Determination of the number of sources in blind source separation

Mahieddine M. Ichir and Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes, CNRS-Supélec-UPS, Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette, France

Abstract. The determination of the number \( n \) of unobserved sources is an important issue in Blind Source Separation (BSS) of linear and instantaneous mixtures. However BSS is already a difficult task, so we generally assume that this number \( n \) is known and a priori fixed. In this paper, we address this issue as a Bayesian model selection problem and view the determination of this number \( n \) as a hypothesis testing problem via comparison of Bayes factors and study the computation of these factors by two numerical approximations: importance sampling from the posterior and simulated annealing sampling.

Seeking for a general solution may be tricky to this problem, so we will be interested in blind separation of sparse sources modeled by a double exponential prior \( \pi(.) \propto \exp(-\lambda |.|) \).

Keywords: Blind source separation, Bayesian model selection, Bayes factors, MCMC.

PACS: 02.50.Sk, 02.50.Ng, 02.50.Le

INTRODUCTION

Blind source separation (BSS) of linear and instantaneous mixtures consists in estimating sources \( s_{1:T} \) from a set of their linear and instantaneous mixtures \( x_{1:T} \):

\[
x_{1:T} = As_{1:T} + e_{1:T}
\]

where \( x_t \) is an \( m \)-column vector, \( s_t \) is an \( n \)-column vector at time \( t \) and \( A \) is the matrix of the mixing elements (the mixing matrix) of dimension \( m \times n \). The blind source separation problem consists in jointly estimating \( A \) and \( s_{1:T} \) (“blind” qualifies the non-knowledge of \( A \)). \( e_t \) is an \( m \)-column vector at time \( t \) of the observation or modeling error. In the following we assume that it is independent and identically Gaussian distributed with a diagonal covariance matrix \( R_e = \text{diag}(\sigma^2_1, \ldots, \sigma^2_m) \).

\[
e_t \sim \mathcal{N}(0, R_e), \quad \forall t
\]

If the number of unobserved sources \( n \) was known, the problem of blind source separation would have been to estimate \{ \( A, s_{1:T} \) \} \( \in \{ \mathbb{R}^{(m \times n)} \times \mathbb{R}^{(n \times T)} \} \). However if \( n \) is not known the only observation of \{ \( x_{1:T} \in \mathbb{R}^{(m \times T)} \) \} makes the problem extremely difficult and we may wonder: “Can we estimate the number \( n \) of unobserved sources from \( x_{1:T} \)?”.

This issue turns out to be very difficult in general and seeking for a “universal” answer could be, to our point of view, tricky. In this paper, we consider a very particular problem: blind separation of sparse sources. In a Bayesian estimation framework, these sources are a priori modeled by a double exponential prior. The problem of sparse sources arises in many problems as in transform domain like wavelets [6, 7].
Bayesian Prior Modeling

Sources Prior. As stated earlier, the problem considered in this paper is that of blind separation of centered sparse sources modeled by a double exponential prior:

\[ \pi(s_{i,t} | \lambda_i) = \frac{\lambda_i}{2} e^{-\lambda_i |s_{i,t}|} \]  

(3)

for \((i, t) \in (1, \ldots, n) \times (1, \ldots, T)\). These kind of priors lead to non tractable posterior distributions for the sources and the resulting optimization problem is difficult. In this paper Bayes factors are essentially obtained by Monte Carlo approximations and the reader can refer to [7] for an efficient sampling method based on the Hastings-Metropolis algorithm with a relatively low rejection ratio.

Mixing Matrix and Hyper-parameters Priors. The mixing matrix is a priori assumed Gaussian with a zero mean matrix and "flat" covariance matrix of the form:

\[ \pi(A[i,j]) = \mathcal{N}(\mu[i,j], \sigma^2_{ij}) \]  

(4)

In this paper the hyper-parameters consist essentially in the scale parameters \(\lambda_1, \ldots, \lambda_n\), and the noise inverse variances \(\sigma_1^{-2}, \ldots, \sigma_m^{-2}\). These scale parameters are a priori modeled by a standard \(\chi^2\) distribution.

In a Bayesian estimation framework, the posterior density function of the unknowns is given by the Bayes theorem:

\[ p(\theta | x_{1:T}, n) = Z_n^{-1} l(x_{1:T} | \theta) \pi(\theta | n) \]  

(5)

In this equation \(\theta\) represents all the unknowns: \(\theta = \{s_{1:T}, A, \sigma_1^{-2}, \ldots, \sigma_m^{-2}, \lambda_1, \ldots, \lambda_n\}\). \(p(.)\) and \(\pi(.)\) are respectively the posterior and prior probability density functions\(^1\) of the unknowns and \(l(.)\) is the likelihood function of the data. \(Z_n\) is the normalization constant, commonly called evidence\(^2\) or model predictive probability. It is given by:

\[ Z_n = p(x_{1:T} | n) = \int_{\Theta} l(x_{1:T} | \theta) \pi(\theta | n) \, d\theta \]  

(6)

Note that we have explicitly written the number of sources \(n\) in the expression of the posterior (5) and the evidence (6).

Bayesian Model Selection

The evidence or model predictive probability \(Z_n\) of equation (6) is the key point of Bayesian model selection problems in general and Bayesian model selection in BSS in particular. The evaluation of that constant “should be a standard part of rational

---

\(^1\) In this work we do not use “improper” priors for the unknown parameters.
Bayesian model comparison is generally done by comparing the ratios of evidence emanating from different models, these ratios are known as Bayes factors.

Bayes Factors

When estimating unobserved sources in a blind source separation problem from the set of observations \( x_{1:T} \), we have \( n = 0 \) to \( n = n_{\text{max}} \) competing models. As we are doing source separation, we consider that the occurrences \( n = 0 \) (no source) and \( n = 1 \) (reduces to a denoising problem) are absurdity. Due to the difficulty of the problem, we assume that \( n_{\text{max}} = m \) (\( m \) being the number of sensors or observations). The posterior marginal probabilities of each competing model is given by:

\[
p(n|x_{1:T}) = \frac{p(x_{1:T}|n)\pi(n)}{\sum_{n'=2}^{n_{\text{max}}} p(x_{1:T}|n')\pi(n')}, \quad n = 2, \ldots, n_{\text{max}}
\]  

(7)

where \( \pi(n) \) is the prior probability of model \( n \) and \( p(x_{1:T}|n) \) its evidence \( Z_n \). The posterior ratio (odds ratio) of two competing models \( n_1 \) and \( n_2 \) is given by:

\[
\frac{p(n_1|x_{1:T})}{p(n_2|x_{1:T})} = \frac{\pi(n_1)p(x_{1:T}|n_1)}{\pi(n_2)p(x_{1:T}|n_2)} = \frac{\pi(n_1)}{\pi(n_2)}B_{12}
\]

posterior odds \( = \) prior odds \( \times \) Bayes factor

(8)

where \( B_{12} = Z_{n_1}/Z_{n_2} \) is the Bayes factor of model \( n_1 \) to model \( n_2 \). In our work, we consider that the competing models are a priori equi-probable, that is \( \pi(n_1) = \pi(n_2) = 1/(n_{\text{max}}-1) \), therefore comparing posterior marginals ratios reduces to comparing Bayes factors of competing models:

\[
\text{posterior odds} = \text{Bayes factors}, \quad \text{given} \quad \pi(n) = \frac{1}{(n_{\text{max}}-1)}
\]  

(9)

In blind source separation, we are faced to a multiple hypotheses testing problem and we have found more convenient to test binary hypotheses of the form:

\[
\begin{align*}
H_n: \text{the number of sources is } n, \\
\bar{H}_n: \text{the number of sources is different from } n.
\end{align*}
\]

for \( n = 2, \ldots, n_{\text{max}} \). For that purpose we define the following quantities:

\[
\tilde{B}_n = \frac{p(N=n|x_{1:T})}{p(N \neq n|x_{1:T})} = \frac{p(N=n|x_{1:T})}{\sum_{i \neq n} p(N=i|x_{1:T})} = \frac{Z_n}{\sum_{i \neq n} Z_i}
\]

(10)
which can be rewritten:

\[
\tilde{B}_{n}^{-1} = \sum_{i=2}^{n_{\text{max}}} B_{ni}^{-1} - 1 = (n_{\text{max}} - 1) \mathcal{H}(B_{ni}) - 1
\]  

(11)

where \( \mathcal{H}(B_{ni}) \) is the harmonic mean of the Bayes factors \( B_{ni} \) corresponding to the hypothesis \( N = n \) vs. \( N = i \), for \( i = 2, \ldots, n_{\text{max}} \) with \( B_{nn} = 1 \).

In order to determine these factors values, we definitely need the numerical value of the evidence for each model order. However its evaluation is not so trivial, especially in high dimensionality problems like the BSS problem and even an analytic approximation (like the Laplace approximation) is not possible. Numerical methods are generally used to numerically evaluate it and in the following we will study two methods and compare them for the BSS problem.

**Importance Sampling from the posterior**

The evidence for a given model (equation 6) can be rewritten as the expectation of the likelihood function with respect to the prior:

\[
Z_{n} = \int l(x_{1:T}|\theta) \pi(\theta|n) \, d\theta = \mathbb{E}_{\pi(\theta|n)}[l(x_{1:T}|\theta)]
\]  

(12)

where \( \theta \) is the set of all the unknowns. Monte Carlo approximation suggests to approach this expectation by:

\[
Z_{n} \approx \frac{1}{I} \sum_{i=1}^{I} l(x_{1:T}|\theta^{i}), \quad \theta^{i} \sim \pi(\cdot|n)
\]  

(13)

converging at \( O(I^{-1/2}) \). However the drawback of this approximation is that most of samples \( \theta^{i} \) has small likelihood values if major of the posterior is well concentrated relative to the prior. An alternative would be to use important sampling:

\[
Z_{n} = \mathbb{E}_{\pi^{*}(\theta|n)} \left[ \frac{l(x_{1:T}|\theta) \pi(\theta|n)}{\pi^{*}(\theta|n)} \right] 
\]

\[
\approx \frac{1}{I} \sum_{i=1}^{I} w_{i} \frac{l(x_{1:T}|\theta^{i})}{\sum_{i=1}^{I} w_{i}}, \quad \theta^{i} \sim \pi^{*}(\cdot|n)
\]  

(14)

where \( w_{i} = \pi(\theta^{i}|n) / \pi^{*}(\theta^{i}|n) \). For a posterior choice of the importance function: \( \pi^{*}(\theta|n) = p(\theta|x_{1:T}, n) \), the evidence is given by:

\[
Z_{n} \approx \left[ \frac{1}{I} \sum_{i=1}^{I} l^{-1}(x_{1:T}|\theta^{i}) \right]^{-1}, \quad \theta^{i} \sim p(\cdot|x_{1:T}, n)
\]  

(15)

where sampling from a distribution known up to a normalization constant (here the evidence) can be always possible.
The expression of equation (15) is the harmonic mean of the likelihood values which may suffer from occasional occurrences of very small likelihood valued samples and hence result in high variance estimates. However, in many cases, it can lead to very accurate approximations of the evidence and it is generally appreciated because of its simplicity.

**Simulated Annealing sampling**

Simulated annealing uses fractional powers of the likelihood \( l^{\gamma} \) to integrate gradually from the prior (\( \beta = 0 \)) to the posterior (\( \beta = 1 \)):

\[
\log Z_n = \int_0^1 \mathbb{E}_{\pi^*_\beta} [\log l(x_{1:T}|\theta)] \, d\beta
\]

where \( \pi^*_\beta(\theta|x_{1:T},n) \propto l^{\beta}(x_{1:T}|\theta) \pi(\theta|n) \). This last equation is known as the “thermodynamic integration formula”. For likelihood functions in the exponential family (as it is the case here), the fractional power \( l^{\beta} \) still belongs to the exponential family and thus sampling from \( \pi^*_\beta \) from different values of \( \beta \) is similar to sampling from the posterior (\( \beta = 1 \)).

However a “suitable” grid for \( \beta \) should be chosen in order to achieve good numerical approximations. In our work equation (16) is approximated by means of Monte Carlo methods as:

\[
\log Z_n \approx \frac{1}{I_1 I_2} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \log l(x_{1:T}|\theta^j),
\]

where \( \theta^j \sim \pi^*_\beta_i (\cdot|x_{1:T},n) \) and \( \beta_i \sim U[0,1] \).

**SIMULATIONS**

We have considered a simulated case where 4 sparse sources \( s_{1:T} \) have been generated with \( T = 2000 \) samples. Their respective scatter plots are presented on FIGURE 1. The sources have been mixed with a \((8 \times 4)\)-mixing matrix:

\[
A = \begin{bmatrix} I_{(4 \times 4)} \otimes (4 \times 4) \\ O_{(8 \times 4)} \end{bmatrix} + U_{(8 \times 4)}, \quad U[i,j] \sim \mathcal{N}(0,\eta)
\]

so that we obtain 8 observations \( x_{1:T} \). A Gaussian noise level has been considered with an “observation-signal to noise ratio” of about 30dB.

In order to estimate the Bayes factors of equation (10) “importance sampling from the posterior” and “simulated annealing sampling” have been done for each allowed source number \( (n = 2, \ldots, 8) \).

**Importance sampling from the posterior.** Since this numerical approximation can result in high variance estimators, 4 parallel chains have been run in order to evaluate
FIGURE 1. Scatter plots of the original sources: $s_i$ (row $i$) as a function of $s_j$ (column $j - 1$)

log $Z_n$ varying the prior “flatness” at each run (this also aims to study roughly the sensitivity of this numerical approximation to the prior). The corresponding estimates of log $Z_n$ for $n = 2, \ldots, n_{\text{max}} = 8$ are presented on FIGURE 2.-a while on FIGURE 2.-b the corresponding box plots are shown.

FIGURE 2. numerical approximation of log $Z_n$ by “importance sampling from the posterior” of equation (15) for 4 different MCMC chains (left), corresponding Box-Whisker plot (right).

Simulated annealing sampling. The evaluation of the evidence using the numerical approximation of equation (16) has been considered with different choices of $\beta$ grids and slightly varying the prior from a run to the other as in the importance sampling approximation. The obtained numerical values of the log evidence are represented on FIGURE 3.-a with the corresponding box plots shown on FIGURE 3.-b.

On TABLE 1. the numerical results of the factors defined in equation (10) and (10) are given for each model order (source number $n$) and for the two approximations (Importance sampling from the posterior and Simulated annealing sampling).
FIGURE 3. Numerical approximation of $\log Z_n$ by “simulated annealing sampling” of equation (16) for 4 different MCMC chains (left), corresponding Box-Whisker plot (right).

TABLE 1. Hypothesis testing table presenting the factors of equation (10) for each source number $n$ summarizing the hypothesis: “$N = n$” against “$N \neq n$”.

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log B_n$ (Important Sampling)</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
<td>208.3264</td>
<td>$-208.3264$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td></td>
<td>$-\infty$</td>
<td>$-\infty$</td>
<td>96.1423</td>
<td>$-96.1423$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
</tbody>
</table>

DISCUSSION

Though the final decision for the two considered approximations of the evidence (Important sampling and Simulated annealing) led to the same “answer”: the number of sources ($n$) inherent in the mixture $x_{1:T}$ is determined with no ambiguity in this particular problem of separation of sparse sources, the returned approximations of the evidence are very different from a method to the other. We did not succeed to explain exactly such a difference between the two estimates of the evidence: we would say that sampling from the posterior in the important sampling approach would miss most of the prior if there is big cross entropy between the posterior and the prior (when the likelihood is peaky) and thus the important sampling approximation is a bad approximation, however if the posterior is peaky, most of the posterior mass is concentrated therein and this should not result in such big errors? The annealing sampling approach is however more accurate and we could rely on such a method. On FIGURE 4, we plotted the expected value of the log likelihood $<\log L>$ with respect to the annealed posterior $L^\beta \pi$ as a function of the annealing parameter $\beta$ where we observe that the problem is a convex problem increasing smoothly from $\beta > 0$ to $\beta = 1$ with a singularity at $\beta = 0$.

Prior sensitivity. We noticed that the annealing approximation is sensitive to the choice of the prior as shown on FIGURE 3.-b: large data-extent bars. While the importance sampling approximation is less sensitive to the choice of the prior as shown on FIGURE 2.: smaller data-extent bars.
FIGURE 4. Evolution of the expected log Likelihood “$\langle \log l(x_{1:T}|\theta,n) \rangle$” for $n = 4$ with respect to the fractional posterior “$L^\beta \pi$” as a function of the fractional parameter $\beta$ (see equation 16).

ACKNOWLEDGMENTS

We would like to thank John Skilling for his very helpful suggestions, guidelines and comments. We also would like to thank the Edwin T. Jaynes foundation for its support.

REFERENCES