Teoria da Informação

Charles Casimiro Cavalcante

charles@gtel.ufc.br

Grupo de Pesquisa em Telecomunicações Sem Fio – GTEL
Programa de Pós-Graduação em Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://www.gtel.ufc.br/∼charles

"A principal função de um sistema de comunicação é reproduzir, exatamente ou de forma aproximada, uma informação proveniente de outro ponto diferente."

Claude Shannon, 1948

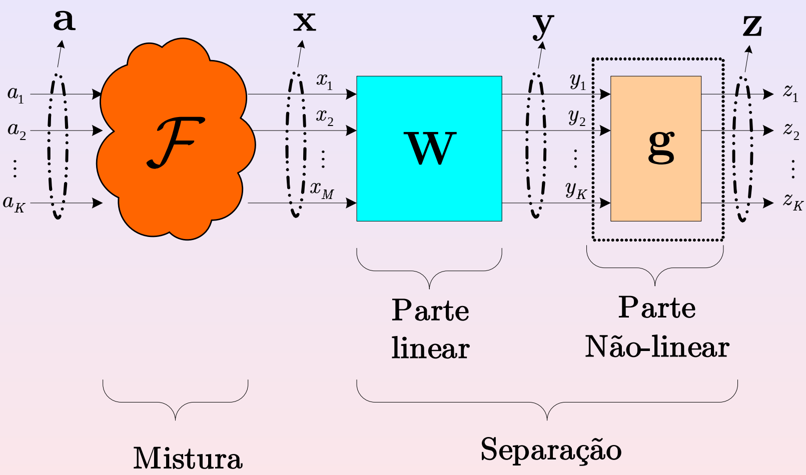
Conteúdo do curso

1. Revisão de probabilidade
2. Informação e Entropia
3. Codificação de fontes
4. Codificação e capacidade de canal
5. Complexidade de Kolmogorov
6. Funções de otimização
7. *Independent Component Analysis*

Parte VII

Independent Component Analysis

Modelo geral



Modelagem matemática

Modelo

$$\mathbf{x}(n) = \mathcal{F}\left(\mathbf{a}(n), \mathbf{v}(n), n\right) \longleftarrow \text{Mapeamento}$$

$$\mathbf{a}(n) = \begin{bmatrix} a_1(n) & a_2(n) & \cdots & a_K(n) \end{bmatrix}^T \longleftarrow K \text{ fontes}$$

$$\mathbf{v}(n) = \begin{bmatrix} v_1(n) & v_2(n) & \cdots & v_V(n) \end{bmatrix}^T \longleftarrow V \text{ sinais de ruído}$$

$$\mathbf{x}(n) = \begin{bmatrix} x_1(n) & x_2(n) & \cdots & x_M(n) \end{bmatrix}^T \longleftarrow M \text{ sensores.}$$

Considerações usuais

- $\mathcal{F}$ é linear e invariante no tempo
- Fontes mutuamente independentes e independentes do ruído
- $V = M$
- $M \geq K$ (mais sensores que fontes no mínimo)

Resultando...

Mistura

$$\mathbf{x}(n) = \mathbf{H}\mathbf{a}(n) + \mathbf{v}(n)$$

Separação

$$\mathbf{y}(n) = \mathbf{W}^H\mathbf{x}(n) = \widehat{\mathbf{a}}(n)$$

Características

1. Indeterminação em relação a permutação e escalonamento

   $$\mathbf{y}(n) = \mathbf{P}\mathbf{D}\mathbf{a}(n)$$

   $\mathbf{P}$ é uma matriz de permutação de ordem $K \times K$ e $\mathbf{D}$ é uma matriz diagonal e inversível de ordem $K \times K$.
2. Possível inserção de não-linearidade após $\mathbf{W}$

Equações: 2 fontes e 2 sensores

$$\boxed{\mathbf{x} = \mathbf{Ha}}$$

$$\begin{cases} x_1 = h_{11}a_1 + h_{12}a_2 \\ x_2 = h_{21}a_1 + h_{22}a_2 \end{cases}$$

Duas (K) variáveis e duas (M) equações!

Separando as fontes...

- Caso sem ruído: *separar é possível se* $\mathbf{W} = \mathbf{H}^{-1}$

Questão 1

Como identificar $\mathbf{H}$ para projetar $\mathbf{W}$?

Questão 2

Quais (e quantas) estatísticas são necessárias para prover a separação?

Resposta para questão 1: Como identificar $\mathbf{H}$?

- Observando a matriz de autocorrelação do vetor de sinais recebidos:

$$\mathbf{R}_x = \mathbb{E}\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\} = \mathbf{H}\mathbf{R}_a\mathbf{H}^T = \mathbf{H}\mathbf{H}^T,$$

- Observar que $\mathbf{H}\mathbf{Q}^T$, em que $\mathbf{Q}$ é uma matriz ortogonal também soluciona equação
- Conclui-se que: $\mathbf{H} = \mathbf{R}_x^{\frac{1}{2}}$
- Extração de raiz quadrada de matrizes: através de decomposição em valores singulares (SVD, Singular Value Decomposition)

Identificando $\mathbf{H}$: possibilidades

- Escrevendo
$$\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\mathcal{V}}^T,$$
em que $\mathbf{U}$ e $\boldsymbol{\mathcal{V}}$ são matrizes retangulares de ordem $K \times M$, tais que $\mathbf{U}\mathbf{U}^T = \boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^T = \mathbf{I}_M$ e $\mathbf{U}^T\mathbf{U} = \boldsymbol{\mathcal{V}}^T\boldsymbol{\mathcal{V}} = \mathbf{I}_K$
- Então...
$$\begin{aligned}\mathbf{R}_x &= \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T.\end{aligned}$$
- Resultado: através da matriz de autocorrelação (estatística de ordem 2) só é possível identificar as matrizes $\mathbf{U}$ e $\boldsymbol{\Lambda}$
- Estatísticas de ordem 2 (SOS, Second Order Statistics) não resolvem o problema por completo
- **Resposta parcial da questão 2: estatísticas necessárias para separação!**

Máximo possível com SOS

Processamento

Projeção dos dados na direção da inversa da matriz de autocorrelação: $\mathbf{T} = \mathbf{R}_x^{-\frac{1}{2}}$

- Assim, tem-se o seguinte conjunto de dados:

$$\bar{\mathbf{x}}(n) = \mathbf{T}\mathbf{x}(n)$$

de tal forma que

$$\boxed{\begin{aligned} \mathbf{R}_{\bar{x}} &= \mathbf{T}\mathbf{R}_x\mathbf{T}^T \\ &= \mathbf{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{U}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U} \\ &= \mathbf{I}_K \end{aligned}}$$

- **Branqueamento!** - em separação cega de fontes é denominado de *esferatização*

Esferatização

- Projeção dos dados na direção dos principais autovetores de $\mathbf{R}_x$
- Técnica de análise por componentes principais (PCA, Principal Component Analysis): projeção dos dados nas direções definidas pelos principais autovetores de $\mathbf{H}$
- Matriz de branqueamento, ou *transformação de Mahalanobis*, é escrita como

$$\mathbf{T} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}_x^T.$$

- Redução do problema para uma matriz de mistura ortogonal

$$\bar{\mathbf{x}}(n) = \mathbf{T}\mathbf{H}\mathbf{a}(n) = \mathbf{Q}\mathbf{a}(n)$$

em que $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$.

Encontrando a matriz $\mathcal{V}$

Considerar um atraso arbitrário ℓ para o qual, não haja duas fontes, i e j, com a mesma autocorrelação

$$\mathbb{E}\left\{a_i(n)a_i(n-\ell)\right\} \neq \mathbb{E}\left\{a_j(n)a_j(n-\ell)\right\} \qquad \forall i \neq j,$$

Diferentes densidades espectrais!

$$\boxed{\begin{aligned}
\mathbf{R}_{\bar{x}}(k) &= \mathbb{E}\left\{\bar{\mathbf{x}}(n)\bar{\mathbf{x}}^T(n-\ell)\right\} \\
&= \mathbf{THR}_a(\ell)\mathbf{HT}^T \\
&= \underbrace{\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T}_{\mathbf{T}}\underbrace{\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathcal{V}^T}_{\mathbf{H}}\mathbf{R}_a(\ell)\underbrace{\mathcal{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^T}_{\mathbf{H}^T}\underbrace{\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}}_{\mathbf{T}^T} \\
&= \mathcal{V}^T\mathbf{R}_a(\ell)\mathcal{V}.
\end{aligned}}$$

Informações restantes de $\mathbf{H}$ – II

- Fontes independentes $\Rightarrow \mathbf{R}_a(\ell)$ diagonal
- Autovalores distintos: fontes com diferentes autocorrelações
- Conclusão: *uso somente de SOS permite separar fontes com espectros diferentes!*
- Algoritmo clássico: *Algorithm for Multiple Unknown Signals Extraction* (AMUSE)
- Na prática, fontes com espectros similares (embora diferentes) não são separáveis
- Impossibilidade de separar fonte brancas ou i.i.d.

Como separar fontes com mesmo espectro?

Proposição de uma nova técnica

- Não procura-se estimar $\mathbf{H}$ para projetar $\mathbf{W}$
- Fator decisivo: suposição de independência das fontes!
- Projetar $\mathbf{W}$ de maneira que sejam obtidas fontes *o mais independentes possível* na saída do dispositivo de separação
- *Independent Component Analysis (ICA)*
- Restrição: no máximo uma fonte pode ser gaussiana
  - Teorema Central do Limite: mistura de gaussianas é gaussiana!
- Interpretação: ponto chave é a não-gaussianidade das fontes.

Questão

Como medir a não-gaussianidade das fontes e utilizar o fato de que as mesmas são estatisticamente independentes para separá-las?

Medidas de independência/gaussianidade – I

Indepedência estatística

$$p_y(\mathbf{y}) \triangleq \prod_{i=1}^{K} p_{y_i}(y_i)$$

$$\mathbb{E}\{y_1 \cdot y_2 \cdots y_K\} = \mathbb{E}\{y_1\} \cdot \mathbb{E}\{y_2\} \cdots \mathbb{E}\{y_K\}$$

Descorrelação estatística

$$\mathbb{E}\{y_1 \cdot y_2 \cdots y_K\} - \mathbb{E}\{y_1\} \cdot \mathbb{E}\{y_2\} \cdots \mathbb{E}\{y_K\} = 0.$$

Independência $\underset{\not\Leftarrow}{\Rightarrow}$ Descorrelação

Entropia e informação mútua

$$\mathcal{H}(\mathbf{x}) \triangleq -\mathbb{E}\{\ln[p_x(\mathbf{x})]\} = -\int_{-\infty}^{\infty} p_x(\mathbf{x}) \cdot \ln[p_x(\mathbf{x})]\, d\mathbf{x}$$

entropia diferencial

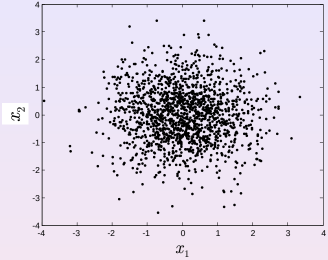
$$\mathcal{H}(\mathbf{x}|\mathbf{y}) = -\mathbb{E}\left\{\ln\left[p_{x|y}(\mathbf{x}|\mathbf{y})\right]\right\} = \int p_{x,y}(\mathbf{x},\mathbf{y}) \cdot \ln\left[p_{x|y}(\mathbf{x}|\mathbf{y})\right] d\mathbf{x}d\mathbf{y}$$

entropia condicional

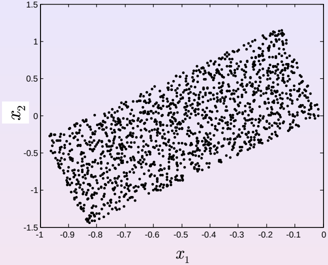
$$\mathcal{I}(\mathbf{x},\mathbf{y}) = \mathcal{H}(\mathbf{x}) - \mathcal{H}(\mathbf{x}|\mathbf{y})$$

informação mútua entre $\mathbf{x}$ e $\mathbf{y}$

E se as fontes forem gaussianas?



Mistura de fontes gaussianas



Mistura de fontes uniformes

Não há direções preferenciais!

Divergência de Kullback-Leibler (KLD)

$$\boxed{D\left(p_x(\mathbf{x})||g_x(\mathbf{x})\right) \triangleq \int_{-\infty}^{\infty} p_x(\mathbf{x}) \cdot \ln\left[\frac{p_x(\mathbf{x})}{g_x(\mathbf{x})}\right] dx}$$

em que $p(x)$ e $g(x)$ são duas funções estritamente positivas

Propriedades:

1. é sempre de valor positivo ou zero; KLD é zero para o caso específico de $p_x(\mathbf{x}) = g_x(\mathbf{x})$.
2. é invariante com relação as seguintes mudanças nas componentes do vetor $\mathbf{x}$;
   - permutação de ordem
   - escalonamento de amplitude
   - transformação monotônica não-linear

Medidas de independência/gaussianidade – IV

Usando a KLD...

$$\mathcal{I}(\mathbf{x},\mathbf{y}) = \int p_{x,y}(\mathbf{x},\mathbf{y}) \cdot \ln\left[\frac{p_{x,y}(\mathbf{x},\mathbf{y})}{p_x(\mathbf{x})\, p_y(\mathbf{y})}\right] d\mathbf{x}d\mathbf{y}$$
$$= D\left(p_{x,y}(\mathbf{x},\mathbf{y}) \,\|\, p_x(\mathbf{x})\, p_y(\mathbf{y})\right)$$

ou ainda

$$\mathcal{I}(\mathbf{y}) = D\left(p_y(\mathbf{y}) \,\|\, \widetilde{p}_y(\mathbf{y})\right)$$
$$= \int \cdots \int p_y(y_1, \ldots, y_K) \cdot \ln\left[\frac{p_y(y_1, \ldots, y_K)}{\prod_{i=1}^{K} p_{y_i}(y_i)}\right] dy_1 \cdots dy_K$$

Medidas de independência/gaussianidade – V

Finalmente...

$$\mathcal{I}(\mathbf{y}) = -\mathcal{H}(\mathbf{y}) + \sum_{i=1}^{K} \mathcal{H}(y_i),$$

o que significa que minimizar a informação mútua entre os componentes do vetor $\mathbf{y}$ significa tornar a entropia de $\mathbf{y}$ o mais próximo possível da soma de suas entropias marginais.

Negentropia - I

- Medida de "não-gaussianidade" baseado na medida de entropia diferencial
- Diferença entre a entropia da v.a. $\mathbf{y}$ e a entropia de uma v.a. $\mathbf{y}^G$ de distribuição gaussiana e com os mesmos momentos de ordem um e dois (média e variância) de $\mathbf{y}$

$$N_G(\mathbf{y}) \triangleq \mathcal{H}(\mathbf{y}^G) - \mathcal{H}(\mathbf{y}).$$

ou

$$N_G(\mathbf{y}) \triangleq D\left(p_y(\mathbf{y}) \| p_{y^G}(\mathbf{y})\right)$$

- Problema: requer conhecimento ou estimativa da densidade de probabilidade de $\mathbf{y}$!

Medidas de independência/gaussianidade – VII

Negentropia - II

- Possibilidade de expressar estatísticas necessárias para separação
- Utilizando KLD e informação mútua pode-se escrever:

$$\mathcal{I}\left(\mathbf{y}\right) = \mathcal{I}\left(\mathbf{y}^{G}\right) + \left(N_{G}\left(\mathbf{y}\right) - \sum_{i=1}^{K} N_{G}\left(y_{i}\right)\right).$$

1. Primeiro termo utiliza SOS
2. Segundo termo mede não-gaussianidade do sinal

Medidas de independência/gaussianidade – VIII

Kurtosis

- Cumulante de quarta ordem

$$\mathcal{K}\left\{y\right\} \triangleq \mathbb{E}\left\{y^4\right\} - 3 \cdot \left(\mathbb{E}\left\{y^2\right\}\right)^2 .$$

- Faixa de valores:
  - Distribuição gaussiana: $\mathcal{K}\left\{y\right\} = 0$
  - Distribuição sub-gaussiana: $\mathcal{K}\left\{y\right\} \leq 0$
  - Distribuição super-gaussiana: $\mathcal{K}\left\{y\right\} \geq 0$
- Propriedades

$$\mathcal{K}\left\{y_1 + y_2\right\} = \mathcal{K}\left\{y_1\right\} + \mathcal{K}\left\{y_2\right\}$$
$$\mathcal{K}\left\{\alpha \cdot y\right\} = \alpha^4 \cdot \mathcal{K}\left\{y\right\}$$

Funções de contraste

- Uma função $\Psi(\cdot)$, no espaço de K fdps (distintas ou não) é dita ser um *contraste* se respeita as seguintes condições:
  1. $\Psi(p_y)$ é invariante a permutações:
  $$\Psi(p_{Py}) = \Psi(p_y) \text{ para qualquer matriz de permutação } \mathbf{P}$$
  2. $\Psi(p_y)$ é invariante a mudanças de escala:
  $$\Psi(p_{Dy}) = \Psi(p_y) \text{ para qualquer matriz diagonal } \mathbf{D}$$
  3. Se $\mathbf{y}$ possui componentes independentes, então:
  $$\Psi(p_{Wy}) \leq \Psi(p_y) \text{ para qualquer matriz inversível } \mathbf{W}$$

Algumas funções de contraste

$$\Psi_{\mathsf{ICA}}\left(p_y\right) = -\mathcal{I}\left(\mathbf{y}\right)$$

ou ainda em relação à matriz de separação

$$\Psi_{\mathsf{ICA}}\left(\mathbf{W}\right) = \ln\left[\left|\det\left(\mathbf{W}\right)\right|\right] - \mathbb{E}\left\{\ln\left[\prod_{i=1}^{K} p_{y_i}\left(y_i\right)\right]\right\}$$

Aproximações da negentropia

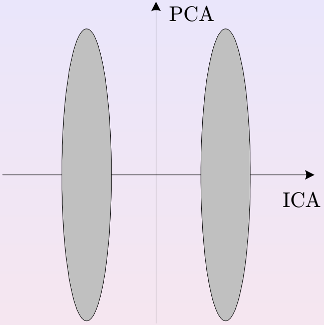
- Necessário para evitar estimações da fdp
- Momentos de ordem superior (clássica)

$$N_G(\mathbf{y}) \approx \frac{1}{12}\left[\mathbb{E}\left\{\mathbf{y}^3\right\}\right]^2 + \frac{1}{48}\left[\mathcal{K}\left\{\mathbf{y}\right\}\right]^2,$$

- Classe de aproximações (*FastICA*)

$$\boxed{\begin{gathered} N_G(\mathbf{y}) \approx \sum_{i=1}^{p} \varrho_i \cdot \left[\mathbb{E}\left\{g_i(\mathbf{y})\right\} - \mathbb{E}\left\{g_i(\nu)\right\}\right]^2 \\ g_1(u) = \frac{1}{a_1}\log\left[\cosh\left(a_1 u\right)\right] \text{ e} \\ g_2(u) = -\exp\left(-\frac{u^2}{2}\right) \end{gathered}}$$

Diferença ICA × PCA



1. PCA: projeta numa dimensão menor preservando *máxima variância* dos dados
2. ICA: projeta numa dimensão menor preservando *estrutura* dos dados

MaxEnt/InfoMax - I

- Separação é obtida quando estimativas das fontes são independentes $\Rightarrow I(\mathbf{y}) = 0$

$$\mathcal{I}(\mathbf{y}, \mathbf{x}) = \mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y}|\mathbf{x})$$

- Mapeamento é determinístico, logo $\mathcal{H}(\mathbf{y}|\mathbf{x}) = 0$

$$\mathcal{I}(\mathbf{y}, \mathbf{x}) = \mathcal{H}(\mathbf{y})$$

- Entropia de $\mathbf{y}$ não é limitada, então

$$\mathbf{z} = \mathbf{g}(\mathbf{W}\mathbf{x})$$

MaxEnt/InfoMax - II

- Funções $g_i(\cdot)$ monotonicamente crescentes, limitadas de tal forma que $g_i(-\infty) = 0$ e $g_i(\infty) = 1$
- Se $g_i(\cdot)$ for igual à função de distribuição cumulativa (fdc) da i-ésima fonte
$$p_z\left(\mathbf{z}\right) = U\left[0,1\right], \quad \text{para}$$
- Adaptação
$$\Delta\mathbf{W} \propto \left(\mathbf{W}^{-T}\right)^{-1} - 2 \cdot \tanh\left(\mathbf{W}\mathbf{x}\right)\mathbf{x}^T,$$
- Função contraste
$$\max_{\mathbf{W}} \left(\Psi_{\mathsf{InfoMax}}(\mathbf{W}) \triangleq \ln\left[\left|\mathsf{det}(\mathbf{W})\right|\right] - \mathbb{E}\left\{\ln\left[\prod_{i=1}^{K} g_i'\left(y_i\right)\right]\right\}\right).$$

Estratégias de separação cega de fontes - III

Máxima verossimilhança - I

- Entropia

$$\mathcal{H}(\mathbf{z}) = -\int p_z(\mathbf{z}) \ln \left[\frac{p_z(\mathbf{z})}{\prod_{i=1}^{K} U\left(z_i\right)} \right] d\mathbf{z} \quad = -D\left(p_z(\mathbf{z}) \| U_N(\mathbf{z})\right)$$

- Modelo paramétrico $\boldsymbol{\Theta}$

$$\mathcal{L}_Q\left(\boldsymbol{\Theta}\right) = \frac{1}{Q} \ln \left[\prod_{q=1}^{Q} p_x\left(\mathbf{x}(q) | \boldsymbol{\Theta}\right) \right] = \frac{1}{Q} \sum_{q=1}^{Q} \ln \left[p_x\left(\mathbf{x}(q) | \boldsymbol{\Theta}\right) \right].$$

Máxima verossimilhança - II

- Lei dos grandes números

$$\mathcal{L}_Q(\boldsymbol{\Theta}) \xrightarrow{Q\to\infty} \mathcal{L}(\mathbf{W}) \triangleq \int p_y(\mathbf{y}|\mathbf{W}) \ln\left[p_{\tilde{a}}(\mathbf{y})\right] d\mathbf{y} + \ln\left[\left|\det(\mathbf{W})\right|\right].$$

- Usando KLD

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= -D\left(p_x(\mathbf{y}|\mathbf{W})\,\|p_{\tilde{a}}(\mathbf{y})\right) - \mathcal{H}(\mathbf{y}|\mathbf{W}) + \ln\left[\left|\det(\mathbf{W})\right|\right] \\ &= -D\left(p_x(\mathbf{y}|\mathbf{W})\,\|p_{\tilde{a}}(\mathbf{y})\right) - \mathcal{H}(\mathbf{x})\end{aligned}.$$

- Função contraste MV

$$\Psi_{\mathsf{MV}}(\mathbf{W}) = -D\left(\mathbf{WHa}\|\tilde{\mathbf{a}}\right).$$

Máxima verossimilhança - III

- Interpretação da MV

$$-\Psi_{\mathsf{MV}}(\mathbf{W}) = -\Psi_{\mathsf{ICA}}(\mathbf{W}) + \sum_{i=1}^{K} D\left(\widetilde{z}_i \| \widetilde{a}_i\right)$$

ou seja

$$\begin{pmatrix}\text{Desvio}\\ \text{total}\end{pmatrix} = \begin{pmatrix}\text{Desvio da}\\ \text{independência}\end{pmatrix} + \begin{pmatrix}\text{Desvio}\\ \text{marginal}\end{pmatrix}.$$

e ainda

$$\Psi_{\mathsf{ICA}}(\mathbf{W}) = \max_{\widetilde{a}}\left(\Psi_{\mathsf{MV}}(\mathbf{W})\right).$$

Estratégias de separação cega de fontes - VI

Critério "universal"

$$J_{\mathsf{BSS}}(\mathbf{W}) = \ln\left[|\mathsf{det}|(\mathbf{W})\right] - \ln\left[\prod_{i=1}^{K}\phi_i(y_i)\right]$$

em que $\mathbf{y} = \mathbf{W}\mathbf{x}$ e, idealmente, as funções $\phi_i(\cdot)$ são as fdps das fontes.

Questão

Como encontrar ou estimar as fdps das fontes?

Resposta possível

Expansões polinomiais através dos cumulantes.