Análise de Séries Temporais
VII. Métodos estatísticos no domínio da frequência

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Julho, 2023

O problema da discriminação entre explosões de mineração e terremotos é um representante razoável para o problema de discriminação.

Esse problema é de fundamental importância para o monitoramento de um tratado abrangente de proibição de testes.

Os problemas de classificação de séries temporais não se restringem a aplicações geofísicas, mas ocorrem sob muitas e variadas circunstâncias em outros campos.

Tradicionalmente, a detecção de um sinal incorporado em uma série de ruídos é analisada na literatura de engenharia por técnicas estatísticas de reconhecimento de padrões.

As abordagens históricas para o problema da discriminação entre diferentes classes de séries temporais podem ser divididas em duas categorias distintas.

A abordagem de otimização, como encontrada na literatura de engenharia e estatística, faz suposições Gaussianas específicas sobre as funções de densidade dos grupos separados e, em seguida, desenvolve soluções que satisfazem critérios de erros mínimos bem definidos.

Uma segunda classe de técnicas, que pode ser descrita como uma abordagem de extração de recursos, prossegue-se de maneira mais heurística, observando quantidades que tendem a ser bons discriminadores visuais para populações bem separadas e que têm alguma base na teoria ou intuição física.

Para séries temporais multivariadas mais longas que possam ser consideradas estacionárias após a subtração da média comum, a abordagem do domínio da frequência será mais fácil computacionalmente porque o vetor $n \times p$ dimensional será reduzido para separar os cálculos feitos nas transformadas discretas de Fourier (DFT's) p-dimensionais.

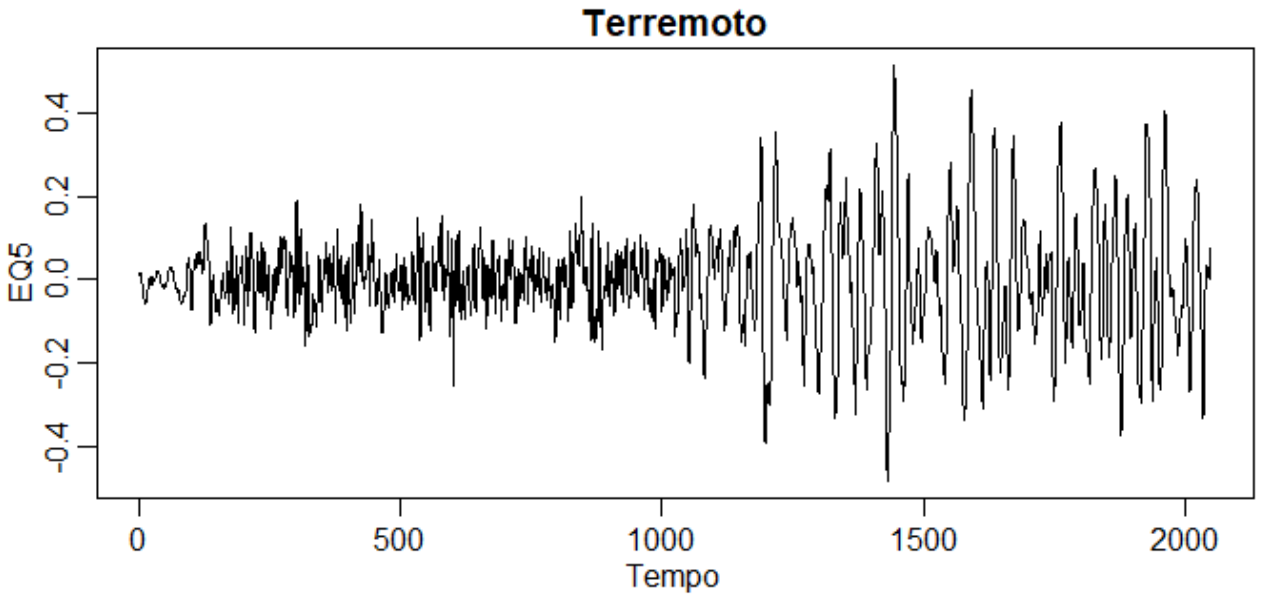
Isso acontece devido à independência aproximada das transformadas discretas de Fourier, $X(\omega_k)$, $0 \leq \omega \leq 1$, uma propriedade que usamos frequentemente anteriormente.

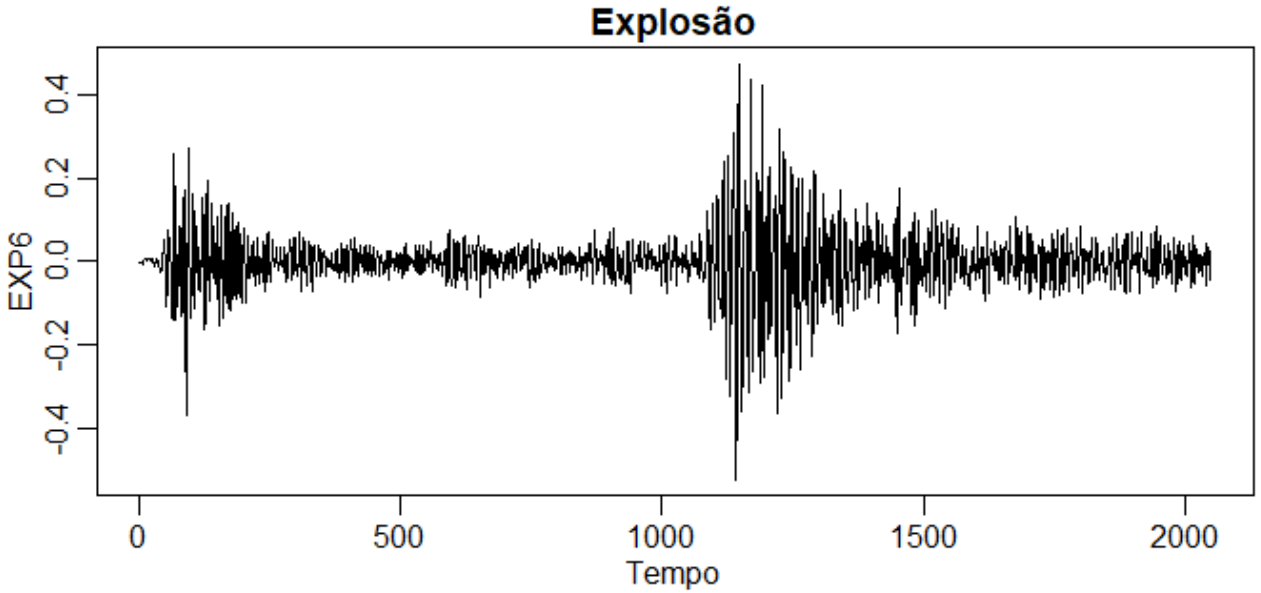
Por fim, as propriedades de agrupamento de medidas como informações sobre discriminação e estatísticas baseadas na verossimilhança podem ser usadas para desenvolver medidas de disparidade para agrupar séries temporais multivariadas.

O problema geral de classificação de uma série temporal vetorial ocorre da seguinte maneira. Observamos uma série temporal X conhecida por pertencer a uma das g populações, denotadas por $Pi_1, \Pi_2, \cdots, \Pi_g$.

O problema geral é atribuir ou classificar essa observação em um dos g grupos da melhor maneira possível.

Um exemplo pode ser as $g = 2$ populações de terremotos e explosões. Gostaríamos de classificar o evento desconhecido, mostrado no gráfico a seguir, como pertencente às populações de terremotos Π_1 ou explosões Π_2. Exemplo I.7 Terremotos e explosões.


Terremoto
EQ5
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Tempo


Explosão
EXP6
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Tempo

Para resolver esse problema, precisamos de um critério de otimização que leve a uma estatística $T(X)$ que possa ser usada para atribuir o evento às populações de terremotos ou explosões.

Para medir o sucesso da classificação, precisamos avaliar os erros que podem ser esperados no futuro em relação ao número de terremotos classificados como explosões (alarmes falsos) e o número de explosões classificadas como terremotos (sinais perdidos).

O problema pode ser formulado assumindo que a série observada X tem uma função de densidade $p_i(x)$ quando a série observada é da população Π_i para $i = 1, \cdots, g$. Em seguida, particione o espaço abrangido pelo processo $n \times p$-dimensional x em g regiões mutuamente exclusivas $R_1, R_2, \cdots, R_g$ tal que, se x cair em R_i, atribuímos x à população Π_i.

À probabilidade de classificação incorreta é definida como a probabilidade de classificar a observação na população Π_j quando ela pertence a Π_i, para $j \neq i$ e seria dada pela expressão

$$P(j \mid i) = \int_{R_j} p_i(x)\mathrm{d}x \cdot .$$

A probabilidade do erro total depende também das probabilidades a priori, por exemplo, $\pi_1, \pi_2, \cdots, \pi_g$ de x pertencer a um dos g grupos. Por exemplo, a probabilidade de uma observação x se originar de Π_i e depois ser classificada em Π_j é obviamente $\pi_i P(j \mid i)$ e a probabilidade total de erro se torna

$$P_\epsilon = \sum_{i=1}^{g} \pi_i \sum_{j \neq i} P(j \mid i) \cdot$$

Embora os custos não tenham sido incorporados na expressão acima, é possível fazê-lo multiplicando $P(j \mid i)$ por $C(j \mid i)$, o custo de atribuir uma série da população Π_i a Π_j.

O erro geral P_ϵ é minimizado classificando x em Π_i se

$$\frac{p_i(x)}{p_j(x)} > \frac{\pi_j}{\pi_i},$$

para todos os $j \neq i$ (ver, por exemplo, Anderson, 1984).

Uma quantidade de interesse, da perspectiva Bayesiana, é a probabilidade a posteriori de uma observação pertencer à população Π_i, condicionada à observação de x, digamos

$$P(\Pi_i \mid x) = \frac{\pi_i p_i(x)}{\sum_j \pi_j p_j(x)}.$$

O procedimento que classifica x na população Π_i para a qual a probabilidade a posteriori é maior é equivalente ao critério implícito anterior. As probabilidades a posteriori dão uma idéia intuitiva plausível das chances relativas de pertencer a cada uma das populações.

Muitas situações ocorrem, como na classificação de terremotos e explosões, nas quais existem apenas $g = 2$ populações de interesse. Para duas populações, o Lema de Neyman – Pearson implica, na ausência de probabilidades a priori, classificar uma observação em Π_1 quando

$$\frac{p_1(x)}{p_2(x)} > K,$$

minimiza cada uma das probabilidades de erro para um valor fixo do outro.

A regra é idêntica à regra de Bayes do erro geral P_ϵ, o qual é minimizado classificando x em Π_i quando $p_i(x)/p_j(x) > \pi_j/\pi_i$, para todos os $j \neq i$. A regra acima é idêntica à regra de Bayes quando $K = \pi_2/\pi_1$.

A teoria apresentada acima assume uma forma simples quando o vetor x tem uma distribuição normal p-variada com vetores de média μ_j e matrizes de covariância Σ_j sob Π_j para $j = 1, 2, \cdots, g$.

Nesse caso, basta usar

$$p_j(x) = (2\pi)^{-p/2}|\Sigma_j|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j)\right\}.$$

As funções de classificação são expressas por quantidades proporcionais aos logaritmos das densidades, por exemplo,

$$g_j(x) = -\frac{1}{2}\ln\left(|\Sigma_j|\right) - \frac{1}{2}x^\top\Sigma_j^{-1}x + \mu_j^\top\Sigma_j^{-1}x - \frac{1}{2}\mu_j^\top\Sigma_j^{-1}\mu_j + \ln\left(\pi_j\right).$$

Nas expressões que envolvem a log-verossimilhança, em geral ignoraremos os termos que envolvem a constante $-2\ln(2\pi)$. Podemos então atribuir uma observação x à população Π_i sempre que

$$g_i(x) > g_j(x)$$

para $i \neq j$, $i = 1, 2, \cdots, g$ e a probabilidade a posteriori tem a forma

$$p(\Pi_i \,|\, x) = \frac{\exp\left(g_i(x)\right)}{\sum_j \exp\left(g_j(x)\right)}.$$

Uma situação comum que ocorre nas aplicações é a classificação sob a premissa de normalidade multivariada e matrizes de covariância iguais; ou seja, $\Sigma_1 = \Sigma_2 = \Sigma$.

Então, o critério de atribuição de uma observação x à população Π_i pode ser expresso em termos da função discriminante linear como

$$\begin{aligned} d_i(x) &= g_1(x) - g_2(x) = (\mu_1 - \mu_2)^\top \Sigma^{-1} x - \\ &\quad \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 + \mu_2) + \ln(\pi_1/\pi_2), \end{aligned}$$

onde classificamos em Π_1 ou Π_2 de acordo com $d_i(x) \geq 0$ ou $d_i(x) < 0$.

A função discriminante linear é claramente uma combinação de variáveis normais e, para o caso $\pi_1 = \pi_2 = 0.5$, terá média $D^2/2$ sob Π_1 e média $-D^2/2$ sob Π_2, com variâncias dadas por D^2 nas duas hipóteses, em que

$$D^2 = (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)$$

é a distância de Mahalanobis entre os vetores médios μ_1 e μ_2. Nesse caso, as duas probabilidades de classificação incorreta são

$$P(1|2) = P(2|1) = \Phi(-D/2),$$

e o desempenho está diretamente relacionado à distância de Mahalanobis.

Para o caso de matrizes de covariâncias diferentes, a função discriminante assume uma forma diferente, com a diferença $g_1(x) - g_2(x)$ na forma

$$\begin{aligned} d_q(x) &= -\frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - \frac{1}{2}x^\top(\Sigma_1^{-1} - \Sigma_2^{-1})x + \\ &\quad (\mu_1^\top\Sigma_1^{-1} - \mu_2^\top\Sigma_2^{-1})x + \ln\left(\frac{\pi_1}{\pi_2}\right), \end{aligned}$$

para $g = 2$ grupos.

Essa função discriminante difere do caso de covariâncias iguais no termo linear e em um termo quadrático não linear envolvendo as diferentes matrizes de covariância.

A distribuição teórica não é tratável para o caso quadrático, portanto, nenhuma expressão conveniente está disponível para as probabilidades de erro da função discriminante quadrática.

Uma dificuldade em aplicar a teoria acima a dados reais é que os vetores médios μ_j e as matrizes de covariância Σ_j raramente são conhecidos.

Alguns problemas de engenharia, como a detecção de um sinal no ruído branco, assumem que os vetores médios e os parâmetros de covariância são conhecidos exatamente e isso pode levar a uma solução ótima.

Na situação multivariada clássica, é possível coletar uma amostra de N_i vetores de treinamento do grupo Π_i, por exemplo, $x_{i,j}$, para $j = 1, \cdots, N_i$ e usá-los para estimar os vetores médios e as matrizes de covariância para cada um dos grupos $i = 1, 2, \cdots, g$; ou seja, basta escolher $\overline{x}_i$ e

$$S_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{i,j} - \overline{x}_i\right)\left(x_{i,j} - \overline{x}_i\right)^\top$$

como estimadores para μ_i e Σ_i, respectivamente. Caso as matrizes de covariância sejam consideradas iguais, basta usar o estimador agrupado

$$S = \frac{1}{\sum_i N_i - g} \sum_i (N_i - 1) S_i.$$

Para o caso de uma função discriminante linear, podemos usar

$$\widehat{g}_i(x) = \overline{x}_i^\top S^{-1} x - \frac{1}{2}\overline{x}_i^\top S^{-1}\overline{x}_i + \log\left(\pi_i\right),$$

como um estimador simples para $g_i(x)$.

Para amostras grandes, $\overline{x}_i$ e S convergem para μ_i e Σ em probabilidade, de modo que $\widehat{g}_i(x)$ converge em distribuição para $g_i(x)$ nesse caso.

O procedimento funciona razoavelmente bem para o caso em que cada N_i; $i = 1, \cdots, g$ são grandes, em relação ao comprimento da série n, um caso relativamente raro na análise de séries temporais. Por esse motivo, recorreremos ao uso de aproximações espectrais para o caso em que os dados são fornecidos como séries temporais longas.

O desempenho das funções discriminantes amostrais podem ser avaliadas de várias maneiras diferentes. Se os parâmetros populacionais forem conhecidos, a distância de Mahalanobis e as probabilidades de classificação incorreta poderão ser avaliadas diretamente. Se os parâmetros forem estimados, a distância estimada de Mahalanobis $\widehat{D}^2$ pode ser substituída pelo valor teórico em amostras muito grandes.

Outra abordagem é calcular as taxas de erro aparentes usando o resultado da aplicação do procedimento de classificação às amostras de treinamento. Se $n_{i,j}$ denota o número de observações da população Π_j classificadas em Π_i, as taxas de erro amostral podem ser estimadas pela razão

$$\widehat{P}(i \mid j) = \frac{n_{i,j}}{\sum_i n_{i,j}}, \qquad i \neq j.$$

Se as amostras de treinamento não forem grandes, esse procedimento pode ser tendencioso e uma opção de reamostragem, como validação cruzada ou o bootstrap, pode ser empregada.

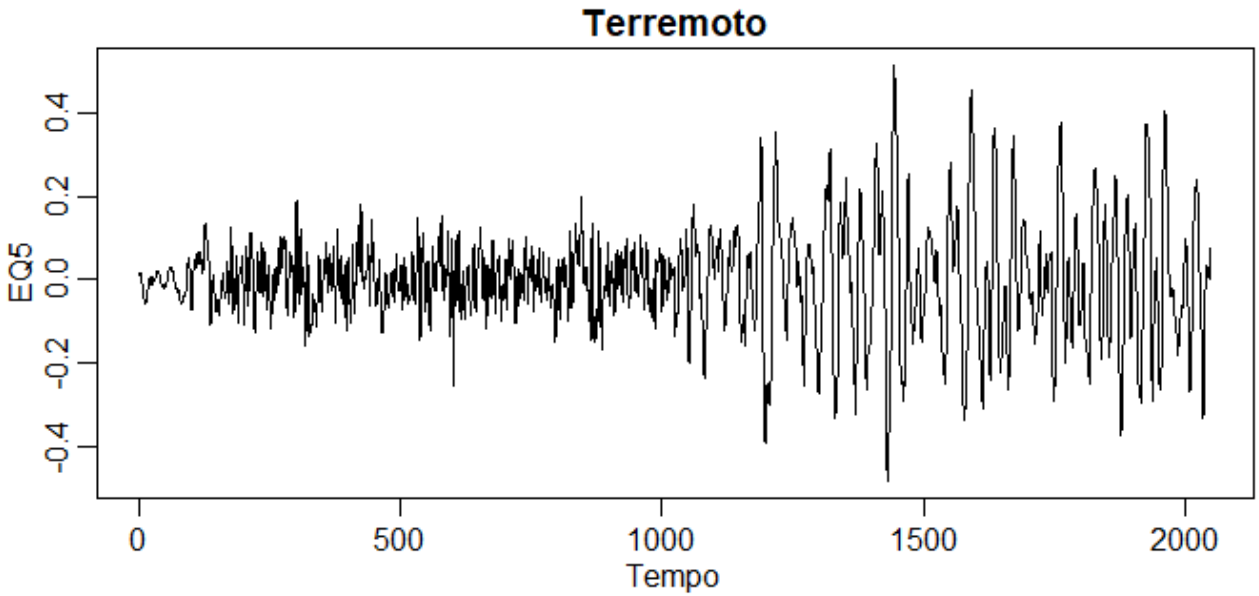
Uma versão simples da validação cruzada é o jackknife proposto por Lachenbruch and Mickey (1968), que sustenta a observação a ser classificada, derivando a função de classificação das demais observações.

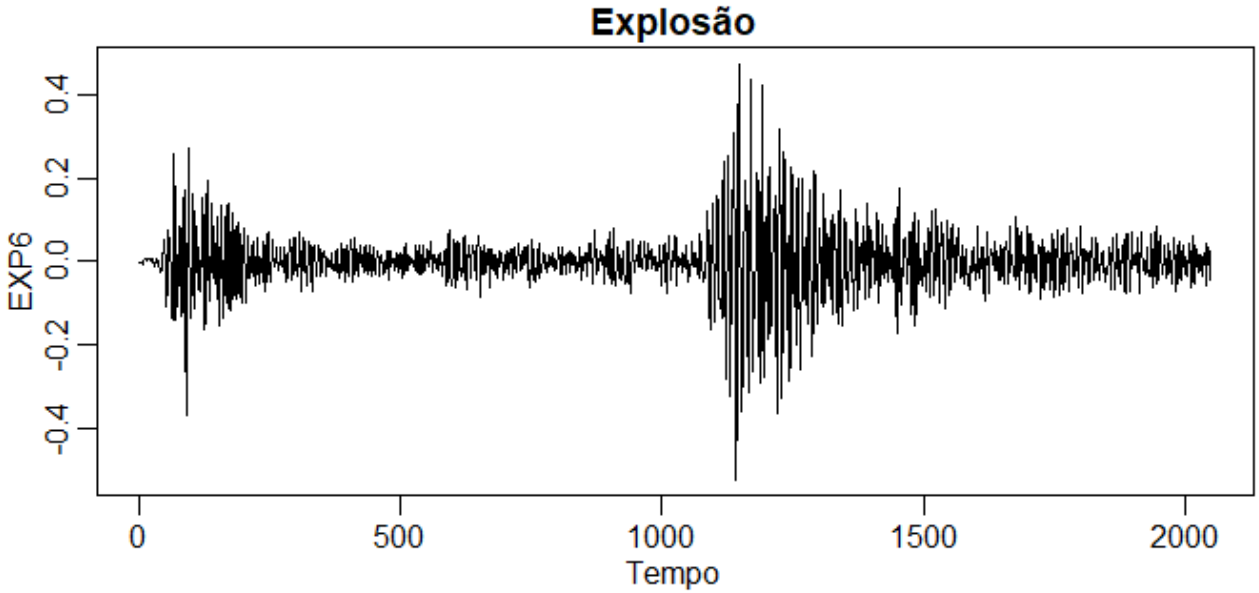
A repetição deste procedimento para cada um dos membros da amostra de treinamento e o cálculo da taxa de erro amostral acima para as amostras de validação levam a melhores estimadores das taxas de erro.

Exemplo VII.10. Análise discriminante usando amplitudes.

Os logaritmos (base 10) das amplitudes máximas pico a pico dos componentes P e S, denotados por $\log_{10}(P)$ e $\log_{10}(S)$, podem ser considerados vetores de característica bidimensionais, por exemplo, $x = (x_1, x_2)^\top = \big(\log_{10}(P), \log_{10}(S)\big)^\top$, de uma população normal bivariada com médias e covariâncias diferentes.

Os dados originais, de Kakizawa et al. (1998), são mostrados na figura anterior. Os dados incluem o evento Novaya Zemlya (NZ) de origem desconhecida. A tendência dos terremotos de terem valores mais altos para $\log_{10}(S)$, em relação ao $\log_{10}(P)$, foi notada por muitos e o uso do logaritmo da razão, ou seja, $\log_{10}(P) - \log_{10}(S)$ em algumas referências, por exemplo, Lay (1997) é um indicador tácito de que uma função linear dos dois parâmetros será um discriminante útil.


Terremoto
EQ5
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Tempo


Explosão
EXP6
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Tempo

As médias amostrais foram

$$\bar{x}_1 = (0.3477384, 1.0244672)$$

e

$$\bar{x}_2 = (0.9222803, 0.9930151)$$

e as matrizes de covariância amostrais foram

$$S_1 = \begin{pmatrix} 0.025907196 & -0.007064702 \\ -0.007064702 & 0.010202513 \end{pmatrix}$$

e

$$S_2 = \begin{pmatrix} 0.025197761 & -0.000760573 \\ -0.000760573 & 0.010342222 \end{pmatrix}$$

com a matriz de covariância agrupada dada por

$$S = \begin{pmatrix} 0.025552479 & -0.003912638 \\ -0.003912638 & 0.010272368 \end{pmatrix}.$$

Embora as matrizes de covariância não sejam iguais, tentamos a função discriminante linear que produz, com probabilidades a priori iguais $\pi_1 = \pi_2 = 0.5$, as funções discriminantes amostrais

$$\widehat{g}_1(x) = 29.68321x_1 + 106.1547x_2 - 59.53698$$

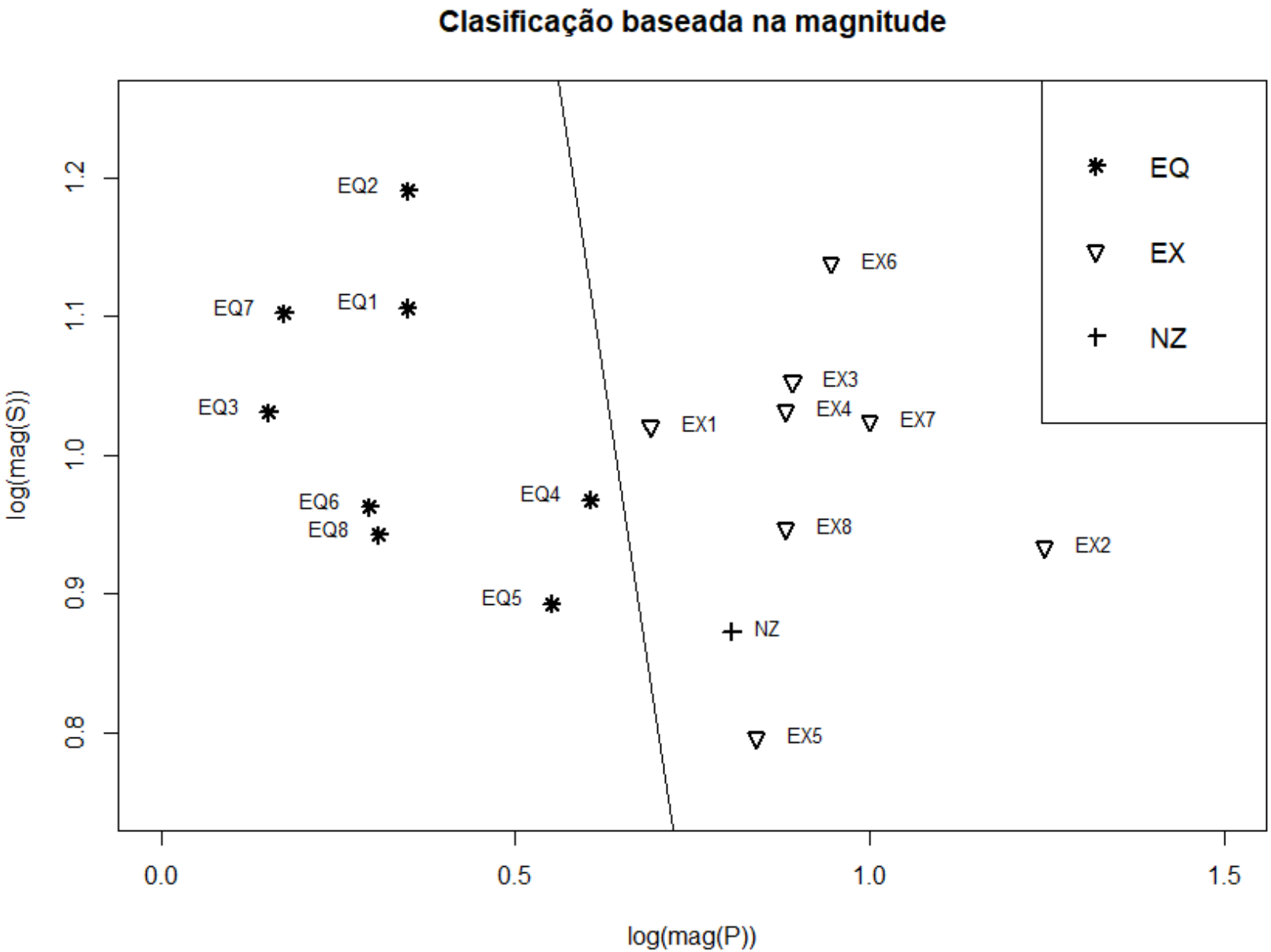
e

$$\widehat{g}_2(x) = 51.95275x_1 + 112.8352x_2 - 80.5039$$

com a função discriminante linear estimada

$$\widehat{g}_2(x) = -23.3795x_1 - 5.843191x_2 + 20.74047.$$

As probabilidades a posteriori jackknife de ser um terremoto para o grupo de terremotos variavam de 0.621 a 1.000, enquanto as probabilidades de explosão para o grupo de explosão variavam de 0.717 a 1.000. O evento desconhecido, NZ, foi classificado como explosão, com probabilidade a posteriori 0.960.


Clasificação baseada na magnitude
log(mag(S))
log(mag(P))
0.0
0.5
1.0
1.5
0.8
0.9
1.0
1.1
1.2
EQ2
EQ7
EQ1
EQ3
EQ6
EQ8
EQ4
EQ5
EX6
EX3
EX4
EX7
EX1
EX8
EX2
NZ
EX5
EQ
EX
NZ

A abordagem de extração de recursos geralmente funciona bem para discriminar entre classes de séries univariadas ou multivariadas quando existe um vetor de baixa dimensão simples que parece capturar a essência das diferenças entre as classes.

Ainda parece sensato, no entanto, desenvolver métodos ótimos de classificação que explorem as diferenças entre as médias multivariadas e as matrizes de covariância no caso de séries temporais.

Tais métodos podem basear-se na aproximação de Whittle à log-verossimilhança.

Nesse caso, assume-se que os DFTs vetoriais, $X(\omega_k)$, sejam aproximadamente normais, com médias $M_j(\omega_k)$ e matrizes espectrais $f_j(\omega_k)$ para a população Π_j nas frequências $\omega_k = k/n$, para $k = 0, 1, \cdots, [n/2]$ e são aproximadamente não correlacionados em diferentes frequências, digamos, ω_k e ω_l, para $k \neq l$.

Em seguida, escrevendo as densidades normais complexas, leva ao critério,

$$g_j(X) = \ln(\pi_j) - \sum_{0<\omega_k<1/2} \Big(\ln(|f_j(\omega_k)|) + X^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k) - 2M_j^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k) + M_j^*(\omega_k) f_j^{-1}(\omega_k) M_j(\omega_k) \Big),$$

onde a soma ultrapassa as frequências nas quais $|f_j(\omega_k) \neq 0|$.

A regra de classificação é como anteriormente mencionado.

No caso de séries temporais, é mais provável que a análise discriminante envolva assumir que as matrizes de covariâncias são diferentes e as médias são iguais. Por exemplo, os testes, mostrados na figura do exemplo anteior, implicam que, para os terremotos e explosões, as diferenças primárias estão nas matrizes espectrais bivariadas e as médias são essencialmente as mesmas. Para este caso, será conveniente escrever a aproximação de Whittle para o logaritmo da verossimilhança na forma

$$\ln\left(p_j(X)\right) = \sum_{0<\omega_k<1/2} \left(-\ln\left(|f_j(\omega_k)|\right) - X^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k)\right),$$

onde omitimos as probabilidades a priori da equação.

O detector quadrático, neste caso, pode ser escrito na forma

$$\ln\left(p_j(X)\right) = \sum_{0<\omega_k<1/2} \left(-\ln\left(|f_j(\omega_k)|\right) - \mathrm{tr}\left(I(\omega_k)f_j^{-1}(\omega_k)\right)\right),$$

onde

$$I(\omega_k) = X(\omega_k)X^*(\omega_k)$$

denota a matriz do periodograma. Para probabilidades a priori iguais, podemos atribuir uma observação x na população Π_i sempre que

$$\ln\left(p_j(x)\right) > \ln\left(p_j(x)\right),$$

para $j \neq i, j = 1, 2, \cdots, g$.

Numerosos autores consideraram várias versões de análise discriminante no domínio da frequência. Por exemplo, Shumway and Unger (1974), Alagón (1989), Dargahi-Noubary and Laycock (1981), Taniguchi et al. (1994) e Shumway (1982).

Medidas de discrepância

Antes de prosseguir com os exemplos de análise discriminante e de agrupamento será útil considerar a relação com a informação de discriminação de Kullback-Leibler (K-L).

Usando a aproximação espectral e observando que a matriz do periodograma tem esperança aproximada

$$\mathrm{E}_j\big(I(\omega_k)\big) \;=\; f_j(\omega_k),$$

podemos escrever as informações aproximadas de discriminação como

$$I(f_1; f_2) \;=\; \frac{1}{n}\mathrm{E}_1\left(\ln\left(\frac{p_1(X)}{p_2(X)}\right)\right).$$

A classificação, neste caso, prossegue simplesmente escolhendo a população Π_j que minimiza $I(\widehat{f}; f_j)$, ou seja, atribuindo x à população Π_i sempre que

$$I(\widehat{f}; f_i) < I(\widehat{f}; f_j),$$

para $j \neq i, j = 1, 2, \cdots, g$. Kakizawa et al. (1998) propuseram o uso da medida de informação de Chernoff (CH) (Chernoff, 1952, Renyi, 1961), definida como

$$B_\alpha(1; 2) = -\ln\left(\mathrm{E}_2\left(\frac{p_2(x)}{p_1(x)}\right)^\alpha\right),$$

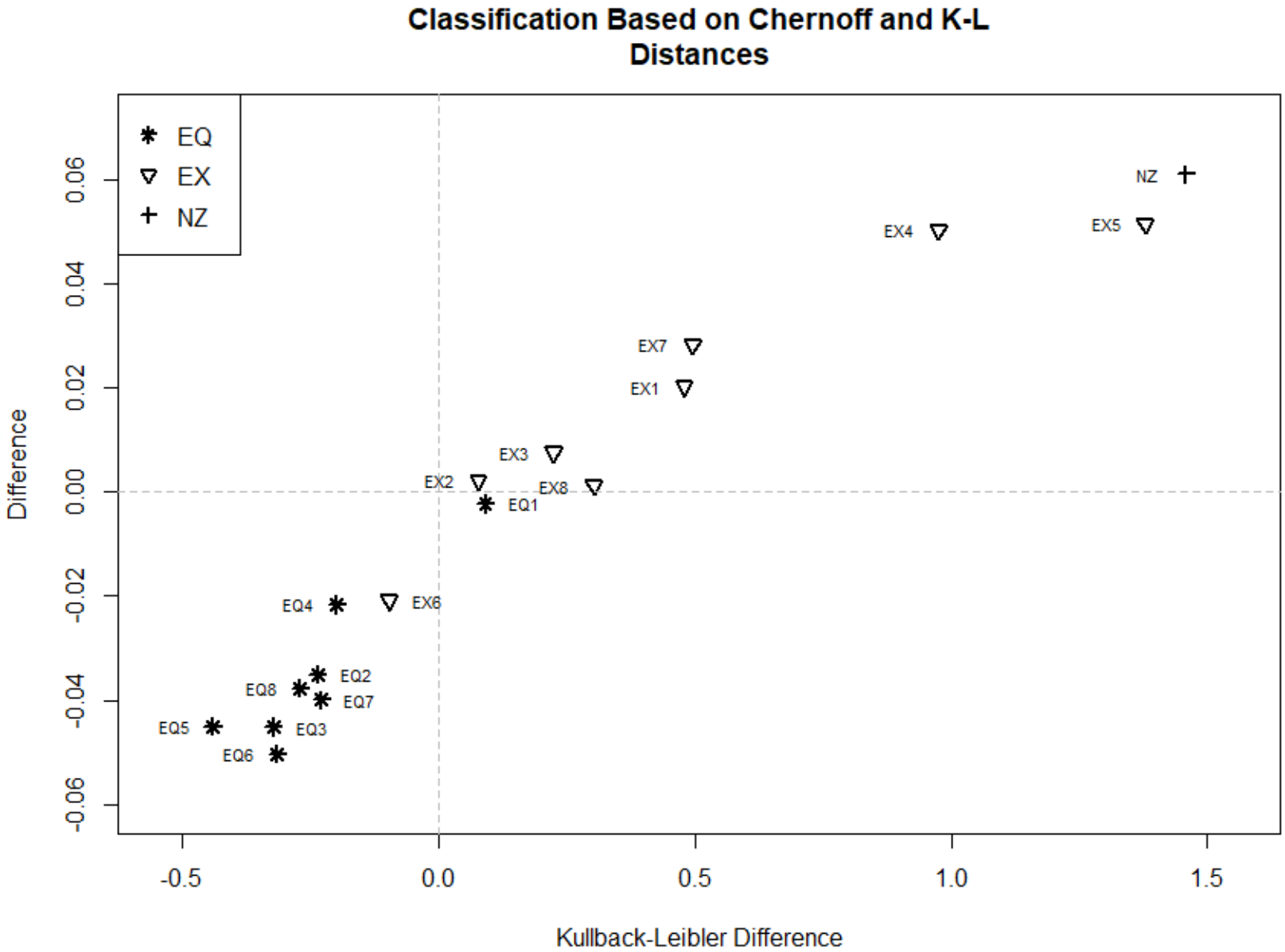
onde a medida é indexada por um parâmetro de regularização α, para $0 < \alpha < 1$. Quando $\alpha = 0.5$, a medida de Chernoff é a divergência simétrica proposta por Bhattacharya (1943). Para o caso multivariado,

$$B_\alpha(1; 2 : x) = -\ln\left(\mathrm{E}_2\left(\frac{p_2(x)}{p_1(x)}\right)^\alpha\right).$$

Exemplo VII.11. Análise discriminante em dados sísmicos.

Um esforço considerável foi despendido no uso de várias razões espectrais envolvendo as fases bivariadas P e S como características de discriminação. Kakizawa et al. (1998) mencionam uma série de medidas que têm sido usadas na literatura sismológica como recursos. Esses recursos incluem relações de potência para as duas fases e relações de componentes de potência em bandas de alta e baixa frequência.

O uso de tais características do espectro sugere que um procedimento ótimo baseado na discriminação entre as matrizes espectrais de dois processos estacionários seria razoável. O fato de que a hipótese de que as matrizes espectrais eram iguais, também foi amplamente rejeitada sugere o uso de uma função discriminante baseada em diferenças espectrais.


Classification Based on Chernoff and K-L Distances
EQ
EX
NZ
Difference
Kullback-Leibler Difference
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.5
0.0
0.5
1.0
1.5
NZ
EX4
EX5
EX7
EX1
EX3
EX2
EX8
EQ1
EQ4
EX6
EQ2
EQ8
EQ7
EQ5
EQ3
EQ6

Para fins de agrupamento, pode ser mais útil considerar uma medida de disparidade simétrica e apresentamos a medida J-Divergência

$$J(f_1; f_2) = I(f_1; f_2) + I(f_2; f_1)$$

e o número de Chernoff simétrico

$$JB_\alpha(f_1; f_2) = B_\alpha(f_1; f_2) + B_\alpha(f_2; f_1)$$

para esse propósito.

Neste caso, definimos a disparidade entre a matriz espectral amostral de um único vetor x, e a população Π_j como

$$J(\widehat{f}; f_j) = I(\widehat{f}; f_j) + I(f_j; \widehat{f})$$

e

$$JB_\alpha(\widehat{f}; f_j) = B_\alpha(\widehat{f}; f_j) + B_\alpha(f_j; \widehat{f}),$$

respectivamente e use-as como quase distâncias entre o vetor e a população Π_j.

As medidas de disparidade podem ser usadas para agrupar séries temporais multivariadas. As medidas simétricas de disparidade, conforme definidas acima, garantem que a disparidade entre f_i e f_j seja igual à disparidade entre f_j e f_i.

Portanto, consideraremos as formas simétricas mostradas acima como quase distâncias com o propósito de definir uma matriz de distância para entrada em um dos procedimentos de agrupamento padrão (ver Johnson and Wichern, 1992).

Em geral, podemos considerar métodos de agrupamento hierárquico ou particionado usando a matriz de quase distância como entrada.

Para fins de ilustração, podemos usar a divergência simétrica $J(\widehat{f}; f_j)$, que implica que a quase-distância entre as séries amostrais com matrizes espectrais estimadas $\widehat{f}_i$ e $\widehat{f}_j$ seria $J(\widehat{f}_i; \widehat{f}_j)$; ou seja,

$$J(\widehat{f}_i; \widehat{f}_j) = \frac{1}{n} \sum_{0<\omega_k<1/2} \left(\text{tr}\left(\widehat{f}_i(\omega_k)\widehat{f}_j^{-1}(\omega_k)\right) + \text{tr}\left(\widehat{f}_j(\omega_k)\widehat{f}_i^{-1}(\omega_k)\right) - 2p \right)$$

para $i \neq j$.

Também podemos usar a forma comparável para a divergência de Chernoff, mas podemos não querer fazer uma suposição para o parâmetro de regularização α.

No agrupamento hierárquico, começamos agrupando os dois elementos da população que minimizam a medida de disparidade $J(\widehat{f}_i; \widehat{f}_j)$. Então, esses dois itens formam um cluster e podemos calcular distâncias entre itens não agrupados como antes.

A distância entre elementos não agrupados e um agrupamento atual é definida aqui como a média das distâncias aos elementos no agrupamento. Novamente, combinamos objetos que estão mais próximos. Também podemos calcular a distância entre os itens não agrupados e os itens agrupados como a distância mais próxima, em vez da média. Quando uma série está em um cluster, ela permanece lá. Em cada estágio, temos um número fixo de clusters, dependendo do estágio de fusão.

Alternativamente, podemos pensar no agrupamento como uma partição da amostra em um número pré-especificado de grupos. MacQueen (1967) propôs isso usando agrupamento de k-médias, usando a distância de Mahalonobis entre uma observação e os vetores de média do grupo.

Em cada estágio, uma redesignação de uma observação em seu grupo de afinidade mais próximo é possível. Para ver como este procedimento se aplica ao contexto atual, considere uma partição preliminar em um número fixo de grupos e defina a disparidade entre a matriz espectral da observação, digamos, $\widehat{f}$ e a matriz espectral média do grupo, digamos, $\widehat{f}_i$, como $J(\widehat{f}; \widehat{f}_i)$, como $J(\widehat{f}; \widehat{f}_i)$, onde a matriz espectral do grupo pode ser estimada por

$$\widehat{f}_i(\omega_k) = \frac{1}{N_i} \sum_{j=1}^{N_i} \widehat{f}_{ij}(\omega_k).$$

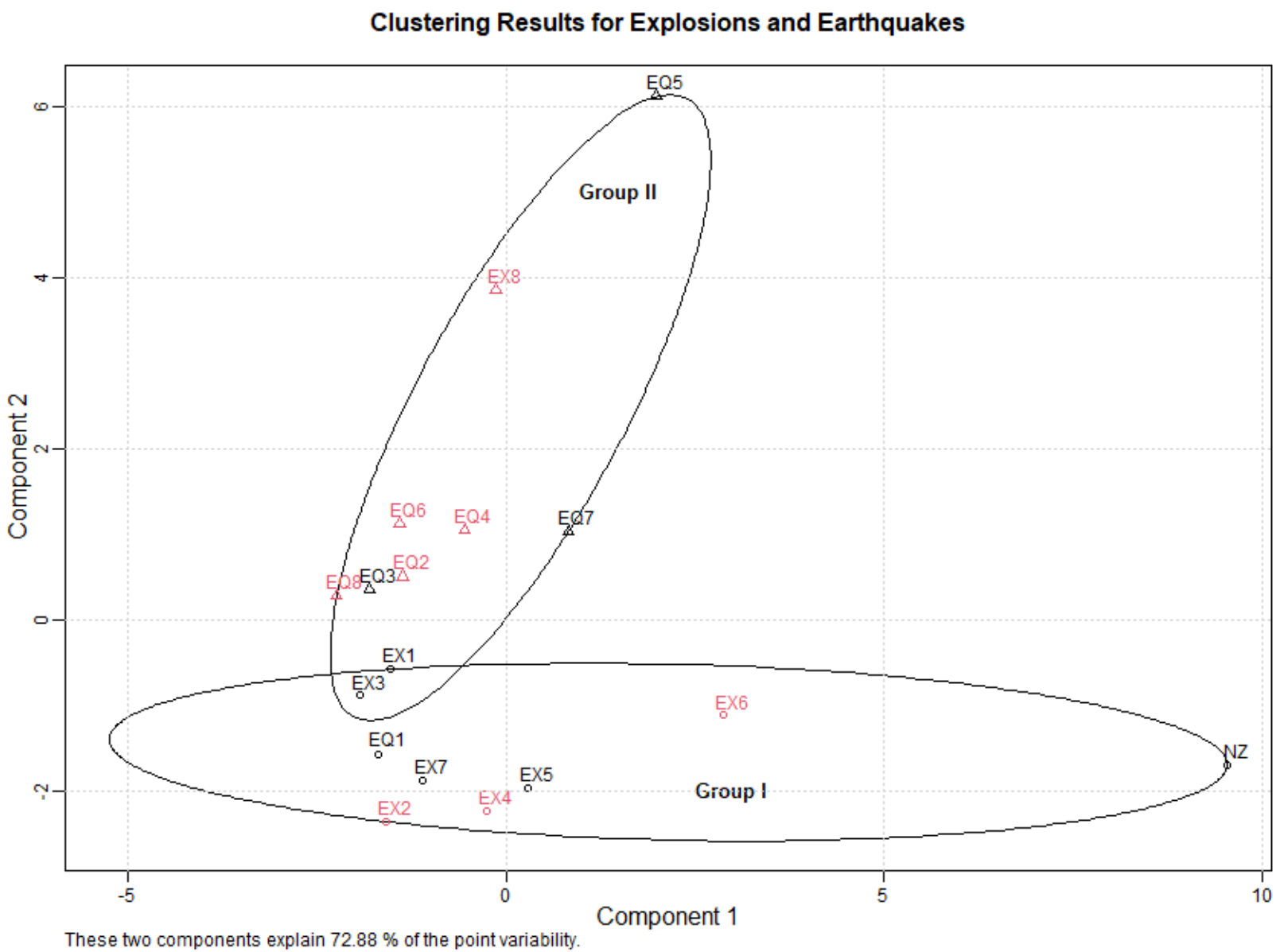
Em qualquer passagem, uma única série é reatribuída ao grupo para o qual sua disparidade é minimizada.

O procedimento de reatribuição é repetido até que todas as observações fiquem em seus grupos atuais. Obviamente, o número de grupos deve ser especificado para cada repetição do algoritmo de particionamento e uma partição inicial deve ser escolhida. Essa atribuição pode ser aleatória ou escolhida a partir de um agrupamento hierárquico preliminar, conforme descrito acima.

Exemplo VII.12. Análise de cluster em dados sísmicos.

É instrutivo tentar um procedimento de agrupamento na população de terremotos e explosões conhecidos. A figura abaixo mostra os resultados da aplicação do algoritmo de agrupamento Partitioning Around Medoids (PAM), que é essencialmente uma robustez do procedimento k-means, sob a suposição de que dois grupos são apropriados.

A partição de dois grupos tende a produzir uma parti¸ão final que concorda intimamente com a configuração conhecida com o terremoto 1 (EQ1) e a explosão 8 (EX8) sendo classificados incorretamente; como nos exemplos anteriores, o evento NZ é classificado como uma explosão.


Clustering Results for Explosions and Earthquakes
EQ5
Group II
EX8
EQ6
EQ4
EQ7
EQ2
EQ3
EQ8
EX1
EX3
EX6
EQ1
EX7
EX5
EX4
EX2
NZ
Group I
Component 2
Component 1
6
4
2
0
-2
-5
0
5
10
These two components explain 72.88 % of the point variability.