Variational Bayesian Approximation methods for inverse problems

Ali Mohammad-Djafari

Bayesian Inference Tools for Inverse Problems

Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes,
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
SUPELEC, Plateau de Moulon, 3 rue Juliot-Curie, 91192 Gif-sur-Yvette, France

Abstract. In this paper, first the basics of the Bayesian inference for linear inverse problems are presented. The inverse problems we consider are, for example, signal deconvolution, image restoration or image reconstruction in Computed Tomography (CT). The main point to discuss then is the prior modeling of signals and images. We consider two classes of priors: *simple* or *hierarchical with hidden variables*. For practical applications, we need also to consider the estimation of the hyper parameters. Finally, we see that we have to infer simultaneously the unknowns, the hidden variables and the hyper parameters.

Very often, the expression of the joint posterior law of all the unknowns is too complex to be handled directly. Indeed, rarely we can obtain analytical solutions to any point estimators such the Maximum A posteriori (MAP) or Posterior Mean (PM). Three main tools can then be used: Laplace approximation (LAP), Markov Chain Monte Carlo (MCMC) and Bayesian Variational Approximations (BVA).

To illustrate all these aspects, we will consider a deconvolution problem where we know that the input signal is *sparse* and propose to use a Student-t distribution for that. Then, to handle the Bayesian computations with this model, we use the property of Student-t which is modelling it via an infinite mixture of Gaussians, introducing thus hidden variables which are the variances. Then, the expression of the joint posterior of the input signal samples, the hidden variables (which are here the inverse variances of those samples) and the hyper-parameters of the problem (for example the variance of the noise) is given. From this point, we will present the joint maximization by alternate optimization and the three possible approximation methods. Finally, the proposed methodology is applied in different applications such as mass spectrometry, spectrum estimation of quasi periodic biological signals and X ray computed tomography.

INTRODUCTION

In many generic inverse problems in signal and image processing, the problem is to infer on an unknown signal $f(t)$ or an unknown image $f(r)$ with $r=(x,y)$ through an observed signal $g(t')$ or an observed image $g(r')$ related between them through an operator $\mathcal{H}$ such as convolution $g=h*f$ or any other linear or nonlinear transformation $g=\mathcal{H}f$. When this relation is linear and we have discretized the problem, we arrive to the relation: $g=Hf+\varepsilon$ where $f=[f_1,\cdots,f_n]'$ represents the unknowns, $g=[g_1,\cdots,g_m]'$ the observed data, $\varepsilon=[\varepsilon_1,\cdots,\varepsilon_m]'$ the errors of modelling and measurement and H the matrix of the system response.

The Bayesian inference approach is based on the posterior law:

$$p(f|g,\theta)=\frac{p(g|f,\theta)\,p(f|\theta)}{p(g|\theta)}\propto p(g|f,\theta)\,p(f|\theta) \tag{1}$$

where the sign $\propto$ stands for "proportional to", $p(g|f,\theta)$ is the likelihood, $p(f|\theta)$ the prior probability law model and $p(g|\theta)$ is called the evidence of the model. θ is called the vector of the hyper-parameters of the problem. It can be divided in two independent parts $\theta = (\theta_1, \theta_2)$ where θ_1 is only appearing in the likelihood term and θ_2 in the prior term. Then, we can write $p(f|g,\theta) \propto p(g|f,\theta_1)\, p(f|\theta_2)$.

When the parameters θ have to be estimated too, a prior probability law $p(\theta|\theta_0)$ with fixed values for θ_0 is assigned to them in such a way to obtain the joint posterior law:

$$p(f,\theta|g,\theta_0) = \frac{p(g|f,\theta)\, p(f|\theta)\, p(\theta|\theta_0)}{p(g|\theta_0)} \tag{2}$$

which can be used to infer them jointly. From this point (omitting θ_0 for simplicity of notations), at least three main approaches are commonly used to infer both unknowns f and θ:

- Joint MAP estimation:

$$(\widehat{f},\widehat{\theta}) = \arg\max_{(f,\theta)} \{p(f,\theta|g)\} \tag{3}$$

  which is done in general by an alternate optimization algorithm such as:

$$\begin{cases} \widehat{f} = \arg\max_f \left\{ p(f,\widehat{\theta}|g) \right\} \\ \widehat{\theta} = \arg\max_\theta \left\{ p(\widehat{f},\theta|g) \right\} \end{cases} \tag{4}$$

- Marginalization over f to obtain:

$$p(\theta|g) = \int p(f,\theta|g)\ \mathrm{d}f \tag{5}$$

  which can be used to infer θ which is then used for the estimation of f. Unfortunately, in general, the marginalization step can not be done analytically. The Expectation-Maximization (EM) algorithm is used to obtain $\widehat{\theta}$. The EM algorithm can be summarized as follows:

$$\begin{cases} Q(\theta,\widehat{\theta}) = \displaystyle\int p(f|\widehat{\theta},g)\, \ln p(f,\theta|g)\, \mathrm{d}f = \langle \ln p(f,\theta|g) \rangle_{p(f|\widehat{\theta},g)} \\ \widehat{\theta} = \arg\max_\theta \left\{ Q(\theta,\widehat{\theta}) \right\} \end{cases} \tag{6}$$

- Variational Bayesian Approximation (VBA):
  The main idea here is to approximate $p(f,\theta|g)$ by a separable one $q(f,\theta) = q_1(f)\, q_2(\theta)$ which can then be used for inferring on f or θ [3, 1, 13, 18, 2, 15, 17, 12, 9].

As we will see more in detail in the next section, $q_1(f)$ and $q_2(\theta)$ are obtained via the following iterative algorithms:

$$\begin{cases} q_1(f) \propto \exp\left\{ -\langle \ln p(f,\theta,g) \rangle_{q_2(\theta)} \right\} \\ q_2(\theta) \propto \exp\left\{ -\langle \ln p(f,\theta,g) \rangle_{q_1(f)} \right\} \end{cases} \tag{7}$$

The two first approaches are very well known. The last one is less and is explained in more details in the next section. These approaches are compared in Figure 1.

Joint MAP:

Main idea:

$$\boxed{p(f,\theta|g)} \longrightarrow \boxed{\begin{array}{c}\text{Joint}\\ \text{Maximum}\\ \text{A Posteriori}\end{array}} \begin{array}{l}\longrightarrow \widehat{f}\\ \longrightarrow \widehat{\theta}\end{array}$$

Algorithm:

$$\begin{array}{ccccccc}
\theta^{(0)} & \longrightarrow & \widehat{\theta} & \longrightarrow & \boxed{\widehat{f} = \arg\max_f \left\{ p(f|\widehat{\theta},g) \right\}} & \longrightarrow \widehat{f} & \longrightarrow \widehat{f} \\
 & & \uparrow & & & \downarrow & \\
\widehat{\theta} & \longleftarrow & \widehat{\theta} & \longleftarrow & \boxed{\widehat{\theta} = \arg\max_\theta \left\{ p(\theta|\widehat{f},g) \right\}} & \longleftarrow \widehat{f} &
\end{array}$$

Marginalization:

Main idea:

$$\boxed{p(f,\theta|g)} \longrightarrow \boxed{p(\theta|g)} \longrightarrow \widehat{\theta} \longrightarrow \boxed{p(f|\theta,g)} \longrightarrow \widehat{f}$$

Algorithm:

$$\begin{array}{ccccccc}
 & & \theta^{(0)} & \longrightarrow \widehat{\theta} \longrightarrow & \boxed{Q(\theta,\widehat{\theta}) = \langle \ln p(f,\theta|g) \rangle_{p(f|\widehat{\theta},g)}} & \longrightarrow Q(\theta,\widehat{\theta}) & \\
 & & & \uparrow & & \downarrow & \\
\widehat{f} & \longleftarrow \boxed{\widehat{f} = \arg\max_f \left\{ p(f|\widehat{\theta},g) \right\}} & \longleftarrow \widehat{\theta} \longleftarrow & \boxed{\widehat{\theta} = \arg\max_\theta \left\{ Q(\theta,\widehat{\theta}) \right\}} & \longleftarrow Q(\theta,\widehat{\theta}) &
\end{array}$$

VBA:

Main idea:

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

Algorithm:

$$\begin{array}{ccccccc}
\widehat{q}_2^{(0)} & \longrightarrow \widehat{q}_2 \longrightarrow & \boxed{q_1(f) \propto \exp\left\{ \left\langle \ln p(g,f,\theta;\mathcal{M}) \right\rangle_{q_2} \right\}} & \longrightarrow \widehat{q}_1(f) \longrightarrow \widehat{f} \\
 & \uparrow & & \downarrow & \\
\widehat{\theta} \leftarrow \widehat{q}_2(\theta) \longleftarrow & & \boxed{q_2(\theta) \propto \exp\left\{ \left\langle \ln p(g,f,\theta;\mathcal{M}) \right\rangle_{q_1} \right\}} & \longleftarrow \widehat{q}_1 &
\end{array}$$

Figure 1: Three different approaches for inferring both unknowns f and θ.

As we will see, the main inconvenient of the first approach is that we are summarizing the joint posterior law $p(f,\theta|g)$ by only its mode. Also, for obtaining this mode, in general an iterative alternate optimization is used, where at each iteration, only the

values of the estimates at previous iterations are used without accounting for their corresponding uncertainties. In the second approach, first θ is estimated and then it is used for the estimation of f, again without accounting for its uncertainty. In the third approach, as we will see, the estimation of f depends on the approximated law $q_2(\theta)$ and the estimation of θ depends on the approximated law $q_1(f)$, thus accounting for uncertainties in both steps.

A simple prior probability law is often not enough for modeling signals and images, in particular for non stationary signals or non homogeneous images. We then may use hierarchical models with hidden variables z which may represent, for example, the class labels in mixture models. In those cases, the prior probability model contains two parts $p(f|z,\theta_2)$ and $p(z|\theta_3)$ and we will have:

$$p(f,z,\theta|g,\theta_0) \propto p(g|f,\theta_1)\, p(f|z,\theta_2)\, p(z|\theta_3)\, p(\theta|\theta_0) \tag{8}$$

and then again different approaches can be used to infer the unknowns f, z and θ.

In this paper, first the general VBA method is detailed for the inference on inverse problems with hierarchical prior models. Then, two particular classes of prior models (Student-t and mixture of Gaussians) are considered and the details of BVA algorithms are given for them.

BAYESIAN VARIATIONAL APPROXIMATION WITH HIERARCHICAL PRIOR MODELS

When a hierarchical prior model $p(f|z,\theta)$ is used and when the estimation of the hyper-parameters θ has to be considered, the joint posterior law of all the unknowns becomes:

$$p(f,z,\theta|g) \propto p(g|f,\theta_1)\, p(f|z,\theta_2)\, p(z|\theta_3)\, p(\theta) \tag{9}$$

which can also be written as $p(f,z,\theta|g) = p(f|z,\theta,g)\, p(z|\theta,g)\, p(\theta|g)$ where

$$p(f|z,\theta,g) = p(g|f,\theta)\, p(f|z,\theta)/p(g|z,\theta) \text{ with } p(g|z,\theta) = \int p(g|f,\theta)\, p(f|z,\theta)\, \mathrm{d}f \tag{10}$$

and

$$p(z|\theta,g) = \frac{p(g|z,\theta)\, p(z|\theta)}{p(g|\theta)} \text{ with } p(g|\theta) = \int p(g|z,\theta)\, p(z|\theta)\, \mathrm{d}z \tag{11}$$

and finally

$$p(\theta|g) = \frac{p(g|\theta)\, p(\theta)}{p(g)} \text{ with } p(g) = \int p(g|\theta)\, p(\theta)\, \mathrm{d}\theta. \tag{12}$$

In general, common chooses for $p(g|f,\theta_1)$ and $p(f|z,\theta_2)$ are Gaussians and for $p(z|\theta_3)$ and $p(\theta)$ are Bernoulli or Binomial (for discrete valued z) or Gamma for inverse of the variances. Thus, the first term

$$p(f|z,\theta,g) \propto p(g|f,\theta) p(f|z,\theta) \tag{13}$$

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 $p(f,z,\theta|g)$ by a separable one, for example

$$q(f,z,\theta|g) = q_1(f)\,q_2(z)\,q_3(\theta) \tag{14}$$

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

$$\mathrm{KL}(q:p) = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q . \tag{15}$$

It is then easy to show that $\mathrm{KL}(q:p) = \ln p(g|\mathcal{M}) - \mathcal{F}(q)$ where $p(g|\mathcal{M})$ is the likelihood of the model

$$p(g|\mathcal{M}) = \iiint p(f,z,\theta,g|\mathcal{M})\,\mathrm{d}f\,\mathrm{d}z\,\mathrm{d}\theta \tag{16}$$

with $p(f,z,\theta,g|\mathcal{M}) = p(g|f,\theta)\,p(f|z,\theta)\,p(z|\theta)\,p(\theta)$ and $\mathcal{F}(q)$ is the free energy associated to q defined as

$$\mathcal{F}(q) = \left\langle \ln \frac{p(f,z,\theta,g|\mathcal{M})}{q(f,z,\theta)} \right\rangle_q . \tag{17}$$

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 $\ln p(g|\mathcal{M})$.

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

$$\left\{\begin{array}{l} q_1(f) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q(z)q(\theta)}\right\} \\ q_2(z) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q(f)q(\theta)}\right\} \\ q_3(\theta) \propto \exp\left\{-\langle \ln p(f,z,\theta,g)\rangle_{q(f)q(z)}\right\} \end{array}\right. \tag{18}$$

Note that these relations represent an implicit solution for $q_1(f)$, $q_2(z)$ and $q_3(\theta)$ which need, at each iteration, the expression of the expectations in the right hand of exponentials. If $p(g|f,z,\theta_1)$ is a member of an exponential family and if all the priors $p(f|z,\theta_2)$, $p(z|\theta_3)$, $p(\theta_1)$, $p(\theta_2)$, and $p(\theta_3)$ are conjugate priors, then it is easy to see that these expressions leads to standard distributions for which the required expectations are easily evaluated. In that case, we may note

$$q(f,z,\theta|g) = q_1(f|\widetilde{z},\widetilde{\theta},g)\,q_2(z|\widetilde{f},\widetilde{\theta},g)\,q_3(\theta|\widetilde{f},\widetilde{z},g) \tag{19}$$

where the tilded quantities $\widetilde{z}$, $\widetilde{f}$ and $\widetilde{\theta}$ are, respectively functions of $(\widetilde{f},\widetilde{\theta})$, $(\widetilde{z},\widetilde{\theta})$ and $(\widetilde{f},\widetilde{z})$ and where the alternate optimization with respect to q_1, q_2 and q_3 becomes alternate updating of the parameters $(\widetilde{z},\widetilde{\theta})$ for q_1, the parameters $(\widetilde{f},\widetilde{\theta})$ 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

$$\begin{aligned}\mathcal{F}(q) &= \left\langle \ln p(f,z,\theta,g|\mathcal{M})\right\rangle_q + \left\langle -\ln q(f,z,\theta)\right\rangle_q \\ &= \left\langle \ln p(g|f,z,\theta)\right\rangle_q + \left\langle \ln p(f|z,\theta)\right\rangle_q + \left\langle \ln p(z|\theta)\right\rangle_q \\ &+ \left\langle -\ln q(f)\right\rangle_q + \left\langle -\ln q(z)\right\rangle_q + \left\langle -\ln q(\theta)\right\rangle_q\end{aligned} \tag{20}$$

where all the expectations are with respect to q.

Other decompositions are also possible:

$$q(f,z,\theta|g) = \prod_j q_{1j}(f_j|\widetilde{f}_{(-j)},\widetilde{z},\widetilde{\theta},g) \prod_j q_{2j}(z_j|\widetilde{f},\widetilde{z}_{(-j)},\widetilde{\theta},g) \prod_l q_{3l}(\theta_l|\widetilde{f},\widetilde{z},\widetilde{\theta}_{(-l)},g) \tag{21}$$

or

$$q(f,z,\theta|g) = q_1(f|\widetilde{z},\widetilde{\theta},g) \prod_j q_{2j}(z_j|\widetilde{f},\widetilde{z}_{(-j)},\widetilde{\theta},g) \prod_l q_{3l}(\theta_l|\widetilde{f},\widetilde{z},\widetilde{\theta}_{(-l)},g) \tag{22}$$

In the following section, we consider this case and give some more details with the hierarchical model of Infinite Mixture model of Student-t which is used for example for modeling the distributions of sparse signals or images [14].

JMAP AND BAYESIAN VARIATIONAL APPROXIMATION WITH STUDENT-T PRIORS

The Student-t model is:

$$p(f|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \text{ with } \mathcal{S}t(f_j|\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1+f_j^2/\nu\right)^{-(\nu+1)/2} \tag{23}$$

Knowing that

$$\mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0,1/z_j)\, \mathcal{G}(z_j|\nu/2,\nu/2)\, \mathrm{d}z_j \tag{24}$$

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

$$\left\{\begin{array}{ll} p(f|z) & = \prod_j p(f_j|z_j) = \prod_j \mathcal{N}(f_j|0,1/z_j) \propto \exp\left\{-\frac{1}{2}\sum_j z_j f_j^2\right\} \\ p(z_j|\alpha,\beta) & = \mathcal{G}(z_j|\alpha,\beta) \propto {z_j}^{(\alpha-1)} \exp\left\{-\beta z_j\right\} \text{ with } \alpha=\beta=\nu/2 \end{array}\right. \tag{25}$$

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

Now consider this prior model for the unknowns f of a linear inverse problem with the linear forward model $g = Hf + \varepsilon$ and assign a Gaussian law to the noise ε which results to $p(g|f,v_\varepsilon) = \mathcal{N}(g|Hf,v_\varepsilon I)$. We also assign a prior $p(\tau_\varepsilon|\alpha_0,\beta_0) = \mathcal{G}(\tau_\varepsilon|\alpha_0,\beta_0)$ to $\tau_\varepsilon = 1/v_\varepsilon$. Figure 2 shows the graphical representation of this model.

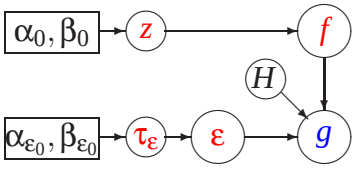

Figure 2: The graphical representation of the proposed model with the Student-t equivalent hierarchical prior.

In the following, we summarize all the equations related to this modeling and inference scheme.

- Forward probability laws:

$$\begin{cases} p(g|f,\tau_\varepsilon) = \mathcal{N}(g|Hf,(1/\tau_\varepsilon)I), & p(\tau_\varepsilon|\alpha_{\varepsilon 0},\beta_{\varepsilon 0}) = \mathcal{G}(\tau_\varepsilon|\alpha_{\varepsilon 0},\beta_{\varepsilon 0}), \\ p(f|z) = \prod_j \mathcal{N}\left(f_j|0,1/z_j\right), & p(z|\alpha_0,\beta_0) = \prod_j \mathcal{G}(z_j|\alpha_0,\beta_0). \end{cases} \tag{26}$$

- Joint posterior laws:

$$\begin{aligned} p(f,z,\tau_\varepsilon|g,\alpha_0,\beta_0,\alpha_{\varepsilon 0},\beta_{\varepsilon 0}) &\propto p(g|f,\tau_\varepsilon)\, p(f|z)\, p(z|\alpha_0,\beta_0)\, p(\tau_\varepsilon|\alpha_{\varepsilon 0},\beta_{\varepsilon 0}) \\ &\propto {\tau_\varepsilon}^{-M/2} \exp\left\{-\tfrac{1}{2}\tau_\varepsilon \|g-Hf\|^2\right\} \quad \prod_j {z_j}^{-1/2} \exp\left\{-\tfrac{1}{2} z_j f_j^2\right\} \\ &\quad \prod_j {z_j}^{-\alpha_0+1} \exp\left\{-\beta_0 z_j\right\} \quad {\tau_\varepsilon}^{-\alpha_{\varepsilon 0}+1} \exp\left\{-\beta_{\varepsilon 0}\tau_\varepsilon\right\}. \end{aligned} \tag{27}$$

- Joint MAP alternate maximization algorithm:
  The objective of the JMAP optimization is:

$$(\widehat{f},\widehat{z},\widehat{\tau_\varepsilon}) = \arg \max_{(f,z,\tau_\varepsilon)} \left\{p(f,z,\tau_\varepsilon|g,\alpha_0,\beta_0,\alpha_{\varepsilon 0},\beta_{\varepsilon 0})\right\}. \tag{28}$$

  The alternate optimization is an iterative optimization, respectively with respect to f, z and τ:

$$\begin{cases} \widehat{f} = \arg\min_f \left\{\widehat{\tau_\varepsilon}\|g-Hf\|^2 + \sum_j \widehat{z_j} f_j^2\right\}, \\ \widehat{z} = \arg\min_z \left\{\frac{N+2\alpha_0-2}{2} \ln z_j + \sum_j z_j \left(\frac{1}{2}\widehat{f_j}^2 + \beta_0\right)\right\}, \\ \widehat{\tau_\varepsilon} = \arg\min_{\tau_\varepsilon} \left\{(\frac{M}{2}+\alpha_{\varepsilon 0}-1)\ln\tau_\varepsilon + \left(\frac{1}{2}\|g-H\widehat{f}\|^2 + \beta_{\varepsilon 0}\right)\right\}. \end{cases} \tag{29}$$

  The first optimization can be done either analytically or using any gradient based algorithm. The second and the third optimizations have analytical expressions:

$$\begin{cases} \widehat{f} = \widehat{\tau_\varepsilon}\widehat{\Sigma}H'g \text{ with } \widehat{\Sigma} = \left(\widehat{\tau_\varepsilon}H'H + \widehat{Z}\right)^{-1} \text{ where } \widehat{Z}^{-1} = \operatorname{diag}[\widehat{z}], \\ \widehat{z_j} = \left(\frac{1}{2}\widehat{f_j}^2 + \beta_0\right)/(\frac{M}{2}+\alpha_{\varepsilon 0}-1), \\ \widehat{\tau_\varepsilon} = \left(\frac{1}{2}\|g-H\widehat{f}\|^2 + \beta_{\varepsilon 0}\right)/(\frac{M}{2}+\alpha_{\varepsilon 0}-1). \end{cases} \tag{30}$$

One iteration of this algorithm is shown in Figure 3.

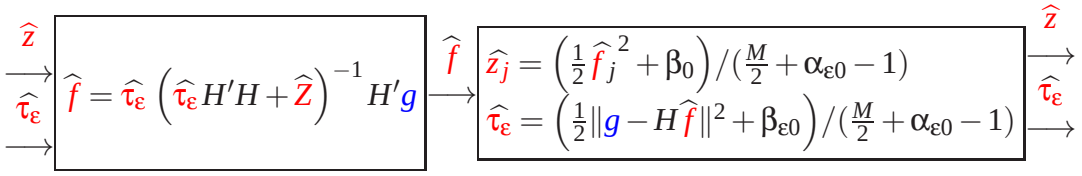

Figure 3: One iteration of the JMAP algorithm.

The main drawback of this method is that the uncertainties of the solution at each step is not accounted for for the next step.

- VBA posterior laws:

$$\left\{\begin{array}{l} q_1(f|\widetilde{\mu},\widetilde{\Sigma}) = \mathcal{N}(f|\widetilde{\mu},\widetilde{\Sigma}), \quad \widetilde{\mu} = \widetilde{\tau}\widetilde{\Sigma}H'g,\ \widetilde{\Sigma} = \left(\widetilde{\tau}H'H + \widetilde{Z}\right)^{-1} \text{with } \widetilde{Z}^{-1} = \text{diag}\,[\widetilde{z}], \\ q_{2j}(z_j) = \mathcal{G}(z_j|\widetilde{\alpha}_j,\widetilde{\beta}_j), \quad \widetilde{\alpha}_j = \alpha_0 + \frac{1}{2}, \quad \widetilde{\beta}_j = \beta_0 + < f_j^2 > /2, \\ q_3(\tau_\varepsilon) = \mathcal{G}(\tau_\varepsilon|\widetilde{\alpha}_\varepsilon,\widetilde{\beta}_\varepsilon) \\ \widetilde{\alpha}_\varepsilon = \alpha_{\varepsilon 0} + (n+1)/2, \quad \widetilde{\beta}_\varepsilon = \beta_{\varepsilon 0} + \frac{1}{2}[\|g\|^2 - 2\langle f\rangle' H'g + H' \langle f f'\rangle H]. \end{array}\right. \tag{31}$$

with

$$< f >= \widetilde{\mu},\ < f f' >= \widetilde{\Sigma} + \widetilde{\mu}\widetilde{\mu}',\ < f_j^2 >= [\widetilde{\Sigma}]_{jj} + \widetilde{\mu}_j^2,\ \widetilde{\tau} = \frac{\widetilde{\alpha}_{\tau_\varepsilon}}{\widetilde{\beta}_{\tau_\varepsilon}} \text{ and } \widetilde{z}_j = \frac{\widetilde{\alpha}_j}{\widetilde{\beta}_j}. \tag{32}$$

The expression of the free energies can be obtained as follows:

$$\begin{array}{l} \mathcal{F}(q) = \left\langle \ln \frac{p(f,z,\tau,g|\mathcal{M})}{q(f,z,\tau)} \right\rangle = \\ \langle \ln p(g|f,z,\tau)\rangle + \langle \ln p(f|z,\tau)\rangle + \langle \ln p(z|\tau)\rangle + \langle -\ln q(f)\rangle + \langle -\ln q(z)\rangle + \langle -\ln q(\tau)\rangle \end{array} \tag{33}$$

where

$$\langle \ln p(g|f,\tau_\varepsilon)\rangle = \frac{n}{2}(< \ln \tau_\varepsilon > - \ln(2\pi)) - \frac{1}{2}\{\langle\lambda\rangle g'g - 2 < f >' H'g + H' < f f' > H\}$$

$$\begin{array}{ll} \langle -\ln p(f|z)\rangle & = -\frac{n+1}{2}\ln(2\pi) - \frac{1}{2}\left\{\sum_j < \ln\alpha_j > < \alpha_j > < f_j^2 >\right\} \\ \langle -\ln p(z)\rangle & = -(n+1)\alpha_{\varepsilon_0}\ln\beta_{\varepsilon_0} + (\alpha_{\varepsilon_0} - 1)\sum_j < \ln\alpha_j > -\beta < \alpha_j > -(n+1)\ln\Gamma(\alpha) \\ \langle p(\tau_\varepsilon))\rangle & = c\ln d + (c-1) < \ln\tau_\varepsilon) > -d\langle\lambda\rangle - \ln\Gamma(c) \\ \langle -\ln q(f)\rangle & = -\frac{n+1}{2}(1 + \ln(2\pi)) - \frac{1}{2}\ln|\Sigma_f| \\ \langle -\ln q(z)\rangle & = -\sum_j \left[\widetilde{\alpha}_j \ln(\widetilde{\beta}_j) + (\widetilde{\alpha}_j - 1) < \ln\widetilde{\alpha}_j > -\widetilde{\beta}_j < \alpha_j > -\ln\Gamma(\widetilde{\alpha}_j)\right] \\ \langle q(\tau_\varepsilon)\rangle & = \tilde{c}\ln\tilde{d} + (\tilde{c}-1) < \ln\tau) > -\tilde{d}\langle\lambda\rangle - \ln\Gamma(\tilde{c}). \end{array}$$

In these equations,

$$\begin{cases} < \ln a_j >= \psi(\tilde{a}_j) - \ln \tilde{b}_j, \\ < \ln \tau >= \psi(\tilde{c}) - \ln \tilde{d}, \\ \psi(a) = \frac{\partial \ln \Gamma(a)}{\partial a}. \end{cases} \tag{34}$$

The three steps of this algorithm is shown in Figure 4.

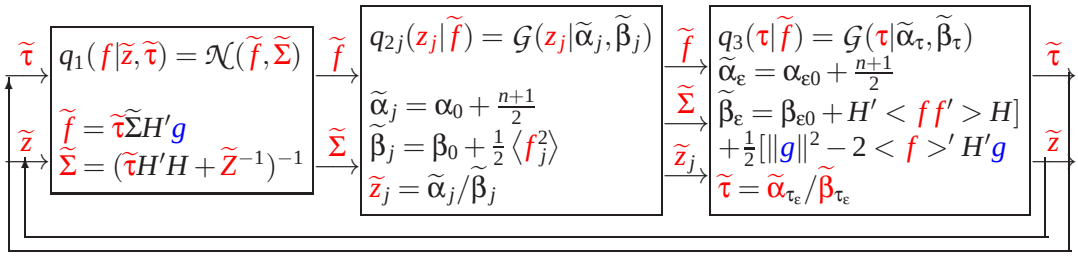

Figure 4: The three steps of the Bayesian Variational Approximation Algorithm.

BAYESIAN VARIATIONAL APPROXIMATION WITH MIXTURE OF GAUSSIANS PRIORS

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

$$p(f|\lambda, v_1, v_0) = \prod_j \left(\lambda \mathcal{N}(f_j|0,v_1) + (1-\lambda)\mathcal{N}(f_j|0,v_0)\right) \tag{35}$$

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

$$\begin{cases} p(f|z) = \prod_j p(f_j|z_j) & = \prod_j \mathcal{N}\left(f_j|0,v_{z_j}\right) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_{z_j}}\right\} \\ P(z_j = 1) = \lambda, & P(z_j = 0) = 1-\lambda \end{cases} \tag{36}$$

In general $v_1 >> v_0$ and λ measures the sparsity ($0 < \lambda << 1$) [11]. In this case also all the equations are very similarly can be obtained. Here, we do not have enough place to write them.

CONCLUSIONS

In this paper, a VBA method is proposed for doing Bayesian computations for inverse problems where a hierarchical prior modeling is used for the unknowns. In particular, two prior models are considered: the Student-t and the mixture of Gaussian models. In both cases, these priors can be written via hidden variables which gives the model a hierarchical structure which is used to do the factorization. For some applications see for example [19, 7, 4, 6, 5, 10, 8, 16] and two other related papers in this volume.

REFERENCES

1. M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
2. Sotirios Chatzis and Theodora Varvarigou. Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributionsclassification of binary random patterns. *IEEE Trans. on Fuzzy Systems*, 17:505–517, 2009.
3. R. A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, University of Oxford, 2002.
4. N. Chu, A. Mohammad-Djafari, and J. Picheral. Two robust super-resolution approaches with sparsity constraint and sparse regularization for near-field wideband extended aeroacoustic source imaging. In *Berlin Beamforming Conference 2012 (BeBeC2012)*, number 29, Berlin, Germany, Feb.22-23,2012.
5. N. Chu, A. Mohammad-Djafari, and J. Picheral. Bayesian sparse regularization in near-field wideband aeroacoustic imaging for wind tunnel test. In *11th CongrÃĺs FranÃğais d'Acoustique and 2012 IOA annual meeting*, Nantes, France, Apr.23-27,2012.
6. N. Chu, A. Mohammad-Djafari, and J. Picheral. A bayesian sparse inference approach in near-field wideband aeroacoustic imaging. In *2012 IEEE International Conference on Image Processing*, Orlando, USA, Sep.30-Oct.4, 2012.
7. N. Chu, J. Picheral, and A. Mohammad-Djafari. A robust super-resolution approach with sparsity constraint for near-field wideband acoustic imaging. In *IEEE International Symposium on Signal Processing and Information Technology*, pages 286–289, Bilbao, Spain, Dec.14-17,2011.
8. Mircea Dumitru and Ali Mohammad-Djafari. Estimating the period of a signal through inverse problem modeling and bayesian inference with sparsity enforcing prior. In *32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, IPP, Garching near Munich, Germany,15-20 July 2012.
9. Aurélia Fraysse and Thomas Rodet. A gradient-like variational Bayesian algorithm. In *SSP 2011*, number S17.5, pages 605–608, Nice, France, jun 2011.
10. Leila Gharsalli, Ali Mohammad-Djafari, Aurélia Fraysse, and Thomas Rodet. Variational bayesian approximation with scale mixture prior for inverse problems: a numerical comparison between three algorithms. In *32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, IPP, Garching near Munich, Germany,15-20 July 2012.
11. J. Rao. H. Ishwaran. Spike and Slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, 2005.
12. L. He, H. Chen, and L. Carin. Tree-Structured Compressive Sensing With Variational Bayesian Analysis. *IEEE Signal. Proc. Let.*, 17(3):233–236, 2010.
13. A. C. Likas and N. P. Galatsanos. A variational approach for bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 2004.
14. A Mohammad-Djafari. Bayesian approach with prior models which enforce sparsity in signal and image processing. *EURASIP Journal on Advances in Signal Processing*, Special issue on Sparse Signal Processing (1 Mars 2012):2012:52, 2012.
15. T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 2008.
16. R. Pérenon, A. Mohammad-Djafari, E. Sage1, L. Duraffourg, S. Hentz1, A. Brenac, R. Morel, and P. Grangeat. Mcmc-based bayesian estimation algorithm dedicated to NEMS mass spectrometry. In *32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, IPP, Garching near Munich, Germany,15-20 July 2012.
17. M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 2001.
18. John Winn, Christopher M. Bishop, and Tommi Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.
19. Sha Zhu, Ali Mohammad-Djafari, Hongqiang Wang, Bin Deng, Xiang Li, and Junjie Mao. Parameter estimation for sar micromotion target based on sparse signal representation. *Eurasip Journal of Signal Processing*, special issue "sparse approximations in signal and image processing", 2012.