Source Separation with Mixture Densities

Ali MOHAMMAD-DJAFARI and Hichem SNOUSSI

Laboratoire des Signaux et Systèmes

CNRS-ESE-UPS

Supélec, Plateau de Moulon

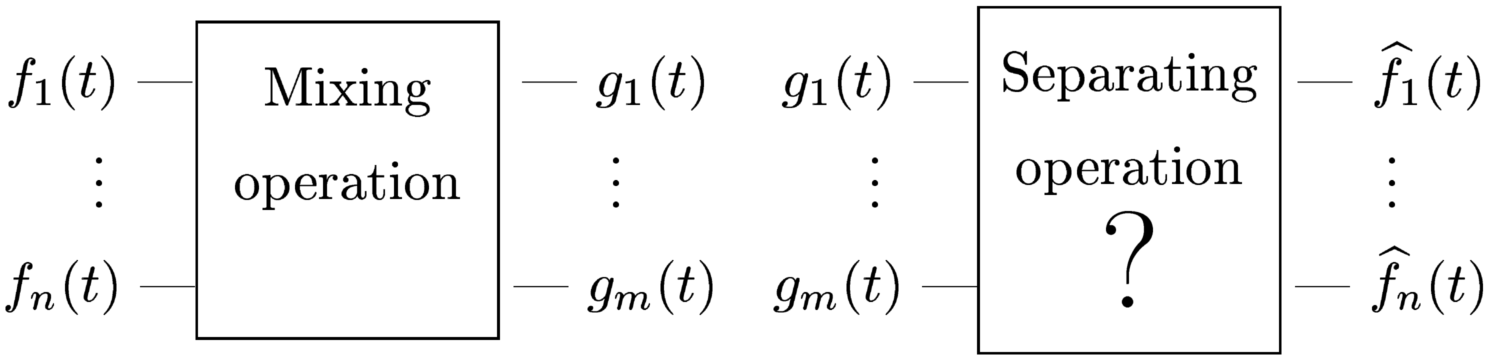
91192 Gif-sur-Yvette, FRANCE.

djafari@lss.supelec.fr

http://www.supelec.fr/lss/perso/adjafari/index.htm

1 Introduction

Source separation problem:



- Linear mixing:

$$\boldsymbol{g}(t) = \int \boldsymbol{A}(t, t')\, \boldsymbol{f}(t')\, \mathrm{d}t'$$

- Convolutive mixing:

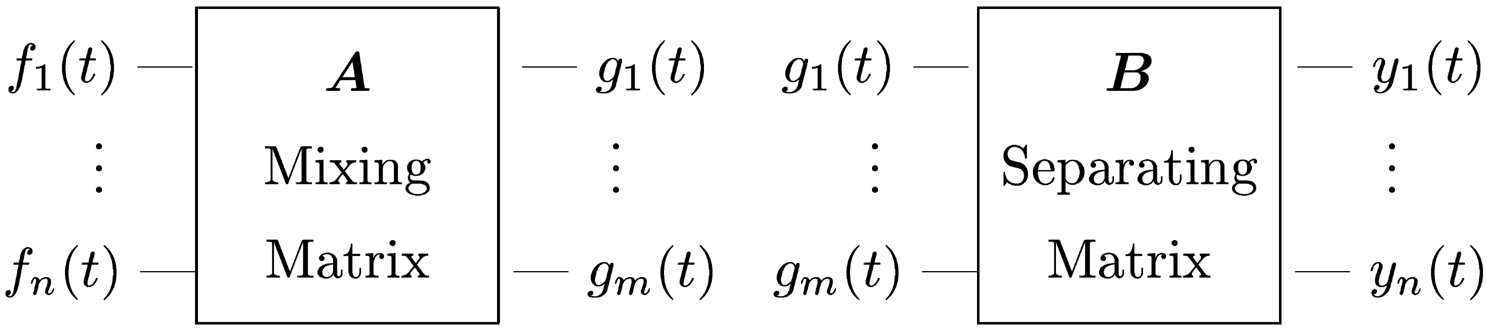
$$\boldsymbol{g}(t) = \int \boldsymbol{A}(t - t')\, \boldsymbol{f}(t')\, \mathrm{d}t'$$

- Instantaneous mixing:

$$\boldsymbol{g}(t) = \boldsymbol{A}\boldsymbol{f}(t)$$

Convolutive mixing $\longrightarrow$ Blind multichannel deconvolution

Linear Instantaneous mixing $\longrightarrow$ Source separation:



- Fondamental underdetermination of the problem: $\longrightarrow \boldsymbol{B} = \boldsymbol{\Sigma}\,\boldsymbol{\Lambda}\,\boldsymbol{A}^{-1}$ where $\boldsymbol{\Sigma}$ is a permutation matrix and $\boldsymbol{\Lambda}$ a diagonal scaling matrix.

- Main assumption : $f_1(t), \ldots, f_n(t)$ are uncorrelated (PCA) or independent (ICA).

- Main classical approaches: Infomax, Contrast function minimization, Higher order statistics (HOS), M-estimation, Maximum likelihood (ML)

Principal Component Analysis (PCA)

- **Main hypothesis:** $f_1(t), \ldots, f_n(t)$ are white and spatially uncorrelated.

$$\boldsymbol{R}_{ff} = \mathrm{E}\left\{\boldsymbol{f}\boldsymbol{f}^t\right\} = \boldsymbol{\Lambda}$$

$$\boldsymbol{R}_{gg} = \mathrm{E}\left\{\boldsymbol{g}\boldsymbol{g}^t\right\} = \mathrm{E}\left\{\boldsymbol{A}\boldsymbol{f}\boldsymbol{f}^t\boldsymbol{A}^t\right\} = \boldsymbol{A}\boldsymbol{R}_{ff}\boldsymbol{A}^t = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^t$$

- **Algorithm:**
  - Estimate

$$[\boldsymbol{R}_{gg}]_{kl} = \sum_t g_k(t)\, g_l(t)$$

  - Singular Value Decomposition (SVD):

$$\boldsymbol{R}_{gg} = \boldsymbol{A}\Lambda\boldsymbol{A}^t \longrightarrow \widehat{\boldsymbol{f}}(t) = (\Lambda^+)^{1/2}\boldsymbol{A}^t\boldsymbol{g}$$

- $\boldsymbol{A}$ can be determined to a rotation matrix factor:

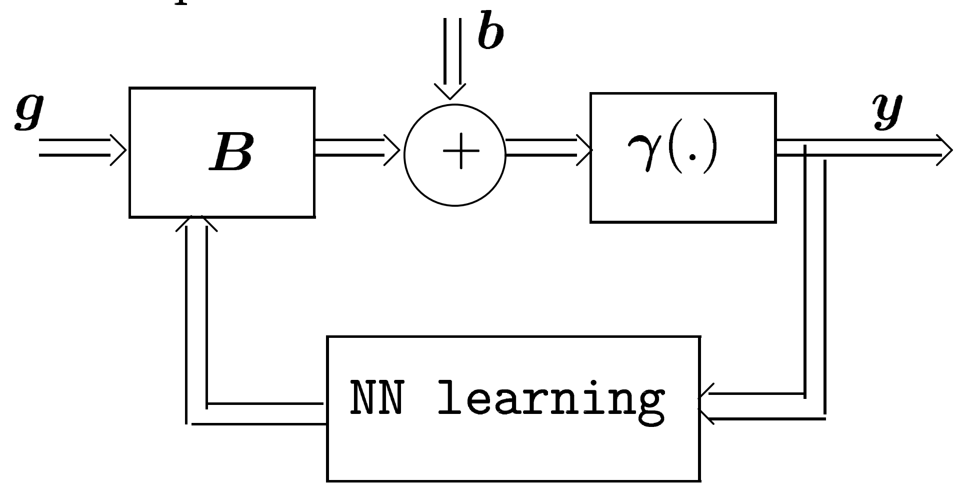
$$\boldsymbol{A} \longrightarrow \boldsymbol{A}\boldsymbol{\Theta} \quad \longrightarrow \quad \boldsymbol{A}\boldsymbol{\Theta}\Lambda\boldsymbol{\Theta}^t\boldsymbol{A}^t = \boldsymbol{A}\Lambda\boldsymbol{A}^t \quad \text{with} \quad \boldsymbol{\Theta} \quad \text{any orthogonal matrix}$$

Independent Components Analysis (ICA)

- Impose a structure for estimation: $\widehat{f}_i = \gamma_i([\boldsymbol{B}\boldsymbol{g} + \boldsymbol{b}]_i) \longrightarrow \widehat{\boldsymbol{f}} = \boldsymbol{\gamma}(\boldsymbol{B}\boldsymbol{g} + \boldsymbol{b})$
- Use the entropy of $\widehat{\boldsymbol{f}}$ as a measure of independence:

$$S = -\sum_i p_i(\widehat{f}_i) \ln p_i(\widehat{f}_i) = -\sum_i p_i(\gamma_i([\boldsymbol{B}\boldsymbol{g}]_i + b_i)) \ln p_i(\gamma_i([\boldsymbol{B}\boldsymbol{g}]_i + b_i))$$

- Optimize S: $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{b}}) = \arg\max_{(\boldsymbol{B},\boldsymbol{b})} \{S(\boldsymbol{B}, \boldsymbol{b})\}$
- Neural Network optimization techniques.



Contrast function minimization

Define a contrast function $c(\boldsymbol{y}) = c\,(\boldsymbol{B}\boldsymbol{g})$ which takes its extremal value when $\boldsymbol{B}$ is a separating matrix . Example:

$$c(\boldsymbol{B}) = KL\left(p(\boldsymbol{y}) : \prod_i p_i(y_i)\right) = \int p(\boldsymbol{y}) \ln \frac{p(\boldsymbol{y})}{\prod_i p_i(y_i)}\, \mathrm{d}\boldsymbol{y}$$

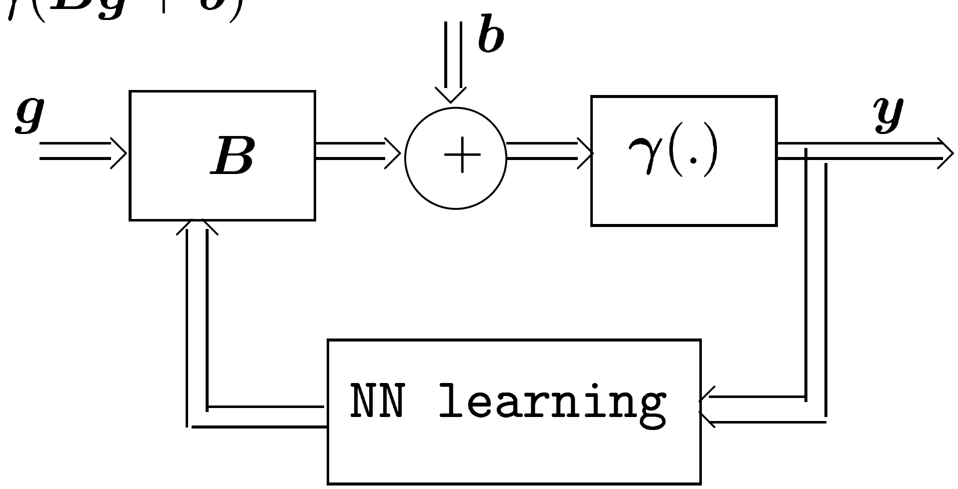
Higher order statistics (HOS)

Approximation of $p_i(y_i)$ by its cumulants = Taylor serie developpement of the caracteristique function.

$$\frac{\partial c(\boldsymbol{B})}{\partial \boldsymbol{B}} \propto \mathrm{E}\left\{\frac{p_i'(y_i)}{p_i(y_i)}\right\} \propto \text{cumulants of } (y_i)$$

Main limitations of classical techniques

- None of these techniques consider the possible errors on the model or the measurement (sensor) noises;
- All these methods assume that the mixing matrix $\boldsymbol{A}$ is invertible.
- All these methods assume that the sources are independent and temporally white. $\longrightarrow$ Whitenning before ICA
- All these techniques fixe the structure of seperating operation as:
  $\widehat{\boldsymbol{f}} = \boldsymbol{B}\gamma(\boldsymbol{g}+\boldsymbol{b})$ or $\widehat{\boldsymbol{f}} = \gamma(\boldsymbol{B}\boldsymbol{g}+\boldsymbol{b})$



2 Bayesian approach

- Main idea: use not only the likelihood $p\left(\boldsymbol{g}_{1..T}|\boldsymbol{A},\boldsymbol{f}_{1..T}\right)$ but also some prior knowledge about the sources $\boldsymbol{f}$ and the mixing matrix $\boldsymbol{A}$ through the assignment of prior probabilities $p(\boldsymbol{f})$ and $p(\boldsymbol{A})$.

$$\ln p\left(\boldsymbol{A},\boldsymbol{f}_{1..T}|\boldsymbol{g}_{1..T}\right) = \ln p\left(\boldsymbol{g}_{1..T}|\boldsymbol{A},\boldsymbol{f}_{1..T}\right) + \ln p(\boldsymbol{A}) + \ln p(\boldsymbol{f}) + cte,$$

Examples of $p(\boldsymbol{A})$:

$$p(\boldsymbol{A}) \propto |\det(\boldsymbol{A})|^{-1}, \qquad p(\boldsymbol{A}) \propto \exp\left[-\lambda\|\boldsymbol{A}\|^2\right],$$

$$p(\boldsymbol{A}) \propto \exp\left[-\frac{1}{2\sigma_a^2}\|\boldsymbol{I}-\boldsymbol{A}\|^2\right],$$

$$p(\boldsymbol{A}) \propto \exp\left[-\frac{1}{2\sigma_a^2}\|\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^t\|^2\right], \quad \text{or} \quad \exp\left[-\frac{1}{2\sigma_a^2}\|\boldsymbol{I}-\boldsymbol{A}^t\boldsymbol{A}\|^2\right].$$

Exact invertible model and independent white sources

$$\boldsymbol{g}(t) = \boldsymbol{A}\boldsymbol{f}(t) = \boldsymbol{B}^{-1}\boldsymbol{f}(t), \quad t = 1, \ldots, T$$

$$p(\boldsymbol{f}) = \prod_i p_i(f_i) \longrightarrow p(\boldsymbol{g}|\boldsymbol{B}) = |\det(\boldsymbol{B})| \prod_i p_i\left([\boldsymbol{B}\boldsymbol{g}]_i\right)$$

$$\ln p\left(\boldsymbol{g}_{1..T}|\boldsymbol{B}\right) = \ln|\det(\boldsymbol{B})|^T + \sum_t \sum_i p_i\left(y_i(t)\right), \quad \text{with} \quad \boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{g}(t)$$

$$J(\boldsymbol{B}) = -\ln p\left(\boldsymbol{B}|\boldsymbol{g}_{1..T}\right) = -T\ln|\det(\boldsymbol{B})| - \sum_t \sum_i \ln p_i\left(y_i(t)\right) + \ln p(\boldsymbol{B}) + cte.$$

MAP estimate:

$$\frac{\partial J(\boldsymbol{B})}{\partial \boldsymbol{B}} = -\sum_t \boldsymbol{H}\left(\boldsymbol{y}(t)\right) \quad \text{with} \quad H(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{B}}\left[\sum_i \ln p_i\left(y_i\right) + \ln|\det(\boldsymbol{B})| + \ln p(\boldsymbol{B})\right]$$

Particular case: $p(\boldsymbol{B})$ uniform $\longrightarrow$ Maximum Likelihood

$$\boldsymbol{H}\,(\boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{y})\,\boldsymbol{y}^\dagger - \boldsymbol{I},$$

with

$$\phi_i(y_i) = -\frac{p_i'(y_i)}{p_i(y_i)}.$$

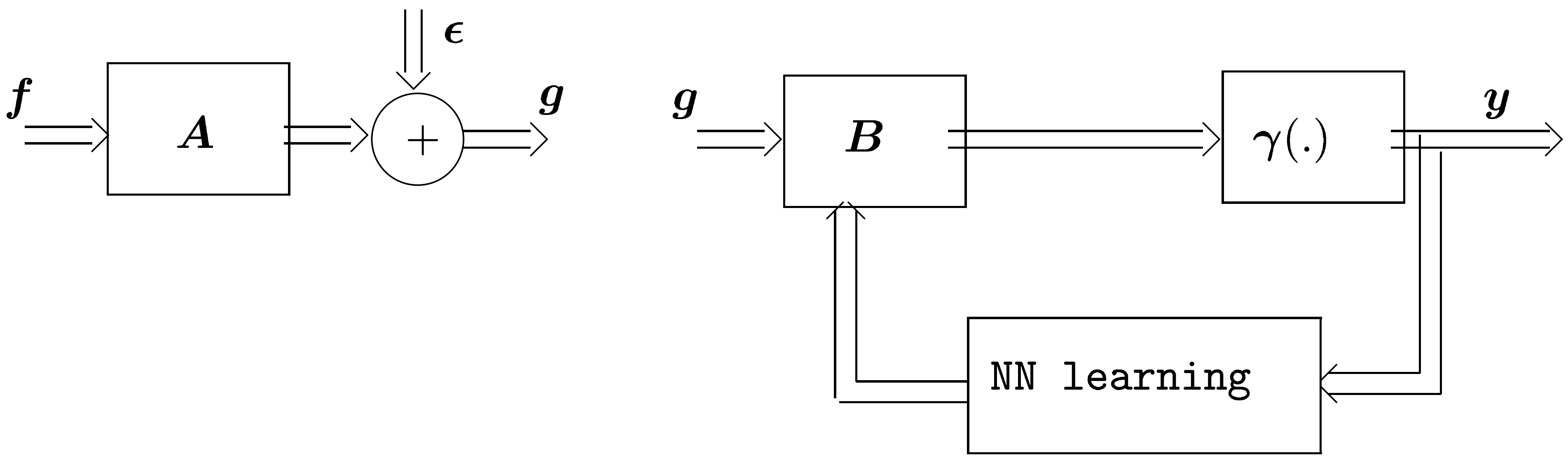
Gauss	$p(z) \propto \exp\left[-\alpha z^2\right]$	$\phi(z) = 2\alpha z$
Laplace	$p(z) \propto \exp\left[-\alpha\|z\|\right]$	$\phi(z) = \alpha \text{sign}(z)$
Cauchy	$p(z) \propto \dfrac{1}{1+(z/\alpha)^2}$	$\phi(z) = \frac{2z/\alpha^2}{1+(z/\alpha)^2}$
sub-Gaussian law	$p(z) \propto \exp\left[-\dfrac{1}{2}z^2\right]\text{sech}^2(z)$	$\phi(z) = z + \tanh(z)$
Mixture of Gaussians	$p(z) \propto \exp\left[-\frac{1}{2}(z-\alpha)^2\right] + \exp\left[-\frac{1}{2}(z+\alpha)^2\right]$	$\phi(z) = \alpha z - \alpha\tanh(\alpha z)$

Link with Neural Network

$$\boldsymbol{H}(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{B}} \left[\sum_i \ln p_i\,(y_i) - \ln |\det(\boldsymbol{B})| \right].$$

Gradient or "Natural gradient"

$$\Delta \boldsymbol{B} \propto H(\boldsymbol{y}) \quad \text{or} \quad \boldsymbol{A}^t \boldsymbol{A} \boldsymbol{H}(\boldsymbol{y}) = [\boldsymbol{I} - \boldsymbol{\phi}(\boldsymbol{y}) \boldsymbol{y}^t] \boldsymbol{B}$$



Accounting for errors

$$\boldsymbol{g}(t) = \boldsymbol{A}\,\boldsymbol{f}(t) + \boldsymbol{\epsilon}(t), \quad t = 1, \ldots, T.$$

$$\ln p(\boldsymbol{\epsilon}(1), \ldots, \boldsymbol{\epsilon}(T)) = \sum_t \sum_i \ln p_i\left(\epsilon_i(t)\right).$$

From this assumption, we obtain

$$\ln p\left(\boldsymbol{g}_{1..T} | \boldsymbol{A}, \boldsymbol{f}_{1..T}\right) = \sum_t \sum_i q_i\left(g_i(t) - [\boldsymbol{A}\boldsymbol{f}]_i(t)\right)$$

with $q_i(.) = \ln p_i(.)$.

$$\begin{aligned}
\ln p\left(\boldsymbol{A}, \boldsymbol{f}_{1..T} | \boldsymbol{g}_{1..T}\right) &= \ln p\left(\boldsymbol{g}_{1..T} | \boldsymbol{A}, \boldsymbol{f}_{1..T}\right) + \ln p(\boldsymbol{f}_{1..T}) + \ln p(\boldsymbol{A}) + cte \\
&= \sum_t \sum_i q_i\left(g_i(t) - [\boldsymbol{A}\boldsymbol{f}]_i(t)\right) + \ln p(\boldsymbol{f}_{1..T}) + \ln p(\boldsymbol{A}) + cte.
\end{aligned}$$

Three directions (depending on applications):

- Integrate out $\boldsymbol{A}$ to obtain $p(\boldsymbol{f}_{1..T}|\boldsymbol{g}_{1..T})$ and estimate $\boldsymbol{f}_{1..T}$ by

$$\widehat{\boldsymbol{f}}_{1..T} = \arg\max_{\boldsymbol{f}_{1..T}} \{p(\boldsymbol{f}_{1..T}|\boldsymbol{g}_{1..T})\}$$

- Integrate out $\boldsymbol{f}_{1..T}$ to obtain $p(\boldsymbol{A}|\boldsymbol{g}_{1..T})$, estimate $\boldsymbol{A}$ by

$$\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \{p(\boldsymbol{A}|\boldsymbol{g}_{1..T})\},$$

  and estimate $\boldsymbol{f}$ by $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{A}}\boldsymbol{g}$;

- Optimize $p(\boldsymbol{A}, \boldsymbol{f}_{1..T}|\boldsymbol{g}_{1..T})$ simultaneously with respect to both $\boldsymbol{f}_{1..T}$ and $\boldsymbol{A}$

$$\left\{\begin{array}{lcl} \widehat{\boldsymbol{f}}^{(k)}_{1..T} & = & \arg\max_{\boldsymbol{f}_{1..T}} \left\{p\left(\widehat{\boldsymbol{A}}^{(k-1)}, \boldsymbol{f}_{1..T}|\boldsymbol{g}_{1..T}\right)\right\} \\ \widehat{\boldsymbol{A}}^{(k)} & = & \arg\max_{\boldsymbol{A}} \left\{p\left(\boldsymbol{A}, \widehat{\boldsymbol{f}}^{(k-1)}_{1..T}|\boldsymbol{g}_{1..T}\right)\right\} \end{array}\right.$$

Independent and white sources

$$\ln p(\boldsymbol{f}_{1..T}) = \sum_t \sum_j r_j(f_j(t))$$

$$p(\boldsymbol{A}) \propto \exp\left[-\frac{1}{2\sigma_a^2} \sum_k \sum_l a_{kl}^2\right],$$

$$\ln p\left(\boldsymbol{A}, \boldsymbol{f}_{1..T} | \boldsymbol{g}_{1..T}\right) = \sum_t \sum_i q_i\left(g_i(t) - y_i(t)\right) + \sum_t \sum_j r_j(f_j(t)) + \frac{1}{2\sigma_a^2} \sum_k \sum_l a_{kl}^2 + cte.$$

Summation over t can be omitted. Simultaneous optimization with respect to $\boldsymbol{f}$ and $\boldsymbol{A}$ becomes:

$$\left\{\begin{array}{rcl} \widehat{\boldsymbol{f}}^{(k)} & = & \arg\max_{\boldsymbol{f}} \left\{ \sum_i q_i\left(g_i - y_i\right) + \sum_j r_j(f_j) \right\} \\ \widehat{\boldsymbol{A}}^{(k)} & = & \arg\max_{\boldsymbol{A}} \left\{ \sum_i q_i\left(g_i - y_i\right) + \frac{1}{2\sigma_a^2} \sum_k \sum_l a_{kl}^2 \right\} \end{array}\right.$$

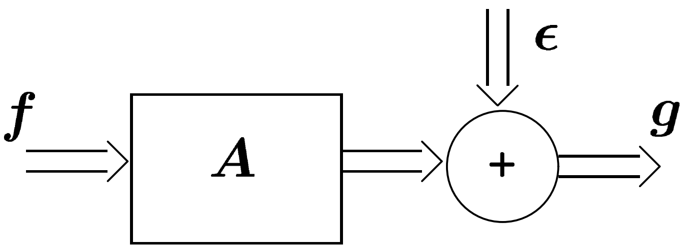
- **Gaussian laws** for noise and sources:

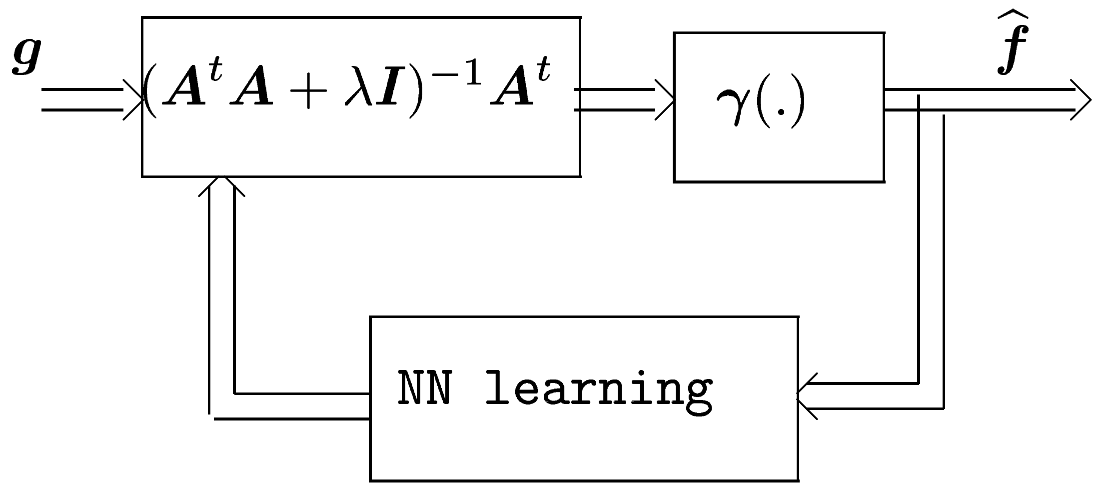
$$\begin{cases} \boldsymbol{f} &= (\boldsymbol{A}^t\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{g} \\ \boldsymbol{A} &= \boldsymbol{g}\boldsymbol{f}^t(\boldsymbol{f}\boldsymbol{f}^t+\mu\boldsymbol{I})^{-1} \end{cases}$$

where $\lambda = \sigma_n^2/\sigma_s^2$ and $\mu = \sigma_n^2/\sigma_a^2$.

- **Non Gaussian law for $\boldsymbol{f}$:**

$$\begin{cases} \boldsymbol{y} &= (\boldsymbol{A}^t\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{g} \\ \boldsymbol{f} &= \gamma(\boldsymbol{y}) \\ \boldsymbol{A} &= \boldsymbol{g}\boldsymbol{f}^t(\boldsymbol{f}\boldsymbol{f}^t+\mu\boldsymbol{I})^{-1} \end{cases}$$





Spatially correlated sources

$$\boldsymbol{f} = (\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{g} \quad \longrightarrow \quad \boldsymbol{f} = (\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{A}^t\boldsymbol{g}$$

Spatially independant but colored sources

- Main difficulty: How to model $p(f_j(1), \ldots, f_j(T))$
  - First order markov chain model:

$$\ln p(f_j(1), \ldots, f_j(T)) = \sum_t \ln p(f_j(t)|f_j(t-1))$$

  - Gaussian case:

$$\boldsymbol{f}(t) = (\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\left[\operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}\,\boldsymbol{f}(t-1) + \boldsymbol{A}^t\boldsymbol{g}(t)\right]$$

Marginalization

$$\boldsymbol{g} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{A}, \boldsymbol{f}|\boldsymbol{g}) \propto \exp\left[-\frac{1}{2\sigma_n^2} J(\boldsymbol{A}, \boldsymbol{f})\right] \quad \text{with} \quad J(\boldsymbol{A}, \boldsymbol{f}) \|\boldsymbol{g} - \boldsymbol{A}\boldsymbol{f}\|^2 + \lambda\phi(\boldsymbol{f}) + \mu\psi(\boldsymbol{A})$$

$$p(\boldsymbol{A}|\boldsymbol{g}) = \int p(\boldsymbol{A}, \boldsymbol{f}|\boldsymbol{g}) \,\mathrm{d}\boldsymbol{f}$$

- Second order approximation:

$$-\ln p(\boldsymbol{A}|\boldsymbol{g}) \propto -\ln \left|\det(\widehat{\boldsymbol{P}}_s^{-1})\right| - J(\boldsymbol{A}, \widehat{\boldsymbol{f}})$$

$$\widehat{\boldsymbol{f}} = (\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{g} \quad \text{and} \quad \widehat{\boldsymbol{P}}_s = (\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1};$$

$\longrightarrow$ We obtain the same Neural Network like algorithm :

$$\Delta\boldsymbol{A} \propto \boldsymbol{A}^t(\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1} + \boldsymbol{g}\boldsymbol{f} + \mu\psi'(\boldsymbol{A})$$

- Expectation-Maximizatio (EM) algorithem:
  A standard iterative algorithem to obtain $\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \{p(\boldsymbol{A}|\boldsymbol{g})\}$ using $(\boldsymbol{g}, \boldsymbol{f})$ as complete data and $(\boldsymbol{g})$ as incomplete data.

Mixture of Gaussians and Hidden variables

$$p(f_j) = \sum_{i=1}^{q_j} \alpha_{ji} \mathcal{N}(m_{ji}, \sigma_{ji}^2)$$

- Hidden variable: $z_j \in \mathcal{Z}_j = (1, \ldots, q_j)$ with $\alpha_{ji} = p(z_j = i)$

$$p(\boldsymbol{f}|\boldsymbol{z}) = \mathcal{N}(m_{ji}, \sigma_{ji}^2)$$

  – Marginal *a priori* law $\quad p(\boldsymbol{f}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} p(\boldsymbol{z})\, p(\boldsymbol{f}|\boldsymbol{z})$

  – Marginal *a posteriori* law $\quad p(\boldsymbol{f}|\boldsymbol{g}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{g})\, p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{z})$

- Estimation procedure:

  – Estimate $\boldsymbol{z}$ using $p(\boldsymbol{z}|\boldsymbol{g})$;

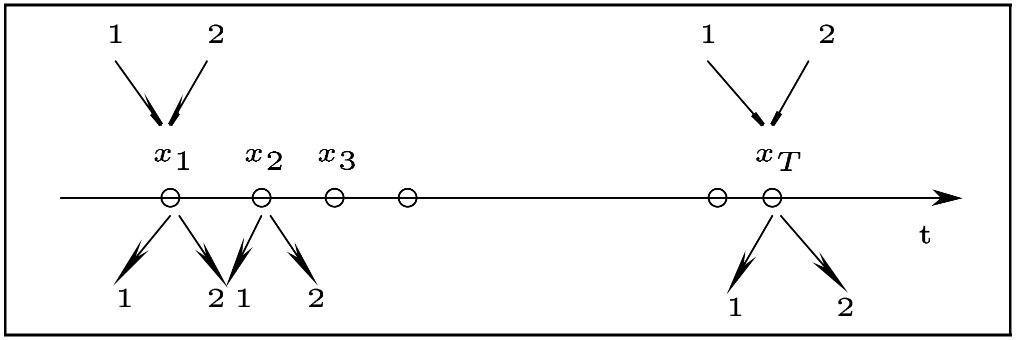
  – Given $\boldsymbol{z}$ using $p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{z})$ is a mixture of Gaussians;

  – Estimate the hyperparameters (the means and variances of Gaussians).

Hyperparameters estimation

- Partition of the set $\mathcal{T} = [1, \ldots, T]$:

$$\mathcal{T}_z = \{\, t \,|\, z_j(t) = z \} \ , \ z \in \mathcal{Z}_j = (1, \ldots, q_j)$$



- Estimation of m_{jz} and σ_{jz} taking account only $\mathcal{T}_z$

  $\longrightarrow$ Hyperparameters estimation of Gaussian distribution using the marginal *a posteriori* laws $p(\theta_{jz} \,|\, \boldsymbol{g}, \, \boldsymbol{A}, \, z_j)$ with

  - Uniform prior laws for means: $\quad \widehat{m}_{jz}^{MAP} = \dfrac{\sum_{t \in \mathcal{T}_z} m_j(t)}{T_z}$
  - Inverted Gamma $\mathcal{G}(\alpha, \, \beta)$ for variances:
    - Conjugate prior;
    - Eliminates the degeneracy of likelihood function [Ridolfi, Idier99]

JMAP algorithm:

Joint estimation of parameters and hyperparameters

- **An iterative algorithem in 5 steps:**

1. Classification of data $\longrightarrow$ Estimation of partitions:
$$\widehat{\mathcal{T}_z} = \left\{ t \,|\, \widehat{(z_j)}^{MAP}(t) = z \right\}$$
2. Estimation of hyperparameters: $\widehat{\psi}_{jz}^{MAP}$ and $\widehat{m}_{jz}^{MAP}$
3. Estimation of hidden variables: $(\widehat{z}_j)_{1..T}^{MAP}$
4. Estimation of sources: $(\widehat{\boldsymbol{f}})_{1..T}^{MAP}$
5. Estimation of mixing matrix: $\widehat{\boldsymbol{A}}^{MAP}$

Link and comparison with EM algorithm

- $\boldsymbol{A}$ mixing matrix, $\boldsymbol{\theta}$ hyperparameters:

$$V(\boldsymbol{A},\, \boldsymbol{\theta}) = \log p(\boldsymbol{g},\, \boldsymbol{f} \,|\, \boldsymbol{A},\, \boldsymbol{\theta})$$

- EM algorithm:

1. E-Step (expectation):

$$Q\left(\boldsymbol{A},\, \boldsymbol{\theta} \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right) = E\left[\log p(\boldsymbol{g},\, \boldsymbol{f} \,|\, \boldsymbol{A},\, \boldsymbol{\theta}) \,|\, \boldsymbol{g},\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right]$$

2. M-Step (maximization):

$$\left(\widehat{\boldsymbol{A}},\, \widehat{\boldsymbol{\theta}}\right) = \arg\max_{(\boldsymbol{A},\boldsymbol{\theta})} \left\{Q\left(\boldsymbol{A},\, \boldsymbol{\theta} \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right)\right\}$$

- Main drawbacks:

  - $V(\boldsymbol{A},\, \boldsymbol{\theta})$ unbounded (degeneracy) [Ridolfi, Idier99]
  - Very sensitive to initial conditions
  - No *a priori* for $\boldsymbol{A}$, computational cost too high

Penalized EM: *a priori* on A and on hyperparameters

- The 2 steps become:

1. E-Step (expectation):

$$Q\left(\boldsymbol{A},\,\boldsymbol{\theta}\,|\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right) = E\left[\log\, p(\boldsymbol{g},\,\boldsymbol{f},\,|\,\boldsymbol{A},\,\boldsymbol{\theta}) + \log\, p(\boldsymbol{A}) + \log\, p(\boldsymbol{\theta})\,|\,\boldsymbol{g},\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right]$$

2. M-Step (maximization):

$$\left(\widehat{\boldsymbol{A}},\,\widehat{\boldsymbol{\theta}}\right) = \arg\max_{(\boldsymbol{A},\boldsymbol{\theta})}\left\{Q\left(\boldsymbol{A},\,\boldsymbol{\theta}\,|\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right)\right\}$$

- $E\left[\,f\,(\boldsymbol{f})\,|\,\boldsymbol{g},\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right] = \displaystyle\sum_{z_1,\ldots,\,z_n} E\left[\,f\,(\boldsymbol{f})\,|\,\boldsymbol{g},\boldsymbol{z} = (z_1,\ldots,\,z_n),\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right]\, p(\boldsymbol{z}\,|\,\boldsymbol{g},\,\boldsymbol{A}',\,\boldsymbol{\theta}')$.

- Very high computational cost $\longrightarrow$ Classification:

$$E\left[\,f\,(\boldsymbol{f})\,|\,\boldsymbol{g},\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right] = E\left[\,f\,(\boldsymbol{f})\,|\,\boldsymbol{g},\boldsymbol{z} = \widehat{\boldsymbol{z}}^{MAP},\,\boldsymbol{A}',\,\boldsymbol{\theta}'\right].$$

Comparison JMAP/EM

- Estimation of hyperparameters:
  - JMAP: Classification + Estimation
  - EM: No classification
- Estimation of $\boldsymbol{A}$:
  - JMAP: With classification + jointly with $\boldsymbol{f}$
  - EM: With classification + Marginal in $\boldsymbol{A}$
- EM sensible to initial conditions, JMAP more robust.

3 Conclusion and perspectives

❏ Specificities of Bayesian approach:

– Taking account of noise and model incertitude

– Introduction of *a priori* for mixing matrix and hyperparameters

❏ Perspectives:

– Spatially correlated and colored sources

- Markov chain modeling for classification labels.
- Other hierarchical modeling

– Number of sources unknown, number of Gaussians in the mixture unknown.

...

4 References

– A. Mohammad-Djafari,"A Bayesian approach to source separation," in MaxEnt99 Proceedings, Kluwer, 1999.

– E. Moulines, J. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in ICASSP-97, (Munich, Germany), Apr. 1997.

– A. Mohammad-Djafari,"Model selection for inverse problem: Best choice of basis function and model order selection," in MaxEnt99 Proceedings, Kluwer, 1999.

– O. Bermond, Mthodes statistiques pour la sparation de sources. PhD thesis, Ecole Nationale Suprieure des Tlcommunications, Jan. 2000.

– A. Ridolfi and J. Idier,"Penalized maximum likelihood estimation for univariate normal mixture distributions," in 17, (Vannes, France), pp. 259-262, Sept. 1999.

– O. Macchi and E. Moreau,"Adaptive unsupervised separation of discrete sources," in Signal Processing 73, pp. 49-66, 1999.

5 Simulation results

Algorithm:

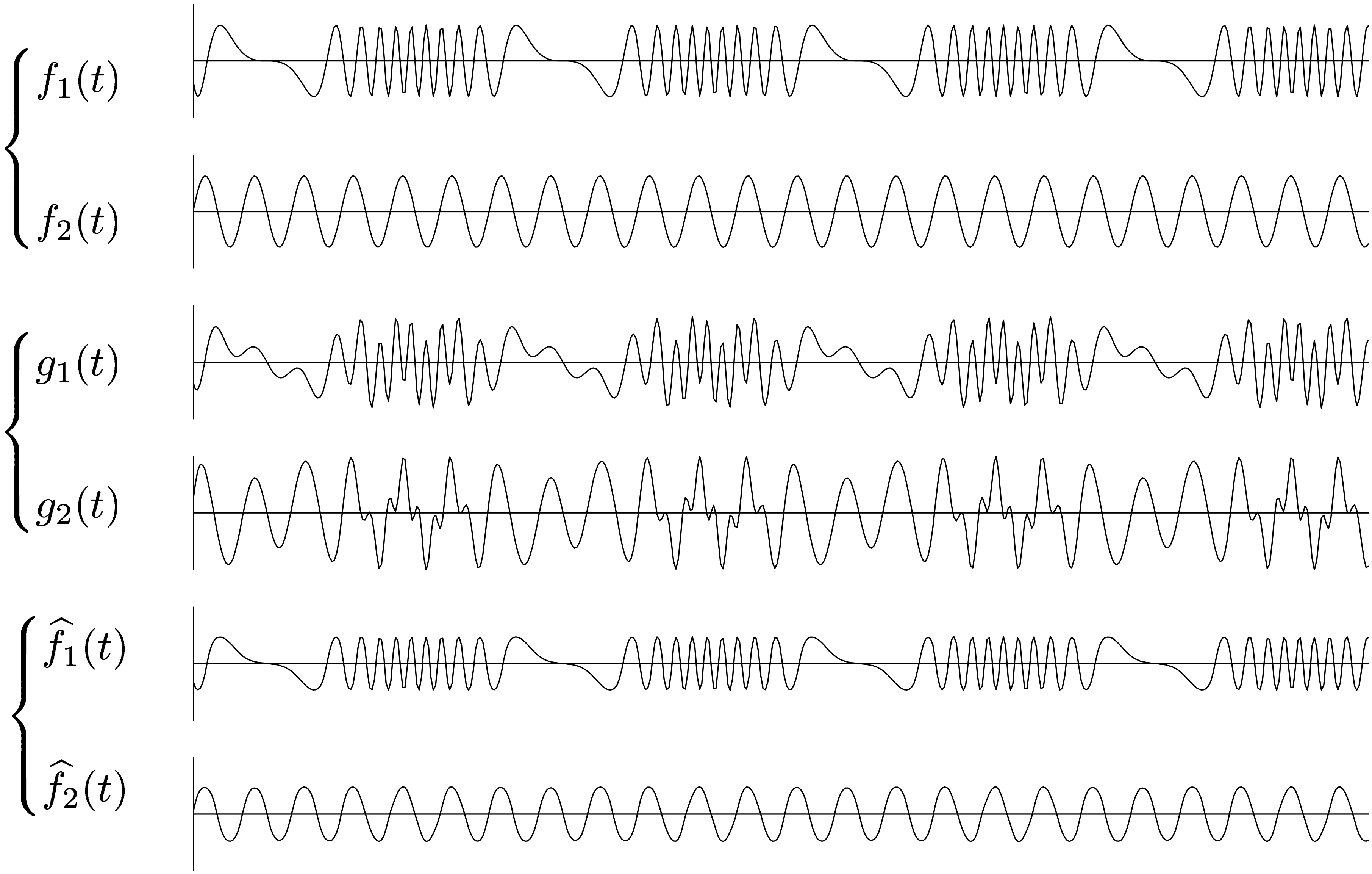
$$\begin{cases} \boldsymbol{y} & = & (\boldsymbol{A}^t\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{g} \\ \boldsymbol{f} & = & \gamma(\boldsymbol{y}) \\ \Delta\boldsymbol{A} & \propto & \boldsymbol{A}^t(\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{I})^{-1} + \boldsymbol{g}\boldsymbol{f} + \mu\psi'(\boldsymbol{A}) \end{cases}$$

$$\lambda = \mu = .1,\ N = 100,\ \text{Appropriate } \boldsymbol{g}$$

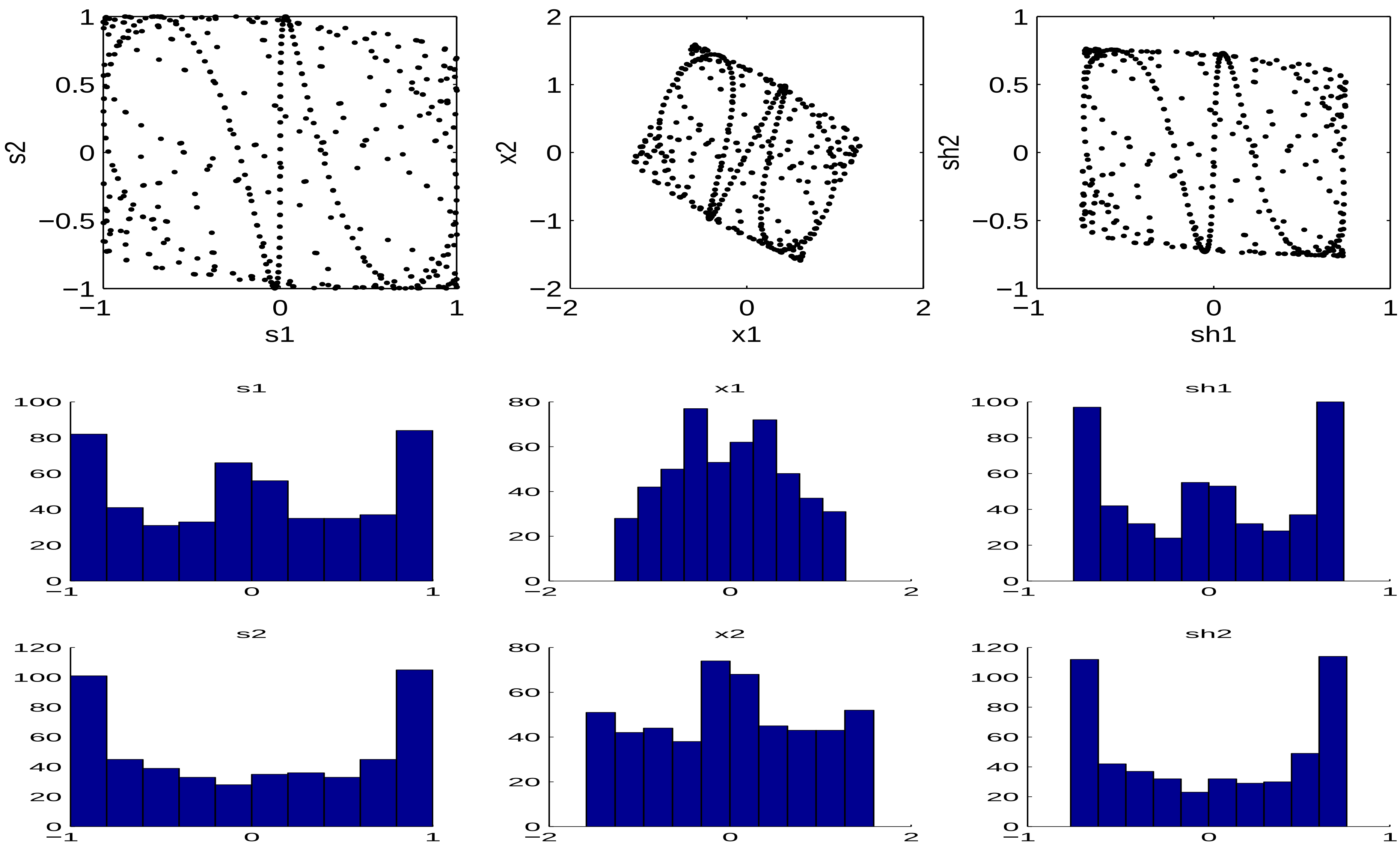
Example 1: 2 sources, 2 sensors

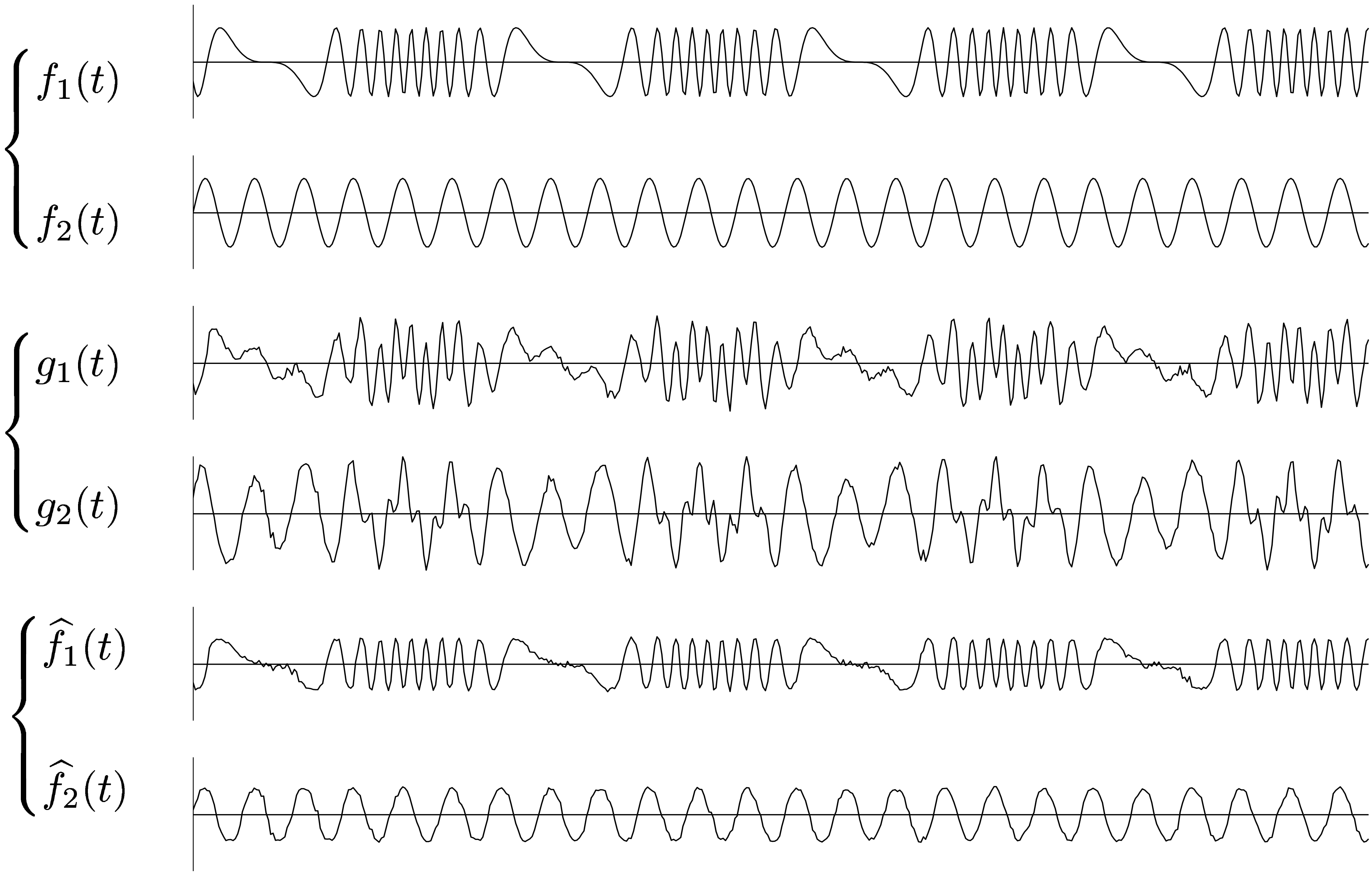
$$\begin{cases} f_1(t) & = & \sin(500t + 10\cos(50t)) \\ f_2(t) & = & \sin(300t) \end{cases}, \quad t = [0 : .001 : .499].$$

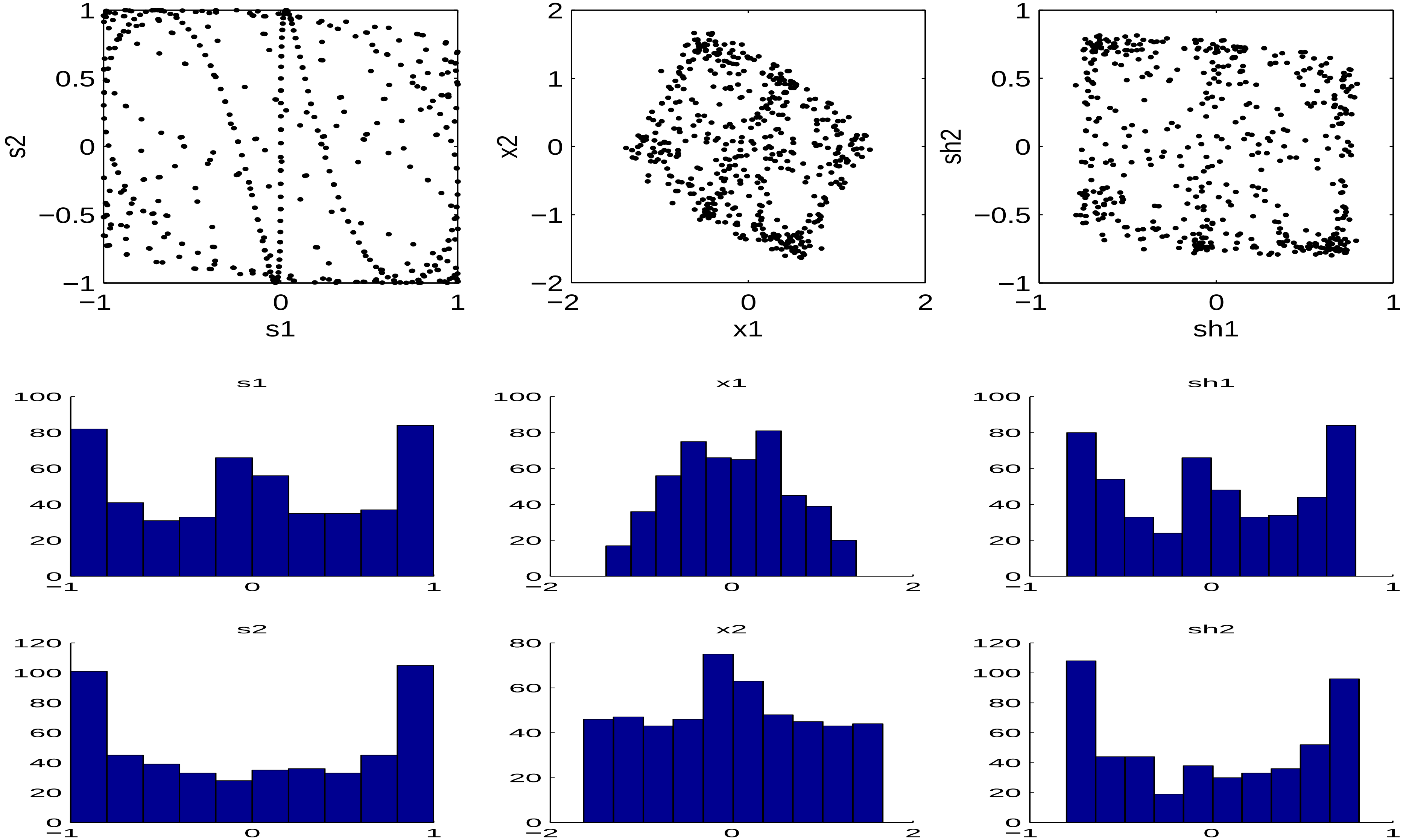
$$\boldsymbol{A} = \begin{pmatrix} 1 & .4 \\ -.6 & 1 \end{pmatrix}$$



$f_1(t)$
$f_2(t)$
$g_1(t)$
$g_2(t)$
$\widehat{f}_1(t)$
$\widehat{f}_2(t)$


s1
s2
x1
x2
sh1
sh2


$f_1(t)$
$f_2(t)$
$g_1(t)$
$g_2(t)$
$\widehat{f_1}(t)$
$\widehat{f_2}(t)$

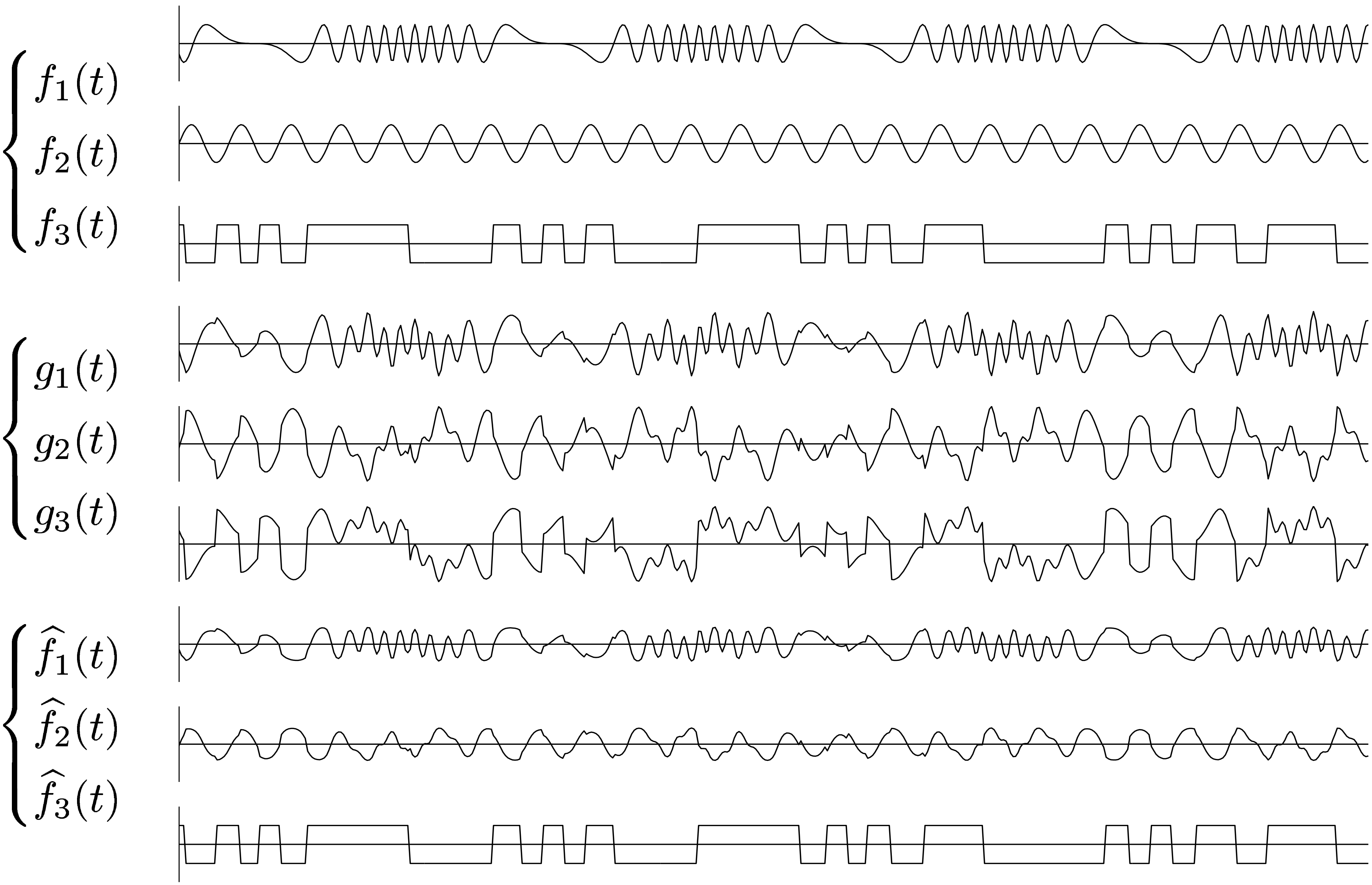


s1
s2
x1
x2
sh1
sh2

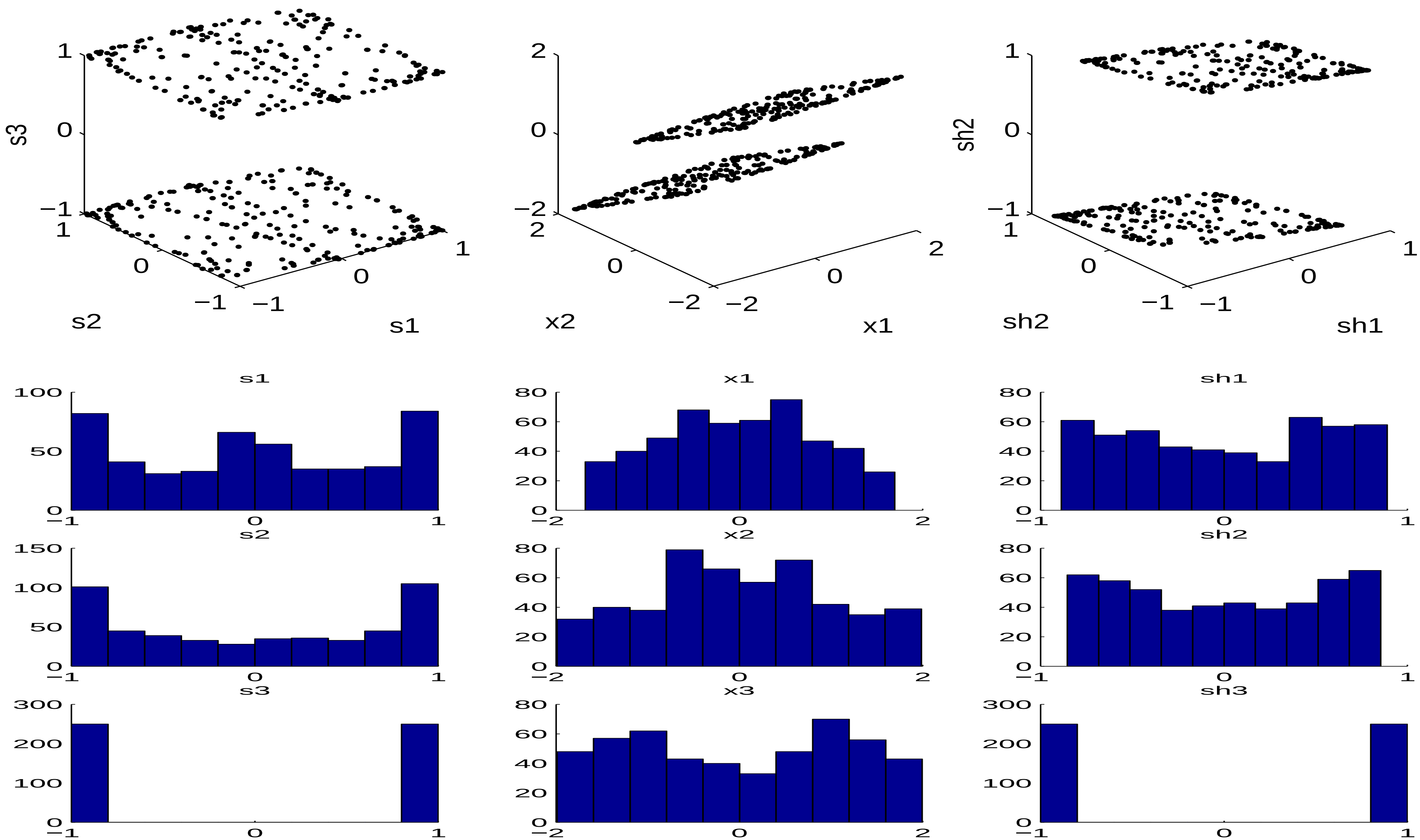
Example 2: 3 sources, 3 sensors

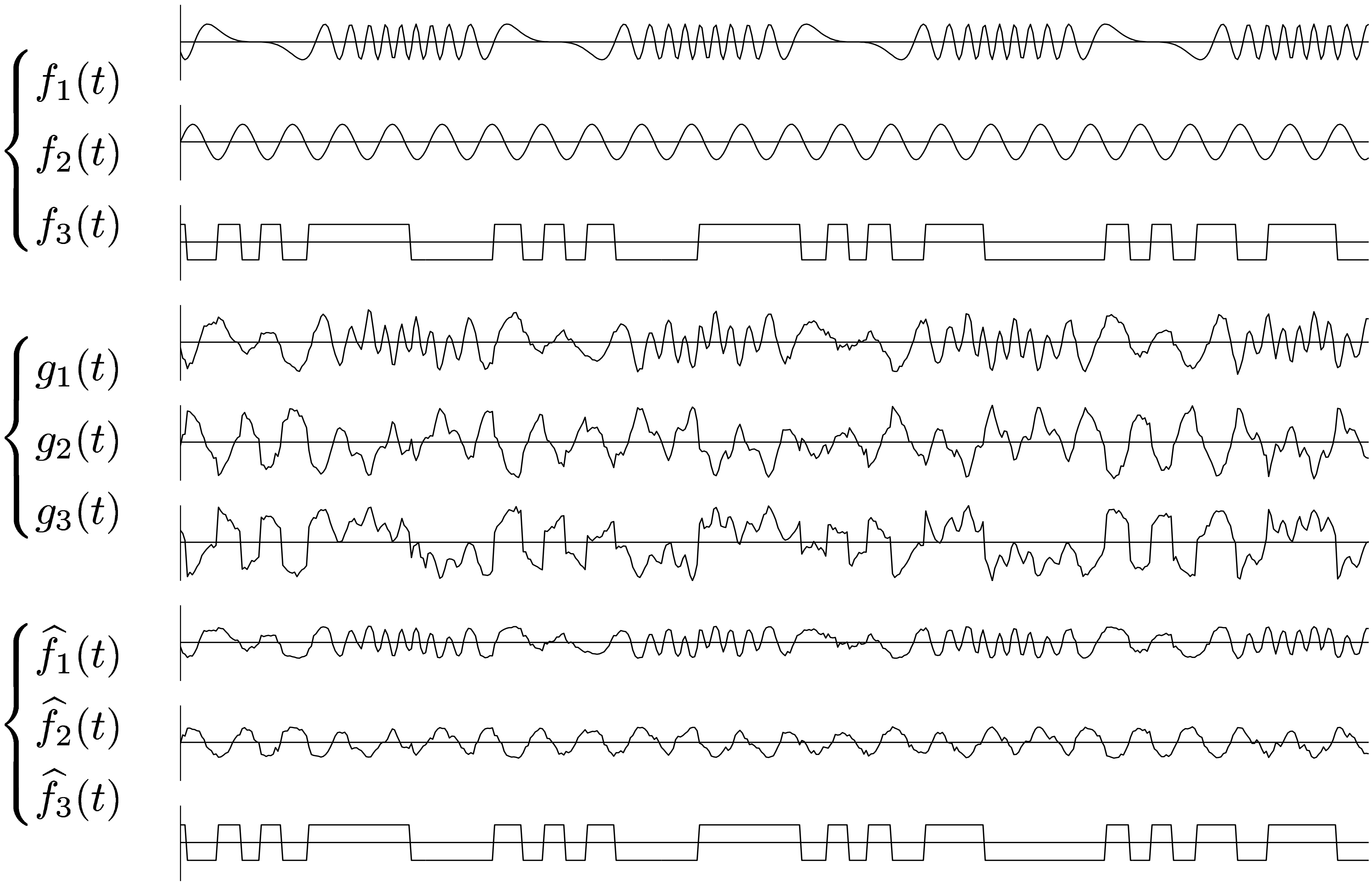
$$\left\{\begin{array}{lcl} f_1(t) & = & \sin(500t + 10\cos(50t)) \\ f_2(t) & = & \sin(300t) \\ f_3(t) & = & \mathrm{sign}(\cos(120t - 5\cos(50t))) \end{array}\right. , \quad t = [0 : .001 : .499].$$

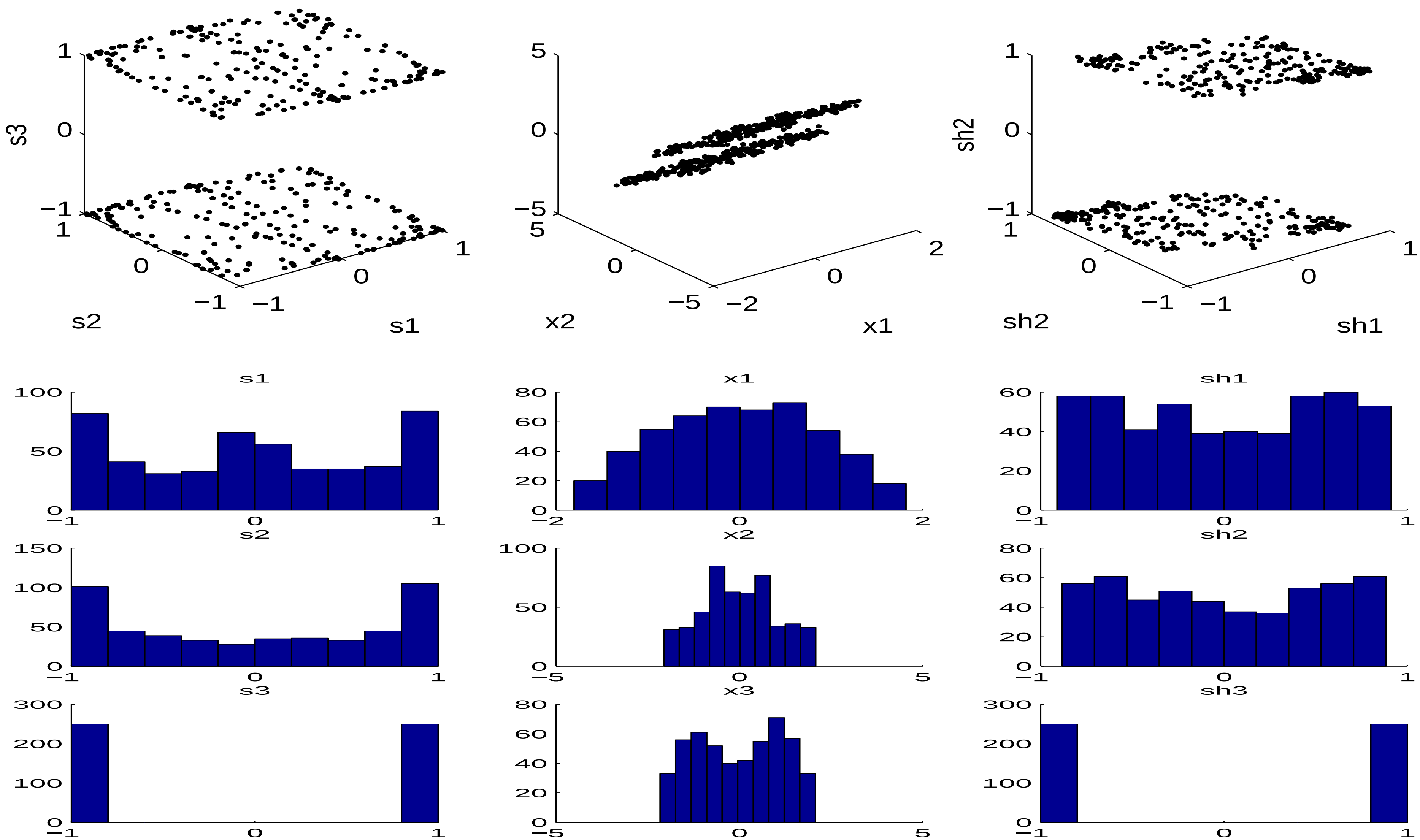
$$\boldsymbol{A} = .3 * \begin{pmatrix} 1. & -.5 & .2 \\ -.5 & 1. & -.5 \\ .5 & -.5 & 1. \end{pmatrix}$$



$f_1(t)$
$f_2(t)$
$f_3(t)$
$g_1(t)$
$g_2(t)$
$g_3(t)$
$\widehat{f}_1(t)$
$\widehat{f}_2(t)$
$\widehat{f}_3(t)$


s3
s2
s1
x2
x1
sh2
sh1
s1
x1
sh1
s2
x2
sh2
s3
x3
sh3


$f_1(t)$
$f_2(t)$
$f_3(t)$
$g_1(t)$
$g_2(t)$
$g_3(t)$
$\widehat{f}_1(t)$
$\widehat{f}_2(t)$
$\widehat{f}_3(t)$

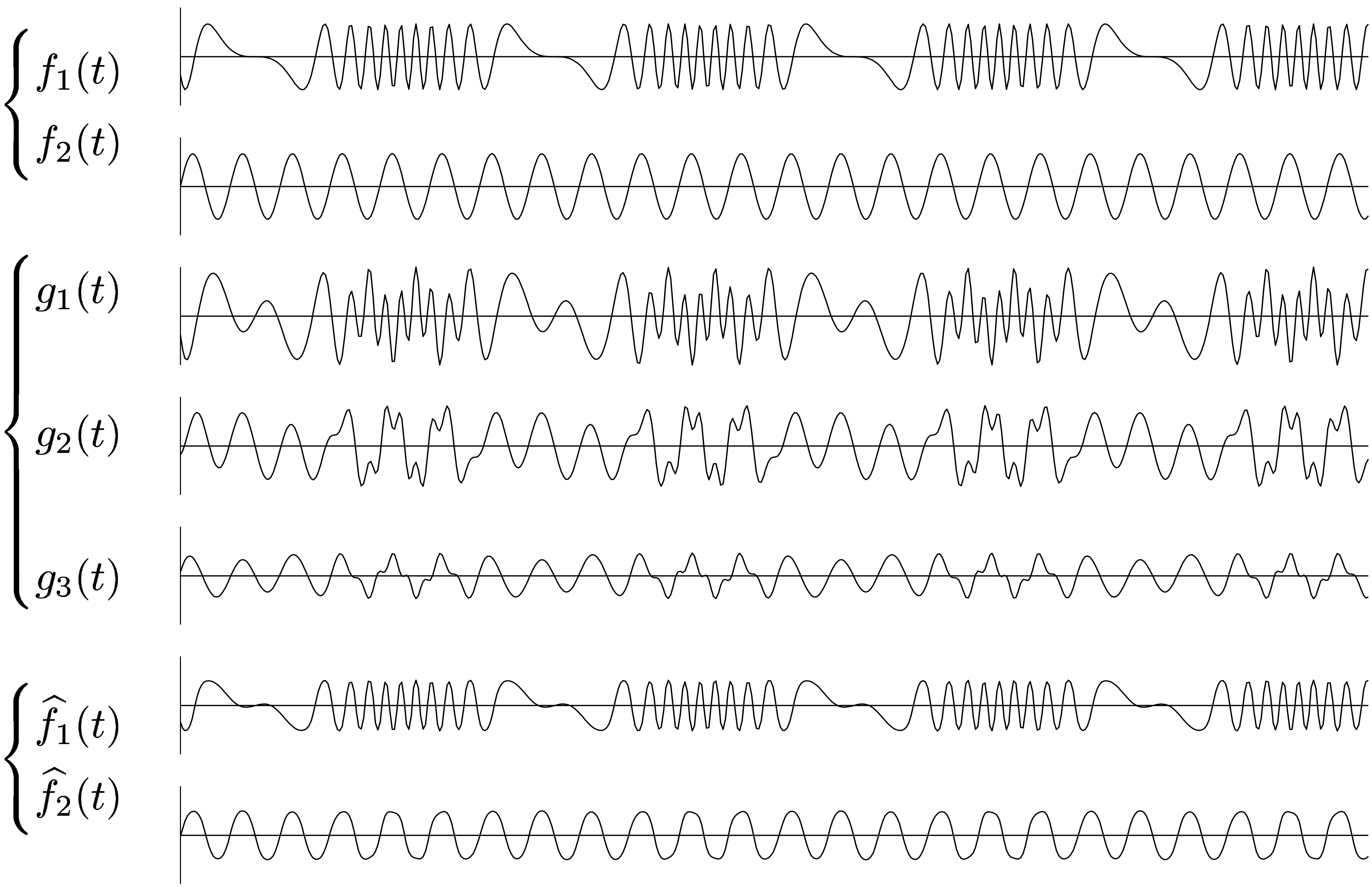


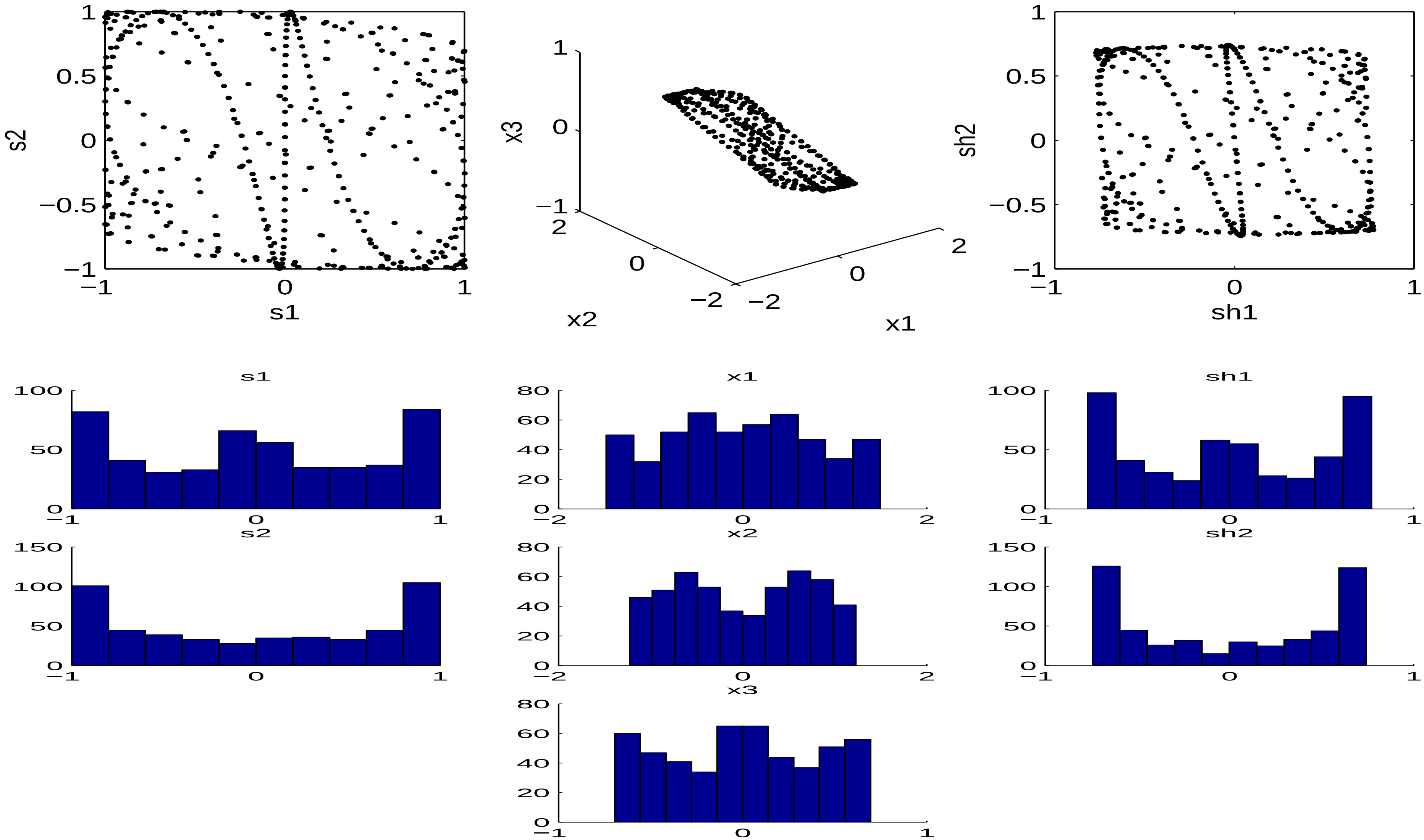
s3
s2
s1
x2
x1
sh2
sh1
s1
x1
sh1
s2
x2
sh2
s3
x3
sh3

Example 3: 2 sources, 3 sensors

$$\begin{cases} f_1(t) = \sin(500t + 10\cos(50t)) \\ f_2(t) = \sin(300t) \end{cases}, \quad t = [0 : .001 : .499].$$

$$\boldsymbol{A} = \begin{pmatrix} 1. & -.5 \\ .5 & 1. \\ -.2 & .5 \end{pmatrix}$$

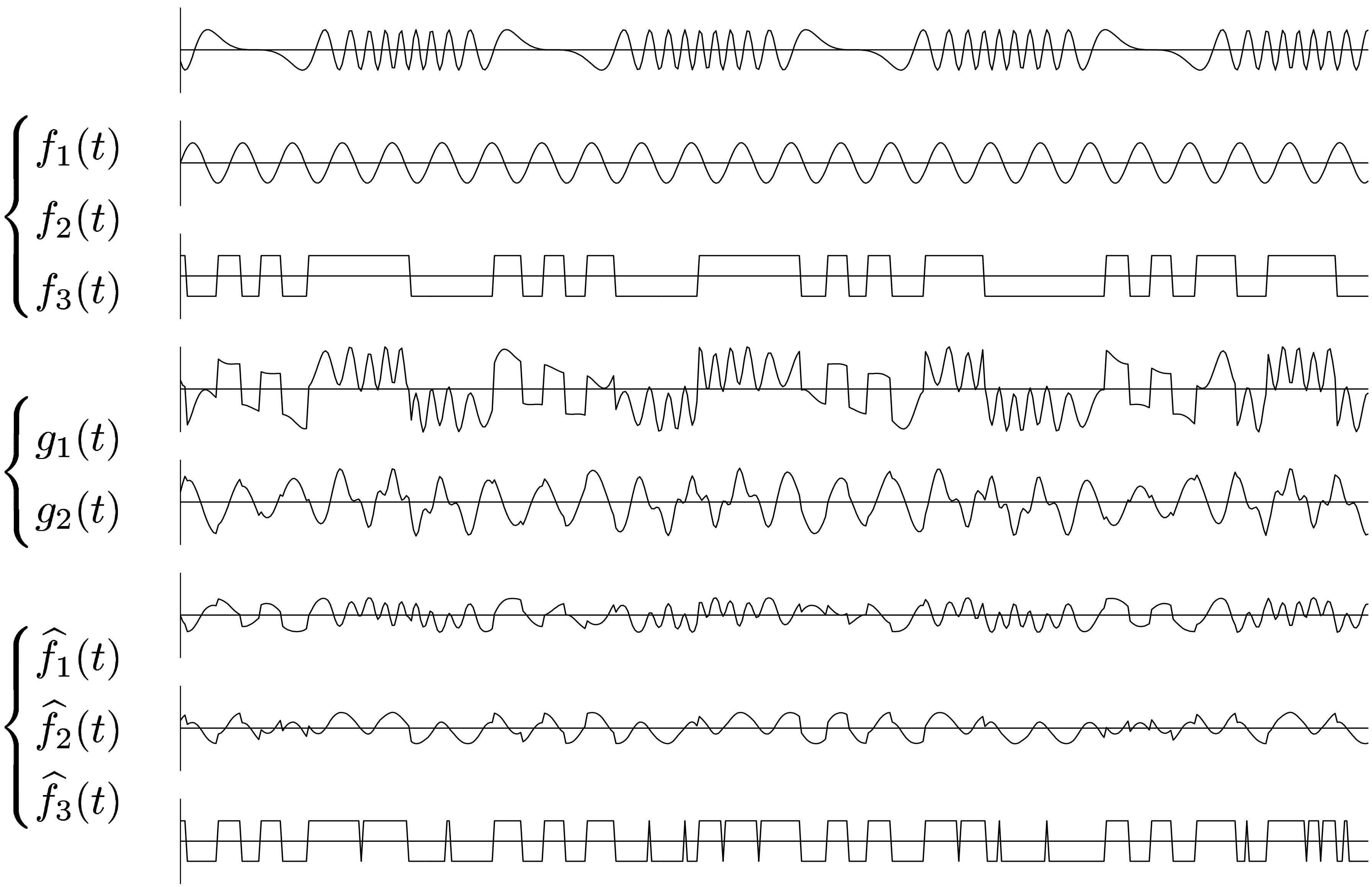

$f_1(t)$
$f_2(t)$
$g_1(t)$
$g_2(t)$
$g_3(t)$
$\widehat{f}_1(t)$
$\widehat{f}_2(t)$


s1
s2
x1
x2
x3
sh1
sh2

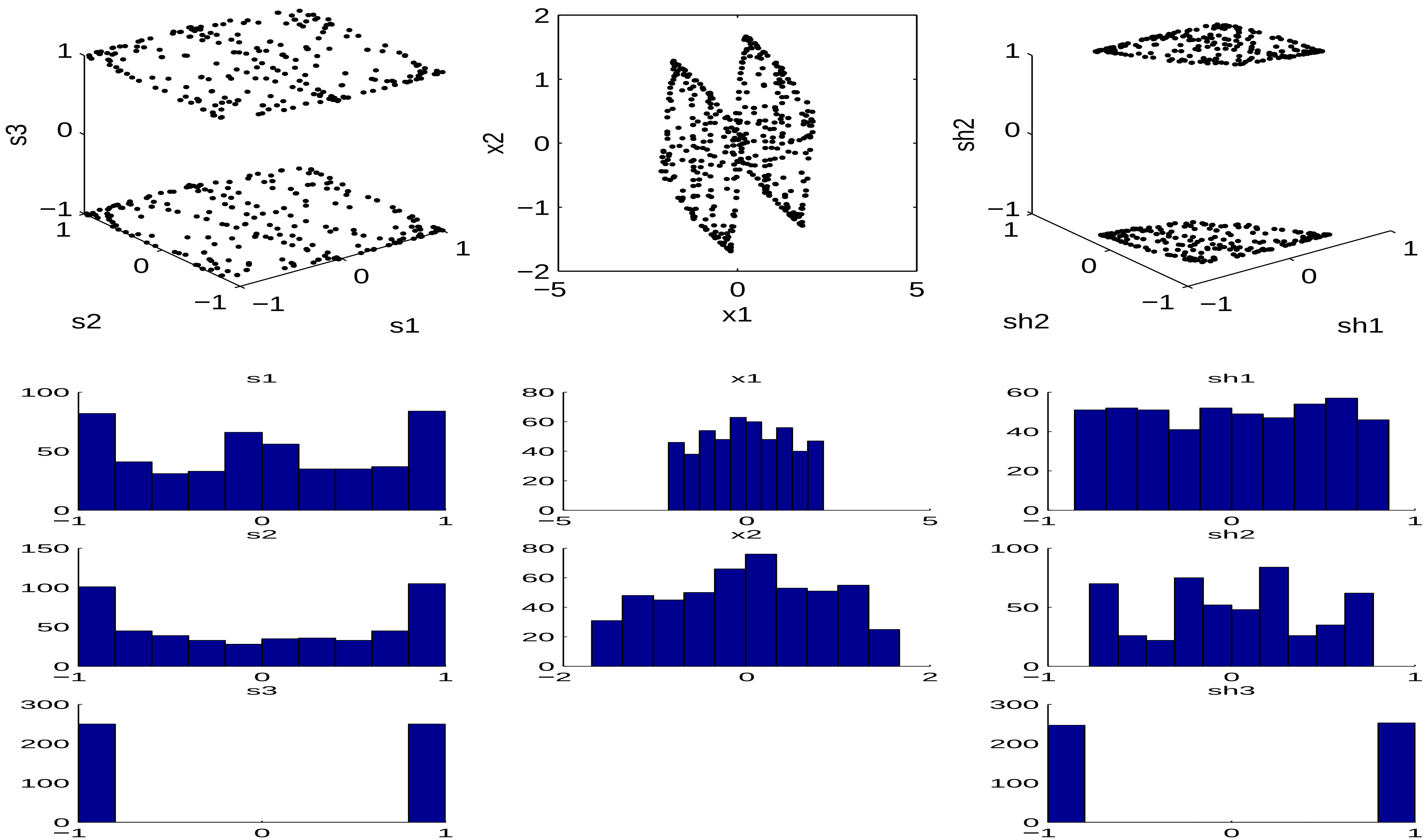
Example 4: 3 sources, 2 sensors

$$\left\{\begin{array}{lcl} f_1(t) & = & \sin(500t + 10\cos(50t)) \\ f_2(t) & = & \sin(300t) \\ f_3(t) & = & \mathrm{sign}(\cos(120t - 5\cos(50t))) \end{array}\right. , \quad t = [0 : .001 : .499].$$

$$\boldsymbol{A} = \begin{pmatrix} 1. & .2 & 1 \\ -.5 & 1. & .2 \end{pmatrix}$$



$f_1(t)$
$f_2(t)$
$f_3(t)$
$g_1(t)$
$g_2(t)$
$\widehat{f}_1(t)$
$\widehat{f}_2(t)$
$\widehat{f}_3(t)$


s3
s2
s1
x2
x1
sh2
sh1
s1
x1
sh1
s2
x2
sh2
s3
sh3

6 Conclusions

- Source separation is an important and difficult problem.
- Classical techniques assume $\boldsymbol{A}$ invertible and do not account for the errors (uncertainty on the model or the noise on the data).
- Bayesian approach can help to push farther the limits of the classical techniques.
- In classical techniques, the structure of learning scheme is fixed in an ad hoc way. In Bayesian approach, this structure comes out from the expressions of the estimators.
- This work is not finished. We have to quantify the benefits of the Bayesian approach compared to the classical ones on real data.

7 References

S. J. Roberts, "Independent component analysis: Source assessment, and separation, a Bayesian approach," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 145, no. 3, 1998.

J. J. Rajan and P. J. W. Rayner, "Decomposition and the discrete karhunen-loeve transformation using a bayesian approach," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 144, no. 2, pp. 116–123, 1997.

K. Knuth, "Bayesian source separation and localization," in *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, San Diego, CA* (A. Mohammad-Djafari, ed.), pp. 147–158, July 1998.

K. Knuth and H. Vaughan JR., "Convergent Bayesian formulation of blind source separation and and electromagnetic source estimation," in *MaxEnt 98 Proceedings: Int. Workshop on Maximum Entropy and Bayesian methods, Garching, Germany* (F. R. von der Linden W., Dose W. and P. R., eds.), p. in press, 1998.

K. Knuth, "A Bayesian approach to source separation," in *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA'99, Aussios, France* (C. J. J.-F. Cardoso and P. Loubaton, eds.), pp. 283–288, 1999.

T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representation," *IEEE Signal Processing Letters*, p. in press, 1999.

A. Mohammad-Djafari, "A Bayesian approach to source separation," presented at Int. Workshop on Maximum Entropy and Bayesian methods (MaxEnt 99), Boise, Idaho, USA