

Análise de Componentes Independentes (ICA)

Índice

1	Introdução	2
2	Descorrelação <i>versus</i> Independência	3
3	Análise de Componentes Independentes	9
3.1	Crítério de Mínima Informação Mútua.....	12
3.2	Não-Gaussianidade e Curtose	15
3.3	Não-Gaussianidade e Negentropia.....	18
3.4	Crítério Infomax.....	19
4	ICA e Separação Cega de Fontes.....	22
5	Referências bibliográficas	30

1 Introdução

- O estudo da metodologia conhecida como análise de componentes independentes (ICA, do inglês *independent component analysis*) é, sem dúvida, um dos tópicos mais relevantes da moderna teoria de processamento não-supervisionado de sinais. Uma das razões para o interesse em ICA é que ela pode ser vista, do ponto de vista de extração de características / redução de dimensionalidade, como uma extensão da clássica abordagem de análise de componentes principais (PCA) no sentido de incorporar uma estrutura estatística mais rica que o arcabouço de segunda ordem edificado por meio da correlação / covariância. Outro motivo é que ICA corresponde a um caminho viável para realizar separação cega de fontes (BSS, do inglês *blind source separation*), uma tarefa crucial em inúmeras aplicações de tratamento da informação.
- Do ponto de vista de nosso curso, parece mais natural apresentar a metodologia como uma extensão da abordagem de PCA, e é exatamente esse o caminho que

percorremos, deixando para uma etapa seguinte a exploração de conexões com BSS.

2 Descorrelação *versus* Independência

- A chave para que se compreendam as diferenças existentes entre ICA e PCA é atentar para o contraste entre os alcances dos termos “descorrelação” e “independência”. Basearemos nossa exposição, por uma questão de simplicidade, no caso de duas variáveis aleatórias – X e Y – embora todos os conceitos abordados possam ser estendidos para o caso geral de N variáveis.
- Recordando o que foi visto no tópico 3, duas variáveis aleatórias são descorrelacionadas quando vale:

$$E[XY] = E[X] \cdot E[Y] \quad (1)$$

- Vê-se em (1) que, quando duas variáveis aleatórias são descorrelacionadas, *a esperança do produto entre elas se iguala ao produto de suas esperanças (ou seja, médias estatísticas), ou seja, sua covariância é zero*. Como, em muitos casos práticos, trabalha-se com variáveis aleatórias de média nula, pode-se simplificar (1), que passa a ser uma condição de *ortogonalidade*:

$$E[XY] = 0 \quad (2)$$

- Quando discutimos PCA no tópico 5, tivemos a chance de mencionar que a condição de ortogonalidade da construção das direções de projeção se vinculava a uma ortogonalidade dos componentes principais, ou seja, à validade da idéia expressa em (2). Como, no contexto de PCA, sempre lidamos com dados de média nula, isso significa que os componentes principais também obedecem à condição (1), ou seja, *os componentes principais são descorrelacionados*.

- Temos portanto, uma visão de PCA como sendo uma metodologia para gerar componentes de máxima variância (ou menor erro quadrático médio de reconstrução) que respeitem (1) e (2).
- Se não buscarmos realizar compressão, ou seja, se realizarmos PCA com um número de componentes igual à dimensão do vetor de dados, obteremos uma versão *branqueada* (*whitened*) desses dados, ou seja, uma versão dos dados com elementos descorrelacionados. A associação entre os termos “branco” e “descorrelacionado” vem da teoria clássica de processos estocásticos, e não deve causar estranheza ao leitor. Quando se analisam em frequência de processos temporalmente “brancos”, vincula-se “resposta plana” no sentido de densidade espectral de potência com o conteúdo da luz branca (que possui várias frequências que podemos separar, por exemplo, com um prisma).
- O que dissemos acima significa que a metodologia de PCA pode ser usada também para branquear dados, ou seja, *descorrelacioná-los*.

- Embora tenhamos, às vezes, o impulso de pensar que sinais descorrelacionados “não têm nenhuma relação entre si”, isso não é correto. Por exemplo, o fato de $E[XY]$ ser igual a $E[X]E[Y]$ nada nos diz, em geral, sobre o valor de $E[X^2Y^3]$ ou, digamos, $E[\ln(X)Y^4]$. É por isso que temos de saber que, via de regra, a ocorrência de descorrelação “linear” (no sentido de (1)) não implica ocorrência de descorrelação “não-linear” (e.g. $E[X^2Y^3] = E[X^2]E[Y^3]$).
- Interessantemente, no caso de variáveis aleatórias gaussianas, é possível mostrar que a condição apresentada em (1) é extremamente forte, pois, se houver descorrelação desse tipo, haverá também descorrelação *em todos os possíveis contextos não-lineares*. No entanto, isso é algo peculiar à gaussiana, que é totalmente definida por média e covariância, ou seja, pode ser perfeitamente descrita com estatísticas que não excedem a ordem da correlação linear.

- No caso geral, para garantir plena desconexão não-linear, é preciso trabalhar com outro conceito, o de independência. Conforme discutimos no capítulo 3, duas variáveis aleatórias contínuas¹ são independentes se e somente se

$$p_{XY}(x,y) = p_X(x)p_Y(y) \quad (3)$$

ou seja, se a densidade de probabilidade conjunta for o produto das densidades marginais.

- Se vale a expressão (3), vale sempre que $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$. Isso não é difícil de mostrar. Começemos com:

$$E[f(X)g(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(y)p_{XY}(x,y)dxdy \quad (4)$$

¹ Neste tópico, trabalharemos com variáveis aleatórias contínuas, embora as noções expostas, *mutatis mutandis*, sejam aplicáveis ao caso discreto.

Se valer (3), podemos reescrever (4) da seguinte forma:

$$E[f(X)g(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(y)p_X(x)p_Y(y)dxdy = \int_{-\infty}^{\infty} f(x)p_X(x)dx \int_{-\infty}^{\infty} g(y)p_Y(y)dy \quad (5)$$

ou seja,

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)] \quad (6)$$

- Portanto, independência estatística acarreta toda e qualquer desconexão não-linear da forma exposta em (6), o que mostra a completude estatística dessa medida. Se tivermos condição de obter componentes independentes, teremos obtido componentes linear e não-linearmente desconexados. Naturalmente, há um preço: para quantificar independência nos termos de (3), é preciso trabalhar

diretamente com as densidades de probabilidade das variáveis envolvidas, não sendo possível lidar, a princípio, apenas com estatísticas de determinadas ordens.

- Conforme já apontamos acima, o caso de variáveis aleatórias gaussianas é especial, pois, nesse caso, *descorrelação linear implica independência*, ou seja, (1) implica (3).

3 Análise de Componentes Independentes

- Conforme vimos na seção anterior, o uso da independência estatística leva a uma condição bastante mais abrangente, em geral, que a condição de descorrelação. No contexto de compressão de dados, surge, de forma natural, a idéia de estender a abordagem de PCA no sentido de buscar projeções que garantam independência: isso leva diretamente à idéia de *análise de componentes independentes (ICA)*.
- Tomemos o seguinte ponto de partida: temos um conjunto de dados N -dimensionais $\mathbf{x}(n)$ que desejamos analisar, e, do ponto de vista estatístico,

consideramos que esse conjunto é gerado por um conjunto M -dimensional de componentes independentes $\mathbf{s}(n)$.

- A relação entre $\mathbf{x}(n)$ e $\mathbf{s}(n)$ é, a princípio, genérica:

$$\mathbf{x}(n) = F[\mathbf{s}(n)] \quad (7)$$

sendo $F[\cdot]$ um mapeamento potencialmente não-linear, com memória etc. Pode ainda haver um ruído aditivo da forma $\mathbf{r}(n)$ que se sobrepõe à aplicação do mapeamento.

- Por uma questão de tratabilidade, no caso mais clássico, supõe-se que o modelo que gera os dados é linear e sem memória, e supõe-se ainda que não há ruído. Sob tais condições, (7) se torna:

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) \quad (8)$$

sendo \mathbf{A} uma matriz de mistura $N \times M$.

- Poderíamos então pensar em uma primeira perspectiva do problema de ICA: tendo os dados $\mathbf{x}(n)$ em mãos, como podemos obter a matriz \mathbf{A} que gera o modelo? Perceba que se trata, fundamentalmente, de um problema de identificação *não-supervisionada*, já que não dispomos de amostras de $\mathbf{s}(n)$ para guiar a estimação. A principal informação disponível é o fato de que os componentes do vetor $\mathbf{s}(n)$ são estatisticamente independentes.
- Outra possibilidade, mais próxima da visão de PCA que construímos, é obter projeções lineares dos dados que se aproximem ao máximo das componentes independentes. Nesse caso, tem-se que:

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) \quad (9)$$

sendo \mathbf{W} a matriz de projeção $M \times N$ e $\mathbf{y}(n)$ um vetor que deve possuir elementos tão mutuamente independentes quanto possível. Note que isso nos leva, mais uma vez, ao familiar cenário de compressão.

- Estabelecido o problema, é chegada a hora de passarmos a uma análise de algumas abordagens clássicas para a realização de ICA. Nossa exposição se baseará fundamentalmente na segunda das formulações de ICA, ou seja, avaliaremos modos de obter a matriz \mathbf{W} em (9) que leve à obtenção dos componentes independentes de uma massa de dados.

3.1 Critério de Mínima Informação Mútua

- Conforme discutimos no tópico 9, a informação mútua é uma grandeza que quantifica o grau de dependência estatística entre variáveis aleatórias. Essa grandeza é sempre não-negativa e vale zero se e somente se vigora uma condição de independência.
- Essa propriedade indica de modo bastante direto que a informação mútua pode fazer o papel de função custo para ajuste da matriz \mathbf{W} . Nesse caso, a idéia é

encontrar a matriz que minimiza $I(\mathbf{y})^2$, ou seja, que produz a menor informação mútua possível entre as projeções. Caso seja possível atingir uma situação em que os componentes gerados sejam estritamente independentes, teremos $I(\mathbf{y}) = 0$. Caso essa situação não seja atingível, teremos, ao menos, uma situação tão próxima do ideal quanto possível.

- É possível mostrar que, sendo válida a expressão (9), a informação mútua é dada por [Suyama, 2007]:

$$I(\mathbf{y}) = \sum_{i=1}^N H(y_i) - H(\mathbf{y}) = \sum_{i=1}^N H(y_i) - H(\mathbf{x}) - \ln |\det(\mathbf{W})| \quad (10)$$

sendo $H(y_i)$, $i = 1, \dots, N$, a i -ésima entropia marginal do vetor de componentes. Como $H(\mathbf{x})$ não depende de \mathbf{W} , minimizar a informação mútua equivale a minimizar a função custo a seguir:

² Em alguns casos, abandonaremos os índices temporais dos vetores para tornar a notação menos carregada. Por esse motivo, escrevemos $I(\mathbf{y})$ e não $I[\mathbf{y}(n)]$.

$$J_{IM}(\mathbf{W}) = \sum_{i=1}^N H(y_i) - \ln |\det(\mathbf{W})| \quad (11)$$

- Trabalhar com a informação mútua impõe um problema bastante relevante de estimação. No caso da função apresentada em (11), o ponto com que se deve lidar é a obtenção das entropias marginais $H(y_i)$. Há vários caminhos para isso, sendo uma possibilidade interessante o uso de estatísticas de ordem [Duarte, 2006].
- O problema de otimização exposto em (11) pode ser complexo, havendo o potencial de existência de mínimos locais e uma significativa dificuldade no que diz respeito à manipulação da função custo. Isso explica por que meta-heurísticas vêm sendo consideradas para resolvê-lo em alguns casos de interesse [Suyama, 2007].

3.2 Não-Gaussianidade e Curtose

- Embora a informação mútua seja uma forma inegavelmente sólida de construir um critério para ICA, é muito importante ressaltar que há outros caminhos de interesse teórico e prático. Um deles é usar a noção de *não-gaussianidade* para obter as componentes independentes.
- A base para o uso de não-gaussianidade em ICA linear é o *teorema central do limite* [Papoulis, 1991]. Fundamentalmente, esse teorema diz que, sob certas condições, a soma de variáveis aleatórias independentes e identicamente distribuídas tende a possuir uma densidade de probabilidade gaussiana. Ora, se supusermos que vale o modelo generativo (8), ou seja, que os dados são compostos de somas ponderadas de componentes independentes, podemos concluir, à luz do teorema, que *os elementos de $\mathbf{x}(n)$ tendem a ser mais gaussianos que os elementos de $\mathbf{s}(n)$* . Portanto, tendo em mãos os dados $\mathbf{x}(n)$, podemos construir a matriz buscando projeções de *máxima não-gaussianidade*. Para que as

projeções sejam realmente diversificadas, é importante forçar uma restrição adicional de ortogonalidade entre direções de projeção (num espírito similar ao que foi feito em PCA quando discutimos a idéia de projeções de máxima variância). Isso pode ser feito iterativamente, por exemplo, com o auxílio da metodologia de Gram-Schmidt [Hyvarinen et al., 2001].

- Uma forma de quantificar o grau de não-gaussianidade de uma variável aleatória Y é utilizar a *curtose*, definida como:

$$\kappa(Y) = E[Y^4] - 3E^2[Y^2] \quad (12)$$

Como a curtose vale zero para variáveis aleatórias gaussianas e é não-nula para praticamente todas as outras variáveis aleatórias, uma forma de obtermos uma projeção de máxima não-gaussianidade é buscar uma direção de projeção \mathbf{w}_i que satisfaça o seguinte critério:

$$\mathbf{w}_i = \arg \max |\kappa(y_i)|, \quad (13)$$

sendo $y_i = \mathbf{w}_i^T \mathbf{x}$. Vale registrar que variáveis aleatórias com curtose negativa são chamadas de variáveis *subgaussianas* e variáveis aleatórias com curtose positiva são chamadas de *supergaussianas*.

- Caso se deseje obter vários componentes independentes, pode-se aplicar um procedimento de estimação “um a um” com diversas aplicações de (13) ou buscar a otimização de (13) para direções, como já foi dito, sujeitas a uma restrição de ortogonalidade.
- Uma dificuldade de lidar com a curtose é que a estimação de (12) com médias temporais é bastante sensível a *outliers*, já que o termo de quarta potência pode engendrar valores bastante elevados.

3.3 Não-Gaussianidade e Negentropia

- Outra forma de quantificar o grau de não-gaussianidade faz uso de um resultado que expusemos no tópico 9. Nesse tópico, tivemos a chance de ver que, para uma média μ e uma variância σ^2 dadas e fixas, a densidade de probabilidade de máxima entropia diferencial será a gaussiana com esses momentos.
- Essa propriedade dá origem a uma função custo de *negentropia* para a obtenção da projeção de máxima não-gaussianidade:

$$J_{\text{NEG}}(\mathbf{w}) = H(y_{\text{gauss}}) - H(y) \quad (14)$$

sendo $y = \mathbf{w}^T \mathbf{x}$ a componente obtida e y_{gauss} uma variável aleatória gaussiana de mesma média e variância que y . A referida propriedade da entropia diferencial garante que $J_{\text{NEG}}(\mathbf{w})$ é sempre não-negativa e nula apenas se y for gaussiana.

Portanto, se maximizarmos a negentropia, tenderemos, mais uma vez, a obter um componente independente se vigorar o modelo descrito em (8).

- O uso da negentropia, embora sólido do ponto de vista teórico, exige, analogamente ao que vimos para a informação mútua, estimação de entropia. Para evitar a realização dessa tarefa, alguns autores propõem o uso de expressões aproximadas obtidas, por exemplo, por meio da expansão de densidades de probabilidade de Gram-Charlier [Hyvarinen et al., 2001]. Uma dessas versões aproximadas fundamenta um dos algoritmos iterativos mais usados para realizar ICA, o FastICA [Hyvarinen et al., 2001].

3.4 Critério Infomax

- O critério Infomax teve origem no importante trabalho de Bell e Sejnowski [Bell e Sejnowski, 1995], embora seja fundamental citar o nome de Linsker, que, ainda

nos anos 80, formulou a idéia de máxima preservação de informação numa rede neural [Haykin, 1994].

- A base para exposição do critério Infomax é uma rede neural com arquitetura como a mostrada na Fig. 1. Perceba que, de certa forma, a rede corresponde a uma estrutura que gera projeções não-lineares correspondentes à aplicação de mapeamentos $g_i(\cdot)$ – tipicamente monotonicamente crescentes com $g_i(-\infty) = 0$ e $g_i(+\infty) = 1$ – a projeções lineares obtidas por meio de uma matriz \mathbf{W} .

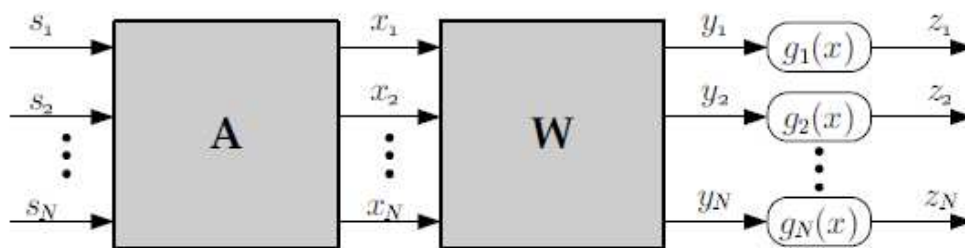


Figura 1: Arquitetura para Aplicação do Método Infomax (com $M = N$)

- O princípio de ajuste dos pesos da rede é obter uma condição de máxima transferência de informação da camada de entrada para a camada de saída. Perceba desde já o leitor o quanto isso é relevante se considerarmos que a rede atua no contexto de compressão.
- Para obter os pesos sinápticos, Bell e Sejnowski mostram que a maximização da informação mútua entre os sinais da camada de entrada e os sinais da saída equivale, no caso em que $M = N$, a obter a condição de máxima entropia da saída. Embora pareça estranho lidar com $M = N$, trata-se de uma situação bastante usual no contexto de separação cega de fontes, na qual a meta não é realizar compressão, mas sim obter os componentes independentes.
- Na condição $M = N$, em que é possível inverter perfeitamente o mapeamento realizado em (8), mostra-se que a estrutura da Fig. 1 recupera, na condição de máxima entropia da saída, exatamente os componentes independentes, ou seja, realiza ICA.

4 ICA e Separação Cega de Fontes

- Embora o modelo exposto em (8) possa ser apenas visto como um modelo generativo baseado em fatores independentes, é também possível interpretá-lo como um modelo efetivo de dados colhidos por vários sensores. Em tal situação, $\mathbf{s}(n)$ corresponderia a um vetor de *fontes* independentes, sinais de informação linearmente combinados segundo uma matriz \mathbf{A} e captados por sensores, formando um vetor de misturas $\mathbf{x}(n)$. O problema de separação cega de fontes (BSS, do inglês *blind source separation*) seria então o de obter estimativas sólidas dos sinais que compõem $\mathbf{s}(n)$ tendo por base as misturas que formam $\mathbf{x}(n)$ e conhecimento estatístico de cunho genérico, como o conhecimento acerca da hipótese de independência entre as fontes.
- Um exemplo clássico para ilustrar o problema de BSS é o chamado *cocktail party problem* [Hyvarinen et al., 2001], que é ilustrado na Fig. 2. Basicamente, tem-se um conjunto de microfones que captam versões misturadas de vários sons do

ambiente, e, assim como ocorre em nosso cérebro, deseja-se contar com a capacidade de separar cada um desses sons.

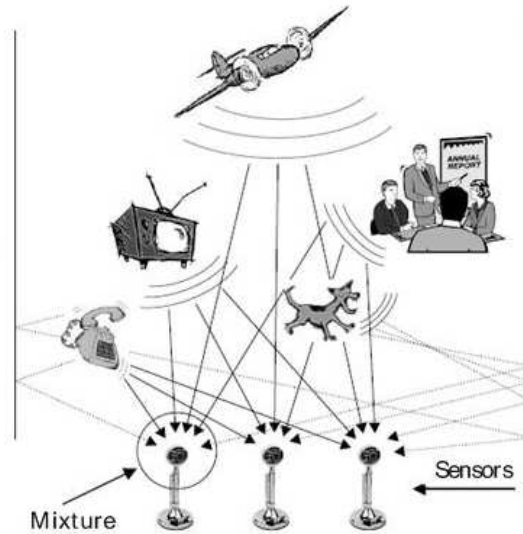


Figura 2: Ilustração do *Cocktail Party Problem*³

³ Os autores são gratos ao Prof. Ricardo Suyama, da UFABC, por esta figura e pelas figuras seguintes.

- A temática de BSS, atualmente, é uma das que mais atraem a atenção dos pesquisadores da área de processamento de sinais, o que se justifica, em grande medida, por sua significativa aplicabilidade. Essa temática abrange tarefas tão diversas quanto tratamento de sinais de ECG, EEG e fMRI, detecção multiusuário em comunicações sem fio, análise de sinais provenientes de sensores químicos, extração de informação de sinais de astrofísica, modelamento de processos do córtex visual humano, tratamento de sinais de áudio etc.
- Embora seja possível lidar até certo ponto com separação de fontes no caso em que há menos sensores que fontes, uma hipótese convencional é a de que deve haver um número de sensores maior que ou igual ao número de fontes. Do ponto de vista deste tópico, restringir-nos-emos ao caso padrão e particularmente tratável em que há igualdade entre o número de misturas e fontes.

- Em tal caso, é possível, a princípio, achar uma matriz de separação \mathbf{W} que, aplicada no contexto delineado pela equação (9), conduza à recuperação do vetor de fontes $\mathbf{s}(n)$. Aliás, do ponto de vista de BSS, são consideradas perfeitamente aceitáveis duas ambigüidades que vale a pena discutir. A primeira delas é que não é preciso recuperar no vetor $\mathbf{y}(n)$ as fontes na mesma ordem em que elas se encontram no vetor original $\mathbf{s}(n)$. A segunda é que a amplitude das fontes recuperadas não precisa ser igual à amplitude das fontes originais, ou seja, são admissíveis fatores de escala. A junção de ambas dá origem à seguinte condição ideal de separação:

$$\mathbf{y}(n) = \mathbf{P}\mathbf{D}\mathbf{s}(n) \quad (15)$$

sendo \mathbf{P} uma matriz de permutação e \mathbf{D} uma matriz diagonal. Se juntarmos a informação trazida pelas equações (8), (9) e (15), podemos reescrever a condição da seguinte maneira:

$$\mathbf{W}\mathbf{A} = \mathbf{P}\mathbf{D} \quad (16)$$

- É importante ressaltar que, no cenário em que há igual número de fontes e misturas, existem trabalhos que fundamentam solidamente os critérios de informação mútua, curtose e negentropia expostos anteriormente, ou seja, que revelam que eles, se devidamente otimizados, geram soluções que obedecem à condição (16). Isso é muito importante, pois revela que a metodologia de ICA é, de fato, uma ferramenta sólida para realizar separação cega de fontes [Comon, 1994].
- Um ponto crucial é que todos os critérios de ICA se baseiam, explicita ou implicitamente, em estatísticas de ordem superior a dois. De fato, não é possível obter de modo não-supervisionado condições ideais de separação usando PCA, ou seja, estatísticas de ordem dois. Essa limitação é ilustrada na Fig. 3.

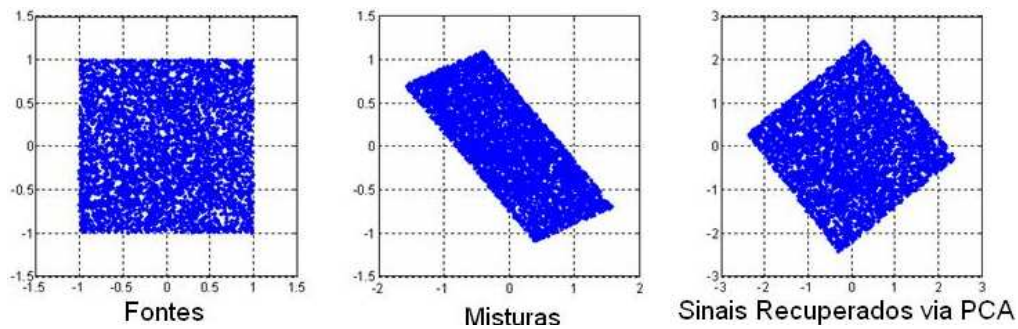


Figura 3: Ilustração das Limitações de PCA em BSS

- Na figura, as fontes são independentes e possuem densidade uniforme definida entre -1 e 1 . Após a aplicação da matriz \mathbf{A} de mistura, os sinais passam a ser dependentes e correlacionados. Aplicando PCA com $M = N$, obtemos um conjunto de sinais descorrelacionados. No entanto, como mostra a figura, *os sinais obtidos não são independentes!*

- Uma forma intuitiva de ver isso é notar que o fato de o sinal obtido via PCA presente no eixo x da Fig. 3 possuir valor ligeiramente maior que dois implica que o outro sinal obtido terá um valor ligeiramente menor que zero, o que indica que um sinal traz informação sobre o outro. É perceptível que a metodologia de PCA, ao realizar o branqueamento, *resolveu o problema a menos de uma rotação da distribuição dos sinais*. Para obter a rotação que conduz à distribuição original, *é preciso lançar mão de estatísticas de ordem superior, exatamente como é feito em ICA, que é capaz de resolver o problema como um todo!*
- Uma vez que independência implica descorrelação, uma metodologia de branqueamento como PCA pode ser vista como um passo de pré-processamento das misturas, “facilitando o trabalho” da técnica de ICA escolhida, que precisará apenas obter uma matriz de rotação apropriada. Essa estratégia “em duas partes”, muito usual em BSS, é ilustrada, para o mesmo exemplo de duas fontes uniformes, na Fig. 4.

- A impossibilidade de realizar ICA apenas com estatísticas de segunda ordem explica por que *a metodologia não pode ser usada em BSS se mais de uma fonte for gaussiana*, já que essa distribuição é plenamente caracterizada por média e covariância (ou correlação). No caso da gaussiana, a recuperação de independência equivalerá a uma condição de perfeito branqueamento, não permitindo que a ambigüidade de rotação exposta seja devidamente resolvida.

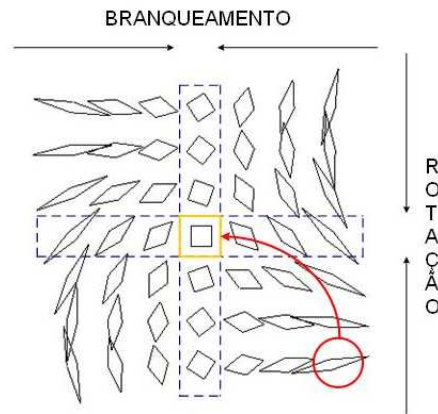


Figura 4: Efeitos dos Processos de Branqueamento e Rotação

5 Referências bibliográficas

- A. J. BELL & T. J. SEJNOWSKI, “An Information-Maximisation Approach to Blind Separation and Blind Deconvolution”, *Neural Computation*, Vol. 7, No. 6, pp. 1129-1159, 1995.
- P. COMON, “Independent Component Analysis: a New Concept?”, *Signal Processing*, Vol. 36, No. 3, pp. 287-314, 1994.
- L. T. DUARTE, “Um Estudo sobre Separação Cega de Fontes e Contribuições ao Caso de Misturas Não-Lineares”, Tese de Mestrado, FEEC/Unicamp, 2006.
- S. HAYKIN, “Neural Networks: a Comprehensive Foundation”, MacMillan, 1994.
- A. HYVARINEN, J. KARHUNEN & E. OJA, “Independent Component Analysis”, Wiley Interscience, 2001.
- A. PAPOULIS, “Probability, Random Variables and Stochastic Processes”, McGraw-Hill, 1991.
- R. SUYAMA, “Proposta de Métodos de Separação Cega de Fontes para Misturas Convolutivas e Não-Lineares”, Tese de Doutorado, FEEC/Unicamp, 2007.