorithms in many different applications such as X-ray computed tomography [35,36], optical diffraction tomography [37-39], positron emission tomography [40], Microwave imaging [41,42], Sources separation [43-46], spectrometry [47,48], Hyper spectral imaging [49], super resolution [50-52], image fusion [53], image segmentation [54], synthetic aperture radar (SAR) imaging [29]. To save the place and be very synthetic, we did not give here any simulation results or any results on different applications of these methods. These can be found in different articles just referenced.

Acknowledgements

This study had been partially founded by the C5Sys project (Circadian and Cell cycle Clock systems in Cancer) of ERASYSBIO+. http://www.erasysbio.net/index.php?index=272

Competing interests

The author declares that they have no competing interests.

Received: 18 January 2012 Accepted: 1 March 2012
Published: 1 March 2012

References

1. A Tikhonov, Regularization of incorrectly posed problems. Soviet Math Dokl. **4**, 1624–1627 (1963)
2. A Tikhonov, V Arénine, Méthodes de Résolution de Problémes Mal Posés, (MIR, Moscu, Russia, 1976). ?É?ditions
3. I Daubechies, M Defrise, CD Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm Pure Appl Math. **57**, 1413–1457 (2004). doi:10.1002/cpa.20042
4. DL Donoho, Compressive sampling. IEEE Trans Inf Theory. **52**(4), 1289–1306 (2006)
5. JA Tropp, AC Gilbert, MJ Strauss, Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Processing, special issue "sparse approximations in signal and image processing". **86**, 572–588 (2006)
6. JA Tropp, Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. Signal Process (special issue "Sparse approximations in signal and image processing"). **86**, 589–602 (2006)
7. R Zass, A Shashua, in *Nonnegative Sparse PCA*, vol. 19. (Cambridge, MA: MIT Press, 2007), pp. 1561–1568
8. EJ Candés, M Wakin, S Boyd, Enhancing sparsity by reweighted l1 minimization. J Fourier Anal Appl. **14**, 877–905 (2008). doi:10.1007/s00041-008-9045-x
9. DM Witten, R Tibshirani, T Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. **10**(3), 515–534 (2009). doi:10.1093/biostatistics/kxp008
10. S Vaiter, G Peyré, C Dossal, J Fadili, Robust Sparse Analysis Reg-ularization. Tech rep, preprint Hal-00627452 http://hal.archives-ouvertes.fr/hal-00627452/ (2011)
11. G Peyré, J Fadili, Learning Analysis Sparsity Priors. Proc of Sampta'11 http://hal.archives-ouvertes.fr/hal-00542016/ (2011)
12. P Williams, Bayesian regularization and pruning using a Laplace prior. Neural Comput. **71**, 117–143 (1995)
13. T Mitchell, J Beauchamp, Bayesian variable selection in linear regression. J Am Stat Assoc. **83**(404), 1023 (1988). doi:10.2307/2290129
14. N Polson, J Scott, Shrink globally, act locally: sparse Bayesian regulariza-tion and prediction. Bayesian Stat. **9**, 1–24 (2010)
15. A Mohammad-Djafari, On the estimation of hyperparameters in Bayesian approach of solving inverse problems, in *Proc IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-93*, vol. 5. (Minneapolis, MN, USA, 1993), pp. 495–498
16. A Mohammad-Djafari, Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems, in *IEEE Int Conf on Image Processing (ICIP), IEEE ICIP 96*, vol. II. (Lausanne, Swisse, 1996), pp. 473–477
17. R Neal, G Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants. Learn Graph Models. **89**, 355–368 (1998)
18. A Doucet, P Duvaut, Bayesian estimation of state-space models applied to deconvolution of Bernoulli-Gaussian processes. Signal Process. **57**(2), 147–161 (1997). doi:10.1016/S0165-1684(96)00192-2
19. T Park, G Casella, The Bayesian Lasso. J Am Stat Assoc. **103**(482), 681–686 (2008). doi:10.1198/016214508000000337
20. M Tipping, Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. **1**, 211–244 (2001)
21. C Févotte, S Godsill, A Bayesian aproach for blind separation of sparse source. IEEE Trans Audio Speech Lang Process. **14**, 2174–2188 (2006)
22. F Caron, A Doucet, Sparse Bayesian nonparametric regression, in *International Conference on Machine Learning*, pp. 88–95 (2008)
23. J Griffin, P Brown, Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. **5**, 171–188 (2010)
24. H Snoussi, J Idier, Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures. IEEE Trans Signal Process. **54**, 3257–3269 (2006)
25. H Ishwaran, JS Rao, Spike and slab variable selection: frequentist and Bayesian strategies. Ann Stat. **33**(2), 730–733 (2005). doi:10.1214/009053604000001147
26. S Chatzis, T Varvarigou, Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributionsclassification of binary random patterns. IEEE Trans Fuzzy Syst. **17**, 505–517 (2009)

27. CA Bouman, KD Sauer, A generalized Gaussian image model for edge-preserving MAP estimation. IEEE Trans Image Process. **2**(3), 296–310 (1993). doi:10.1109/83.236536
28. H Zou, T Hastie, Regularization and variable selection via the elastic net. J Royal Stat Soc, Ser, B. **67**(2), 301–320 (2005). doi:10.1111/j.1467-9868.2005.00503.x
29. S Zhu, A Mohammad-Djafari, H Wang, B Deng, X Li, J Mao, Parameter estimation for SAR micromotion target based on sparse signal representation. Eurasip special issue "Sparse approximations in signal and image processing". **13** (2012)
30. J Idier, Approche Bayésienne Pour les Problémes Inverses,(Traité IC2, Série traitement du signal et de l'image, Hermés, Paris, 2001)
31. F Champagnat, Y Goussard, J Idier, Unsupervised deconvolution of sparse spike trains using stochastic approximation. IEEE Trans Signal Process. **44**(12), 2988–2998 (1996). doi:10.1109/78.553473
32. D Ge, J Idier, E Le Carpentier, A new MCMC algorithm for blind Bernoulli-Gaussian deconvolution, Proceedings of EUSIPCO: Septembre 2008, (Lausanne, Suisse, 2008)
33. JJ Kormylo, JM Mendel, Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes. IEEE Trans Inf Theory. **28**, 482–488 (1982). doi:10.1109/TIT.1982.1056496
34. M Lavielle, Bayesian deconvolution of Bernoulli-Gaussian processes. Signal Process. **33**, 67–79 (1993)
35. A Mohammad-Djafari, Gauss-Markov-Potts priors for images in computer tomography resulting to joint optimal reconstruction and segmentation. Int J Tomography Stat. **11**(W09), 76–92 http://djafari.free.fr/pdf/IJTS 08.pdf (2008)
36. N Gac, A Vabre, A Mohammad-Djafari, F Buyens, GPU implementation of a 3D bayesian CT algorithm and its application on real foam reconstruction, Proceedings of the first International Conference on Image Formation in X-Ray Computed Tomography,(Salt Lake City, Utah, USA, 2010)
37. H Ayasso, B Duchne, A Mohammad-Djafari, Bayesian inversion for optical diffraction tomography. J Modern Opt. **57**(9), 765–776 (2010). doi:10.1080/09500340903564702
38. H Ayasso, A Mohammad-Djafari, Joint NDT image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation. IEEE Trans. Image Process. **19**(9), 2265–2277 http://dx.doi.org/10.1109/TIP.2010.2047902 (2010)
39. H Ayasso, B Duchne, A Mohammad-Djafari, Optical diffraction tomography within a variational Bayesian framework. Inverse Probl Sci Eng. **iFirst**, 1–15 (2011)
40. MD Fall, ?É? Barat, C Comtat, T Dautremer, T Montagu, A Mohammad-Djafari, A discrete-continuous Bayesian model for emission tomography, IEEE International Conference on Image Processing (ICIP),(Bruxell, Belgium, 2011), pp. 1401–1404
41. O Féron, B Duchêne, A Mohammad-Djafari, Microwave imaging of inhomogeneous objects made of a finite number of dielectric and conductive materials from experimental data. Inverse Probl. **21**(6), 95–115 http://djafari.free.fr/pdf/ (2005). doi:10.1088/0266-5611/21/6/S08
42. O Féron, B Duchêne, A Mohammad-Djafari, Microwave imaging of piece-wise constant objects in a 2D-TE configuration. Int J Appl Electromag Mech. **26**(6), 167–174 http://djafari.free.fr/pdf/jae00905.pdf (2007)
43. H Snoussi, A Mohammad-Djafari, Fast joint separation and segmentation of mixed images. J Electron Imag. **13**(2), 349–361 http://djafari.free.fr/pdf/ (2004). doi:10.1117/1.1666873
44. H Snoussi, A Mohammad-Djafari, Bayesian unsupervised learning for source separation with mixture of Gaussians prior. J VLSI Signal Process Syst. **37**(2/3), 263–279http://djafari.free.fr/pdf/VLSIpapar.pdf (2004)
45. A Mohammad-Djafari, Bayesian source separation: beyond PCA and ICA, in ESANN 2006, (Belgium, 2006) http://djafari.free.fr/pdf/
46. M Ichir, A Mohammad-Djafari, Hidden Markov models for wavelet-based blind source separation. IEEE Trans Image Process. **15**(7), 1887–1899 (2006)
47. A Mohammad-Djafari, J Giovannelli, G Demoment, J Idier, Regulariza-tion, maximum entropy and probabilistic methods in mass spectrometry data processing problems. Int J Mass Spectrom. **215**(1-3), 175–193 http://djafari.free.fr/pdf/maspec12013.pdf (2002). doi:10.1016/S1387-3806(01)00562-0
48. S Moussaoui, D Brie, A Mohammad-Djafari, C Carteret, Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. IEEE Trans Signal Process. **54**(11), 4133–4145 (2006)
49. N Bali, A Mohammad-Djafari, Bayesian approach with hidden Markov modeling and mean field approximation for hyperspectral data analysis. IEEE Trans Image Process. **17**(2), 217–225 (2008)
50. F Humblot, A Mohammad-Djafari, Super-resolution using hidden Markov model and bayesian detection estimation framework, in *EURASIP J Appl Signal Process*, vol. 16. (Special number on Super-Resolution Imaging: Analysis, Algorithms, and Applications, 2006) http://www.hindawi.com/GetArticle.aspx?doi=10.1155/ASP/2006/36971. Article ID 36971
51. A Mohammad-Djafari, Super-resolution: a short review, a new method based on hidden Markov modeling of HR image and future challenges. Comput J http://djafari.free.fr/pdf/bxn005v1.pdf (2008)
52. M Mansouri, A Mohammad-Djafari, Joint super-resolution and segmentation from a set of low resolution images using a Bayesian approach with a Gauss-Markov-Potts prior. Int J Signal Imag Syst Eng. **3**(4), 211–221 (2010). doi:10.1504/IJSISE.2010.038017
53. O Féron, A Mohammad-Djafari, Image fusion and joint segmentation using an MCMC algorithm. J Electron Imag. **14**(2) http://arxiv.org/abs/physics/0403150 (2005). paper no. 023014
54. P Brault, A Mohammad-Djafari, Unsupervised Bayesian wavelet domain segmentation using a Potts-Markov random field modeling. J Electron Imag. **14**(4) http://djafari.free.fr/pdf/ (2005). 043011-1-043011-16

doi:10.1186/1687-6180-2012-52
Cite this article as: Mohammad-Djafari: **Bayesian approach with prior models which enforce sparsity in signal and image processing.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:52.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com