

## A Análise do Espectro Singular

---

*Carlos Dias Maciel\* \*\*  
Wagner Coelho de Albuquerque Pereira\**

### RESUMO

**O presente trabalho mostra uma nova abordagem da técnica de Componentes Principais. Essa nova abordagem procura extrair informações de periodicidade do sinal em estudo de forma que se possa realizar uma decomposição do sinal original em diversos sub-espacos com propriedades bem definidas.**

### INTRODUÇÃO

**A** Análise do Espectro Singular – SSA (*Singular Spectrum Analysis*) é um método usado em processamento de sinais digitais que é projetado para se extrair a máxima quantidade possível de informações de uma série temporal pequena e ruidosa sem o conhecimento da dinâmica por detrás do sistema físico.

O método é uma forma de análise de Componentes Principais no domínio do tempo, procurando descrever a varia-

bilidade de uma série em termos de sua estrutura de covariância.

Um exemplo do uso da SSA é na análise do sinal irradiado por um navio em movimento captado por um sistema sonar passivo. Este sinal é composto basicamente de ruídos gerados pelas máquinas deste navio e pela cavitação de seu hélice. O espectro deste sinal pode ser descrito como uma série de linhas espectrais, devidas principalmente às máquinas, sobrepondo-se a um espectro contínuo dominado pelo ruído de cavitação do hélice e ruído do ambiente.

A SSA poder ser usada, neste cenário descrito para:

---

\* Programa de Engenharia Biomédica – COPPE/UFRJ.

\*\* Centro Univ. Mocy S. Bastos – RJ.

1. identificar e extrair componentes periódicas ou quase-periódicas;
2. identificar e extrair variabilidades de baixa-frequência e tendências;
3. estimar o espectro das componentes discretas e contínuas.

O significado de termo “espectro de singularidade” vem da análise de sistemas dinâmicos e é representado pelo gráfico dos autovalores  $\{\lambda_k\}$  da matriz de covariância ordenados de forma decrescente. As formulas que serão apresentadas são aplicações do teorema da expansão bio-ortogonal Karhunen-Loève (KL) que é muito utilizada para compressão de informações e melhoria da relação sinal/ruído.

O presente trabalho é composto das seguintes partes: uma revisão teórica onde se pretende apresentar a natureza da decomposição espectral, o espectro singular, simulação, e discussão e conclusão.

## REVISÃO TEÓRICA

Os processos de estimação de frequências para senoides somadas à ruído branco é um problema clássico em estimação espectral. Começando pela análise por DFT (*Discrete Fourier Transform*) muitos métodos foram propostos baseados em abordagens paramétricas. Para um pequeno número de dados amostrados, todos os métodos paramétricos fornecem uma melhor solução que os métodos baseados em DFT, embora em cenários de baixa relação sinal-ruído os métodos paramétricos muitas vezes falhem.

O processo apresentado é robusto em baixos valores de relação sinal-ruído com resultados comparados aos demais métodos existentes, JOHNSON, (1982). Esse algoritmo é baseado na modelagem de dados usando a separação do sinal em um sub-espaço de sinais e ruído.

### O espectro de singularidades

A SSA é um método de análise de dados aleatórios baseado na técnica de Componentes Principais (CP). A análise clássica de CP, [BROOMHEAD & KING (1986), é usada em séries temporais fornecendo as coordenadas da seqüência com respeito a uma nova base ortogonal ( $E^k$ ,  $1 \leq k \leq M$ ) conhecida como EOF's (*empirical orthogonal function*):

$$E\{E^i E^j\} = \lambda_i \delta_{ij}, 1 \leq i \leq M, 1 \leq j \leq M, \quad \text{eq. 1}$$

onde  $E\{.\}$  é o operador valor esperado;

$E^k$  são vetores colunas.

Os autovalores  $\lambda_k$  são não negativos para o caso de estacionariedade dos dados de entrada e a fórmula de decomposição espectral assegura que:

$$R_x \equiv \sum_{k=1}^M \lambda_k E^k E^{kT} \quad \text{eq. 2}$$

onde  $R_x$  é a matriz de correlação da seqüência  $x$ ;

T indica o transposto do vetor;

M ordem da matriz de correlação.

As Componentes Principais fornecem uma representação para a série temporal em seu novo sistema de coordenadas cuja principal característica é possuir a maior parte de sua informação representada ao longo das  $p$  primeiras coordenadas ( $p < M$ ). Uma interpretação geométrica para o conjunto  $\{(\lambda_k, E^k), 1 \leq k \leq M\}$  é a representação de uma elipsoide  $M$ -dimensional com os vetores  $E^k$  informando a direção do eixo, e os valores  $\lambda_k$  o comprimento dos eixos principais do elipsoide.

Os vetores  $E^k$  são os autovetores da matriz de covariância da seqüência  $x_i$ . Para uma série escalar de valores denotados por  $x_i$ ,  $1 \leq i \leq N$ , a expansão é equivalente a

$$x_{i+j} = \sum_{k=1}^M a_i^k E^k_j, \quad 1 \leq j \leq M. \quad \text{eq. 3}$$

Os coeficientes de projeção  $a_i^k$  são chamados de CP da série  $x$ .

A estrutura da matriz de covariância  $R_x$  é Toeplitz para os casos de estacionariedade e próxima a Toeplitz para quasi-estacionariedade. Nestes casos, as EOF's são ortogonais e seus autovetores são simétricos ou anti-simétricos com respeito a  $M/2$ . Esses resultados são condições vindas da estacionariedade dos dados, embora o processo seja robusto à falta desta. O processo de CP é baseado na decomposição KL que não necessita da estacionariedade dos dados.

VAUTARD & GHIL (1989) mostraram que uma componente periódica somada a  $x$ , é geralmente caracterizada por autovalores aos pares. Assim uma componente periódica na série temporal (ou linha espectral) será representada por duas EOF's com autovalores próximos. Tipicamente, duas EOF's são necessárias para se extrair o módulo e fase de cada componente periódica. Desta forma, a SSA divide entre os dois autovalores correspondentes a energia do processo. VAUTARD & GHIL (1989) usaram a SSA para predizer se esses valores eram próximos e extrair possíveis componentes periódicas em uma decomposição do sinal em estudo.

Um cuidado adicional são com os erros oriundos da solução numérica para o cálculo dos autovalores. Este pode levar a autovalores praticamente iguais em estruturas não-periódicas. No ruído branco, p. ex., todas as componentes possuem autovalores pequenos e estatisticamente indistintos. Desta forma, devem ser adotados critérios para indicar os autovalores correlacionados a estruturas periódicas.

## A estimação da ordem M

O principal problema em determinar a dimensão M adequada é que dois parâmetros estão envolvidos  $\tau_s$  (período de amostragem) e  $\tau_w$  (tamanho da janela). Considera-se um conjunto de N amostras da série contínua amostrada com período  $\tau_s$  e contaminada com um ruído branco  $w(t)$  com variância  $\sigma_w^2$ . Assume-se que N é grande o suficiente para fornecer uma boa estimativa da matriz  $R_x$  do processo  $x$ .

De acordo com as referências [BROOMHEAD & KING, 1986; VAUTARD & GHIL, 1989], a matriz de covariância para a série discreta é dada por:

$$R_x = R^{(T)} + R^{(w)} \quad \text{eq. 4}$$

onde  $R^{(T)}$  é a matriz de covariância real ou intrínseca do sinal amostrado e  $R^{(w)}$  é a matriz diagonal de ruído com elementos  $\sigma_w^2$ . Os autovalores estimados para esse problema são:

$$\lambda_k = \lambda_k^{(T)} + (1/M)\sigma_w^2 \quad \text{eq. 5}$$

onde  $\lambda_k^{(T)}$  representa a estimativa real de autovalor para o sinal amostrado, convergindo para um autovalor do sinal contínuo assim que M tende para infinito, mantendo  $\tau_s$  fixo,  $M = \tau_w / \tau_s$ . Como  $\lambda_k^{(T)}$  só depende do sinal, a estimativa de  $\{\lambda_k\}$ ,  $k = 1, 2, \dots, M$  fica polarizada pela componente de ruído. Pela eq. 5 o nível de ruído ocorre em  $\sigma_w^2/M$  e depende do período de amostragem  $\tau_s$  para uma janela fixa de comprimento  $\tau_w$ .

Observa-se com isso, que o número de autovalores significantes também depende do período de amostragem. Reduzindo o período de amostragem, mantendo um  $\tau_w$  fixo, resulta em uma quantidade maior de autovalores acima do nível de ruído.

Na prática, os dados não devem ser amostrados a taxas elevadas pois existe uma escala de tempo abaixo da qual o ruído não é decorrelacionado.

Considerando que  $\tau_s$  está fixo e as amostras das diversas medidas sucessivas disponíveis, o problema então se reduz à escolha de  $\tau_w$ , ou de forma equivalente, a de M. O período de amostragem  $\tau_s$  e o tempo total da série temporal  $N\tau_s$  impõem limites na largura de banda do sinal disponível para a análise.

Se M é um valor muito pequeno, uma resolução grosseira pode fazer com que diversos picos em uma região do espectro de  $x$  se unam. Por outro lado, quando existir uma oscilação intermitente, representada como um pico largo, valores grandes de M (alta resolução) irão dividir esse pico em muitas componentes de frequências vizinhas.

Denota-se o vetor  $\sim E^k(f)$  à Transformada de Fourier reduzida de  $E^k$ , ou seja:

$$\sim E^k(f) = \sum_{j=1}^M E_j^k e^{2\pi i j f} \quad \text{eq. 6}$$

A expressão  $\sim E^k(f)$  é também a função de resposta de um filtro transformando  $x$  em sua  $k^{\text{ésima}}$  PC.  $\sim E^k(f)$  é uma soma de funções periódicas de frequência  $f$  com períodos  $1, 1/2, 1/3, \dots, 1/M$ . Desta forma, a resolução espectral é  $1/M$ .

A equação eq. 6 também mostra que os filtros são incapazes de isolar picos com frequência inferior a  $1/M$ , ou seja, períodos maiores que  $M$ . Dado um pico no espectro de potência  $P_x(f)$  de  $x$  – com densidade espectral máxima em  $f_0$  e largura  $2\delta f$ , o valor de  $M$  que melhor indicará o processo é relacionado pela expressão abaixo:

$$\frac{1}{f_0} \leq M \leq \frac{1}{2\delta f} \quad \text{eq. 7}$$

O tamanho da janela deve ser escolhido entre o período da oscilação ( $f_0$ ) e o tempo médio da sua duração. Na prática, esses valores não podem ser estimados a priori mas o SSA terá sucesso em isolar oscilações intermitentes com valores tipicamente na faixa de  $(M/5, M)$ . Uma limitação do processo é que so podem ser separadas frequências se o pico esta separado de mais de  $1/M$ .

Do ponto de vista espectral, as EOF's correspondem a filtros de média móvel adaptativos com os dados de entrada. O espectro de potência  $P_k(f)$  das Componentes Principais ( $a^k$ ) é:

$$P_k(f) = P_x(f) \left| \sim E^k(f) \right|^2. \quad \text{eq. 8}$$

O valor de  $P_k(f)$  poderá ser re-escrito para qualquer frequência  $f$ , como a soma dos espectros das componentes principais é idêntica ao espectro de potencia da série  $x$ , ou seja:

$$P_x(f) = \frac{1}{M} \sum_{k=1}^M P_k(f) \quad \text{eq. 9}$$

essa expressão fornece uma informação importante de espectros empilhados (*stack spectra*) pela soma das contribuições das diversas componentes principais.

As restrições de ortogonalidade do problema fornecem a identidade

$$1 = \frac{1}{M} \sum_{k=1}^M \left| \sim E^k(f) \right|^2 \quad \text{eq. 10}$$

## SIMULAÇÃO

O algoritmo será testado com o somatório de duas cossenoides e ruído gaussiano como descrito pelas equações abaixo:

$$x_n = y_n + w_n \quad \text{eq. 11}$$

$$y_n(i) = 2 \cos(\Omega_1 i + \phi_1) + \cos(\Omega_2 i + \phi_2) \quad \text{eq. 12}$$

onde:  $w_n(i)$  = ruído branco gaussiano com variância  $\sigma^2$  e média zero;

$$\Omega_1 = (2/20)\pi \text{ e } \Omega_2 = (2/7)\pi$$

As variâncias do ruído gaussiano para os casos de testes serão adotadas como sendo (Figura 2): a)  $\sigma^2 = 0,5$ ; b)  $\sigma^2 = 2,5$ ; c)  $\sigma^2 = 4,5$ ; d.)  $\sigma^2 = 6,5$ . Para cada análise, o processamento será realizado com 150 pontos. As dimensões para a análise de SSA são com  $M=20$  e  $M=40$ .

A Figura 3 representa os autovalores ordenados de forma decrescente de amplitude. Podem ser observados dois pares de autovalores indicando duas possíveis oscilações no sistema em estudo.

A Figura 4 mostra os autovalores reordenados no eixo x pela frequência máxima do espectro reduzido do autovetor associado. As frequências apresentadas no eixo x correspondem a uma conversão de  $1/2\pi$  para as frequências originais da simulação.

Observa-se na Figura 4 que os sinais significativos estão acima do platô fornecido pela maior parte dos autovalores. Os pares de autovalores além de se aproximarem em amplitude possuem frequências próximas. Este é um forte indicativo de comportamento periódico no sinal em estudo.

Um problema adicional é determinar a quantidade de autovalores significativos sem o conhecimento prévio dos sinais em estudo. Observa-se nas Figura 4 que a medida que os sinais ficam mais ruidosos a distinção entre sinais e ruídos fica mais difícil.

Therrien (1992) cita o trabalho de Wax e Kailath que desenvolveram a formulação do Critério de Informação de Akaike (*Akaike Information Criterion* – AIC) e do Comprimento Descritivo Mínimo (*Minimum Description Length* – MDL) aplicável ao problema. No contexto do problema de sinais imersos em ruído, esses critérios podem ser explícitos nas seguintes expressões :

$$AIC(p) = -2N(M-p)\ln(\rho(p)) + 2P(2M-p) \quad \text{eq. 13}$$

$$MDL(p) = -N(M-p)\ln(\rho(p)) + P(2M-p)\ln(N)/2 \quad \text{eq. 14}$$

onde: N é o número de dados

M é o número total de autovalores p é a quantidade de autovalores significantes

$\rho(p)$  é o quociente entre a média geométrica e aritmética dos autovalores (eq. 15)

$$\rho(p) = \frac{\sqrt[M-p]{(\lambda_{p+1}\lambda_{p+2}\cdots\lambda_M)}}{(\lambda_{p+1} + \lambda_{p+2} + \cdots + \lambda_M) / (M-p)} \quad \text{eq. 15}$$

A Figura 5 apresenta os sinais reconstruídos usando a abordagem desenvolvida no parágrafo anterior, considerando o conjunto de autovalores significantes como o menor dentre os indicados pelos processos AIC e MDL (TABELA 1).

A Figura 5 mostra as reconstruções do sinal original com o uso das primeiras componentes da decomposição. Para o caso de  $\sigma^2 = 6,5$  a reconstrução foi de apenas uma

estrutura periódica sendo que o segundo par de autovalores considerado como pertencente ao conjunto de ruído.

## DISCUSSÃO E CONCLUSÃO

Observa-se que o diagrama da Figura 4 apresenta mais informações que a Figura 3. Essa nova informação no eixo x é um indicativo das principais frequências que ocorrem nos dados em estudo.

Os dados da Tabela 1 fornecem uma aproximação da quantidade de autovalores que compõem o sub-espço de sinais. Observando os valores indicados na tabela, conclui-se que a medida que o ambiente fica mais ruidoso, a capacidade de separar os autovalores para o sub-espço de sinais e ruído diminui. No caso extremo, houve a reconstrução do sub-espço de sinais com apenas um par de autovalores indicando apenas uma oscilação na série inicial.

Em modelos mais simples, como somas de senoides, a separação dos autovalores por uma dado limiar é a forma clássica para classificação. Em ambientes complexos, por exemplo sistemas caóticos, devem ser adotados critérios mais significativos, por exemplo teste de hipótese nula [THEILER & EUBANK (1992) ], para a classificação.

O presente modelo esta sendo aplicado a análise dos sinais de ultra-som utilizados para a investigação do tecido biológico. Em especial, tecidos como o do fígado apresentam uma estrutura morfológica bem definida (hepatócito) e estruturas difusas, produzindo sinais com características periódicas acrescidas de um “ruído” aleatório. Cabe resaltar que esse “ruído” também é uma informação significativa para a caracterização do tecido hepático.

## REFERÊNCIAS BIBLIOGRÁFICAS

- BROOMHEAD, D.S., KING, Gregory (1986) “Extracting Qualitative Dynamics From Experimental Data” *Physica D* 20, North-Holland Amsterdam.
- GHIL, Michael, YIOU, Pascal “Spectral Methods: What They Can and Cannot Do for Climatic Times Series” In D. Anderson and J. Willebrand (Eds.), *Decadal Climatic Variability: Dynamics and Predictability*, Elsevier.
- JOHNSON, Don (1982) “The Application of Spectral Estimation Methods to Bearing Estimation Problems” *Proceeding of the IEEE* vol. 70 no. 9.
- KAY, Steven (1988) “Modern Spectral Estimation” Prentice Hall, Englewood Cliffs, New Jersey.
- THEILER, J., EUBANK, S. (1992) “Testing For Nonlinearity in Times Series: The Surrogate Data” *Physica D* 58, North-Holland Amsterdam.
- TERRIEN, Charles W., (1992) “Discrete Random Signals and Statistical Signal Processing” Prentice Hall Signal Processing Series, Englewood Cliffs, New Jersey.

THOMSON. David (1982) "Spectrum Estimation and Harmonic Analysis" Proceeding of the IEEE vol. 70 no. 9 september .

VAUTARD, R., GHIL, M. (1989) "Singular Spectrum Analysis in Nonlinear Dynamics, With Applications to Paleoclimatic Time Series" Physica D 35, North-Holland Amsterdam.

**ANEXO – Figuras e Tabelas**

Tabela 1 – Tabela com os resultados do cálculo da dimensão do sub-espço de sinais usando os critérios AIC e MDL.

ORDEM	$\sigma^2 = 0,5$		$\sigma^2 = 2,5$		$\sigma^2 = 4,5$		$\sigma^2 = 6,5$	
	M=20	M=40	M=20	M=40	M=20	M=40	M=20	M=40
AIC	9	18	8	10	4	8	4	2
MDL	8	15	7	7	4	4	2	2

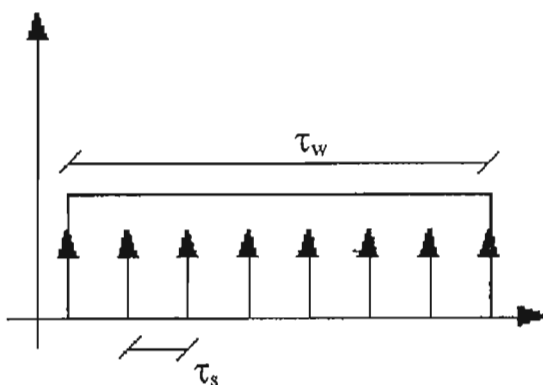


Figura 1 – Representação de  $\tau_s$  para uma janela fixa de comprimento  $\tau_w$ .  
M é o número de pontos dentro da janela  $\tau_w$ .



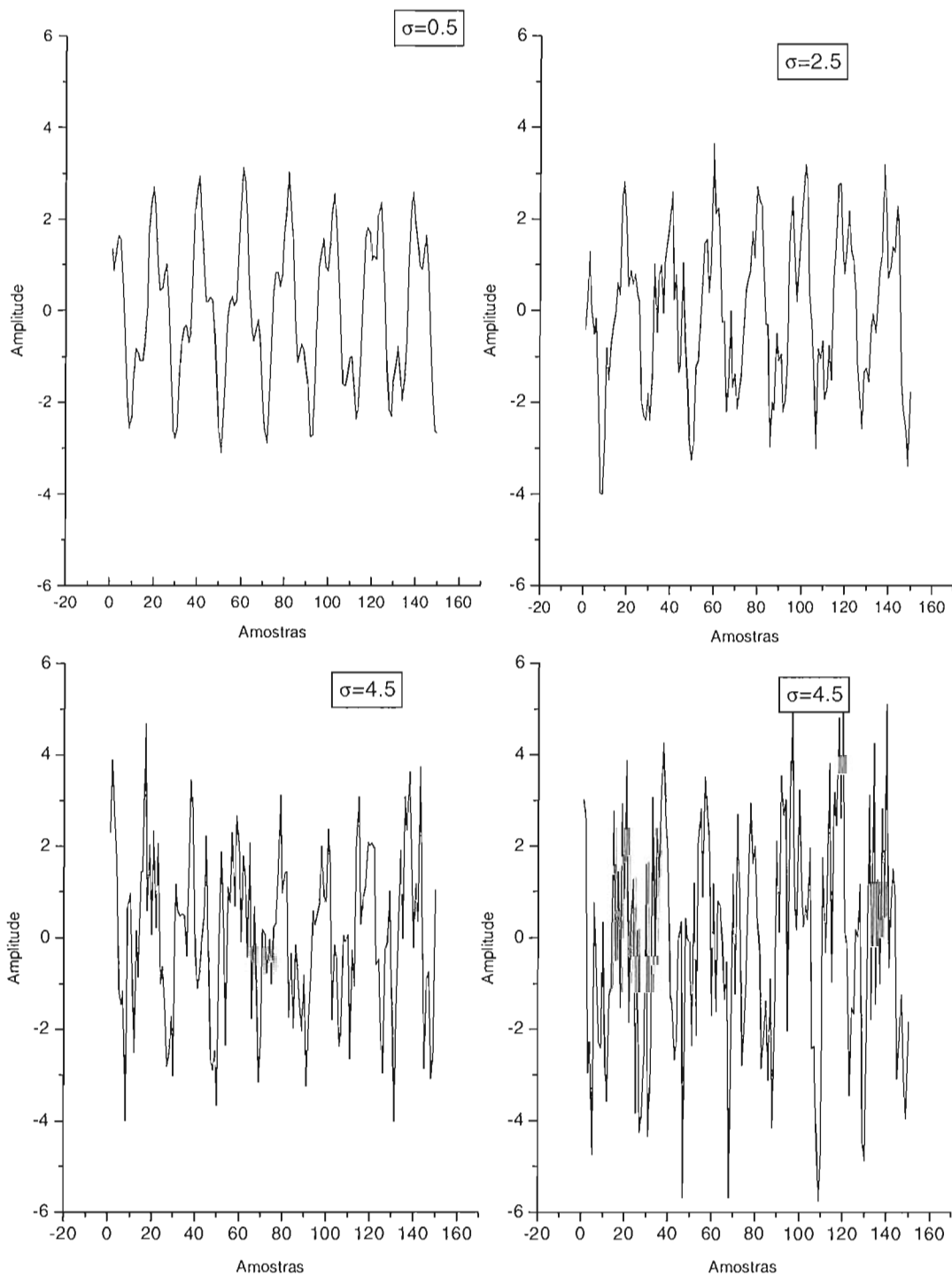


Figura 2 – Cada um dos sinais apresenta níveis de contaminação pelo ruído em valores crescentes.

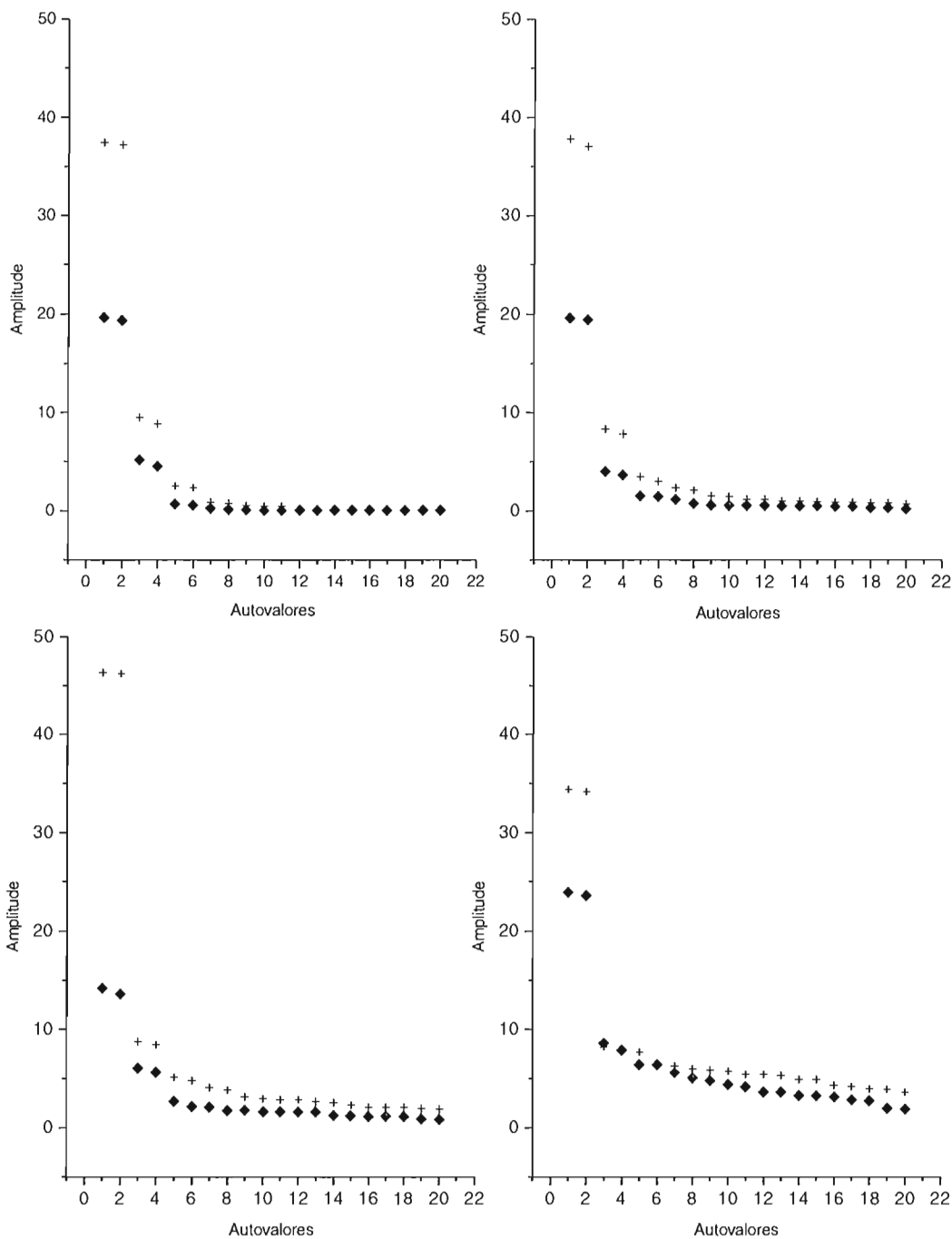


Figura 3 – Espectro singular para os sinais considerados. As marcas ‘+’ correspondem aos autovalores com M=20 e as ‘♦’ com M=40.

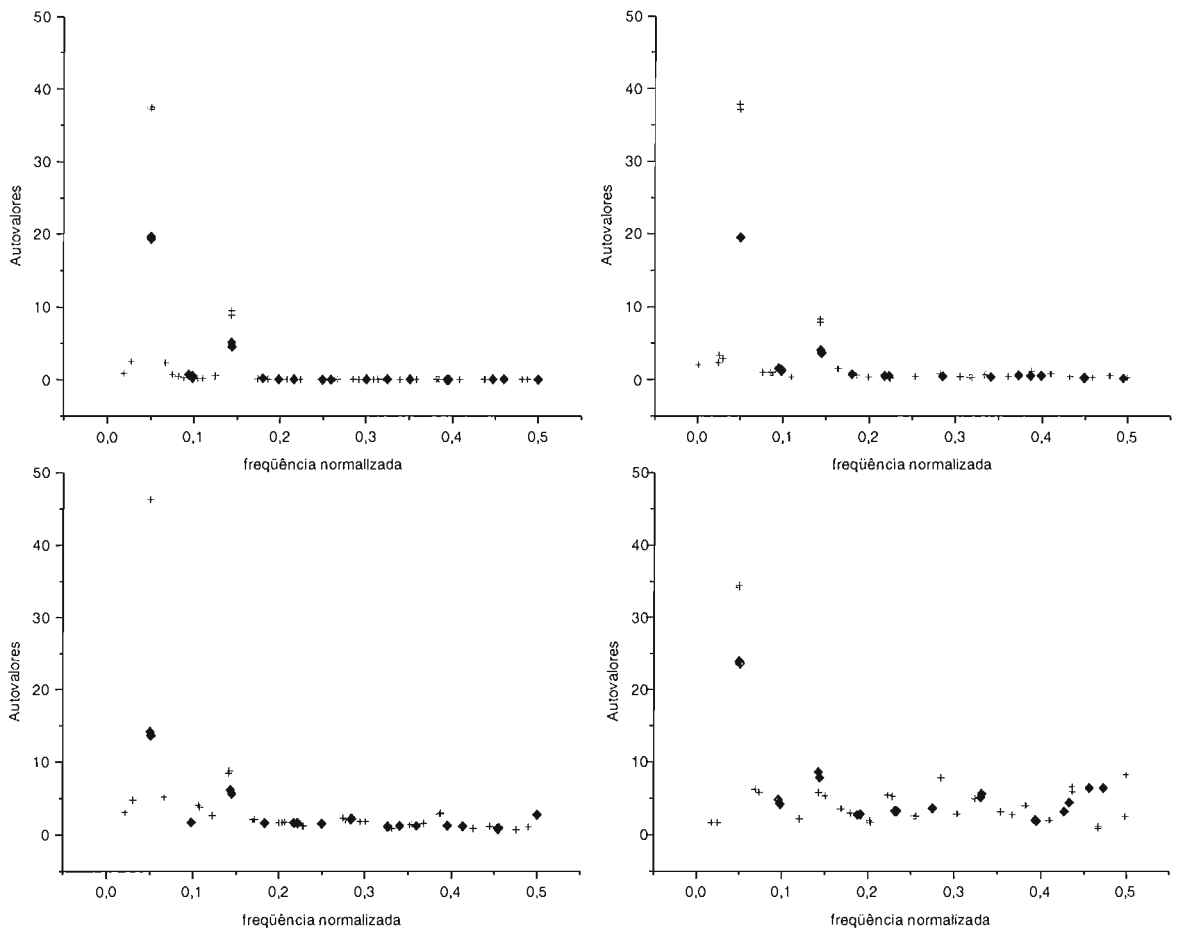


Figura 4 – Os autovalores são ordenados pela frequência máxima do autovetores correspondentes. As marcas '+' correspondem aos autovalores com  $M=20$  e as '◆' com  $M=40$ .

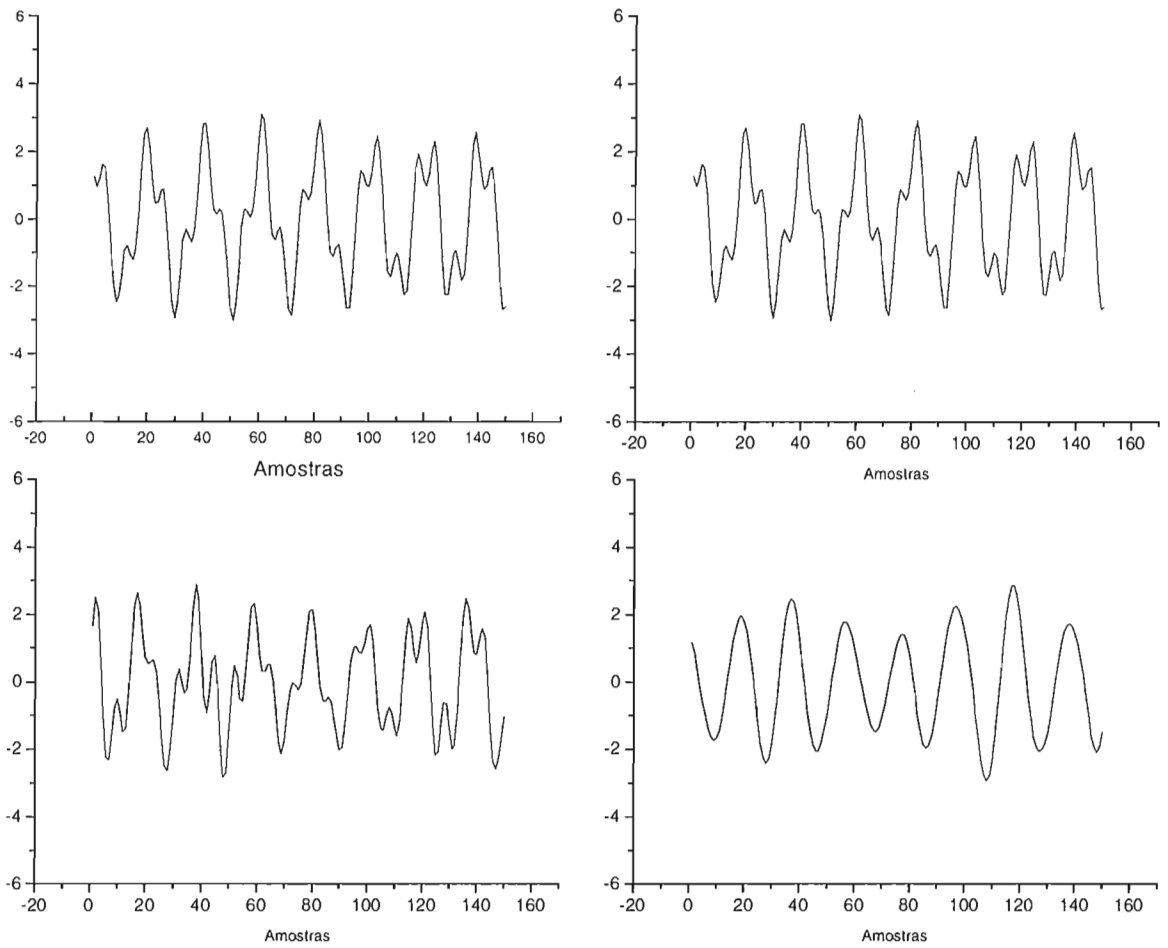


Figura 5 – Sinal reconstruído apartir das informações de ordem AIC e MDL.