Nonparametric Independent Component Analysis for Nonstationary Mixtures

Nima Reyhani[1] and Peter Bickel[2]
[1]Helsinki University of Technology
[2]UC Berkeley, Department of Statistics

September 7, 2009

Independent Component Analysis-set up

- ICA assumption
  - $S = (s_1, \ldots, s_m)^T$, where s_i are independent r.v.s
  - assume model

$$X = AS, \quad A \in \mathbb{R}^{m \times m} \quad X, S \in \mathbb{R}^m$$

    for X observable and S latent r.v.s.

- Question: Given observations $X_1, \ldots, X_n$ estimate the mixing matrix A

Identifiability condition: (for iid) At most one of sources are Normal.

Algorithms:

- fastICA: Non-Gaussianity
- Samarov and Tsybakov: Empirical Fisher Information-like matrix.

ICA with non-iid sources

- Assume $s_i(.), i = 1, \ldots m$, is a time series and $s_i \perp s_j, i \neq j$, i.e. temporal structure.
- Let's define time-lagged covariance matrices (they are the same in iid case)

$$C_i := \mathbb{E}\{X(t)X(t+\tau_i)\}$$

$$\begin{aligned} X(t) = AS(t) \Rightarrow \mathbb{E}\{X(t)X(t+\tau_i)^T \\ = AS(t)S(t+\tau_i)^T A^T = A\Lambda(\tau_i)A^T \\ [\Lambda(\tau_i)]_{j,j} = [\Lambda_i]_{j,j} = \mathbb{E}\{s_j(t)s_j(t+\tau)\}. \end{aligned}$$

- in stationary case, $\|\Lambda_i - \Lambda_j\|_F \neq 0$ for some $i, j \Rightarrow$ contrast
- Estimating A is equivalent to find a matrix simultaneously diagonalizes C_i and C_j

Block Stationary ICA

Block Stationary:
take $[0, T] = [0, t_1) \cup \cdots \cup [t_L, T]$ and assume $s_i(t) \forall i$ stationary time series within each interval.

Sources of contrast:
e.g. temporal fixed structure within each block and different statistical properties between blocks. Also, we usually assume non-Gaussianity.

Block-Stationary ICA

- Solution proposed by Pham & Cardoso:
- log-likelihood of source vector:

$$C_{ML} = -\frac{1}{2}\sum \frac{s_i^2(t)}{\sigma_i^2(t)} + \log(2\pi\sigma_i^2(t)) =$$
$$-\frac{1}{2}\mathrm{tr}[\Sigma^{-2}(t)S(t)S(t)^T] - \frac{1}{2}\log\det[2\pi\Sigma^2(t)] =$$
$$-\frac{1}{2}\mathrm{tr}[\Sigma^{-2}(t)A^{-1}X(t)X(t)^{-T}A^{-T}] - \frac{1}{2}\log\det[2\pi\Sigma^2(t)]$$
$$\Sigma^2(t) = \mathrm{diag}[\sigma_1^2(t),\ldots,\sigma_K^2(t)]$$

assume $\sigma_i^2(t) = \sigma_{i,l}^2$ and set $\Sigma_l^2 = \mathrm{diag}(\sigma_{1,l}^2,\ldots,\sigma_{K,l}^2)$

- With some calculations:

$$C_{ML} = \frac{1}{2}\sum_{l=1}^{L} w_l[\mathrm{tr}(R_l^{-1}\hat{R}_l) - \log\det(R_l^{-1}\hat{R}_l)] + K,\ R_l = A\Sigma_l^2 A^T$$

$$C_{ML} = \sum_{l=1}^{L} off(A^{-1}\hat{R}^l A^{-T}),\quad \hat{R}_l = \frac{1}{\#T_l}\sum X_l(t)X_l(t)^T$$

- *off* measures deviation from diagonality

Proposed Algorithm

Let's consider

$$M^l(\gamma_l) := \mathbb{E}_p\{\gamma_l(X_l)\nabla p(X_l)\nabla^T p(X_l)\} = \sum_{i,j=1}^{m} M^l_{i,j}\beta_i\beta_j^T$$

$$M^l_{ij} = \det(B)^2\mathbb{E}\left\{\gamma_l(X_l)\prod_{m\neq i} p_m(X_l^T\beta_m)\prod_{k\neq j} p_k(X_l^T\beta_k)p_i'(X_l^T\beta_i)p_j'(X_l^T\beta_j)\right\}$$

β_i is the i-th row vector of $B = A^{-1}$, $u_i^l := X_l^T\beta_i$

$$M^l_{ij} = m^l_{ij}\int \gamma_l(X)p_i^2(u_i^l)p_j^2(u_j^l)p_i'(u_i^l)p_j'(u_j^l)\mathrm{d}u_i\mathrm{d}u_j$$

$$\int p_j^2(u)p_j'(u)du = 0, \forall j,\ M^l_{ij} = 0, i\neq j \qquad M^l = B\mathrm{diag}(M^l_{ii})B^T \Rightarrow$$

Replace the covariance matrix with the Fisher-like information matrix.

Empirical estimates

Assume $\gamma \equiv 1$,

$$\widehat{M}^l = \frac{1}{\#T_l} \sum_{i \in T_l} \nabla \hat{p}_{-i}(X_i^l) \nabla^T \hat{p}_{-i}(X_i^l)$$

$$\frac{\partial \hat{p}_{-i}}{\partial x_s}(X_i^l) = \frac{1}{(\#T_l - 1)h^{m+1}} \sum_{j=1, j \neq i} K_1\left(\frac{X_{js}^l - X_{is}^l}{h}\right) \prod_{k=1, k \neq s} K\left(\frac{X_{jk}^l - X_{ik}^l}{h}\right),$$

Joint Diagonalization

- Suppose matrices $C_1, \ldots, C_n$, have a common diagonalizing matrix A

$$C_i = A\Lambda_i A^T,$$

where Λ_i diagonal matrix.

- We are looking for the common diagonalizing matrix A.

$$\underset{A \in \mathbb{R}^{m \times m}, \Lambda_i}{\operatorname{argmin}} \sum_{i=1}^{n} \| C_i - A\Lambda_i A^T \|_F$$

for $\|.\|_F$ Frobenius-norm.

- Optimization technique: Block coordinate descent or Alternating Projection/Least Squares

Conclusion

- Constructing Fisher-like information matrix is more expensive but we get
  - more robustness to the noise
  - faster algorithmic convergence
  - more efficient than using covariance matrices
- Non-stationary is not always bad :-)

Thank you