Bayesian inference with error variable splitting and sparsity enforcing priors for linear inverse problems

Ali Mohammad-Djafari, Mircea Dumitru, Camille Chapdelaine, Nicolas Gac
Laboratoire des signaux et systèmes (L2S), CNRS-CentraleSupélec-Univ Paris Saclay, 91192 Gif-sur-Yvette, France
Email: djafari@lss.supelec.fr

Abstract—Regularization and Bayesian inference based methods have been successfully applied for linear inverse problems. In these methods, often simple Gaussian or Poisson models for the forward model errors have been considered. In this work, we use variable splitting for the errors to model different sources of errors and their possible non-stationarity or impulsive nature using Student-t or other heavy tailed distributions. Also, as a prior model, a sparsity enforcing hierarchical model of Infinite Gaussian Mixture model is introduced. With these prior models, we obtain a complete Bayesian inference framework which can efficiently be implemented for any linear inverse problem. Interestingly, many recent regularization-based algorithms such as Alternating Direction Method of Multipliers (ADMM) as well as more classical Bayesian based methods such as Sparse Bayesian Learning (SBL) are obtained as particular cases. One advantage of the Bayesian approach is the possibility to estimate, jointly with the reconstruction, the hyper-parameters such as the regularization parameter, thus the capability of proposing unsupervised methods. Examples of implementation of the proposed method in different signal and image processing such as deconvolution in mass spectrometry, estimation of periodic components estimation in biological signals and computed tomography are mentioned and referenced.

keywords: Variable splitting, Bayesian inference; Sparsity enforcing, Inverse problems, Approximate Bayesian computation (ABC)

I. INTRODUCTION

The classical single error term forward model for linear inverse problems is:

\[ g = Hf + \epsilon, \]

where all the uncertainties are summarized by \( \epsilon \). For this simple model, nowadays, almost everything has been told, starting by deterministic methods: Least Squares (LS), then Quadratic Regularization (QR), \( L_1 \) regularization, Total Variation and all associated optimization algorithms such as Augmented Lagrangian (AL), ADMM, ISTA, FISTA, etc. [1], [2]. But, also the probabilistic methods and in particular the Bayesian approach with simple Gaussian models for the noise and Gaussian prior model, Double Exponential prior, Student-t prior [3] to much more sophisticated Hierarchical models [4], [5], [6].

However, in many real applications, it is needed to propose forward models which can account for other sources of uncertainties. For example, if we want to distinguish between the measurement noise and model uncertainties, we can propose the following variable splitting model:

\[ g = Hf + \xi + \epsilon \]  

which can also be written as:

\[
\begin{align*}
    g &= g_0 + \epsilon, \\
    g_0 &= Hf + \xi,
\end{align*}
\]

where \( \epsilon \) represents the measurement noise and \( \xi \) represents the modelling errors [7].

In this paper, we consider this noise splitting forward model and propose to model \( \epsilon \) by a centered white Gaussian model \( \mathcal{N}(\epsilon|0,v,\mathbf{I}) \) with unknown variance \( v \) and \( \xi \) with a heavy tailed Student-t model as we want it to represent a sparse error term. The choice of Student-t is motivated by the equivalent Infinite Gaussian Scaled Mixture (IGSM) model of the Student-t.

For the unknown \( f \) also, we propose the following hierarchical model

\[ f = Dz + \zeta \]

where \( D \) is any linear transformation which can be invertible, unitary or represent an overdetermined Dictionary and \( \zeta \) an error term for this prior model. We propose to model it by a Gaussian \( \mathcal{N}(\zeta|0,v_\zeta\mathbf{I}) \) with unknown variance \( v_\zeta \). Finally, as \( z \) is, in general, sparse, we model it again via the Student-t probability distribution with its equivalent hierarchical IGSM model.

The rest of this paper is organized as follows: In section 2, we present the details of the proposed forward model, expressions of the likelihood, priors and the posterior. In Section 3, the MAP estimator and the details of the optimization algorithm are presented. In Section 4, the details of the VBA algorithm are presented. In section 5, some applications where we have implemented the proposed methods are just mentioned, and finally in section 6, the conclusions are described.

II. PROPOSED FORWARD MODEL

The proposed forward model can be summarized in the following three equations:

\[
\begin{align*}
    g &= g_0 + \epsilon, \\
    g_0 &= Hf + \xi, \\
    f &= Dz + \zeta,
\end{align*}
\]
where \( g \) represents the observed data, \( \epsilon \) the measurement errors, \( g_0 \) the ideal no-noise data, \( H \) the forward model, \( \xi \) the model error, \( f \) the interesting real unknowns of the model, \( D \) any linear transformation model, \( z \) the corresponding representation of the unknowns in this transformed domain which is, in general, sparse and finally, \( \zeta \) the errors of this transformation \([8]\).

With this model, a first deterministic regularization based method can be proposed by defining the following criterion

\[
J(f, g_0, z) = \frac{1}{2\nu_\epsilon} \| g - g_0 \|^2_2 + \frac{1}{2\nu_\epsilon} \| g_0 - Hf \|^2_2 + \frac{1}{2\nu_\zeta} \| f - Dz \|^2_2 + \alpha \| z \|_1
\]

(6)

and trying to optimize it. This criterion can be assimilated as the criterion of the Maximum A Posteriori (MAP) estimate in a Bayesian framework

\[
(f, g_0, z) = \arg \max \{ p(f, g_0, z|g) \}
\]

(7)

where \( p(f, g_0, z|g) \propto p(g|g_0)p(g_0|f)p(f|z)\) with the following expressions:

\[
\begin{align*}
    p(\epsilon) &= \mathcal{N}(\epsilon|0, \nu_\epsilon I) \rightarrow p(g|g_0) = \mathcal{N}(g|g_0, \nu_\epsilon I) \\
    p(\xi) &= \mathcal{N}(\xi|0, \nu_\xi I) \rightarrow p(g_0|f) = \mathcal{N}(g_0|f, \nu_\xi I) \\
    p(z) &= \mathcal{N}(z|0, \nu_\zeta I) \rightarrow p(f|z) = \mathcal{N}(f|z, \nu_\zeta I) \\
    p(z) &= \mathcal{D}\mathcal{E}(z|\alpha) \propto \exp[\alpha \| z \|_1].
\end{align*}
\]

(8)

The following particular cases are of great interest:

1) Quadratic regularization is obtained with \( \xi = 0 \) (i.e. \( g_0 = Hf \)) and \( D = 0 \) (i.e. \( f = \zeta \)). The criterion becomes the classical quadratic regularization:

\[
J(f) = \frac{1}{2} \| g - Hf \|^2_2 + \mu \| f \|^2_2, \quad \text{with} \quad \mu = \frac{\nu_\epsilon}{2\nu_\zeta}
\]

(9)

2) Classical \( L_1 \) regularization is obtained with \( \xi = 0 \), \( D = I \) and \( \zeta = 0 \) (i.e. \( f = z \)). The criterion becomes:

\[
J(f) = \frac{1}{2} \| g - Hf \|^2_2 + \mu \| f \|_1, \quad \text{with} \quad \mu = \frac{\nu_\epsilon}{\alpha}
\]

(10)

3) Synthesis \( L_1 \) regularization is obtained with \( \xi = 0 \) and \( \zeta = 0 \) (i.e. \( f = Dz \)). The criterion becomes:

\[
J(z) = \frac{1}{2} \| g - HDz \|^2_2 + \mu \| z \|_1, \quad \text{with} \quad f = Dz
\]

(11)

4) Analysis \( L_1 \) regularization is obtained with \( \xi = 0 \), \( D = \) an invertible (\( f = Dz \) and \( z = D^{-1}f \)) or orthogonal matrix \( (D'D = I, f = Dz \) and \( z = D^Tf \)) and \( \zeta = 0 \). The criterion becomes equivalent to the classical Total Variation (TV):

\[
J(f) = \frac{1}{2} \| g - Hf \|^2_2 + \mu \| D'f \|_1, \quad \text{with} \quad \mu = \frac{\nu_\epsilon}{\alpha}
\]

(12)

5) The case \( \xi = 0 \), i.e \( g = Hf + \epsilon \) and \( \zeta = 0 \) where \( f = Dz \) and \( D \) is an invertible or orthogonal operator, i.e. \( z = D^Tf \) is the last interesting particular case because the solution is defined by the following optimization problem:

\[
J(f, z) = \frac{1}{2\nu_\epsilon} \| g - Hf \|^2_2 + \alpha \| z \|_1 \text{ s.t. } z = D^Tf
\]

(13)

Using the Augmented Lagrangian (AL) method to this linear constraint optimization problem consists in defining:

\[
J(f, z, \lambda) = \frac{1}{2\nu_\epsilon} \| g_0 - Hf \|^2_2 + \alpha \| z \|_1 + \| z - D'f \|^2_2 + \lambda' \| z - D'f \|_1
\]

(14)

and looking for its stationary point via the expressions of its gradients with respect to its arguments or as an alternate optimization algorithm:

\[
\begin{align*}
    f &:= \arg \min_f \left\{ \frac{1}{2\nu_\epsilon} \| g - Hf \|^2_2 + \| z - D'f \|^2_2 + \lambda' \| z - D'f \|_1 \right\} \\
    z &:= \arg \min_z \left\{ \| z - D'f \|^2_2 + \alpha \| z \|_1 + \lambda' \| z - D'f \|_1 \right\}, \\
    \lambda &:= \lambda + (z - D'f)
\end{align*}
\]

(15)

where we can compare it with some particular cases of ADMM.

6) In the previous case, considering \( D = I \), i.e. \( f = z \), we obtain

\[
\begin{align*}
    z &:= z + \alpha_1^{(k)} \left( H'(g - Hz) \right) \\
    f &:= \arg \min_f \left\{ \| f - z \|^2_2 + \alpha \| f \|_1 \right\} = \text{ST}_\alpha(z)
\end{align*}
\]

(16)

where \( \text{ST}_\alpha(z) \) represents the Soft Thresholding (ST) of \( z \) and we thus can compare it to the well known Iterative Soft Thresholding Algorithm (ISTA) \([9, 10, 11]\).

More generally, alternate optimization of this criterion with respect to \( f, g_0 \) and \( z \) becomes:

\[
\begin{align*}
    f^{(k+1)} &= f^{(k)} + \alpha_1^{(k)} \left( H'(g_0 - Hf^{(k)}) - \mu_f f^{(k)} \right), \\
    g_0^{(k+1)} &= g_0^{(k)} + \alpha_2^{(k)} \left( g_0^{(k)} - Hf^{(k)} \right), \\
    z^{(k+1)} &= \arg \min_z \left\{ \frac{1}{2\nu_\epsilon} \| f - Dz \|^2_2 + \alpha \| z \|_1 \right\}
\end{align*}
\]

(17)

can be compared to many special purpose optimization algorithms such as Augmented Lagrangian (AL) \([12, 11]\) and Bregman Duality (BD) \([13, 14, 12, 15, 16, 11, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]\).

III. ADVANTAGES OF BAYESIAN VERSUS REGULARIZATION APPROACHES

The main difficulties in these regularization based methods are twofold:

- How to determine the regularization parameters, and
- How to quantify the remaining uncertainty on the obtained solution.

The Bayesian approach gives these possibilities. To see this, let consider the simplest case where

\[
\begin{align*}
    g &= Hf + \epsilon, \quad \text{with} \quad p(\epsilon) = \mathcal{N}(\epsilon|0, \nu_\epsilon I) \\
    f &= \zeta, \quad \text{with} \quad p(f) = \mathcal{N}(f|0, \nu_f I)
\end{align*}
\]

(18)

For this case, we have:

\[
\begin{align*}
    p(f|g) &= \mathcal{N}(f|\hat{f}, \hat{\Sigma}) \quad \text{with} \\
    \hat{f} &= \arg \min_f \left\{ \| g - Hf \|^2 + \mu \| f \|^2 \right\} \\
    \hat{\Sigma} &= \nu_\epsilon H'H + \mu I
\end{align*}
\]

(19)
In the supervised case where \( v_c \) and \( v_f \) are known, the problem is entirely completed. We have access to the posterior law and any question about \( f \) can be answered and uncertainties quantified.

In practical situations we may not know these two hyper-parameters. In the Bayesian approach, it is possible to include the hyperparameters \( \theta = (v_c, v_f) \) in the estimation process by looking at the joint posterior:

\[
p(f, \theta | g) \propto p(g|f, v_c)p(f|v_f)p(\theta)
\]

with \( \theta = (\theta_1, \theta_2) \) [28], [29]. This joint posterior law can be used to infer \( f \) and \( \theta \) jointly and to quantify the uncertainty on the solution. This can be done, for example, by computing the posterior covariance. However, often, the exact computation of the posterior mean and covariance may not be easy or may be too costly. Hopefully, solutions exist. For example, we can use the Variational Bayesian Approximation (VBA) methods to approximate \( p(f, \theta | g) \) by \( q(f, \theta) = q_1(f)q_2(\theta) \) by choosing appropriate families for \( q_1 \) and \( q_2 \). For example, by choosing the a priori laws:

\[
\begin{align*}
p(v_c) &= \mathcal{IG}(v_c | \alpha_{v_c}, \beta_{v_c}), \\
p(v_f) &= \mathcal{IG}(v_f | \alpha_{v_f}, \beta_{v_f}),
\end{align*}
\]

it is easy to obtain the expression of the joint posterior law:

\[
p(f, v_c, v_f | g) \propto p(g|f, v_c)p(f|v_f)p(v_c)p(v_f)
\]

and approximating it by

\[
q(f, v_c, v_f) = q_1(f)q_2(v_c)q_3(v_f)
\]

with \( q_1(f) = \mathcal{N}(f | \mu_1, \Sigma_1) \), \( q_2(v_c) = \mathcal{IG}(v_c | \alpha_1, \beta_1) \) and \( q_3(v_f) = \mathcal{IG}(v_f | \alpha_2, \beta_2) \) we obtain:

\[
\begin{align*}
q_1(f) &= \mathcal{N}(f | \mu_1, \Sigma_1) \\
q_2(v_c) &= \mathcal{IG}(v_c | \mu_2, \Sigma_2) \\
q_3(v_f) &= \mathcal{IG}(v_f | \mu_3, \Sigma_3)
\end{align*}
\]

where the VBA algorithm becomes an iterative algorithm which updates successively \( (\mu_1, \Sigma_1), (\alpha_1, \beta_1) \) and \( (\alpha_2, \beta_2) \). At the convergence or at the end of the iterations we get the approximate separable \( q(f, v_c, v_f) = q_1(f)q_2(v_c)q_3(v_f) \) which we can use to infer on \( f, v_c \) and \( v_f \). For the expressions of the updating these tilted variables and more details see [30], [4], [31]. Figure 1 shows the generative graphical representations of the supervised and unsupervised models.

IV. FULL BAYESIAN FRAMEWORK FOR PROPOSED FORWARD MODELS

We keep the general forward model (5) with the same Gaussian prior models for \( \epsilon \) and \( \zeta \) but letting their variances to be unknown. For \( \xi \) we choose a Student-t model to be able to control its sparsity degree. For direct sparsity enforcing of \( f \) or for sparsity in the transform domain \( z \), we propose also to use Student-t in place of Double Exponential (DE) or Generalized Gaussian (GG) due to its scaled mixture property which gives us to propose a hierarchical Normal-Inverse Gamma model. In these following two cases are expanded.

A. Direct sparsity case

To summarize, for the direct sparsity of \( f \), we have the following relations:

1) \( g = g_0 + \epsilon, \quad \epsilon \text{ Gaussian:} \)

\[
\begin{align*}
p(g|g_0, \epsilon) &= \mathcal{N}(g|g_0, \epsilon I), \\
p(\epsilon) &= \mathcal{IG}(\epsilon | \alpha_\epsilon, \beta_\epsilon).
\end{align*}
\]

2) \( g_0 = H f + \xi, \quad \xi \text{ Student:} \)

\[
\begin{align*}
p(g_0|f, \xi) &= \mathcal{N}(g_0|H f, \Sigma_f), \\
p(\xi) &= \prod_{i=1}^{M} p(\xi_i) = \prod_{i=1}^{M} \mathcal{IG}(\xi_i | \alpha_\xi, \beta_\xi). \quad (26)
\end{align*}
\]

3) \( f \text{ Sparse:} \)

\[
\begin{align*}
p(f|v_f) &= \mathcal{N}(f | 0, \Sigma_f), \\
p(v_f) &= \prod_{j=1}^{N} p(v_{f_j}) = \prod_{j=1}^{N} \mathcal{IG}(v_{f_j} | \alpha_{v_f}, \beta_{v_f}) \quad (27)
\end{align*}
\]

which results in:

\[
p(f, g_0, \epsilon, \xi, v_f, v_f | g) \propto \exp \left[ -J(f, g_0, \epsilon, \xi, v_f, v_f) \right] \quad (28)
\]

with

\[
J(f, g_0, \epsilon, \xi, v_f, v_f) = \frac{1}{2\nu_c} \| g - g_0 \|^2 + (\alpha_{v_c} + 1) \ln v_c + \frac{\beta_{v_c}}{v_c} \\
+ \frac{1}{2} \| \Sigma^{-1/2}(g_0 - H f) \|^2 + \sum_{i=1}^{M} \left[ (\alpha_{\xi_i} + 1) \ln v_{\xi_i} + \frac{\beta_{\xi_i}}{v_{\xi_i}} \right]
\]

Alternate optimization of this criterion results to the optimization of the following criteria:

1) with respect to \( f \):

\[
J(f) = \frac{1}{2\nu_c} \| V^{-1/2}(g_0 - H f) \|^2 + \frac{1}{2\nu_c} \| f - D z \|^2 
\]

2) with respect to \( g_0 \):

\[
J(g_0) = \frac{1}{2\nu_c} \| g - g_0 \|^2 + \frac{1}{2} \| V^{-1/2}(g_0 - H f) \|^2 
\]

3) with respect to \( \epsilon \):

\[
J(\epsilon) = \frac{1}{2\nu_c} \| g - g_0 \|^2 + (\alpha_{\epsilon} + 1) \ln \epsilon + \frac{\beta_{\epsilon}}{\epsilon} 
\]

4) with respect to \( \xi \):

\[
J(\xi) = \frac{1}{2} \| V^{-1/2}(g_0 - H f) \|^2 + \sum_{i=1}^{M} \left[ (\alpha_{\xi_i} + 1) \ln v_{\xi_i} + \frac{\beta_{\xi_i}}{v_{\xi_i}} \right] 
\]

5) with respect to \( v_f \):

\[
J(v_f) = \frac{1}{2} \| V^{-1/2} f \|^2 + \sum_{j=1}^{N} \left[ (\alpha_{v_f} + 1) \ln v_{f_j} + \frac{\beta_{v_f}}{v_{f_j}} \right] 
\]

which can be compared to ADMM like algorithms with extra advantage which is updating of the hyperparameters which makes the method unsupervised and more robust.

The second main advantage is that we have access to the expressions of all the conditional posterior laws as well as to the joint posterior law up to its normalization constant. All the conditional posterior laws are either Gaussian or Inverse Gamma which can be handled easily and we have access to their means and variances.
Finally, the third advantage is that we can do better than JMAP by using the Variational Bayesian Approximation methods and for example approximate the joint posterior law 
\[ p(f, g_0, v_r, v_\xi, v_f | g) \]
by the following separable one

\[ q(f, g_0, v_r, v_\xi, v_f) = q_1(f)q_2(g_0)q_3(v_r)q_4(v_\xi)q_5(v_f) \]  
(30)

using the Kullback-Leibler divergence

\[ \text{KL}(q : p) = \int q \ln \frac{q}{p} \]  
(31)

which gives rise to the VBA algorithms [32], [33], [4], [34], [35], [36], [37].

The following figure shows the graphical representations of the new forward model.

![Graphical representations of the proposed model.](image)

Fig. 1. Graphical representations of the proposed model.

V. APPLICATIONS

The simple forward model 1 and the more detailed forward model 2 with different prior models for \( f \) (without or with another hierarchical level) and the two proposed models have been applied with succes in different application area:

- 1D biological signal processing and in particular periodic components estimation in short time biological time series [36], [5], [38].
- Signal deconvolution in Mass Spectrometry [39], [37].
- Image restoration in RAMAN Mass Spectrometry or in Radio Astronomy [39], [37].
- Detecting and estimating unknown periodic shapes in a short duration signal [39].
- 2D and 3D Industrial Computed Tomography (CT) for Non Destructive Testing (NDT) application [40], [8], [41], [42], [43], [44], [45].
- Low dose and limited angle CT for biological or medical applications [8], [46], [47], [44].

We refer the readers to the appropriate references mentioned for these applications for more details.

VI. CONCLUSIONS

In this work, we extended the classical single additive noise, linear forward model to account for different uncertainties by variable splitting techniques.

In the first step, we splitted the error term in two parts to distinguish between observation noise and forward model uncertainty and assigning to them different and appropriate prior probability distributions. Then, we examined in parallel the deterministic regularization and the Bayesian MAP approach focusing more on the second. As a by-product, we could see the links between an alternate optimization algorithm for the MAP estimator and the optimization algorithms such as Alternate Descent Minimization Maximization (ADMM) or ISTA or its fast version FISTA for the particular case of double exponential prior. The main advantage of the Bayesian approach are twofold: i) giving the tools to go further by estimating the hyperparameters and ii) being able to quantify the uncertainties of the solutions of the inversion problems. These two points are often crucial in real applications. However, one may be careful about choosing appropriate hyperparameters, the order of the optimization and the convergency of the algorithms. As the problems are in general very ill-posed, we have to carefully choose the priors to guarantee, at least, the local convergency of the algorithms. But this problem is a very general task for any inverse problem. One main question is still open: finding the reason for better performances of these variable splitting algorithms. This goes in the opposite direction of the regularization idea of restricting the space of the solution to obtain a regularized solution. As a final conclusion, we may try to answer the following question: Is it better to restrict the space and define a criterion with a global minimum or, in the opposite, increase the dimension of the unknown space, define a criterion which may have many minima and looking for a local minimum of it?

REFERENCES


[37] Ali Mohammad-Djafari, “Approximate bayesian computation tools for large scale inverse problems and hierarchical models for big data,” in booktitle=37th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Sao Paulo, Brazil, year=2017,


