Reconhecimento de Padrões
Análise de Componentes Independentes

Prof. George von Borries
Departamento de Estatística
Universidade de Brasília

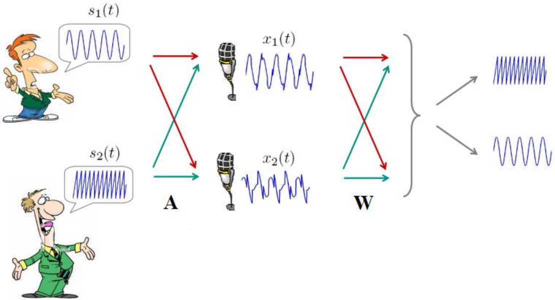
1 - 2024

Objetivos

- **Análise de Componentes Independentes (ICA)**
  projeções/direções independentes e não normais.
- **Busca de Projeção (PP)** - PROJECTION PURSUIT
  projeções/direções não normais.
- **Análise de Componentes Principais (PCA)**
  projeções/direções não correlacionadas.
- PCA com componentes normalmente distribuídas $\Rightarrow$ ICA.
- Semelhante a **Análise Fatorial**, porém com possíveis relações não lineares, diferentes suposições e não busca redução de dimensionalidade.

Motivação

THE COKTAIL-PARTY



Problema da fala. Journée (2008)

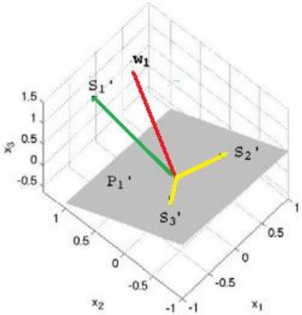
$$
\begin{aligned}
\mathbf{x}_1 &= a_{11}\mathbf{s}_1 + a_{12}\mathbf{s}_2 \\
\mathbf{x}_2 &= a_{21}\mathbf{s}_1 + a_{22}\mathbf{s}_2
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{s}_1 &= w_{11}\mathbf{x}_1 + w_{12}\mathbf{x}_2 \\
\mathbf{s}_2 &= w_{21}\mathbf{x}_1 + w_{22}\mathbf{x}_2
\end{aligned}
$$

Motivação

Geometricamente

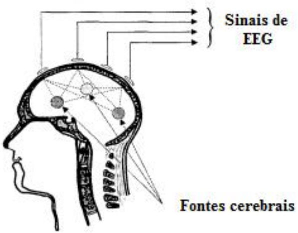
- Estimar **A**, **W**, **S** ;
- Encontar **W** que forneça fontes o mais independentes possível;
- **W** ótimo maximiza a não-gaussianidade de **WX**.



Extração de uma fonte (S_1) a partir de três misturas. Stone (2004)

Motivação

ICA em EEG



Representação das fontes no cérebro. Rajapakse, Cichocki e A. Sanchez (2002)

- Sinais altamente correlacionados.
- Fontes estatisticamente independentes.
- Quantidade de fontes?
- Localização das fontes?

Modelos

1. Modelo ruidoso: $\mathbf{X} = \mathrm{f}(\mathbf{S}) + \mathbf{e}$
2. Modelo sem ruído: $\mathbf{X} = \mathrm{f}(\mathbf{S})$
3. Modelo linear e estacionário: $\mathbf{X} = \mathrm{f}(\mathbf{S}) = \mathbf{AS}$
4. Modelo temporal ou indexado: $\mathbf{X}(t) = \mathbf{AS}(t)$
   (ex. sinais de áudio/música, EEG, MEG)

Em geral,

- $\{\mathbf{s}_j\}$, $j = 1, \ldots, m$, são variáveis latentes independentes e não normais (ou no máximo com uma componente normal),
- $\mathrm{f} : \Re^m \to \Re^p$,
- $\mathrm{E}(\mathbf{e}) = \mathbf{0}$,
- $\mathrm{E}(\mathbf{S}) = \mathbf{0}$,
- $\mathrm{Cov}(\mathbf{S}) = \mathbf{I}_m$.

Identificabilidade do Modelo

A identificabilidade do modelo sem ruído pode ser assegurada se

1. Todas a componentes independentes $\mathbf{s}_j$, com possível exceção de uma componente, são não gaussianas;
2. O número m de misturas observadas for igual ou superior ao número k de componentes independentes;
3. A matriz $\mathbf{A}$ for de posto completo.

Pré-processamento

- Consiste em centralizar e "branquear" as variáveis.
- Seja $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^\top$, com $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ e $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Sigma}$,
    - obter a SVD de $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$
    - Calcular $\mathbf{X} \leftarrow \mathbf{D}^{-1/2}\mathbf{U}^\top(\mathbf{X} - \boldsymbol{\mu})$, que corresponde a obter a PCA dos dados centralizados e fazer cada PC com variância unitária.

Aplicações

- Separação cega de fontes: EEG, ECG, MEG, Processamento de sons e imagens;
- Extração de características;
- Estimação de densidades;
- Regressão;
- Estudos em economia, psicologia, ciências sociais.

Características adicionais

- A ordem das ICs não pode ser determinada: $\mathbf{X} = \mathbf{AS} = \mathbf{AP}^{-1}\mathbf{Ps}$.
- A variância e o sinal das ICs não podem ser determinados:

  $\mathbf{X} = \sum_j \left(\frac{1}{\lambda}\mathrm{a}_j\right)(\mathrm{s}_j\lambda)$

  Estipula-se, portanto, variâncias unitárias às ICs.
- Multiplicar uma IC por -1 não provoca mudanças em $\mathbf{X} = \mathbf{AS}$.

Estimação

Método ICA = Função Objetivo (contraste) + Algoritmo de Otimização

Funções Objetivo podem estimar todo o modelo (MULTI-UNIT CONTRAST FUNCTONS) ou uma componente por vez (ONE-UNIT CONTRAST FUNCTIONS)

- Máxima verossimilhança
- Maximização da não-gaussianidade
  (a) Curtose: $\mathrm{k} = E(x^4) - 3[E(x^2)]^2$
  (b) Negentropia: $\mathrm{J}(\mathbf{x}) = \mathrm{H}(\mathbf{x}_{gaussiana}) - \mathrm{H}(\mathbf{x})$, sendo $\mathrm{H}(\mathbf{x}) = -\int \mathrm{f}(\mathbf{x}) \log \mathrm{f}(\mathbf{x}) dx$ a Entropia Diferencial.
- Cancelamento de correlações não lineares
- Critério de PCA não linear
- Cumulantes de ordem superior

  Cumulantes são constantes da expansão de $\log \hat{\mathrm{f}}(t)$ numa Série de Taylor, em que $\hat{\mathrm{f}}(t) = \mathrm{E}\{\exp(itx)\}$ é a função característica de uma variável aleatória (centralizada) x. A expansão será então,

$$\log \hat{\mathrm{f}}(t) = k_1(it) + k_2(it)^2/2 + \ldots + k(it)^2/r! + \ldots$$

  e k_r são as cumulantes. Exemplos: $k_1 = \mathrm{E}(x)$; $k_2 = \mathrm{E}(x^2)$; $k_3 = \mathrm{E}(x^3)$. A curtose é uma cumulante de quarta ordem.

Estimação

Método ICA = Função Objetivo (contraste) + Algoritmo de Otimização

Algoritmos indicam como a função objetivo escolhida será otimizada.

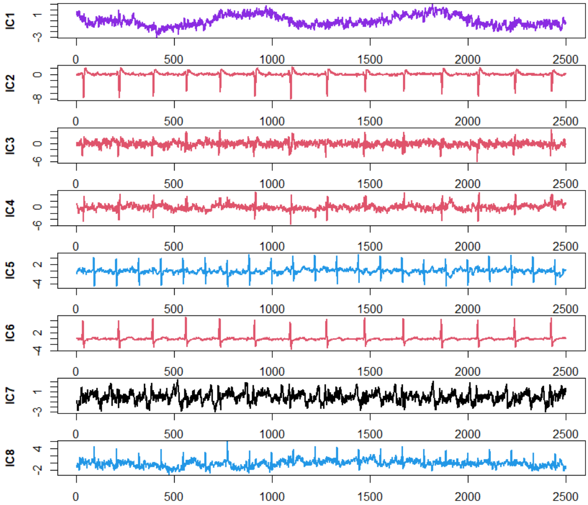
- FastICA[1]: algoritmo neural (paralelo e distribuído)
  Observação: ICASSO é um algoritmo para investigar a estabilidade das estimativas obtidas por FastICA.
- Maximização da verossimilhança
- Maximização da negentropia
- Ortogonalização de vetores

[1] Ver Izenman (2008).

Exemplo: separação de ritmos cardíacos de mãe e bebê.

Ritmos cardíacos obtidos de uma gestante (Izenman, 2008).

Exemplo: utilizando algoritmo FastICA no R



Curvas **vermelhas** (IC2, IC3, IC4, IC6) refletem o ritmo cardíaco da mãe;
Curvas **azuis** (IC5, IC8) refletem o ritmo cardíaco acelerado do feto;
Curva **roxa** (IC1) provavelmente reflete um componente respiratório;
curva **preta** (IC7) reflete ruído.

Referências

- Hyvärinen., A. (1999) Survey on Independent Component Analysis. *Neural Computing Surveys*, 2, 94-128.
- Izenman, A.J. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.
- Stone, J.V. (2004) *Independent Component Analysis. A Tutorial Introduction*. The MIT Press.
- Kock, I. (2014) *Analysis of Multivariate High-Dimensional Data*. Cambridge.