ICA: recent advances and open questions

Aapo Hyvärinen
for the Canberra MLSS 2005

What to do with noise?

- "There is always noise in the real world" (???)
- $x = As + n$
- What should we assume about noise?
  - Gaussian? What kind of covariance?
  - Nongaussian independent?
- Or should noise be in some of the components?
- Are properties of noise known or not?
- What is noise anyway?

Noisy ICA: some methods (1)

- Gaussian noise is not very problematic for cumulant-based methods: cumulants are "immune" to gaussian noise
- But whitening is still a problem if covariance matrix is not known
- Difficult to extend to robust methods

Noisy ICA: some methods (2)

- For non-cumulant-based methods, we can try to reduce bias induced by noise
- Sometimes even possible to remove it (“Gaussian moments”, Hyvärinen 1999)
- But this requires that we know noise covariance matrix and the noise is gaussian

Noisy ICA: maximum likelihood

- It is not difficult to formulate the joint likelihood:

$$\log L(\mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(T)) =$$
$$-\sum_{t=1}^{T} \left[\frac{1}{2} \|\mathbf{A}\mathbf{s}(t) - \mathbf{x}(t)\|^2_{\Sigma^{-1}} + \sum_{i=1}^{n} f(s_i(t)) \right] + C$$

  where Sigma is noise covariance and f is log p of ICs

- Looks nice but does not seem to work

Noisy ICA: EM algorithm

- Very fashionable in some circles
- Can be used if the distributions of the ICs are modelled as mixtures of gaussians
- BUT: computational complexity exponential as a function of dimension!

Estimation of noise covariance

- Crucial in many methods
- Perhaps known a priori?
- Classic factor analysis gives one solution
- ML estimation of both A and noise covariance?

Estimation of noise covariance

- Crucial in many methods
- Perhaps known a priori?
- Classic factor analysis gives one solution
- ML estimation of both A and noise covariance?
- IF noise covariance equals signal covariance, ordinary ICA works!!!!

GLS method

- A new method for estimating both A and noise cov (Shimizu and Kano, 2003)
- Compute the fourth-order moments implied by a given A and noise cov, and try to match these with the observed moments
- Not much practical experience available, but similar methods widely used in some branches of statistics

Noisy ICA: conclusions

- Turns out to be difficult
- Little proof of practical utility
- In practice, often better to use ordinary ICA methods and first reduce noise by
  - Time filtering (lowpass)
  - PCA
  - Any other methods

Nonlinear ICA

Nonlinear is nonparametric?

- Usually, this means very general functions:
- $x=f(s)$ where f is “almost anything”
- Should perhaps be called nonparametric as in statistics
- Can this be solved for a general f?
  No.

Indeterminacy of nonlinear ICA

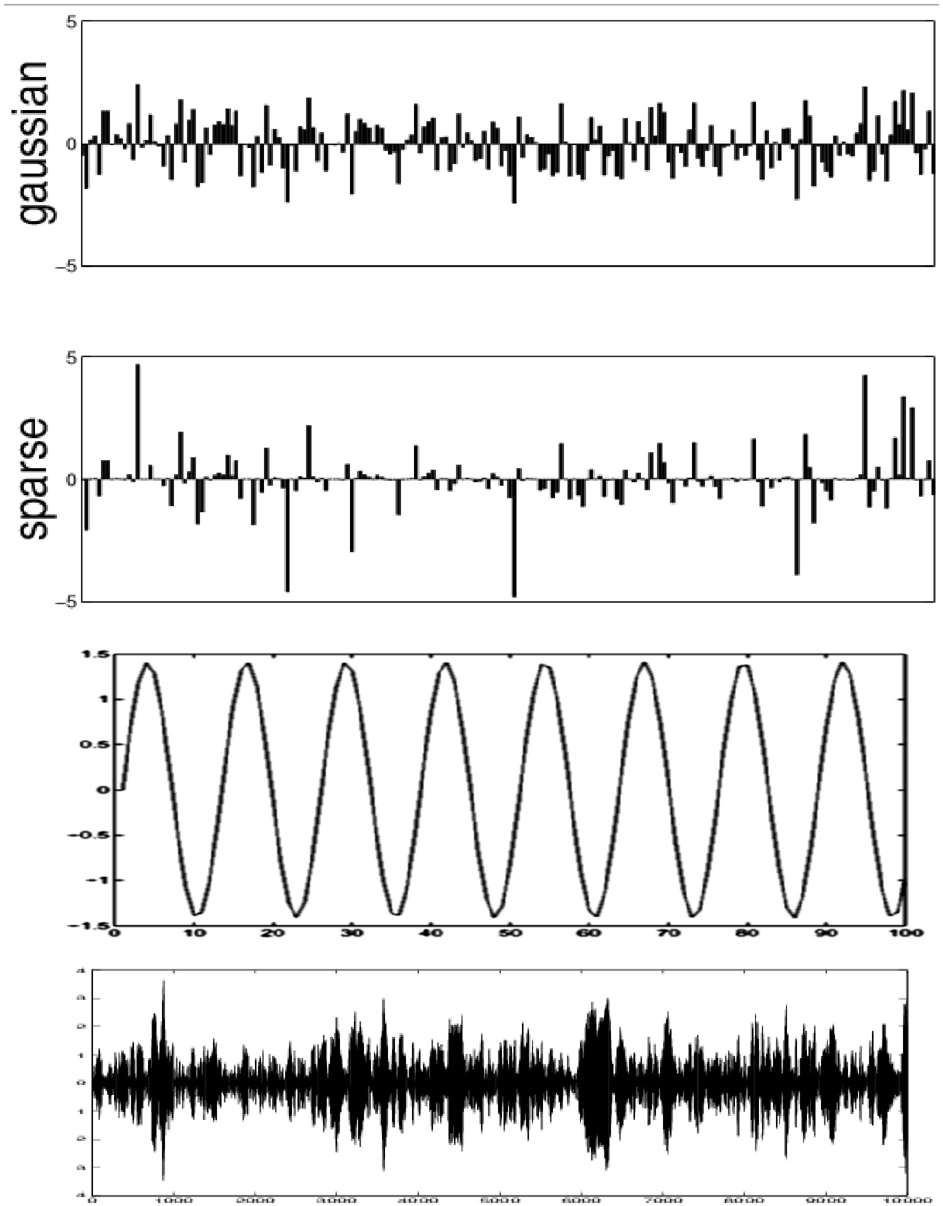
- We can always find an infinity of different nonlinear functions g so that $y=g(x)$ has independent components and these are very different solutions from each other
- We must restrict the problem in some way
- Recently, many people propose $x_i=f(w_i^t x)$ where f is scalar. "Post-nonlinear" mixtures

Nonlinear ICA

- A simple solution? Do nonlinear PCA (e.g. Kohonen map) and then linear ICA (very constrained)
- Another solution: constrain the nonlinearity to be smooth, use this as Bayesian prior (Harri Valpola et al)
- A lot of work needs to be done...

Using time structure of signals

Nongaussianity vs time structure



Linearly correlated signal

“Nonstationary variance”

Separation principles

- All three principles can be used for separation
  - Nongaussianity
  - Linear correlations (with some reservations)
  - Smoothly changing nonstationary variance
- Signals with time structure are much richer than random variables

Linear correlations

- A surprisingly simple algorithm is possible if all the source signals have non-zero autocorrelations $E\{s(t)\, s(t-d)\}$ where d is some delay, e.g. d=1
- Whiten the data and consider $C' = E\{x(t)\, x(t-d)^t\}$
- Do eigenvalue decomposition $C' = UDU^t$
- U gives mixing matrix if EVD is uniquely defined
- BIG IF: the autocorrelations but be distinct for the source signals (i.e. no two are equal)

Nonstationary variance

- Assume the source signals have smoothly changing “activity level”, i.e. variance
- Typical of speech signals, video data
- This enables separation by finding linear combinations that maximize variance nonstationarity
- (Cf.: maximize nongaussianity)

Comparison of separation principles

- These principles are independent from each other: none of them implies the other
- It would be interesting to develop combinations of these – some already exist (Complexity pursuit, Bubbles, see my home page)
- Incorporating as much prior information as possible should increase separation performance
- Unifying framework: finding signals of minimum coding complexity?

Miscellaneous stuff

Too many components

- If there are more ICs than observed variables, things get difficult
- A is no longer invertible, ICs cannot be computed exactly
- Can try to use joint likelihood to compute both A and s (with noise perhaps):

$$\log L(\mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(T)) = -\sum_{t=1}^{T} \left[\frac{1}{2} \|\mathbf{A}\mathbf{s}(t) - \mathbf{x}(t)\|^2_{\Sigma^{-1}} + \sum_{i=1}^{n} f(s_i(t)) \right] + C$$

Dimension reduction

- PCA is just an ad hoc method to use here
- How to properly reduce dimension, taking the ICA model into account
- What is the optimal number of principal components to retain?

How to “evaluate” components

- Which components are really interesting?
- Which components are local minima?
- Which components are due to overlearning?
- A simple approach: run algorithms many times, with randomization (Icasso, see my home page)

Thank you!