Bayesian inference methods for sources separation

Ali Mohammad-Djafari
Laboratoire des signaux et systèmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
SUPELEC, Plateau de Moulon, 91192 Gif-sur-Yvette, France

ABSTRACT

The main aim of this paper is first to present the Bayesian inference approach for sources separation where we want to infer on the mixing matrix, the sources and all the hyper-parameters associated to probabilistic modeling (likelihood and priors). For this purpose, the sources separation problem is considered in four steps: i) Estimation of the sources when the mixing matrix is known; ii) Estimation of the mixing matrix when the sources are known; iii) Joint estimation of sources and the mixing matrix; and finally, iv) Joint estimation of sources and the mixing matrix, hidden variables and hyper-parameters. In all cases, one of the main steps is modeling of sources and the mixing matrix prior laws. We propose to use sparsity enforcing probability laws (such as Generalized Gaussian, Student-t and mixture models) both for the sources and the mixing matrix. For algorithmic and computational aspects, we consider either Joint MAP, MCMC Gibbs sampling or Variational Bayesian Approximation tools. For each class of methods we discuss about their relative costs and performances.

1 Introduction

The general sources separation problem can be viewed as an inference problem where first we provide a model linking the observed data (mixed signals) $g(t)$ to unknown sources $f(t)$ through a forward model. In this paper, we only consider the instantaneous mixing model:

$$g(t) = Af(t) + \varepsilon(t), \quad t \in [1, \cdots, T] \tag{1}$$

where $\varepsilon(t)$ represents the errors of modeling and measurement. A is called mixing matrix and when it is invertible, its inverse $B = A^{-1}$ is called the separating matrix. The second step is to write down the expression of the $p(f|A,g)$ when the mixing matrix A is known, $p(A|f,g)$ when the the sources f are known, or the joint posterior law:

$$p(f,A|g,\theta) = \frac{p(g|f,A,\theta_1)\, p(f|\theta_2)\, p(A|\theta_3)}{p(g|\theta_1,\theta_2,\theta_3)} \tag{2}$$

where $p(g|f,A,\theta_1)$ is the likelihood and $p(f|\theta_2)$ and $p(A|\theta_3)$ are the priors on sources and the mixing matrix and $\theta = (\theta_1,\theta_2,\theta_3)$ represent the hyper-parameters of the problem. In real and effective problems, we also have to estimate them. This can be done, by assigning them a prior $p(\theta)$ often separable $p(\theta) = p(\theta_1)\,p(\theta_2)\,p(\theta_3)$ and writing the expression of the joint posterior:

$$p(f,A,\theta|g) = \frac{p(g|f,A,\theta_1)\,p(f|\theta_2)\,p(A|\theta_3)\,p(\theta)}{p(g)} \tag{3}$$

In this paper, we will consider these two cases with different prior modeling for sources $p(f|\theta_2)$ and different priors for the mixing matrix $p(A|\theta_3)$. In particular, we consider the Generalized Gaussian (GG), Student-t (St), Elastic net (EN) and Mixture of Gaussians (MoG) models. Some of these models are well-known [2, 5–10], some others less. In general, we can classify them in two categories: i) Simple Non Gaussian models with heavy tails and ii) Mixture models with hidden variables z which result to hierarchical models.

The second main step in the Bayesian approach is to do the computations. The Bayesian computations in general can be:

- Joint optimization of $p(f,A,\theta|g)$ which needs optimization algorithms;
- MCMC Gibbs sampling methods which need generation of samples from the conditionals $p(f|A,\theta,g)$, $p(A|f,\theta,g)$ and $p(\theta|f,A,g)$;
- Bayesian Variational Approximation (BVA) methods which approximate $p(f,A,\theta|g)$ by a separable one

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

and then using them for the estimation [11].

The rest of the paper is organized as follows:

In section 2, we review a few prior models which are frequently used in particular when sparsity has to be enforced. For example, the Generalized Gaussian (GG) with two particular cases of Gaussian (G) and Double Exponential (DE) or Laplace, the Student-t model which can be interpreted as an infinite mixture with a variance hidden variable, Elastic net and the mixture models.

In Section 3, we examine in details the estimation of the sources f when the mixing matrix A is known. In section 4 the estimation of the mixing matrix A when the sources f are known is considered. In section 5 we examine the joint estimation of the mixing matrix A and the sources f, and finally, the more realistic case of joint estimation of the mixing matrix A, the sources f, their hidden variables z and the hyper-parameters θ.

In section 6, we give principal practical algorithms which can be used in real applications, and finally, in section V we show some results and real applications and in particular in SAR imaging [12] and Sonar sources localization [3, 4].

2 Prior models enforcing sparsity

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 [1]). Its expression is:

$$p(f|\gamma,\beta) = \prod_j \mathcal{GG}(f_j|\gamma,\beta) \propto \exp\left\{-\gamma\sum_j |f_j|^\beta\right\} \tag{4}$$

where

$$\mathcal{GG}(f_j|\gamma,\beta) = \frac{\beta\gamma}{2\Gamma(1/\beta)}\exp\left\{-\gamma|f_j|^\beta\right\} \tag{5}$$

and where the particular cases of $\beta = 2$ (Gaussian):

$$p(f|\gamma) = \prod_j \mathcal{N}(f_j|0,1/(2\gamma)) \propto \exp\left\{-\gamma\sum_j |f_j|^2\right\} \propto \exp\left\{-\gamma\|f\|_2^2\right\} \tag{6}$$

$\beta = 1$ (Double exponential or Laplace):

$$p(f|\gamma) = \prod_j \mathcal{DE}(f_j|\gamma) \propto \exp\left\{-\gamma\sum_j |f_j|\right\} \propto \exp\left\{-\gamma\|f\|_1\right\} \tag{7}$$

as well as the cases where $0 < \beta < 1$ are of great interest for sparsity enforcing.

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

The second simplest model is the Student-t model:

$$p(f|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \propto \exp\left\{-\frac{\nu+1}{2}\sum_j \log\left(1+f_j^2/\nu\right)\right\} \tag{8}$$

where

$$\mathcal{S}t(f_j|\nu) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1+f_j^2/\nu\right)^{-(\nu+1)/2} \tag{9}$$

Knowing that

$$\mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0,1/\tau_j)\mathcal{G}(\tau_j|\nu/2,\nu/2)\,\mathrm{d}\tau_j \tag{10}$$

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

$$\begin{array}{ll} p(f|\tau) & = \prod_j p(f_j|\tau_j) = \prod_j \mathcal{N}(f_j|0,1/\tau_j) \propto \exp\left\{-\frac{1}{2}\sum_j \tau_j f_j^2\right\} \\ p(\tau_j|\alpha,\beta) & = \mathcal{G}(\tau_j|\alpha,\beta) \propto \tau_j^{(\alpha-1)}\exp\left\{-\beta\tau_j\right\} \text{ with } \alpha=\beta=\nu/2 \end{array} \tag{11}$$

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

$$p(f) = \prod_j \mathcal{C}(f_j) \propto \exp\left\{-\sum_j \log\left(1+f_j^2\right)\right\} \tag{12}$$

2.3 Elastic Net (EN) prior model

A prior model inspired from Elastic Net regression literature [**?**] is:

$$p(f|v) = \prod_j \mathcal{EN}(f_j|v) \propto \exp\left\{-\sum_j (\gamma_1|f_j| + \gamma_2 f_j^2)\right\} \tag{13}$$

where

$$\mathcal{EN}(f_j|v) = \mathcal{N}(0, 1/\gamma_1)\,\mathcal{DE}(\gamma_1) = c\exp\left\{-\gamma_1|f_j| - \gamma_2 f_j^2)\right\} \tag{14}$$

which is a product of a Gaussian and a Double Exponential pdfs.

2.4 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:

$$p(f|\lambda, v_1, v_0) = \prod_j \left(\lambda\mathcal{N}(f_j|0, v_1) + (1-\lambda)\mathcal{N}(f_j|0, v_0)\right) \tag{15}$$

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

$$\left\{\begin{array}{ll} p(f|z) = \prod_j p(f_j|z_j) & = \prod_j \mathcal{N}\left(f_j|0, v_{z_j}\right) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_{z_j}}\right\} \\ P(z_j = 1) = \lambda, & P(z_j = 0) = 1-\lambda \end{array}\right. \tag{16}$$

In general $v_1 >> v_0$ and λ measures the sparsity ($0 < \lambda << 1$).

2.5 Bernoulli-Gaussian (BG) model

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

$$p(f|\lambda, v) = \prod_j p(f_j) = \prod_j \left(\lambda\mathcal{N}(f_j|0, v) + (1-\lambda)\delta(f_j)\right) \tag{17}$$

which can also be written as

$$\left\{\begin{array}{ll} p(f|z) = \prod_j p(f_j|z_j) & = \prod_j \left[\mathcal{N}(f_j|0, v)\right]^{z_j} \prod_j \left[\delta(f_j)\right]^{(1-z_j)} \\ P(z_j = 1) = \lambda, & P(z_j = 0) = 1-\lambda \end{array}\right. \tag{18}$$

This model has also been called *spike and slab* [6].

3 Bayesian inference for sources f when A is known

First let assume the error ε to be centered, Gaussian and white: $\varepsilon \sim \mathcal{N}(\varepsilon|0, v_\varepsilon I)$. Then, using the forward model (1) we have

$$p(g|f,A) = \mathcal{N}(g|Af, v_\varepsilon I) \propto \exp\left\{\frac{1}{2v_\varepsilon}\|g - Af\|^2\right\} \tag{19}$$

Now, we consider different priors for f.

3.1 Simple prior models

Given $p(g|f,A)$ and any simple prior law $p(f)$, the posterior law is written:

$$p(f|g,A) \propto p(g|f,A)\, p(f) \propto \exp\{J(f)\} \tag{20}$$

with

$$J(f) = \frac{1}{2v_\varepsilon}\|g - Af\|^2 + \Omega(f) \tag{21}$$

where $\Omega(f) = -\ln p(f)$ 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. The MAP estimate with this prior can be summarized in the following scheme:

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

Thus, depending on the choice of the prior we obtain different expressions for $\Omega(f)$. For example for the GG model of (4) we get

$$\Omega(f) = \gamma \sum_j |f_j|^\beta). \tag{22}$$

For the Student-t model (8) we get

$$\Omega(f) = \frac{\nu+1}{2}\sum_j \log\left(1 + f_j^2/\nu\right). \tag{23}$$

For the Elastic Net model we get

$$\Omega(f) = \sum_j \left[\gamma_1 |f_j| + \gamma_2 f_j^2\right] \tag{24}$$

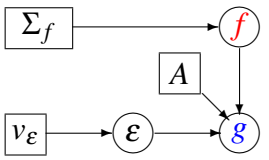
For each of these cases, we may discuss on the unimodality and convexity of the criterion $J(f)$ which depends mainly on its Hessian

$$\Delta J(f) \quad = \left[\frac{\partial^2 J(f)}{\partial f_j \partial f_i}\right] = A'A + \left[\frac{\partial^2 \Omega(f)}{\partial f_i \partial f_j}\right] \tag{25}$$

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

3.2 Gaussian priors

The case of the Gaussian prior is interesting, because, in that case, the posterior is also Gaussian and we have an analytical solution for $\widehat{f}$.



$$p(f|v_f\Sigma_f) = \mathcal{N}(0, v_f\Sigma_f) \propto \exp\left\{-\frac{1}{2v_f}\|Df\|^2\right\} \quad \text{with} \quad \Sigma_f = (D'D)^{-1}$$

The MAP estimate can be computed as:

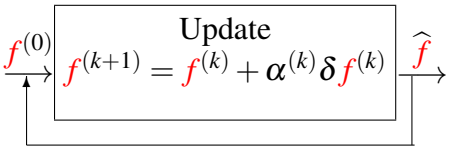
$$\widehat{f} = \arg\min_f \{J(f)\} \text{ with } J(f) = \|g - Af\|^2 + \lambda\|Df\|^2 \tag{26}$$

and the solution is

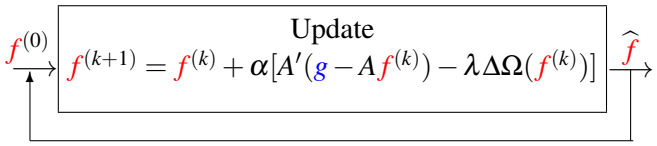
$$\widehat{f} = (A'A + \lambda D^t D)^{-1} A'g \tag{27}$$

3.3 Non Gaussian priors

For other cases, the optimization is done iteratively. The basics of this algorithm is shown in the following scheme:



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



where $\Delta\Omega$ here means the gradient of $\Omega(f)$.

3.4 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 f and the hidden variables: τ in Cauchy model, z in MoG2 and BG models. Let examine these a little in details.

Student-t and Cauchy models

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

$$\begin{array}{l} p(f,\tau) = \prod_j p(f_j|\tau_j)p(\tau_j) = \prod_j \mathcal{N}(f_j|0,1/\tau_j)p(\tau_j) \\ \propto \exp\left\{-\frac{1}{2}\sum_j \tau_j f_j^2 + \alpha \ln \tau_j - \beta \tau_j\right\} \text{ with } \alpha = \beta = \nu/2 \end{array} \tag{28}$$

such that

$$p(f,\tau|g) \propto p(g|f)p(f,\tau) \propto \exp\{-J(f,\tau)\} \tag{29}$$

where

$$J(f,\tau) = \frac{1}{2\nu_\varepsilon}\|g - Af\|^2 + \sum_j \frac{1}{2}\tau_j f_j^2 - \alpha \ln \tau_j + \beta \tau_j \tag{30}$$

$$\boxed{p(f,\tau|g)} \longrightarrow \boxed{\begin{array}{c}\text{Joint Optimization of} \\ J(f,\tau)\end{array}} \begin{array}{l}\longrightarrow \widehat{f} \\ \longrightarrow \widehat{\tau}\end{array}$$

Joint optimization of this criterion, alternatively with respect to f (with fixed τ)

$$\widehat{f} = \arg\min_f \{J(f,\tau)\} = \arg\min_f \left\{\frac{1}{2\nu_\varepsilon}\|g - Af\|^2 + \sum_j \frac{1}{2}\tau_j f_j^2\right\} \tag{31}$$

and with respect to τ (with fixed f)

$$\widehat{\tau} = \arg\min_\tau \{J(f,\tau)\} = \arg\min_\tau \left\{\sum_j \frac{1}{2}\tau_j f_j^2 - \alpha \ln \tau_j + \beta \tau_j\right\} \tag{32}$$

results in the following iterative algorithm:

$$\left\{\begin{array}{l} \widehat{f} = [A'A + \nu_\varepsilon D(\widehat{\tau})]^{-1} A'g \\ \widehat{\tau}_j = \phi(\widehat{f}_j) = \frac{\alpha}{\widehat{f}_j^{\,2} + \beta} \\ D(\widehat{\tau}) = \text{diag}\left[1/\widehat{\tau}_j, j = 1, \cdots, n\right] \end{array}\right. \tag{33}$$

$$\widehat{\tau} \longrightarrow \boxed{[A'A + \nu_\varepsilon D(\widehat{\tau})]^{-1} A'g} \longrightarrow \widehat{f} \longrightarrow \boxed{\phi(\widehat{f}_j)} \longrightarrow \widehat{\tau}$$

Note that τ_j is inverse of a variance and we have $1/\tau_j = \frac{f_j^2+\beta}{\alpha}$. 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 unimodal and its alternate optimization converges to its unique solution.

We may also consider a Gibbs sampling scheme

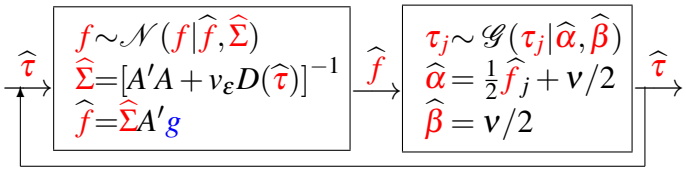
$$\left\{\begin{array}{l} f\sim p(f|\tau,g)\propto p(g|f)p(f|u)=\mathcal{N}(f|\widehat{f},\widehat{\Sigma}) \\ \tau\sim p(\tau|f,g)\propto p(f|\tau)\ p(\tau) \ =\prod_j \mathcal{G}(\tau_j|\widehat{\alpha},\widehat{\beta}) \end{array}\right. \tag{34}$$

where

$$\left\{\begin{array}{l} \widehat{\Sigma}=[A'A+v_\varepsilon D(\tau)]^{-1} \\ \widehat{f}=\widehat{\Sigma}A'g \end{array}\right. \tag{35}$$

and

$$\left\{\begin{array}{ll} \widehat{\alpha}=\frac{1}{2}\widehat{f}_j+\alpha=\frac{1}{2}\widehat{f}_j+\nu/2 \\ \widehat{\beta}=\beta \qquad\quad =\nu/2 \end{array}\right. \tag{36}$$



MoG model

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

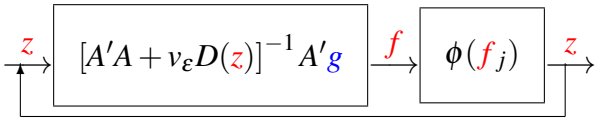
$$p(f,z|g)\propto p(g|f)p(f,z)\propto\exp\{-J(f,z)\} \tag{37}$$

where

$$\begin{array}{rl} J(f,z) & =\frac{1}{2v_\varepsilon}\|g-Af\|^2 \\ & \quad +\sum_j \frac{f_j^2}{2v_{z_j}}+z_j\ln\lambda+(1-z_j)\ln(1-\lambda) \end{array} \tag{38}$$

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

$$\left\{\begin{array}{l} \widehat{f}=[A'A+v_\varepsilon D(z)]^{-1}A'g \\ \widehat{z}_j=\phi(f_j)=\left\{\begin{array}{ll} 1, & \text{if } \widehat{f}_j^{\,2}\geq(v_1-v0)\ln\frac{1-\lambda}{\lambda} \\ 0, & \text{if } \widehat{f}_j^{\,2}<(v_1-v0)\ln\frac{1-\lambda}{\lambda} \end{array}\right. \\ D(\widehat{z})=\text{diag}\left[v_{\widehat{z}_j},j=1,\cdots,n\right] \end{array}\right. \tag{39}$$



Here too, we may also consider a Gibbs sampling scheme

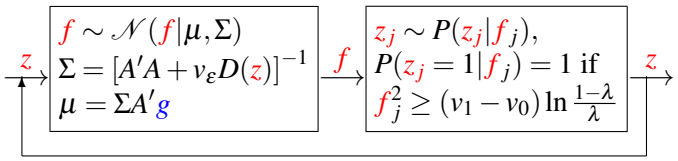
$$\left\{\begin{array}{l} f\sim p(f|z,g)\propto p(g|f)p(f|u)=\mathcal{N}(f|\widehat{f},\widehat{\Sigma}) \\ z\sim p(z|f,g)\propto p(f|z)\ p(z) \ =\prod_j P(z_j=k|f_j) \end{array}\right. \tag{40}$$

where

$$\begin{cases} \widehat{\Sigma}=[A'A+v_\varepsilon D(z)]^{-1} \\ \widehat{f}=\widehat{\Sigma}A'g \end{cases} \tag{41}$$

and

$$\begin{cases} P(z_j=1|f_j)=1, & \text{if } f_j^2 \geq (v_1-v0) \\ P(z_j=0|f_j)=1, & \text{if } f_j^2 < (v_1-v0)\ln\frac{1-\lambda}{\lambda} \end{cases} \tag{42}$$



4 Estimation of the mixing matrix A with known sources f

4.1 Sources separation: a bilinear problem

First, we may note that, the sources separation problem is a bilinear problems: $g=Af=Fa$. To see some details, consider the case of $N=2$ sources and $M=2$ mixtures:

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_1 & 0 & f_2 & 0 \\ 0 & f_1 & 0 & f_2 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix}$$

In the following, we use the following notations:

$$F=f\odot I, \quad a=\text{vec}(A)$$

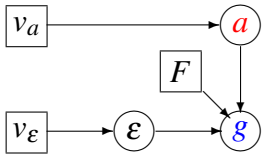
With this introduction and these notations, we may now compare the following cases:

- Estimation of f with known A: $g=Af$;
- Estimation of A with known f: $g=Af=Fa$; and
- Joint Estimation of f and A: $g=Af=Fa$.

We may note that the second and the third problems are more and more under determined.

4.2 The Gaussian case

So, if we follow now the same strategy for $g=Fa$ than the case of $g=Af$, we can summarize the results as follows:



$$\begin{aligned} p(a|v_f) &= \mathcal{N}(0,v_aI) \\ &\propto \exp\left\{-\tfrac{1}{2v_a}\textstyle\sum_j |a_j|^2\right\} \end{aligned}$$

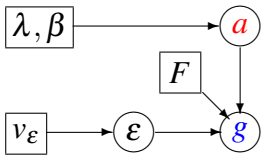
The MAP estimate is given by:

$$\widehat{a} = \arg\min_{a} \{J(a)\} \text{ with } J(a) = \|g - Fa\|^2 + \lambda \sum_j |a_j|^2 \tag{43}$$

which results to:

$$\widehat{a} = (F^t F + \lambda I)^{-1} F^t g \text{ or } \widehat{A} = g f^t (f f^t + \lambda I)^{-1}. \tag{44}$$

4.3 The Generalized Gaussian case



$$\begin{aligned} p(a|\lambda,\beta) &= \mathcal{GG}(\lambda,\beta) \\ &\propto \exp\left\{-\lambda \textstyle\sum_j |a_j|^\beta\right\} \end{aligned}$$

The MAP estimate becomes

$$\widehat{a} = \arg\min_{a} \{J(a)\} \text{ with } J(a) = \|g - Fa\|^2 + \lambda \sum_j |a_j|^\beta \tag{45}$$

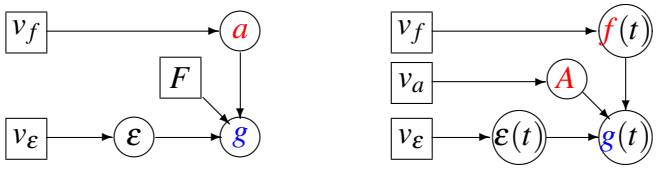
but unfortunately there is no analytic expression for the solution. However, its numerical computation can be done very easily.

5 Joint estimation of the sources f and the mixing matrix A

Now, we consider the main problem, and as in the previous cases, we consider different cases and summarize them as follows:

5.1 Gaussian priors for f and A

Now consider the joint estimation which is illustrated in two forms, one as usual and the second with considering the time index too:



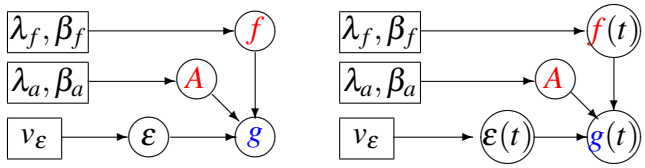
and the equations can be summarized as follows:

$$p(A, f(t)|g(t), v_\varepsilon) \propto p(g(t)|A, f(t), v_\varepsilon)\; p(A|v_a)\; p(f(t)|v_f) \tag{46}$$

$$\begin{cases} \widehat{f}(t) = (\widehat{A}'\widehat{A} + \lambda_f I)^{-1} \widehat{A}' g(t), & \lambda_f = v_\varepsilon / v_f \\ \widehat{A} = \sum_t g(t) \widehat{f}'(t) \left(\sum_t \widehat{f}(t) \widehat{f}'(t) + \lambda_a I\right)^{-1} & \lambda_a = v_\varepsilon / v_a \end{cases} \tag{47}$$

5.2 Gaussian prior for f and Generalized Gaussian prior for A

This case also can be summarized as follows:



$$\begin{aligned} p(f|v_f) &= \mathcal{N}(0, v_f I) \propto \exp\left\{-\tfrac{1}{2v_f}\|f\|^2\right\} \\ p(a|\lambda_a, \beta_a) &= \mathcal{GG}(\lambda_a, \beta_a) \propto \exp\left\{-\lambda_a \textstyle\sum_j |a_j|^\beta\right\} \end{aligned} \tag{48}$$

Joint Posterior:

$$p(A, f|g, \theta) \propto p(g|A, f, v_\varepsilon)\, p(A|v_a)\, p(f|v_f)$$

Integration over f can be down easily

5.3 Gaussian prior for A and Non Gaussian for sources $f(t)$

Let make a comparison between the Gaussian and Non Gaussian priors for $f(t)$ when trying to implement joint MAP computation iteratively.

- For the **Gaussian laws** for the sources, at each iteration we have:

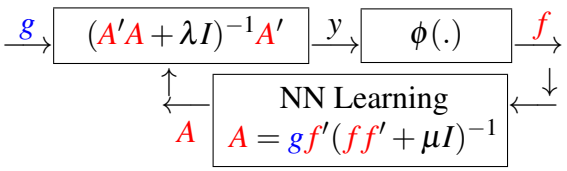
$$\left\{\begin{array}{lcl} f(t) & = & (A'A + \lambda I)^{-1} A' g(t) \\ A & = & \sum_t g(t) f'(t) (\sum_t f(t) f'(t) + \mu I)^{-1} \end{array}\right. \tag{49}$$

  with $\lambda = \sigma_\varepsilon^2/\sigma_s^2$ and $\mu = \sigma_\varepsilon^2/\sigma_a^2$.

- For the **Non Gaussian** sources $f(t)$, we have:

$$\left\{\begin{array}{lcl} y(t) & = & (A'A + \lambda I)^{-1} A' g(t) \\ f(t) & = & \phi(y(t)) \\ A & = & \sum_t g(t) f'(t) (\sum_t f(t) f'(t) + \mu I)^{-1} \end{array}\right.$$

For both cases, we can see the following scheme:



For Gaussian case $\phi(y) = y$ and a nonlinear function for Non Gaussian case.

5.4 Joint estimation: General case

Let consider now the following general notations which can also be used for accounting for time dependent sources (colored or non stationary sources).

$$g(t) = A\,f(t) + \varepsilon(t) \longrightarrow g_{1..T} = A f_{1..T} + \varepsilon_{1..T} \tag{50}$$

If we assume that the noise is iid and white, then

$$p(g_{1..T}|f_{1..T}) = \prod_t \prod_i p_i(g_i(t) - [Af]_i(t)) \tag{51}$$

and if we assume the same for the sources

$$p(f_{1..T}) = \prod_t \prod_j p_j(f_j(t)), \tag{52}$$

and

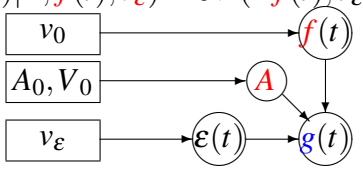
$$p(A) = \prod_i \prod_j p(a_{ij}) \propto \exp\left\{-\frac{1}{2\sigma_a^2}\sum_i\sum_j (a_{ij} - a_{0ij})^2\right\}, \tag{53}$$

then we have:

$$p\left(A, f_{1..T}|g_{1..T}\right) \propto \prod_t \prod_i p_i\left(g_i(t) - [Af]_i(t)\right) + \prod_t \prod_j p_j(f_j(t)) + \prod_i \prod_j p(a_{ij}) \tag{54}$$

Let consider first the Gaussian priors:

$$\begin{array}{l} p(f_j(t)|v_{0j}) = \mathcal{N}(0, v_{0j}) \\ p(f(t)|v_0) \propto \exp\left\{-\frac{1}{2}\sum_j f_j^2(t)/v_{0j}\right\} \\ p(A_{ij}|A_{0ij}, V_{0ij}) = \mathcal{N}(A_{0ij}, V_{0ij}) \\ p(A|A_0, V_0) = \mathcal{N}(A_0, V_0) \\ p(g(t)|A, f(t), v_\varepsilon) = \mathcal{N}(Af(t), v_\varepsilon I) \end{array} \tag{55}$$



We also assume the following:

- All sources a priori same variance v_f and note $v_0 = [v_f, .., v_f]'$.
- All noise terms a priori same variance v_ε and note $v_\varepsilon = [v_\varepsilon, .., v_\varepsilon]'$
- $A_0 = 0, \quad$ and $\quad V_0 = v_a I$

Then, we have:

$$p(f(t)|g(t), A, v_\varepsilon, v_0) = \mathcal{N}(\widehat{f}(t), \widehat{\Sigma}) \text{ with } \left\{ \begin{array}{l} \widehat{\Sigma} = (A'A + \lambda_f I)^{-1} \\ \widehat{f}(t) = (A'A + \lambda_f I)^{-1} A' g(t) \\ \lambda_f = v_\varepsilon / v_f \end{array} \right. \tag{56}$$

$$p(A|g(t),f(t),v_\varepsilon,A_0,V_0)=\mathcal{N}(\widehat{A},\widehat{V}) \text{ with } \begin{cases} \widehat{V}=(F'F+\lambda_f I)^{-1} \\ \widehat{A}=\sum_t g(t)f'(t)\,(\sum_t f(t)f'(t)+\lambda_a I)^{-1} \\ \lambda_a=v_\varepsilon/v_a \end{cases} \tag{57}$$

The joint MAP algorithm becomes:

$$\begin{array}{ccccc} A^{(0)} & \longrightarrow \widehat{A} \longrightarrow & \boxed{\left(\widehat{A}'\widehat{A}+\lambda_f I\right)^{-1}\widehat{A}'g} & \longrightarrow & \widehat{f}(t) \\ & \uparrow & & & \downarrow \\ & \widehat{A} \longleftarrow & \boxed{\sum_t g(t)\widehat{f}'(t)\left(\sum_t \widehat{f}(t)\widehat{f}'(t)+\lambda_a I\right)^{-1}} & \longleftarrow & \widehat{f}(t) \end{array} \tag{58}$$

6 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 θ and write the joint posterior law of all the unknowns:

$$p(f,A,z,\theta|g)\propto p(g|f,A,\theta_1)\,p(f|z,\theta_2)\,p(z|\theta_3)\,p(A|\theta_4)\,p(\theta) \tag{59}$$

where $\theta=(\theta_1,\theta_2,\theta_3,\theta_4)$ represents all the hyper-parameters of the problem.

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

$$q(f,A,z,\theta|g)=q_1(f|g)\,q_2(A|g)\,q_3(z|g)\,q_4(\theta|g) \tag{60}$$

illustrated here:

$$\boxed{p(f,A,z,\theta|g)} \longrightarrow \boxed{\begin{array}{c}\text{Variational}\\ \text{Bayesian}\\ \text{Approximation}\end{array}} \begin{array}{l} \longrightarrow q_1(f|g)\longrightarrow \widehat{f} \\ \longrightarrow q_2(A|g)\longrightarrow \widehat{A} \\ \longrightarrow q_3(z|g)\longrightarrow \widehat{z} \\ \longrightarrow q_4(\theta|g)\longrightarrow \widehat{\theta} \end{array}$$

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

$$\mathrm{KL}(q:p)=\int q\ln p/q=<\ln p/q>_q \tag{61}$$

It is then easy to show that

$$\mathrm{KL}(q:p)=\ln p(g|\mathcal{M})-\mathcal{F}(q) \tag{62}$$

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

$$p(g|\mathcal{M})=\iiiint p(f,A,z,\theta,g|\mathcal{M})\,\mathrm{d}f\,\mathrm{d}A\,\mathrm{d}z\,\mathrm{d}\theta \tag{63}$$

with

$$p(f,A,z,\theta,g|\mathcal{M})=p(g|f,A,\theta_1)\,p(A|\theta_2)\,p(f|z,\theta_3)\,p(z|\theta_4)\,p(\theta) \tag{64}$$

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

$$\mathscr{F}(q) = \left\langle \ln \frac{p(f,A,z,\theta,g|\mathscr{M})}{q(f,A,z,\theta)} \right\rangle_q \tag{65}$$

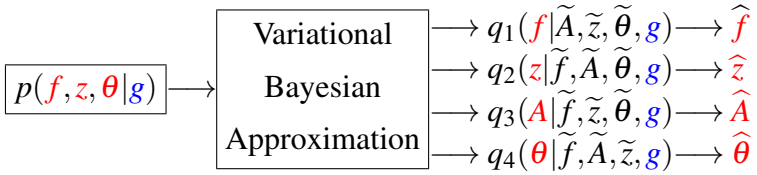
So, for a given model $\mathscr{M}$, minimizing $\mathrm{KL}(q:p)$ is equivalent to maximizing $\mathscr{F}(q)$ and when optimized, $\mathscr{F}(q^*)$ gives a lower bound for $\ln p(g|\mathscr{M})$.

An alternate optimization of $\mathscr{F}(q)$ with respect to q_1, q_2, q_3 and q_4 results in

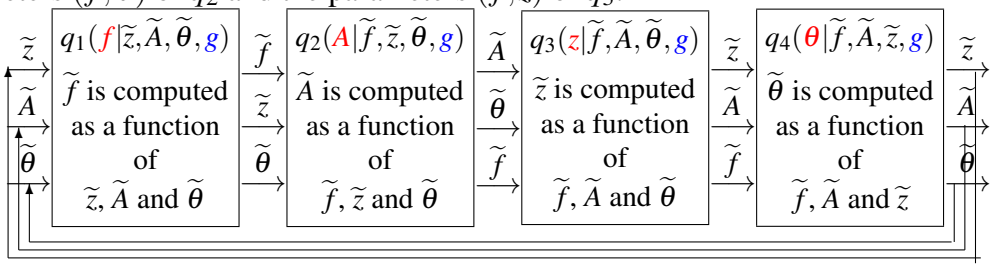
$$\begin{aligned}
q_1(f) &\propto \exp\left\{ -\langle \ln p(f,z,\theta,g)\rangle_{q_2(A)q_3(z)q_4(\theta)} \right\} \\
q_2(A) &\propto \exp\left\{ -\langle \ln p(f,A,z,\theta,g)\rangle_{q_1(f)q_3(z)q_4(\theta)} \right\} \\
q_3(z) &\propto \exp\left\{ -\langle \ln p(f,A,z,\theta,g)\rangle_{q_1(f)q_2(A)q_3(z)q_4(\theta)} \right\} \\
q_4(\theta) &\propto \exp\left\{ -\langle \ln p(f,A,z,\theta,g)\rangle_{q_1(f)q_2(A)q_3(z)} \right\}
\end{aligned}$$

Note that these relations represent an implicit solution for $q_1(f)$,$q_2(A)$, $q_3(z)$ and $q_4(\theta)$. For conjugate conditional distributions we consider, these leads to standard distributions for which the required expectations are easily evaluated. In that case, we may note

$$q(f,A,z,\theta|g) = q_1(f|\widetilde{A},\widetilde{z},\widetilde{\theta};g)\, q_2(A|\widetilde{f},\widetilde{A},\widetilde{\theta};g)\, q_3(z|\widetilde{f},\widetilde{A},\widetilde{\theta};g)\, q_4(\theta|\widetilde{f},\widetilde{A},\widetilde{z};g) \tag{66}$$



where the alternate optimization results to 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

$$\begin{aligned}
\mathscr{F}(q) &= \left\langle \ln \frac{p(f,A,z,\theta,g|\mathscr{M})}{q(f,A,z,\theta)} \right\rangle_q \\
&= \langle \ln p(f,A,z,\theta,g|\mathscr{M})\rangle_q + \langle -\ln q(f,A,z,\theta)\rangle_q \\
&= \langle \ln p(g|f,A,z,\theta)\rangle_q + \langle \ln p(f|z,A,\theta)\rangle_q + \langle \ln p(z|\theta)\rangle_q \\
&\quad + \langle -\ln q(f)\rangle_q + \langle -\ln q(z)\rangle_q + \langle -\ln q(\theta)\rangle_q
\end{aligned} \tag{67}$$

where all the expectations are with respect to q.

Other decompositions are also possible. To illustrate this approach, here we consider only the simple case of sources separation with Gaussian priors. For a more extensive examples in particular for inverse problems, the readers may refer to [7].

6.1 Variational Bayesian Approximation for sources separation

Here the main objective is to approximate $p(f_{1..T}, A|g_{1..T})$ by $q_1(f_{1..T}|A, g_{1..T})\, q_2(A|f_{1..T}, g_{1..T})$.

The solution is:

$$
\begin{aligned}
q_1(f(t)|g(t),A,v_\varepsilon,v_0) = \mathcal{N}(\widehat{f}(t),\widehat{\Sigma}) \quad \text{with} \quad & \left\{ \begin{array}{l} \widehat{\Sigma} = (A'A + \lambda_f \widehat{V})^{-1} \\ \widehat{f}(t) = (A'A + \lambda_f \widehat{V})^{-1} A' g(t), \\ \lambda_f = v_\varepsilon / v_f \end{array} \right. \\
q_2(A|g(t),f(t),v_\varepsilon,A_0,V_0) = \mathcal{N}(\widehat{A},\widehat{V}) \quad \text{with} \quad & \left\{ \begin{array}{l} \widehat{V} = (F'F + \lambda_f \widehat{\Sigma})^{-1} \\ \widehat{A} = \sum_t g(t) f'(t) \left(\sum_t f(t) f'(t) + \lambda_a \widehat{\Sigma} \right)^{-1} \\ \lambda_a = v_\varepsilon / v_a \end{array} \right.
\end{aligned}
\tag{68}
$$

Algorithm:

$$
\begin{array}{ccccc}
A^{(0)} \longrightarrow \widehat{A} \longrightarrow & \boxed{\widehat{f}(t) = \left(\widehat{A}'\widehat{A} + \lambda_f \widehat{V}\right)^{-1} \widehat{A}' g(t)} & \longrightarrow \widehat{f}(t) \\
V^{(0)} \longrightarrow \widehat{V} \longrightarrow & \boxed{\widehat{\Sigma} = (A'A + \lambda_f \widehat{V})^{-1}} & \longrightarrow \widehat{\Sigma} \\
\Uparrow & & \Downarrow \\
\widehat{A} \longleftarrow & \boxed{\widehat{A} = \sum_t g(t) \widehat{f}'(t) \left(\sum_t \widehat{f}(t) \widehat{f}'(t) + \lambda_a \widehat{\Sigma} \right)^{-1}} & \longleftarrow \widehat{f}(t) \\
\widehat{V} \longleftarrow & \boxed{\widehat{V} = (F'F + \lambda_f \widehat{\Sigma})^{-1}} & \longleftarrow \widehat{\Sigma}
\end{array}
$$

We may compare this algorithm with the joint MAP in 58. As we can see, in 58 the mixing matrix and the sources are updated without accounting for the uncertainties, but here the uncertainties of each step of the estimation is also updated and used in the next iteration.

7 Conclusion

The main aim of this paper was first to present the Bayesian inference approach for blind sources separation (BSS) where we want to infer on the mixing matrix, the sources and all the hyper-parameters associated to probabilistic modeling. For this purpose, the sources separation problem was considered in a step by step approach: i) Sources estimation when the mixing matrix is known; ii) Mixing matrix estimation when the sources are known; iii) Joint estimation of the mixing matrix and the sources; and finally, iv) Joint estimation of the mixing matrix, the sources and the associated hidden variables and hyper-parameters. One of the main steps in any of these steps is modeling of sources and the mixing matrix prior laws. Here, we focused on the sparsity enforcing probability laws (such as Generalized Gaussian, Student-t and mixture models) both for the sources and the mixing matrix. For algorithmic and computational aspects, we considered either Joint MAP, MCMC Gibbs sampling or Variational Bayesian Approximation (VBA)

tools. For each class of methods we discuss about their relative costs and performances.

References

[1] C. Bouman and K. Sauer. "A generalized gaussian image model for edge-preserving map estimation." *IEEE Trans. on Medical Imaging*, MI-2(3), 296–310, 1993.

[2] S. Chatzis and T. 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] N. CHU, A. M. 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*, page No.18. Berlin, Germany, Feb.22-23,2012.

[4] 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.

[5] J. Griffin and P. Brown. "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 2010.

[6] J. R. H. Ishwaran. "Spike and Slab variable selection: Frequentist and Bayesian strategies." *Annals of Statistics*, 2005.

[7] 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, 2012.

[8] T. Park and G. Casella. "The Bayesian Lasso." *Journal of the American Statistical Association*, 2008.

[9] N. Polson and J. Scott. "Shrink globally, act locally: sparse Bayesian regularization and prediction." *Bayesian Statistics 9*, 2010.

[10] H. Snoussi and J. Idier. "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures." *IEEE Trans. on Signal Processing*, 2006.

[11] M. Tipping. "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, 2001.

[12] S. Zhu, A. Mohammad-Djafari, H. Wang, B. Deng, X. Li, and J. Mao. "Parameter estimation for sar micromotion target based on sparse signal representation." *EURASIP Journal on Advances in Signal Processing*, Special issue on Sparse Signal Processing, 2012.