Estudo de Séries Temporais na Análise em Componentes Principais e na Análise em Componentes Independentes

Fernando Sebastião

Instituto Politécnico de Leiria, Departamento de Matemática da ESTG e CM-UTAD - fsebast@estg.ipleiria.pt

Irene Oliveira

Universidade de Trás-os-Montes e Alto Douro, Departamento de Matemática e CM-UTAD - ioliveir@utad.pt

Resumo: As séries temporais são constituídas por observações autocorrelacionadas e como tal não podem ser permutadas entre si, contudo a suposição em relação à independência entre observações não é necessária para aplicar as técnicas de Análise em Componentes Principais (ACP) e Análise em Componentes Independentes (ACI) do ponto de vista descritivo. A aplicação destas técnicas multivariadas a séries temporais permite realçar alguns resultados e interpretações que sugerem uma conexão com as inter-relações existentes entre as observações, pelo menos em termos empíricos.

O objectivo deste trabalho é apresentar as técnicas de ACP e ACI, descrevendo com maior ênfase a ACI. Para além disso é apresentado um exemplo de dados meteorológicos em séries temporais assim como os respectivos resultados que salientam as diferenças e semelhanças entre as duas técnicas em análise, nomeadamente ao nível dos domínios do tempo e da frequência, incluindo resultados para a qualidade das reconstruções dos dados originais para cada uma das técnicas.

Palavras–chave: análise em componentes principais, análise em componentes independentes, séries temporais.

Abstract: Time series data set consists of autocorrelated observations that can not be interchanged, but no assumption concerning independence, or not, between observations is necessary to perform the techniques of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) in a descriptive point of view. The application of these multivariate techniques to time series allows enhancing some results and interpretations that suggest a connection with the existent interrelations between observations, at least from an empirical point of view.

The aim of this paper is to present the techniques of PCA and ICA, describing ICA with more detail. In addition we present an example of climatic time data sets and some results to illustrate the differences and similarities between these techniques in time and frequency domains, including results for the quality of the reconstructed patterns of the original data for each one of the techniques.

Keywords: principal component analysis, independent component analysis, time series.

1 Introdução

A Análise em Componentes Principais (ACP) tem desempenhado um papel fundamental na análise de dados multivariados em praticamente todas as áreas do conhecimento, já a Análise em Componentes Independentes (ACI) é uma técnica bastante utilizada em apenas algumas áreas mais específicas tais como cancelamento do ruído auditivo, sinais biomédicos, processamento de imagem, séries temporais econométricas e telecomunicações, entre outras. No estudo de séries temporais pode usar-se a ACP ou a ACI em duas direcções: no domínio do tempo onde se admite que as observações sucessivas possuem algum tipo de relação ao longo deste; e no domínio da frequência onde se analisa a estrutura que resulta de diferentes oscilações em diferentes frequências.

Neste trabalho são apresentadas as técnicas de ACP e ACI, onde a sintetização da informação e a extracção das componentes com a informação principal acerca da autocorrelação entre as observações são os principais objectivos a ter em conta. A ACI é ainda pouco divulgada entre os estatísticos e, como tal, é descrito o modelo geral e são apresentados alguns conceitos básicos tais como centragem e branqueamento de variáveis, maximização da não normalidade e ainda duas formas distintas de obter as componentes independentes através de dois algoritmos - o FastICA (Hyvärinen *et al.*, 2001) e o AMUSE (Tong *et al.*, 1991). Usaremos as técnicas de ACP e ACI para tratar um conjunto de dados acerca da pressão média mensal, ao nível do mar no Norte do Oceano Pacífico, e são apresentados alguns resultados que evidenciam as diferenças e semelhanças entre as referidas técnicas, nomeadamente em termos de correlações e espectros das coordenadas das componentes, assim como ao nível das análises da qualidade das reconstruções dos dados originais.

2 Análise em Componentes Principais

Desde o início do século XX que a ACP se tornou numa das técnicas estatísticas mais utilizadas em variadas áreas (Jolliffe, 2002).

O objectivo da ACP é reduzir a dimensão de um conjunto de dados originais, mantendo tanto quanto possível a sua informação inicial. A partir de um conjunto de dados descrito por uma matriz $\mathbf{X}$, constituída por n observações em p variáveis, pretendem-se encontrar combinações lineares dessas variáveis, não correlacionadas entre si, designadas por componentes principais (CPs), as quais são dispostas de acordo com uma ordem decrescente da variância.

Utiliza-se a decomposição espectral da matriz de variâncias-covariâncias (ou noutros casos a matriz de correlações $\mathbf{R}$) de $\mathbf{X}$, $\mathbf{S} = \mathbf{PLP}^T$, onde $\mathbf{P}$ é a matriz dos vectores próprios ortogonais e $\mathbf{L}$ é a matriz diagonal dos valores próprios. Ao projectar os dados centrados no espaço gerado pelas colunas de $\mathbf{P}$, obtém-se a matriz de projecções $\mathbf{Z}$ cujas colunas representam as coordenadas das CPs. Após se extraírem as CPs, por vezes há interesse em utilizar algumas destas, as primeiras, para obter valores aproximados da matriz dos dados iniciais $\mathbf{X}$, ou

seja, para efectuar a reconstrução da matriz $\mathbf{X}$ obtendo, tanto quanto possível, uma representação fidedigna dos dados. Várias aplicações da ACP relacionam-se com séries temporais cujas observações são autocorrelacionadas e como tal não podem trocar entre si como é por exemplo nos dados climáticos.

3 Análise em Componentes Independentes

A ACI (Hyvärinen *et al.*, 2001) é uma técnica estatística e computacional que foi introduzida por Hérault e Ans (1984) e Hérault, Jutten e Ans (1985) e exposta de uma forma mais clara por Comon (1994).

O objectivo principal deste método é encontrar componentes ou factores escondidos que relacionem conjuntos de variáveis aleatórias, medições ou sinais. No modelo admite-se que as variáveis referentes aos dados são misturas lineares de variáveis latentes desconhecidas, isto é, que não podem ser directamente observadas, as quais são denominadas de componentes independentes e abreviadas por CIs (por vezes também designadas de factores ou fontes).

Este método distingue-se dos outros dentro do género ao admitir que as componentes são estatisticamente independentes e que não seguem uma distribuição Normal. Embora a técnica da ACI esteja relacionada com a ACP, é com a Análise Factorial que apresenta mais afinidades, embora esta não tenha em conta a não normalidade dos dados. Geralmente, a ACI é uma técnica muito mais potente para encontrar os factores escondidos quando os métodos clássicos falham por completo.

3.1 Descrição do modelo

Considere-se o modelo geral constituído por n variáveis aleatórias x_1, ..., x_n, descritas à custa de misturas lineares de n variáveis latentes s_1, ..., s_n, na forma matricial $\mathbf{x} = \mathbf{As}$, onde $\mathbf{x}$ é o vector aleatório cujos elementos são as misturas x_1, ..., x_n, $\mathbf{s}$ é o vector aleatório cujos elementos são as componentes independentes s_1, ..., s_n e $\mathbf{A}$ é a matriz das misturas cujos elementos a_{ij} $(i, j = 1, ..., n)$ são os parâmetros desconhecidos.

Para que se possa estimar o modelo básico de ACI, é preciso admitir que:

- as componentes s_i $(i = 1, ..., n)$ sejam estatisticamente independentes, uma vez que este é o princípio no qual a ACI se baseia;
- pelo menos $n - 1$ das componentes s_i $(i = 1, ..., n)$ não sigam uma distribuição Normal, uma vez que apenas uma pode seguir uma distribuição Normal pelo facto de ser impossível separar várias componentes que sigam distribuições Normais;
- por simplicidade, a matriz $\mathbf{A}$ seja quadrada (embora noutros modelos para além do básico, esta possa não ser quadrada).

Numa primeira etapa de aplicação da ACI há que começar por estimar a matriz $\mathbf{A}$ e o vector $\mathbf{s}$, observando apenas $\mathbf{x}$, tendo em conta os princípios apresentados de seguida que permitirão descrever os algoritmos existentes para a estimação do modelo. Após estimar a matriz $\mathbf{A}$, e admitindo que esta é invertível,

pode determinar-se a sua inversa $\mathbf{W} = \mathbf{A}^{-1}$, e podem obter-se as componentes independentes dadas por $\mathbf{s} = \mathbf{W}\mathbf{x}$. À semelhança da ACP, caso se pretendam comparar os valores observados de $\mathbf{x}$ com os reconstruídos, há que efectuar a reconstrução dos dados originais multiplicando o vector das componentes independentes estimadas pela matriz das misturas estimadas.

3.2 Centragem e branqueamento de variáveis

Antes de aplicar um algoritmo para executar a ACI a um determinado conjunto de dados, geralmente efectuam-se algumas técnicas de pré-processamento, as quais facilitam a estimação dos parâmetros do modelo. A centragem é a mais básica das técnicas, sendo comum centrarem-se os dados em $\mathbf{x}$.

O branqueamento das variáveis observadas é outra técnica de pré-processamento dos dados, usada para estimar as CIs, a qual consiste em transformar linearmente o vector das observações $\mathbf{x}$, multiplicando-o por uma matriz $\mathbf{V}$, de forma a obter-se um outro vector branqueado $\mathbf{z} = \mathbf{V}\mathbf{x}$, cujas componentes sejam não correlacionadas e de variâncias unitárias, ou seja, tal que a sua matriz de covariâncias seja igual à matriz identidade.

3.3 Maximização da não normalidade

Um princípio simples da estimação de um modelo ACI é baseado na maximização da não normalidade. Considera-se como ponto de partida o teorema do limite central e como medida de não normalidade a curtose que corresponde ao cumulante de quarta ordem. Ao contrário de muitas outras variáveis aleatórias, uma variável aleatória (v. a.) que siga uma distribuição Normal apresenta uma curtose de zero. Uma vez que as CIs podem ser encontradas através das direcções nas quais os dados maximizem a não normalidade, esta pode ser medida pela maximização dos valores absolutos da curtose.

Uma vez que nem sempre a curtose é a melhor medida de não normalidade para a estimação do modelo ACI, nomeadamente por esta ser por vezes sensível a *outliers*, surge como alternativa uma outra medida designada por *neguentropia*. Contudo, é mais difícil utilizar a *neguentropia* em termos computacionais, dado que a sua estimação usando a própria definição requereria uma estimativa da função densidade de probabilidade. No entanto, aproximações da *neguentropia* são muito úteis e podem ser utilizadas para obter um eficiente método para ACI. O método clássico de aproximação da *neguentropia*, que utiliza cumulantes de ordem superior, é (para uma v. a. de média zero e variância unitária)

$$J(x) \approx \frac{1}{12}\left[E(x^3)\right]^2 + \frac{1}{48}\left[curt(x)\right]^2, \tag{1}$$

onde $curt(x)$ é a curtose da v. a..

Outras aproximações foram desenvolvidas, de entre as quais se destaca a que é descrita por $J(x) \approx \left(E\left[G(x)\right] - E\left[G(x_{NormalStd})\right]\right)^2$, onde G é uma função não quadrática e $x_{NormalStd}$ é uma v. a. com distribuição Normal estandardizada.

Para o caso em que $G(x) = x^4$ obtém-se uma generalização da curtose. Ao escolher cuidadosamente a função G, obtêm-se aproximações da *neguentropia* que fornecem melhores resultados que a descrita em (1) e ao escolher uma função G que não cresça muito rapidamente, obtêm-se estimadores mais robustos. As funções $G_1(x) = \frac{1}{\alpha}\log(\cosh(\alpha x))$ e $G_2(x) = -\exp(-x^2/2)$, onde $1 \leq \alpha \leq 2$ é uma constante, são bastante utilizadas na aproximação da *neguentropia*.

3.4 Algoritmos utilizados

O algoritmo FastICA, baseado na maximização da não normalidade, converge rapidamente e apresenta certas propriedades que fazem dele um algoritmo mais eficiente em relação aos algoritmos baseados no gradiente, na maioria dos casos. A versão utilizada deste algoritmo (existente no *Package fastICA* do *Software R*), e que usa o método da ortogonalização simétrica, descreve-se sumariamente através dos passos seguintes:

1. Centrar os dados para que estes tenham média nula.
2. Branquear os dados de forma a obter $\mathbf{z}$.
3. Escolher o número m ($m \leq n$) de componentes independentes a estimar.
4. Escolher os valores iniciais para cada vector unitário $\mathbf{w}_i$ ($i = 1, ..., m$). Ortogonalizar a matriz $\mathbf{W}$.
5. Para cada $i = 1, ..., m$, implementar a iteração básica do ponto fixo $\mathbf{w}_i \leftarrow E\left\{\mathbf{z}g(\mathbf{w}_i^T\mathbf{z})\right\} - E\left\{g'\left(\mathbf{w}_i^T\mathbf{z}\right)\right\}\mathbf{w}$, onde g é uma função que serve para obter uma aproximação da *neguentropia*.
6. Após cada iteração ortogonalizar simetricamente a matriz $\mathbf{W}$.
7. Se não convergir, voltar ao passo 5. Estimar a matriz $\mathbf{A}$ dos coeficientes da mistura e o vector $\mathbf{s}$ das componentes independentes.

Um outro algoritmo denominado por AMUSE é uma alternativa para casos em que se admite que as CIs possuem uma dependência temporal, procurando encontrar a matriz que anule as covariâncias e também as covariâncias desfasadas. A versão deste algoritmo que foi utilizada e implementada no *Software R* descreve-se sumariamente através dos passos seguintes:

1. Centrar os dados para que estes tenham média nula.
2. Branquear os dados de forma a obter $\mathbf{z}$.
3. Escolher o número m ($m \leq n$) de componentes independentes a estimar.
4. Seleccionar um desfasamento no tempo, τ, para aplicar a decomposição em valores próprios da matriz $\left[C_\tau^{\mathbf{z}} + (C_\tau^{\mathbf{z}})^T\right]/2$, onde $C_\tau^{\mathbf{z}} = E\left\{\mathbf{z}(t)\mathbf{z}(t-\tau)^T\right\}$ é a matriz de covariâncias dos desfasamentos no tempo para um dado desfasamento τ.
5. A partir da matriz dos vectores próprios obtida em 4, estimar a matriz $\mathbf{A}$ dos coeficientes da mistura e o vector $\mathbf{s}$ das componentes independentes.

4 O caso em estudo

Considere-se um conjunto de 216 valores mensais médios de pressão ao nível do mar observados em cada uma das 8 estações meteorológicas (desde Janeiro de 1979 a Dezembro de 1996) no Norte do Oceano Pacífico (nos estados do Alasca, Califórnia, Havai e Washington). As estações meteorológicas são: 1 - Crescent City; 2 - San Diego; 3 - San Francisco; 4 - Hilo; 5 - Honolulu; 6 - NeahBay; 7- Seldovia e 8 - Sitka. A amplitude temporal (de Jan. de 1979 a Dez. de 1996) é suficiente para examinar as relações dos padrões obtidos pelas técnicas e pela série temporal do Índice de El Niño. A este conjunto de dados aplicaram-se as técnicas de ACP e ACI para os 2 algoritmos descritos (FastICA e AMUSE) e utilizaram-se somente as 3 primeiras CPs e as 2 primeiras CIs por se verificar serem suficientes na obtenção de resultados satisfatórios.

Consideremos as coordenadas para as primeiras componentes como se ilustra na Figura 1.

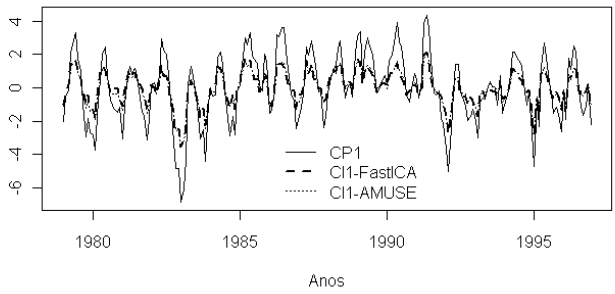

Figura 1: Coordenadas de CP1, CI1-FastICA e CI1-AMUSE.

Para analisar as coordenadas destas componentes em séries temporais, compararam-se as correlações (Tabela 1) entre pares de componentes das duas técnicas e para comparar os coeficientes dos vectores usaram-se os ângulos entre cada par de vectores de componentes, de onde se destacam os cossenos mais elevados (em valor absoluto): cos(CP1, CI1) = 0.906; cos(CP2, CI2) = 0.483 e cos(CP3, CI2) = 0.687, os quais permitem identificar relações existentes entre os vectores de ambas as técnicas. Após a análise das semelhanças existentes entre os coeficientes dos vectores das componentes procedeu-se ao estudo das coordenadas dessas componentes. Na Figura 1 mostra-se a semelhança entre CP1, CI1-FastICA e CI1-AMUSE e na Figura 2 a comparação entre os espectros das coordenadas das componentes. Na Figura 2 observa-se que as diferentes componentes assumem valores espectrais idênticos e em particular os espectros CP1 e CI1-FastICA são muito semelhantes assim como o CP3 e CI2-FastICA.

Tabela 1: Correlações entre as coordenadas das componentes principais em ACP e das componentes independentes em ACI (através do FastICA e do AMUSE).

	$CP1$	$CP2$	$CP3$
CI1-FastICA	0.965	0.001	-0.235
CI2-FastICA	-0.209	0.592	-0.692
CI1-AMUSE	0.986	-0.093	-0.122
CI2-AMUSE	-0.053	0.584	-0.721

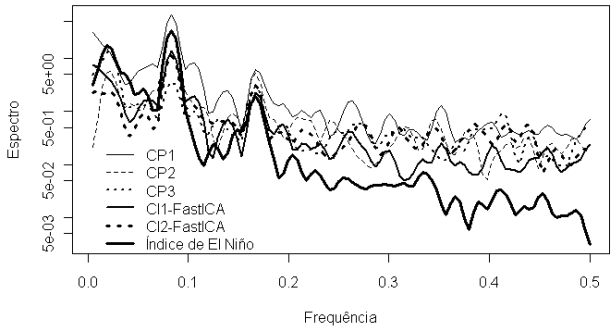

Figura 2: Comparações entre os espectros das coordenadas de CP1, CP2, CP3, CI1-FastICA, CI2-FastICA e a série temporal do Índice de El Niño.

Existe ainda uma relação entre o Índice de El Niño e as componentes em estudo, como se pode observar no gráfico da Figura 2, uma vez que as séries temporais têm estruturas idênticas evidenciando dois principais picos de frequência em 0.083 e 0.167. Decidiu-se não representar no gráfico os espectros para as séries obtidas através do algoritmo AMUSE porque seriam praticamente idênticos aos que foram obtidos através do algoritmo FastICA.

Para cada uma das estações meteorológicas, apresentam-se as somas de quadrados dos resíduos (SQR) na Tabela 2, onde se utiliza um número distinto de componentes da ACP e da ACI (através do FastICA e do AMUSE) na reconstrução dos dados originais. Quando se usa apenas a primeira componente, os valores das somas de SQR são praticamente idênticos entre as reconstruções de CP1 e CI1-AMUSE enquanto que a soma de SQR de CI1-FastICA é um pouco mais elevada. As reconstruções com as duas primeiras componentes apresentam valores de somas de SQR relativamente próximos para ambas as técnicas.

Como era de esperar, a soma de SQR decresce quando o número de componentes intervenientes na reconstrução dos padrões aumenta.

Tabela 2: Soma de quadrados dos resíduos para as reconstruções com a ACP e a ACI (através do FastICA e do AMUSE).

	$CP1$	$CP12$	$CP123$	$^{FastICA}_{CI1}$	$^{FastICA}_{CI12}$	$^{AMUSE}_{CI1}$	$^{AMUSE}_{CI12}$
CCity.CA	0.494	0.419	0.222	0.220	0.220	0.267	0.220
SDie.CA	0.684	0.329	0.160	0.629	0.465	0.717	0.465
SFranc.CA	0.466	0.426	0.057	0.262	0.111	0.392	0.111
Hilo.HW	0.527	0.211	0.132	0.602	0.599	0.613	0.599
Honol.HW	0.771	0.412	0.172	0.840	0.840	0.844	0.840
NBay.WA	1.159	0.331	0.326	1.203	0.688	0.926	0.688
Seld.AK	1.924	1.595	0.971	2.982	0.500	2.275	0.500
Sitka.AK	0.981	0.489	0.229	1.502	0.238	1.077	0.238
Soma de SQR	7.007	4.213	2.270	8.238	3.660	7.111	3.660

5 Conclusões

A informação retirada a partir dos gráficos nos três métodos é essencialmente a mesma. A estrutura temporal contida nos dados permite extrair a informação da autocorrelação nas primeiras componentes, a qual está directamente relacionada com o Índice de El Niño. Nas reconstruções dos dados originais, a ACI revela ser uma melhor alternativa à ACP essencialmente para duas componentes, e o AMUSE sugere melhores resultados comparativamente ao FastICA.

Referências

[1] Comon, P. (1994). Independent Component Analysis, A New Concept? *Signal Processing*, 36, 287-314.

[2] Hérault, J. e Ans, B. (1984). Circuits Neuronaux à Synapses Modifiables: Décodage de Messages Composites par Apprentissage non Supervisé. *Comptes Rendus de l'Académie des Sciences*, 299(III-13), 525-528.

[3] Hérault, J., Jutten, C. e Ans, B. (1985). Détection de Grandeurs Primitives dans un Message Composite par une Architecture de Calcul Neuromimétique en Apprentissage non Supervisé. *Actes du X colloque GRETSI*, 1017-1022, Nice, France.

[4] Hyvärinen, A., Karhunen, J. e Oja, E. (2001). *Independent Component Analysis.* John Wiley & Sons, Inc., U.S.A..

[5] Jolliffe, I. T. (2002). *Principal Component Analysis.* (2nd Edition) Springer-Verlag, New York.

[6] Tong, L., Liu, R.-W., Soon, V. C. e Huang, Y.-F. (1991). Indeterminacy and Identifiability of Blind Identification. *IEEE Transactions on Circuits and Systems*, 38(5), 499-509.