

ANÁLISE EM COMPONENTES INDEPENDENTES NA SEPARAÇÃO CEGA DE FONTES COINCIDENTES NO ESPECTRO

José Maria Côrte-Real da Costa Pereira* e André Teixeira Puga**

* Direcção de Engenharia e Operações de Rede, Vodafone Portugal
e-mail: jose.pereira@vodafone.com

** Secção de Matemática DEMEGI-FEUP & INESC-Porto
Campus da FEUP
Rua Dr. Roberto Frias S/N
4200-465 Porto Portugal.
e-mail: apuga@fe.up.pt, <http://www.fe.up.pt>

Palavras-chave: Entropia de Shannon, Negentropia, Análise em Componentes Independentes, Separação Cega de Fontes, Electrocardiograma Fetal, Angiogénese.

Resumo. *Este artigo recorda os fundamentos da teoria da informação subjacentes à técnica de Análise em Componentes Independentes. Posteriormente, apresenta duas aplicações de separação cega de fontes oriundas do contexto biomédico: a extracção do electrocardiograma fetal e a identificação de microvasos em imagem de patologia médica associada à angiogénese do cancro da mama. Dada a sobreposição espectral das fontes registada nas aplicações aqui consideradas, a eventual utilização de técnicas convencionais de filtragem de Fourier seria completamente desadequada.*

1. INTRODUÇÃO

O problema aqui abordado consiste no processo de recuperar um conjunto de (um ou mais) sinais, com origem em fontes independentes, que foram misturados linearmente. Este problema é particularmente interessante quando existe sobreposição espectral das fontes. Com efeito, nestes casos, as técnicas de filtragem convencional de Fourier são incapazes de separar as fontes com espectros coincidentes.

São precisamente estes problemas que motivaram o aparecimento de um método como a Análise em Componentes Independentes (ACI). No entanto, deve salientar-se que a ACI é um método suficientemente genérico para ser aplicado também em casos onde não existe coincidência espectral dos sinais envolvidos. Aliás, a palavra “cega” no título do presente artigo significa que a ACI não assume nada, ou praticamente nada, sobre os sinais que definem o problema.

No âmbito do presente trabalho foi implementado um método clássico de ACI. Posteriormente, foi testada a aplicação do referido método de ACI a dois casos práticos distintos. No primeiro caso o problema está relacionado com as leituras electro-cardiográficas de um feto, onde a mistura de sinais envolvida, para além do ruído subjacente ao processo de captura, é essencialmente entre o electrocardiograma (ECG) do feto e da sua mãe. No segundo caso o problema estava relacionado com a análise patológica de tecidos procurando identificar microvasos presentes em imagens de angiogénese do cancro da mama.

O presente artigo começa por apresentar, na secção 2, o conceito fundamental da teoria da informação adequado ao manuseio de processos estocásticos espectralmente coincidentes recorrendo ao caso limite. O caso limite de sobreposição espectral é a total inexistência de memória no processo, ou seja, processos completamente brancos em todas as suas componentes. Assim, assumindo também estacionaridade, a entropia diferencial do processo coincide com a entropia diferencial de primeira ordem. Seguidamente, recorrendo à definição de negentropia, alcança-se uma conveniente representação da informação mútua do processo.

Na secção 3, recorrendo à teoria das distribuições, alcança-se uma aproximação computável para a negentropia marginal. É precisamente esta aproximação que está na base dos actuais algoritmos de Análise em Componentes Independentes.

Nas secções 4 e 5 ilustra-se a utilidade desta técnica em duas aplicações biomédicas reais onde os processos se apresentam na forma de componentes espectralmente coincidentes e, consequentemente, intratáveis à luz da teoria convencional de filtragem de Fourier.

2. ANÁLISE ENTRÓPICA DIFERENCIAL DE 1ª ORDEM DE UM PROCESSO SEM MEMÓRIA

A entropia diferencial de um processo estocástico multivariado estacionário \mathfrak{X} é dada por [1]

$$h(\mathfrak{X}) = \lim_{L \rightarrow \infty} \frac{1}{L} h(X_1, X_2, \dots, X_L) \quad (1)$$

onde L traduz o número de realizações temporais. No caso particular de o processo ser desprovido de memória (processo branco) com amostras identicamente distribuídas, então a sua entropia diferencial (1) coincide com a sua entropia diferencial de primeira ordem :

$$h(\mathfrak{X}) = \lim_{L \rightarrow \infty} \frac{1}{L} [h(X_1) + h(X_2) + \dots + h(X_L)] = h(X) \quad (2)$$

Que, por sua vez, é definida [1] por (3) onde $p_x(\xi)$ é a função densidade de probabilidade de X .

$$h(X) = - \int p_x(\xi) \log_2 p_x(\xi) d\xi \quad (3)$$

A mais relevante propriedade de (3) está associada à expressão da entropia diferencial de um processo $Y=MX$ (com M uma transformação não singular) em função da entropia diferencial do processo X [1]:

$$H(Y) = H(X) + \log_2 |\det M| \quad (4)$$

A partir da entropia diferencial de 1ª ordem define-se negentropia [2] como:

$$J(X) = H(G_X) - H(X) \quad (5)$$

Onde $H(G_X)$ é a entropia diferencial da densidade gausseana que partilha os dois primeiros cumulantes com a densidade de X e $H(X)$ é a entropia diferencial do processo. Esta (5) exhibe a importante propriedade (6) de invariância segundo transformação linear não singular [2]:

$$J(MX) = J(X) \quad (6)$$

Considere-se, por último, a informação mútua das componentes do processo X no sentido da divergência de *Kullback-Leibler* [2]:

$$I(X) = D\left(p_X \parallel \prod_{i=1}^n p_{x_i}\right) = \int p_X(\xi) \log_2 \frac{p_X(\xi)}{\prod_{i=1}^n p_{x_i}(\xi_i)} d\xi \quad (7)$$

Então, recorrendo às propriedades (4) e (6), a informação mútua (7) pode ser representada em função das negentropias (marginais e conjunta) e do segundo cumulante (matriz covariância) do processo conforme descrito em (8):

$$I(X) = J(X) - \sum_{i=1}^n J(x_i) + \frac{1}{2} \log_2 \frac{\prod_{i=1}^n c_{x_i}^2}{|\det c_X^{i_1, i_2}|} \quad (8)$$

3. ANÁLISE EM COMPONENTES INDEPENDENTES

Da expressão anterior interessa salientar dois resultados:

- i) se X seguir uma distribuição gausseana então a informação mútua resulta simplifcadamente como:

$$I(X) = \frac{1}{2} \log_2 \frac{\prod_{i=1}^n c_{x_i}^2}{|\det c_X^{i_1, i_2}|} \geq 0 \quad (9)$$

Assim, no caso gausseano a informação mútua nula – isto é, a independência estatística – alcança-se por via de um segundo cumulante diagonal.

- ii) Seja X uma v.a. multivariada esférica (com matriz covariância identidade) que segue distribuição não gausseana. Considere-se ainda M uma rotação plana originando $Y = M X$. Então, de acordo com a propriedade de invariância perante transformações lineares da negentropia (6), tem-se:

$$I(Y) \leq I(X) \Leftrightarrow \sum_{i=1}^N J(y_i) \geq \sum_{i=1}^N J(x_i) \quad (10)$$

Esta última expressão sugere como procedimento para minimização da informação mútua entre as componentes de X – ou equivalentemente obtenção de independência estatística – a maximização da soma das negentropias marginais.

Assim, à luz desta abordagem, são dois os passos sugeridos a dar no sentido da independência estatística ou da minimização da informação mútua (8): na primeira fase anula-se o termo que depende exclusivamente do segundo cumulante, bastando para isso tornar esse processo esférico; na segunda fase, através de rotações, tenta-se maximizar:

$$\sum_{i=1}^N J(y_i) \quad (11)$$

Estes dois passos vão contribuir para que o valor da informação mútua da v.a. multivariada resultante, numa situação ideal, seja nulo. O que equivale a dizer que as componentes dessa v.a. são estatisticamente independentes.

A negentropia marginal pode ser utilizada como uma medida de afastamento da independência. O problema na utilização desta grandeza, é a sua elevada complexidade computacional. Em termos práticos para se utilizar a negentropia teria que ser calculado o integral na definição de entropia diferencial. Esta tarefa não se revela no entanto de grande facilidade pois o integral envolve a fdp. Para além disso os estimadores desta função (fdp), são também eles de complexidade computacional elevada e propensos a erros [3]. Por estas razões a entropia diferencial de primeira ordem e a negentropia são essencialmente conceitos teóricos com reduzida aplicabilidade imediata. Na prática há algumas aproximações destes conceitos que podem ser utilizadas.

A aproximação clássica para a negentropia é um método que recorre a cumulantes de ordem superior. De facto este método baseia-se numa aproximação da fdp, contida no integral da definição de entropia. Esta aproximação parte apenas do princípio que essa densidade $p_x(x)$ se aproxima da densidade gausseana normalizada $G_x(x)$.

A aproximação é dada pela expansão em série de *Edgeworth* [4] ou equivalentemente em série de *Gram-Charlier*. Estas expansões conduzem a resultados semelhantes. Elas utilizam os polinómios de *Chebyshev* (h_i) onde o índice é um inteiro não negativo que define a ordem do polinómio. Os polinómios de *Chebyshev* formam um sistema ortonormal oriundo das sucessivas derivações de $G_x(x)$ [4].

A expansão em série de *Edgeworth* da fdp truncada, para incluir apenas os dois primeiros termos não-constantes, é dada pela expressão:

$$p_x(x) \approx \hat{p}_x(x) = G_x(x) \left[1 + c_x^3 \frac{h_3(x)}{3!} + c_x^4 \frac{h_4(x)}{4!} \right] \quad (12)$$

Esta expansão, como já foi dito, é baseada na ideia de que a fdp é muito semelhante a uma gausseana. A parte não gausseana da fdp é reflectida pelos cumulantes de ordem superior, neste caso de terceira e quarta ordem. Recorde-se que foi assumido que a variável X estaria normalizada pelo que a expansão começa directamente nos cumulantes de ordem superior.

Desta aproximação da fdp, resulta para a negentropia uma aproximação computável.

Considerando que, no caso geral, pode tratar-se de uma fdp simétrica – logo o terceiro cumulante será nulo – estabelece-se [2] uma aproximação entre o quadrado do cumulante de ordem quatro – *kurtosis* – e a negentropia:

$$J(x) \approx \frac{1}{48 \ln 2} (c_x^4)^2 \quad (13)$$

Este é um resultado fundamental que relaciona um elemento da teoria da informação – negentropia – com estatísticas de ordem superior. Por outras palavras, minimizar a informação mútua (7) entre componentes de uma v.a. (X) é equivalente à maximização da soma das negentropias marginais que possui uma relação de proporcionalidade directa com a soma dos quadrados das *kurtosis*.

$$\sum_{i=1}^n J(x_i) \approx \sum_{i=1}^n [(c_{x_i}^4)^2] \quad (14)$$

A concretização algorítmica da maximização do funcional (14) foi alvo, na última década, de imensa investigação da qual resultaram vários algoritmos entretanto largamente divulgados. O algoritmo utilizado no presente trabalho [5] consiste num método de Jacobi (para instanciação de sucessivas rotações planas) originalmente proposto por Comom.

4. EXTRACÇÃO CEGA DE ELCTROCARDIOGRAMA FETAL

A monitorização do comportamento cardíaco de um feto pode ajudar na identificação prematura de problemas de saúde que o afectem no futuro. Os exames de diagnóstico cardíaco (ECG) estão hoje em dia desenvolvidos para adultos ou crianças. O mesmo exame aplicado a um feto assume outros contornos e dificuldades. Um ECG de um feto, feito de uma forma não-invasiva, terá sempre presente outras fontes de ruído. Fontes estas que podem ser mais fortes do que o próprio *FECG*. É na tentativa de separar o *FECG* do *MECG* que se recorre à *ACI*.

Considere-se as oito leituras feitas através de eléctrodos espalhados nas regiões torácica e abdominal de uma mulher grávida e representadas na figura 1.

As primeiras cinco componentes correspondem às leituras efectuadas pelos eléctrodos localizados na região abdominal e as últimas três são leituras obtidas da região torácica.

Ao observar os sinais obtidos pela leitura através de eléctrodos directamente colocados no corpo da mulher (que são as misturas lineares das fontes independentes), não se consegue vislumbrar nenhum sinal com frequência diferente. Todos eles aparentam ter a mesma frequência. Os sinais onde o batimento cardíaco da mulher será mais evidente é nas leituras obtidas pelos eléctrodos da região torácica (representados nos últimos três sinais).

O sinal que aparece no topo da janela parece ter um pico de amplitude mais reduzida entre cada dois picos de amplitude detectado nas outras leituras. Isto é, parece haver aqui

indícios do batimento cardíaco do feto.

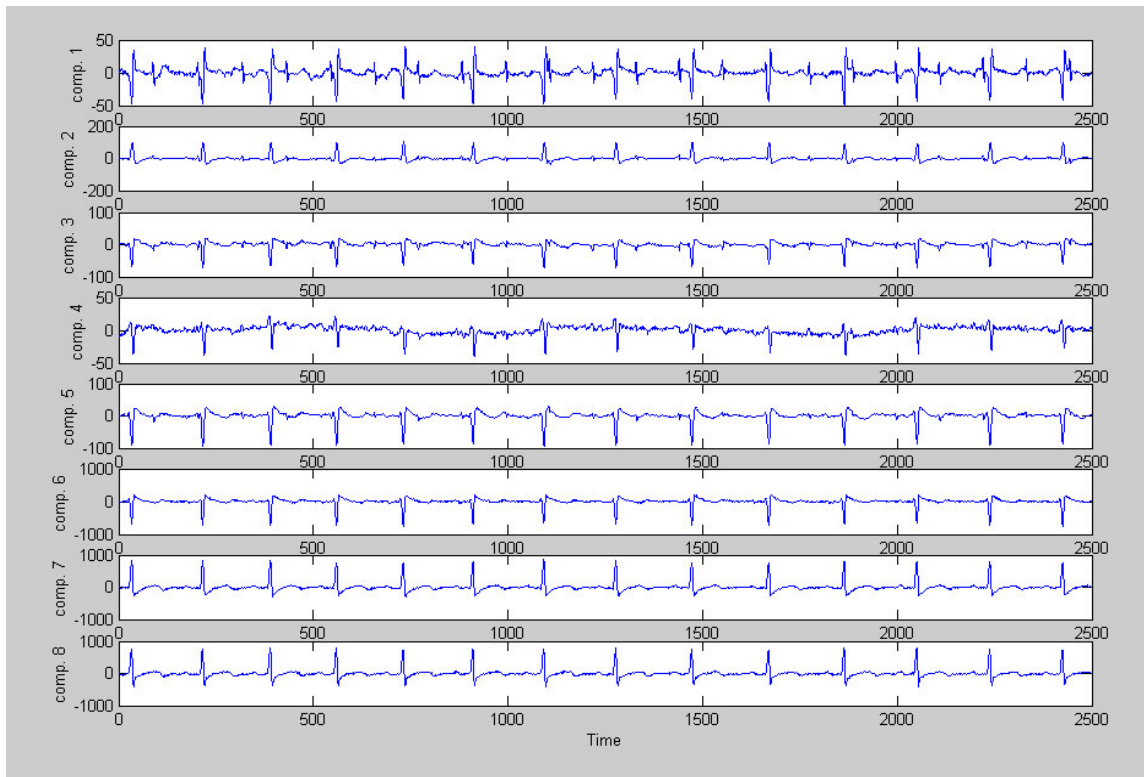


Figura 1. Leituras obtidas pelos oito eléctrodos.

Considerando que a frequência dominante (em todos os sinais obtidos por qualquer eléctrodo) é a da actividade cardíaca da mãe, então uma frequência quase dupla desta pode muito bem ser indicadora de outra fonte: a actividade cardíaca do feto. De facto o batimento do coração de um feto tem uma frequência substancialmente mais elevada – cerca de duas vezes – do que o coração de um adulto [6][7]. O que se observa na leitura do topo da janela parece ser uma mistura de dois sinais um com a frequência dupla do outro. Ainda que esta observação seja muito ténue, e pouco indicadora quanto à actividade cardíaca do feto.

A coincidência espectral das leituras (pode constatar-se na figura 2) inviabiliza qualquer aplicação das técnicas de filtragem de Fourier.

No entanto, a aplicação do algoritmo de análise em componentes independentes ao processo proveniente dos eléctrodos produz a representação do mesmo traduzida na figura 3 onde é bem evidente uma componente traduzindo a actividade electrocardiográfica do feto.

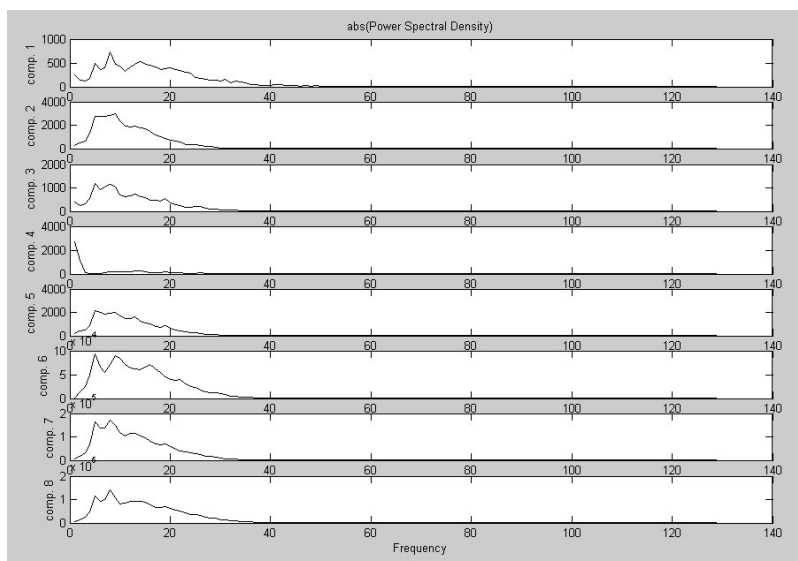


Figura 2 Espectro de frequências das oito leituras do ECG (período corresponde a 256)

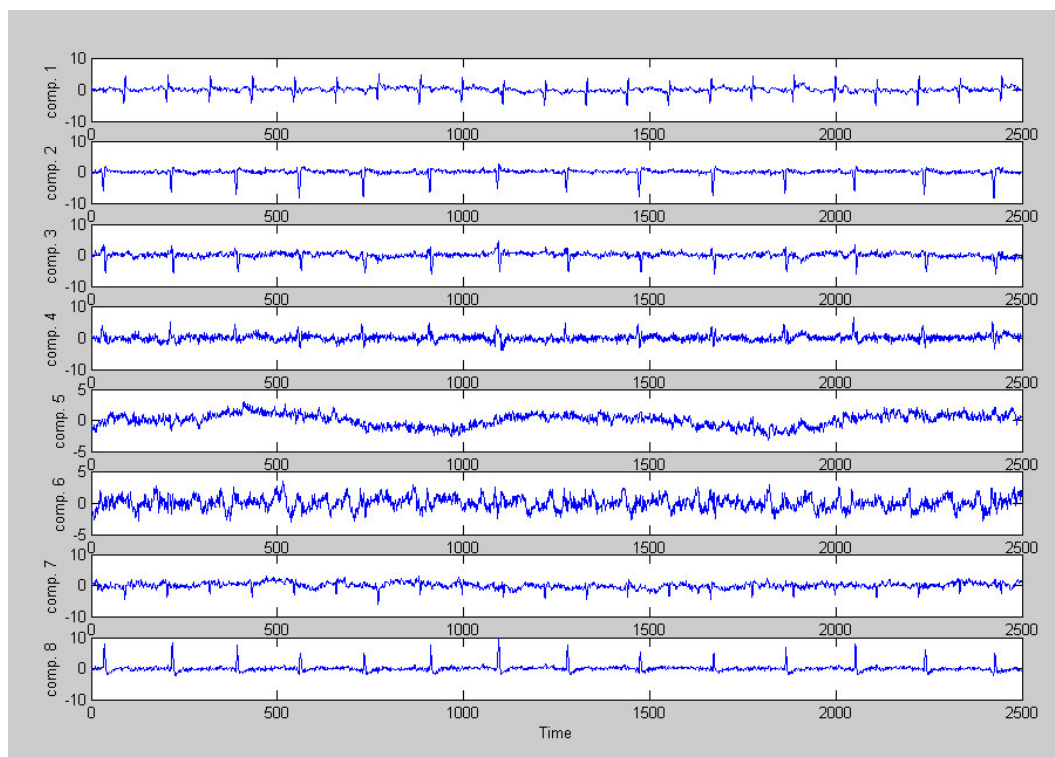


Figura 3 Resultados da aplicação do método de ACI no problema de extração do *FECG*

5. IDENTIFICAÇÃO DE MICROVASOS EM ANGIOGÊNESE DO CANCRO DA MAMA

Para o problema da angiogênese do cancro da mama o patologista considera existirem apenas duas fontes de informação visual contidas nas imagens médicas: microvasos e tudo resto – aqui designado por tecido. Recorrendo a um modelo de ACI – instantâneo e de mistura linear, como se assume nesta investigação – pretende-se identificar e isolar a fonte dos microvasos a partir das imagens médicas utilizadas pelo especialista.

À luz do modelo de ACI descrito, assume-se que a imagem médica é uma mistura com três componentes – Red, Green, Blue (*RGB*) – de três fontes independentes. Os microvasos serão uma dessas fontes. De facto espera-se que a fonte cuja densidade se apresente mais esparsa represente a distribuição dos microvasos. Quanto mais esparsa for a distribuição de uma fonte maior o seu valor de *kurtosis*. Daí que a maximização do *kurtosis* – através da diagonalização do cumulante de ordem quatro implementado pelo método de ACI desenvolvido – seja por si só um método suficiente para extrair a fonte dos microvasos. Considere-se a seguinte imagem de uma amostra de tecido como exemplo.

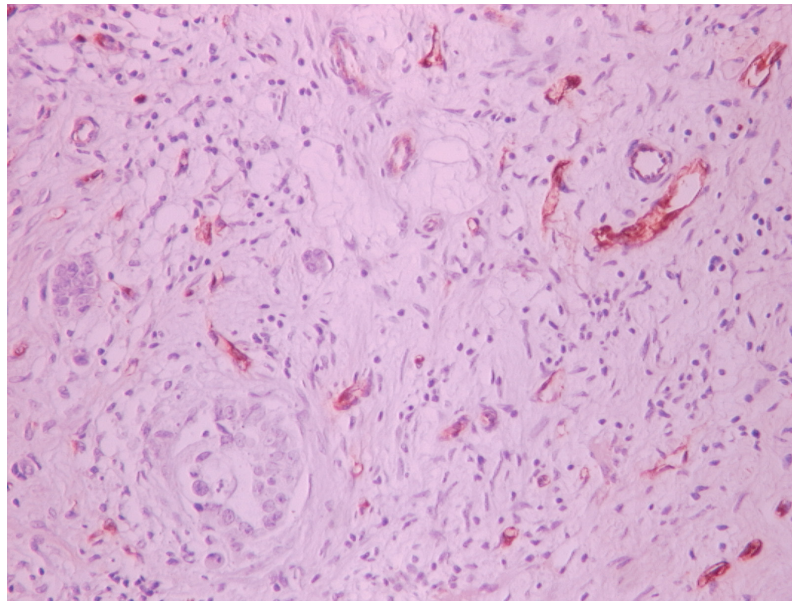


Figura 4 - Imagem médica original

Inicialmente decompõem-se a imagem nas suas três componentes. Obtendo-se uma v.a. multivariada de dimensão três. No exemplo considerado cada componente corresponde a 1918800 leituras. Como se pode ver da figura 5, as três misturas que se obtém são coincidentes no espectro inviabilizando novamente as técnicas de filtragem convencionais.

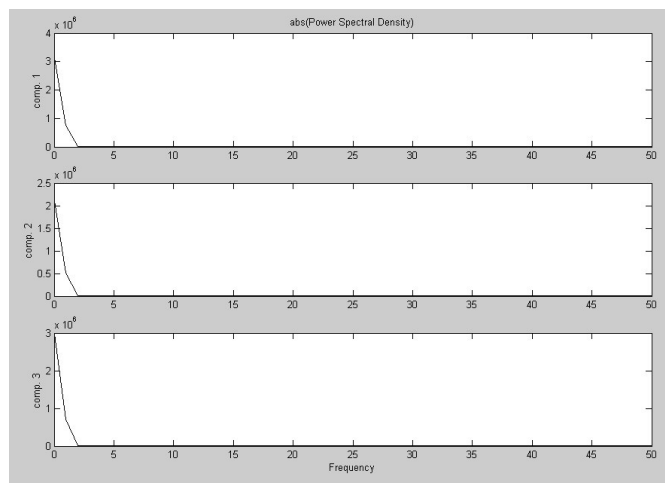


Figura 5 - Espectros de Fourier das três componentes de cor da imagem médica.

No entanto, após uma transformação linear deduzida pelo algoritmo de Análise em Componentes Independentes, obtêm-se novas imagens. Destas, a associada ao maior valor de kurtosis (mais esparsa), está representada na figura 6 traduzindo a localização dos microvasos.

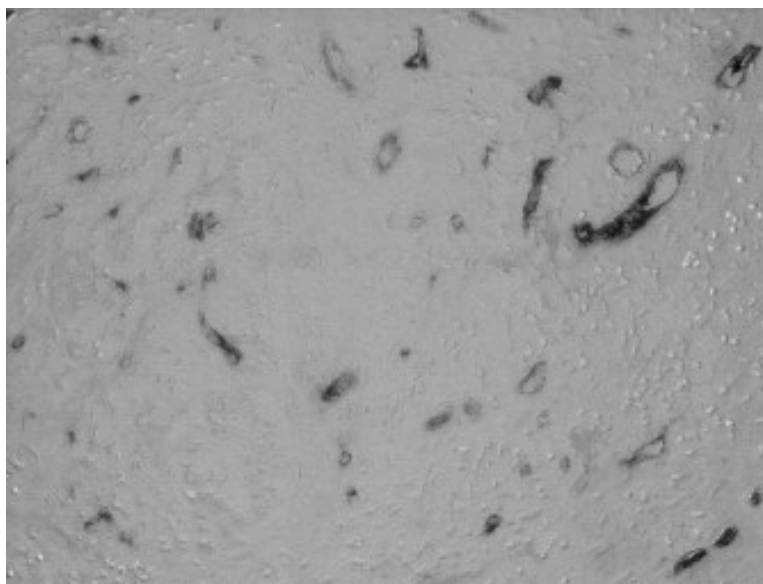


Figura 6 - Imagem monocromática obtida após ACI.

5. CONCLUSÃO

Foram apresentadas duas aplicações biomédicas onde a (até muito recentemente) julgada impossível separação cega de fontes coincidentes no espectro foi alcançada com sucesso e nexos aplicacionais. Esperamos, deste modo, ter estimulado os colegas condicionados pelas teorias convencionais de filtragem a porém em prática a utilização da Análise em Componentes Independentes noutros domínios técnicos e científicos.

REFERÊNCIAS

- [1] T.M. Cover e J.A. Thomas, *Elements of Information Theory*, Wiley, (1991).
- [2] Comon P., *Independent component analysis, A new concept?*. *Signal Processing*, vol. 36, pp 287-314, (1994),
- [3] Hyvärinen A., Karhunen J., Oja E., *Independent Component Analysis*, John Wiley & Sons Inc., (2001).
- [4] Stuart A., Ord J.K., *Kendall's Advanced Theory of Statistics – Distribution Theory*, 6th edition, vol. I, Edward Arnold, (1994).
- [5] J.M. Costa Pereira, *Análise em Componentes Independentes na Separação Cega de Fontes Coincidentes no Espectro*, Tese de Mestrado, Faculdade de Ciências da Universidade do Porto, (2003).
- [6] Zarzoso V., Nandi A.K., Bacharakis E., *Maternal and Foetal ECG Separation Using Blind Source Separation Methods*, *IMA Journal on Mathematics in Medicine & Biology*, vol. 14, pp 207-225, (1997).
- [7] Talmon J.L., *Pattern Recognition of the ECG - A Structured Analysis*, Ph.D. Thesis, Free University of Amsterdam, (1983).