Bayesian approach for blind separation of sparse sources

Cédric Févotte and Simon J. Godsill

cf269@cam.ac.uk

HASSIP Workshop, Cambridge
15th Sept 2004
Blind Source Separation (BSS) consists in estimating $n$ unknown signals (the sources) from the sole observation of $m$ mixtures of them (the observations).
Applications of BSS

Audio: «demix»/«remix» of musical recordings, speech enhancement in noisy environment (mobile phones, hearing aids), etc.

Biomedical: decomposition of ElectroEncephalograms (EEG), ElectroCardioGrams (ECG), ElectroMyoGrams (EMG), etc.

Applications in astronomy, telecommunications, etc.
Several problems

Number $m$ of observations w.r.t number of sources $n$?

- if $m \geq n \rightarrow$ (over)determined problem
  $\Rightarrow$ «Good» conditions.

- if $m < n \rightarrow$ underdetermined problem
  $\Rightarrow$ ill-posed problem ($prior$ information on the sources is required).
Several problems (ctd)

Nature of mixing?

- **Linear Time Invariant mixings:**
  - **Instantaneous** mixings (i.e., "memory less" mixings),
  - **Convolutive** mixings (e.g., reverberation environment).

- **Time-Varying** mixings (e.g., speakers moving in a room).

- **Non-linear** mixings (e.g., distortion due to sensors).
Several problems (ctd)

Nature of the sources?

- Nature of each source:
  - Stationarity/non-stationarity?
  - Sparsity?
  - Positivity?

- Joint information on the sources:
  decorrelation/independence?
Separation of instantaneous mixtures of sources having a sparse representation in an orthogonal basis

Bayesian approach
Outline

(I) Sparsity for source separation
   (a) What is sparsity?
   (b) A simple algorithm

(II) Bayesian approach
   (a) MMSE estimate
   (b) Gibbs sampler
   (c) Conditional distributions

(III) Results
   (a) Underdetermined mixing
   (b) Overdetermined (noisy) mixing
Sparsity for source separation
What is sparsity?

A signal is said to be **sparse on a dictionary** if only “a few” coefficients of its decomposition are “significantly” different from zero.

**E.g:**

- A sum of sines is sparse on the Fourier basis,
- Audio signals are “sparser” on MDCT basis (local DCT).
Ex: MDCT of audio signal

Piano waveform

MDCT (normalised) W=64ms

Histogram of the time samples

Histogram of the coefficients
\[ \forall t \in [0, \ldots, N - 1]: \]
\[ x_t = A s_t + n_t \]

- \( x_t = [x_{1,t}, \ldots, x_{m,t}]^T \): observations vector,
- \( s_t = [s_{1,t}, \ldots, s_{n,t}]^T \): zero-mean sources vector,
- \( A \): full column rank mixing matrix,
- \( n_t \): noise vector.

**Notation:** \( x = [x_0, \ldots, x_{N-1}] \)
Aim

Estimating the sources \( s \) and the mixing matrix \( A \) up to BSS standard indeterminacies on gain and order:

\[
\hat{s} = (P D) s \\
\hat{A} = A (D^{-1} P^T)
\]

where \( D \) is a diagonal matrix and \( P \) is a permutation matrix.
We can equivalently solve our problem in any other (orthogonal) basis.

Let $\Phi$ denote a $N \times N$ matrix such that $\Phi^T \Phi = I_N$.

**Notation:** $\tilde{x} = x \Phi$

\[
\begin{align*}
x &= A s + n \\
\uparrow \\
\tilde{x} &= A \tilde{s} + \tilde{n}
\end{align*}
\]
Let us consider the following simple problem:

\[
\begin{bmatrix}
  x_{1,t} \\
  x_{2,t}
\end{bmatrix} = \begin{bmatrix}
  1 & 1 \\
  1 & -1
\end{bmatrix} \begin{bmatrix}
  s_{1,t} \\
  s_{2,t}
\end{bmatrix}
\]

\[
= s_{1,t} \begin{bmatrix}
  1 \\
  1
\end{bmatrix} + s_{2,t} \begin{bmatrix}
  1 \\
  -1
\end{bmatrix}
\]

where \( s_1 \) and \( s_2 \) are audio signals (piano & guitar).
Sparsity for source separation (ctd)

Clusters directions are colinear to mixing matrix columns
A simple source separation algorithm:

- Compute transformation (e.g., MDCT) of the observations,
- Estimate the mixing matrix columns via clustering algorithm,
- Compute estimates of the sources via

\[
\hat{s} = \hat{A}^{-1} \tilde{x} \\
\hat{s} = \hat{s} \Phi^T
\]
Sparsity for source separation (ctd)

Limitations of the method:

- Sources cannot be estimated for underdetermined mixtures (A is not invertible),
- The method does not take into account the noise (no denoising of the estimates).

A solution to overcome these limitations: use a model of the sources (coefficients).

⇒ we model sources coefficients distributions by Student $t$ distributions and derive a Bayesian approach.
Bayesian approach
Assumptions

1)  ∀i, \( \tilde{s}_i \) is i.i.d with Student \( t \) distribution \( t(\alpha_i, \lambda_i) \):

\[
p(\tilde{s}_{i,k} | \alpha_i, \lambda_i) \propto \left( 1 + \frac{1}{\alpha_i} \left( \frac{\tilde{s}_{i,k}}{\lambda_i} \right)^2 \right)^{-\frac{\alpha_i+1}{2}}
\]

\( \alpha_i \) = degrees of freedom, \( \lambda_i \) = scale parameter.

- for \( \lambda = 1 \) and \( \alpha = 1 \), Student \( t \) = standard Cauchy distribution,

- for \( \lambda = 1 \) and \( \alpha \rightarrow +\infty \), Student \( t \) tends to standard Gaussian.
Assumptions (ctd)

t densities for $\alpha \in \{0.01, 0.1, 1, 10\}$ with equal mode
Student $t$ can be expressed as a Scaled Mixture of Gaussians:

$$p(\tilde{s}_{i,k} | \alpha_i, \lambda_i) = \int_0^{+\infty} \mathcal{N}(\tilde{s}_{i,k} | 0, v_{i,k}) \mathcal{IG} \left( v_{i,k} | \frac{\alpha_i}{2}, \frac{2}{\alpha_i \lambda_i^2} \right) dv_{i,k}$$

with

- $\mathcal{N}(x|0, v)$: normal distribution with mean 0 and variance $v$,
- $\mathcal{IG}(x|\gamma, \beta)$: Inverted Gamma distribution

$$\mathcal{IG}(x|\gamma, \beta) = \frac{x^{-(\gamma+1)} \exp\left(-\frac{1}{\beta x}\right)}{\Gamma(\gamma) \beta^\gamma}, \text{ for } x \geq 0$$
Assumptions (ctd)

Introducing the auxilary random variable $v_{i,k}$

$\sim p(\tilde{s}_{i,k} | \alpha_i, \lambda_i)$ is a marginal density of the joint distribution

$p(\tilde{s}_{i,k}, v_{i,k} | \alpha_i, \lambda_i)$, defined by:

$$p(\tilde{s}_{i,k}, v_{i,k} | \alpha_i, \lambda_i) = p(\tilde{s}_{i,k} | v_{i,k}) p(v_{i,k} | \alpha_i, \lambda_i)$$

with:

1. $p(\tilde{s}_{i,k} | v_{i,k}) = \mathcal{N}(\tilde{s}_{i,k} | 0, v_{i,k})$,

2. $p(v_{i,k} | \alpha_i, \lambda_i) = \mathcal{IG} \left( v_{i,k} | \frac{\alpha_i}{2}, \frac{2}{\alpha_i \lambda_i^2} \right)$
2) \( \{\tilde{s}_1, \ldots, \tilde{s}_n\} \) are mutually independent:

\[
p(\tilde{s}) = \prod_{i=1}^{n} p(\tilde{s}_i)
\]

3) \( \tilde{n} \) (or equivalently \( n \)) is a i.i.d Gaussian zero-mean noise with

\[
E\{\tilde{n}_k \tilde{n}_k^T\} = \sigma^2 I_m
\]
Aim

Compute **Minimum Mean Square Estimate** of \( \theta = \{A, \tilde{s}, \sigma, v, \alpha, \lambda\} \):

\[
\hat{\theta} = \int \theta p(\theta | \tilde{x}) \, d\theta
\]

Integral intractable

\( \leadsto \) Resort to MCMC method to sample \( \{\theta^{(1)}, \ldots, \theta^{(K)}\} \) from \( p(\theta | \tilde{x}) \) and approximate \( \hat{\theta} \) by

\[
\hat{\theta} \approx \frac{1}{K} \sum_{l=1}^{K} \theta^{(l)}
\]
Gibbs sampler

Only requires to sample from posterior distribution of each parameter conditionally upon the data $\tilde{x}$ and the other parameters:

Initialize $\theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_M^{(0)}\}$

for $k = 1 : K + K_{Burnin}$ do

$\theta_1^{(k)} \sim p(\theta_1|\theta_2^{(k-1)}, \ldots, \theta_M^{(k-1)}, \tilde{x})$

$\theta_2^{(k)} \sim p(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_M^{(k-1)}, \tilde{x})$

$\theta_3^{(k)} \sim p(\theta_3|\theta_1^{(k)}, \theta_2^{(k)}, \theta_4^{(k-1)}, \ldots, \theta_M^{(k-1)}, \tilde{x})$

$\vdots$

$\theta_M^{(k)} \sim p(\theta_M|\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{M-1}^{(k)}, \tilde{x})$

end for
Conditional densities

Notation: $\theta_{-y}$ is the set of parameters in $\theta$ except $y$.

$$p(\theta_k | \theta_{-k}, \tilde{x}) = \frac{p(\theta | \tilde{x})}{p(\theta_{-k} | \tilde{x})}$$

$$= \frac{p(\tilde{x} | \theta) p(\theta)}{p(\theta_{-k} | \tilde{x}) p(\tilde{x})}$$

$$\propto p(\tilde{x} | \theta) p(\theta)$$

$\leadsto$ The required conditional distributions are proportional to the likelihood of the data times the priors and the parameters.
Assuming independence of $\mathbf{A}$, $\sigma$ and sources parameters:

\[
p(\theta) = p(\tilde{s}, \mathbf{A}, \sigma, \mathbf{v}, \alpha, \lambda) \\
= p(\tilde{s}, \mathbf{v}, \alpha, \lambda) p(\mathbf{A}) p(\sigma) \\
= p(\tilde{s} | \mathbf{v}) p(\mathbf{v} | \alpha, \lambda) p(\alpha) p(\lambda) p(\mathbf{A}) p(\sigma)
\]
With Gaussian noise and i.i.d assumptions:

\[
p(\tilde{x}|\theta) = p(\tilde{x}|A, \tilde{s}, \sigma) \\
= \prod_{k=0}^{N-1} \mathcal{N}(\tilde{x}_k|A\tilde{s}_k, \sigma^2 I_m) \\
= \frac{1}{(2\pi \sigma^2)^{N\frac{m}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=0}^{N-1} \|\tilde{x}_k - A\tilde{s}_k\|_F^2\right)
\]
\[
p(\tilde{s}|\theta_{\tilde{s}}, \tilde{x}) \propto p(\tilde{x}|A, \tilde{s}, \sigma) p(\tilde{s}|v)
\]

where

\[
\Sigma_{\tilde{s}_k} = \left( \frac{1}{\sigma^2} A^T A + \text{diag} \left( v_k \right)^{-1} \right)^{-1}
\]

\[
\mu_{\tilde{s}_k} = \frac{1}{\sigma^2} \sum_{\tilde{s}_k} A^T \tilde{x}_k
\]
Sampling

\[ p(A|\theta_{-A}, \tilde{x}) \propto p(\tilde{x}|A, \tilde{s}, \sigma) p(A) \]

With \( [r_1 \ldots r_m] = A^T \),

\[
\hat{S}_k = \begin{bmatrix}
\hat{s}_k^T \\
\vdots \\
0
\end{bmatrix}
\quad \text{and} \quad
a = \begin{bmatrix}
r_1 \\
\vdots \\
r_m
\end{bmatrix}
\]

\[
\sim A \hat{s}_k = \hat{S}_k a
\]
Sampling \( A \) (ctd)

With uniform prior \( p(a) \propto 1 \),

\[
p(a|\theta_{-a}, \tilde{x}) = \mathcal{N}(a|\mu_a, \Sigma_a)
\]

\[
\Sigma_a = \sigma^2 \left( \sum_{k=0}^{N-1} \tilde{S}_k^T \tilde{S}_k \right)^{-1}
\]

\[
\mu_a = \frac{1}{\sigma^2} \sum_{k=0}^{N-1} \tilde{S}_k^T \tilde{x}_k
\]

To fix BSS indeterminacies on gain, we set in practice the first row of \( A \) to ones and only estimate the other rows.
\[ p(\sigma|\theta_{-\sigma}, \tilde{x}) \propto p(\tilde{x}|A, \tilde{s}, \sigma) p(\sigma) \]

Using Jeffreys prior \( p(\sigma) = 1/\sigma, \)

\[ p(\sigma|\theta_{-\sigma}, \tilde{x}) \propto \sigma^{-(2\gamma_\sigma+1)} \exp\left(-\frac{1}{\beta_\sigma \sigma^2}\right) \]

with \( \gamma_\sigma = \frac{mN}{2} \) and \( \beta_\sigma = 2/ \sum_{k=0}^{N-1} \|\tilde{x}_k - A \tilde{s}_k\|_F^2. \)

\[ (\sigma^2|\theta_{-\sigma}, \tilde{x}) \sim IG(\gamma_\sigma, \beta_\sigma) \]
\[
p(v | \theta_{-v}, \hat{x}) \propto p(\hat{s} | v) p(v | \alpha, \lambda)
\]

\[
p(v | \theta_{-v}, \hat{x}) = \prod_{k=0}^{N-1} \prod_{i=1}^{n} IG \left( v_{i,k} | \gamma_{v_i}, \beta_{v_i,k} \right)
\]

with

\[
\gamma_{v_i} = (\alpha_i + 1)/2
\]
\[
\beta_{v_i,k} = 2/(\tilde{s}_{i,k}^2 + \alpha_i \lambda_i^2)
\]
\[ p(\alpha|\theta_{-\alpha}, \tilde{x}) \propto p(v|\alpha, \lambda) p(\alpha) \]

With \( p(\alpha) \propto 1 \),

\[
p(\alpha|\theta_{-\alpha}, \tilde{x}) \propto \prod_{i=1}^{n} \frac{P_i^{-\left(\frac{\alpha_i}{2} + 1\right)}}{\Gamma\left(\frac{\alpha_i}{2}\right)^N} \left(\frac{\alpha_i \lambda_i^2}{2}\right)^{\frac{\alpha_i \cdot N}{2}} \exp\left(-\frac{\alpha_i \lambda_i^2}{2} S_i\right)
\]

with \( S_i = \sum_{k=0}^{N-1} \frac{1}{v_{i,k}} \) and \( P_i = \prod_{k=0}^{N-1} v_{i,k}. \)

\[\sim\] in practice we sample \( \alpha \) from a uniform grid of discrete values with probability mass given above.
Sampling $\lambda$

$$p(\lambda|\theta_{-\lambda}, \tilde{x}) \propto p(v|\alpha, \lambda) p(\lambda)$$

With Jeffreys prior $p(\lambda_i) = 1/\lambda_i$,

$$p(\lambda|\theta_{-\lambda}, \tilde{x}) \propto \prod_{i=1}^{n} \lambda_i^{\alpha_i N} \exp \left( -\frac{\alpha_i S_i}{2} \lambda_i^2 \right) \frac{1}{\lambda_i}$$

$$\begin{align*}
(\lambda_i^2|\theta_{-\lambda}, \tilde{x}) &\sim \mathcal{G}(\gamma_{\lambda_i}, \beta_{\lambda_i}) \\
\gamma_{\lambda_i} &= (\alpha_i N)/2 \quad \text{and} \quad \beta_{\lambda_i} = 2/(\alpha_i S_i) \quad \text{and where} \quad \mathcal{G}(\gamma, \beta) \quad \text{is} \\
\text{the Gamma distribution} \\
(\mathcal{G}(x|\gamma, \beta) &= x^{\gamma-1}/(\Gamma(\gamma) \beta^\gamma) \exp(-x/\beta), \quad \text{for} \quad x \geq 0).}
\end{align*}$$
Results
Our method is compared with the one in:

Coefficients distributions modeled by a mixture of 2 Gaussians:

- One with small variance and high probability: state ’off’
- One with high variance and small probability: state ’on’

⇒ EM based estimation of the parameters
We use the performance evaluation criteria designed in E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in Blind Audio Source Separation”, IEEE Trans. Speech and Audio Processing

Given and estimate $\hat{s}$ of $s$, we measure:

- Signal to Distortions Ratio (SDR) (global criterion)
- Signal to Interferences Ratio (SIR)
- Signal to Noise Ratio (SNR)
- Signal to Artifacts Ratio (SAR)
Underdetermined mixture

- 3 sources (8s at 8000Hz \(\sim N = 65536\))
  
  **Source 1** **Source 2** **Source 3**

- 2 mixtures

\[
A = \begin{bmatrix}
1 & 1 & 1 \\
0.8 & 1.3 & -0.9 \\
\end{bmatrix}
\]

- 30 dB noise on each observation \((\sigma = 0.01)\)

**Mix 1** **Mix 2**

- MDCT window length = 64ms
Underdetermined mixture (ctd)
Underdetermined mixture (ctd)

<table>
<thead>
<tr>
<th></th>
<th>$\hat{s}_1$</th>
<th>$\hat{s}_2$</th>
<th>$\hat{s}_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
</tr>
<tr>
<td>$t + \text{MCMC}$</td>
<td>3.3</td>
<td>12.2</td>
<td>4.2</td>
</tr>
<tr>
<td>FMoG + EM</td>
<td>2.3</td>
<td>16.3</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Original matrix

$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0.8 & 1.3 & -0.9 \end{bmatrix}$

Student $t$ prior + MCMC

$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7915 & 1.3016 & -0.9010 \end{bmatrix}$

(FMoG prior + EM)

$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7643 & 1.3250 & -0.8995 \end{bmatrix}$
Overdetermined mixture

- 3 sources (8s at 8000Hz $\rightarrow N = 65536$)
- 2 mixtures
- $9.5$ dB noise on each observation ($\sigma = 0.1$)
- MDCT window length = 64ms

\[
A = \begin{bmatrix}
1 & 1 & 1 \\
0.8 & 1.3 & -0.9 \\
1.2 & -0.7 & 1.1
\end{bmatrix}
\]
Overdetermined mixture (ctd)
Overdetermined mixture (ctd)

<table>
<thead>
<tr>
<th></th>
<th>$\hat{s}_1$</th>
<th></th>
<th>$\hat{s}_2$</th>
<th></th>
<th>$\hat{s}_3$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR</td>
<td>SIR</td>
<td>SAR</td>
<td>SNR</td>
<td>SDR</td>
<td>SIR</td>
</tr>
<tr>
<td>$t + \text{MCMC}$</td>
<td>12.3</td>
<td>35.8</td>
<td>13.8</td>
<td>17.7</td>
<td>15.4</td>
<td>37.1</td>
</tr>
<tr>
<td>FMoG + EM</td>
<td>5.6</td>
<td>36.7</td>
<td>5.6</td>
<td>33.5</td>
<td>6.5</td>
<td>48.4</td>
</tr>
</tbody>
</table>

Original matrix

\[
\mathbf{A} = \begin{bmatrix}
1 & 1 & 1 \\
0.8 & 1.3 & -0.9 \\
1.2 & -0.7 & 1.1
\end{bmatrix}
\]

Student $t$ prior + MCMC

\[
\hat{\mathbf{A}} = \begin{bmatrix}
1 & 1 & 1 \\
0.7933 & 1.3041 & -0.8977 \\
\pm(0.0045) & (\pm0.0051) & (\pm0.0049) \\
1.1939 & -0.7037 & 1.1001 \\
\pm(0.0056) & \pm(0.0048) & \pm(0.0044)
\end{bmatrix}
\]

FMoG prior + EM

\[
\hat{\mathbf{A}} = \begin{bmatrix}
1 & 1 & 1 \\
0.7790 & 1.3181 & -0.8872 \\
1.1864 & -0.7182 & 1.1069
\end{bmatrix}
\]
Conclusions

- $t$ model gives better SAR $\Rightarrow$ sounds more natural
- FMoG model gives better SNRs and better SIRs on underdetermined mixtures
- Strength of MCMC approach: precise estimation of $A$ whatever initialisation (contrary to EM approach)
- Main drawback: heavy computational burden - several hours against a few minutes for EM approach. Can be alleviated using a rough estimation of $A$ for initialisation.
Use of indicator variable $\gamma_{i,k} \in \{0, 1\}$:

$$p(\tilde{s}_{i,k} | \alpha_i, \lambda_i) = (1 - \gamma_{i,k}) \delta_0(\tilde{s}_{i,k}) + \gamma_{i,k} t(\tilde{s}_{i,k} | \alpha_i, \lambda_i)$$

Extension to overcomplete dictionaries $\sim$ improve sparsity.

Adapt the same scheme to other source priors expressed as Scaled Mixtures of Gaussians.