Blind Signal Separation: An Overview of Density-Based Methods

Scott C. Douglas

Department of Electrical Engineering
School of Engineering and Applied Science
Southern Methodist University
Dallas, TX 75205 USA
Outline of Talk

- What is Blind Signal Separation (BSS)?
- BSS for Instantaneous Mixtures
- Translating BSS Algorithms to the Speech Separation Task
- An Example
- Conclusions and Additional Topics
What is Blind Signal Separation?

The goal of signal separation is to estimate \( m \) statistically-independent source signals from \( n \) different (usually linear) measured mixtures of these signals, where \( m \leq n \). Such a process is blind if examples of the source signals, along with their corresponding mixtures, are unavailable for training.
Applications of Blind Signal Separation

- Array processing in wireless communications: Employ diversity in multiple-antenna array to blindly separate signals sent by different transmitters.
- Signal enhancement in medicine: Characterize bodily processes by extracting coherent features in multichannel sensor recordings (EEG, MRI).
  ⇒ Speech separation in acoustics: Recover individual talker’s speech from multiple-microphone room recordings.
Speech Separation
Speech Separation (cont.)

Room Acoustic Response (Physical) -> Blind Signal Separation System (Electronic)
An Important Distinction: Instantaneous vs. Convolutive Mixing Conditions

- Spatial-only
- Narrowband
- Requires \((m \times n)\) demixing matrix

\[ s(k) \rightarrow A \rightarrow x(k) \]

- Spatio-temporal
- Wideband
- Requires \((m \times n)\) multichannel demixing filter

\[ s(k) \rightarrow A(z) \rightarrow x(k) \]

\\Rightarrow Lots of algorithms \hspace{1cm} \Rightarrow Fewer algorithms

©1999, Scott C. Douglas, Southern Methodist University

SOUTHERN
METHODIST
UNIVERSITY

S.C. Douglas
IWAENC’99
Goal of Blind Signal Separation (BSS)

\[ x(k) \xrightarrow{n} W \xrightarrow{m} y(k) \]

- Find a \( W \) such that \( WA = \Phi D \)

\[ x(k) \xrightarrow{n} W(z) \xrightarrow{m} y(k) \]

- Find a \( W(z) \) such that \( W(z)A(z) = \Phi D(z) \)

\( D = \{ \text{diagonal scaling matrix} \}, \quad D(z) = \{ \text{diagonal filtering matrix} \} \)

\( \Phi = \{ \text{permutation (shuffling) matrix} \} \)

Only known fact: Elements of \( s(k) \) are independent of each other
Why is BSS Popular Now?

- Development of a statistical framework for understanding BSS problems (by Comon, Cardoso, Amari, and others)
- Corresponding development of simple, efficient BSS algorithms
- The applications are timely (telecommunications, multimedia, medicine)
- The problem is intriguing right now – the ”Can you really do that?” factor
Approaches to BSS

- **Density-based:** Attempt to make the p.d.f. of $y(k)$ “look like” that of $s(k)$
  - Produces simple, efficient algorithms
  - Must have coarse knowledge of source distributions (e.g., speech is “heavy-tailed”)

- **Contrast-based:** Search a cost function dependent on $y(k)$ whose extrema (maxima and/or minima) correspond to independent sources
  - No knowledge of sources needed
  - More complicated, less efficient algorithms
Density Based Approaches to BSS

- Maximum Likelihood [Pham et al 1992]: Find the most-likely description of the data given a p.d.f. model
- Information Maximization [Bell et al 1995]: Maximize the joint information in $y(k)$ under constraints
- Minimum Mutual Information [Amari et al 1996]: Minimize the mutual information between elements of $y(k)$

$\implies$ Can be unified under a single criterion: Kullback-Leibler divergence (p.d.f. matching) [Cardoso 1997]
Source Signals

Source signal vector:

\[ s(k) = [s_1(k) \ s_2(k) \ \cdots \ s_m(k)]^T \]

where each \( s_i(k) \) is independent of \( s_j(k), i \neq j. \)

Joint Probability Density Function:

\[
\begin{align*}
p_s(s_1, \ldots, s_m) &= p_{s_1}(s_1) \times p_{s_2}(s_2) \times \cdots \times p_{s_m}(s_m) \\
p_{s_i}(s_i) &= \{ \text{marginal p.d.f. of } s_i(k) \}
\end{align*}
\]
Example: Instantaneous Maximum Likelihood BSS

Assumptions: (1) Identical source p.d.f.'s $p_s(s)$, (2) $m = n$

Cost Function: for data block $x(i), 1 \leq i \leq N$,

$$J_{ML}(W) = -\log |\text{det } W| - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \log p_s(y_j(i))$$

Gradient Descent Algorithm: for $y^{(k)}(i) = W(k)x(i),$

$$W(k + 1) = W(k) + \mu(k) \left[ W^{-T}(k) - \frac{1}{N} \sum_{i=1}^{N} f(y^{(k)}(i))x^T(i) \right]$$

$$f(y) = [f(y_1) \cdots f(y_m)]^T, \quad f(y) = -\frac{p'_s(y)}{p_s(y)} \text{ (score fn.)}$$
BSS Using Maximum Likelihood (cont.)

Properties of Gradient Algorithm:

- Can work even when $p_s(s)$ deviates from the true marginal distributions – sufficient condition:

$$E\{s_i^2(k)\}E\{f'(s_i(k))\} - E\{s_i(k)f(s_i(k))\} > 0$$

  $\Rightarrow$ Can use $f(y) = \text{sgn}(y)$ for speech signals

- Converges slowly when $\mathbf{A}$ is ill-conditioned

- Involves matrix inverse — Complex; No simple extension to convolutive mixing case
The Natural Gradient [Amari 1995]

Natural gradient adaptation: a modified gradient search.

Def’n: For the cost function $\mathcal{J}(\mathbf{w})$ and parameter vector $\mathbf{w}$,

$$
\mathbf{w}(k + 1) = \mathbf{w}(k) - \mu(k) \mathbf{G}^{-1}(\mathbf{w}(k)) \frac{\partial \mathcal{J}(\mathbf{w}(k))}{\partial \mathbf{w}}
$$

$$
\mathbf{G}(\mathbf{w}) = \{ \text{Riemannian metric tensor} \}
$$

$$
\frac{\partial \mathcal{J}(\mathbf{w}(k))}{\partial \mathbf{w}} = \{ \text{Standard (Euclidean) gradient} \}
$$
Why Natural Gradient?

Riemannian Geometry: the mathematics of curved space.
Natural Gradient for ML BSS

Algorithm ($N = 1$) [Cichocki et al 1994, Amari et al 1995]

$$W(k + 1) = W(k) + \mu(k) \left[ W^{-T}(k) - f(y(k))x^T(k) \right] W^T(k) W(k)$$

$$= W(k) + \mu(k) \left[ I - f(y(k))y^T(k) \right] W(k)$$

- Simple $\Rightarrow$ no matrix inversion needed
- Performance depends on $C(k) = W(k)A$ $\Rightarrow$ Converges equally fast even for ill-conditioned $A$ (equivariance property)
BSS for Convolution Mixtures

Measured Signal Model:

\[ x(k) = \sum_{j=-\infty}^{\infty} A_j s(k - j) \]

Separated Output Signals:

\[ y(k) = \sum_{j=-\infty}^{\infty} W_j(k) x(k - l) \]

**Goal:** Adjust \( \{W_j(k)\} \) such that

\[ W(z, k) A(z) \rightarrow \Phi D(z) \]

(Aside: If \( [D(z)]_{ii} = d_{ii} z^{-_\Lambda_i} \), this is multichannel deconvolution)
Spatio-Temporal Extensions of Spatial Adaptive Algorithms


1. **Multiplication** of two matrices in the spatial-only case is equivalent to convolution of their associated sequences in the multichannel dispersive case.

2. **Addition** of two matrices in the spatial-only case is equivalent to element-by-element addition of their associated sequences in the multichannel dispersive case.

3. **Transposition** of a matrix in the spatial-only case is equivalent to element-by-element transposition and time-reversal of its associated sequence in the multichannel dispersive case.
Approximation Issues

Problem: Translated algorithms usually cannot be implemented as given; some approximation is required

- $W(z, k)$ is IIR $\Rightarrow$ Truncate to FIR
- Updates are non-causal $\Rightarrow$ Use delayed updates
- Algorithms are complex $\Rightarrow$ Simplify by assuming slowly-varying coefficients

Aside: These same approximations are used in deriving active noise control algorithms
Natural Gradient for Convolutive BSS

ML Algorithm [Amari/Douglas/Cichocki/Yang 1997]:

\[ y(k) = \sum_{p=0}^{L} W_p(k)x(k - p) \]

\[ u(k) = \sum_{q=0}^{L} W^T_{L-q}(k)y(k - q) \]

\[ W_j(k + 1) = W_j(k) + \mu(k)[W_j(k) - f(y(k - L))u^T(k - j)] \]

- Complexity: approximately four multiply-adds/tap ⇒ Simple
- Involves FIR convolution operations ⇒ Can implement directly or in block form using FFTs
  × Some loss in convergence speed due to delayed updates
Getting the Algorithm to Work

- Pseudo-normalized step size for each \((i,j)\)th filter channel

\[
\mu_i(k) = \frac{\mu_0}{\beta + \sum_{p=0}^{L} y_i(k-p) f_i(y_i(k-p))}
\]

- For speech separation, use block-based updates; sample-by-sample updates appear to offer no significant advantage

- Filter initialization: \([W_p]_{ij} = \delta_{ij(p-q)}, 0 < q \ll L/2\).

- Perform individual linear prediction of each \(x_i(k)\) prior to separation ⇒ Use the inverse of these filters on the separated outputs to obtain a “more-natural-sounding” result.
Example: Two-Talker Speech Separation

- Environment: Conference room (5.5m by 8m) with low-level ventilation noise
- Data: 16 seconds at 8kHz sampling rate
- Algorithm: Block-update natural gradient BSS with linear predictive prewhitening
- Parameters: $L = 2048$, $p = 256$, $f(y) = \text{sgn}(y)$, 64-tap predictor
- 10 passes through data, with reduced step size at each pass
Conclusions

Blind signal separation . . .

- is an emerging technology (?)
- has already yielded novel algorithmic ideas
- can be applied to convolutive mixtures (e.g. speech separation)

Density-based methods . . .

- are the easiest to understand
- do have their limitations (e.g. p.d.f. knowledge)
Additional Topics

- Two-stage contrast-based methods
- Equivariant adaptive separation using independence (EASI) [Cardoso 1996]
- Adaptive prewhitening schemes
- Single-channel blind deconvolution
- Adaptive paraunitary filter banks
- Self-stabilized subspace analysis
To Learn More

- Read the October 1998 issue of *Proc. IEEE*.
- Get papers from my web page: www.seas.smu.edu/~douglas
- Take my “director’s cut” tutorial (26 hours/550 slides).