XXI CONGRESSO

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

# ESTATÍSTICA:
# A CIÊNCIA DA INCERTEZA

ATAS DO XXI CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA

**Editores:**
Isabel Pereira • Adelaide Freitas • Manuel Scotto
Maria Eduarda Silva • Carlos Daniel Paulino

SPE
SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

# ESTATÍSTICA:

# A CIÊNCIA DA INCERTEZA

Atas do XXI Congresso Anual da

Sociedade Portuguesa de Estatística

Aveiro, 29 de novembro a 2 de dezembro de 2013

**Editores**

*Isabel Pereira*

*Adelaide Freitas*

*Manuel Scotto*

*Maria Eduarda Silva*

*Carlos Daniel Paulino*

Dezembro, 2014

**Edições SPE**

**Editores:** Isabel Pereira, Adelaide Freitas, Manuel Scotto, Maria Eduarda Silva, Carlos Daniel Paulino

**Título:** Estatística: A Ciência da Incerteza.
Atas do XXI Congresso da Sociedade Portuguesa de Estatística

**Editora:** Sociedade Portuguesa de Estatística

**Conceção Gráfica da Capa:** Carina Sousa

**Impressão:** Instituto Nacional de Estatística

**Tiragem:** 250 Exemplares

**ISBN:** 978-972-8890-35-3

**Depósito Legal:** 385476/14

## Prefácio

O conteúdo deste volume é o resultado de uma seleção criteriosa de artigos apresentados no XXI Congresso da Sociedade Portuguesa de Estatística (SPE) e submetidos a apreciação para publicação nas respetivas atas.

O congresso realizado no Hotel Meliá Ria, em Aveiro, entre 29 de novembro e 2 de dezembro de 2013, contou com cerca de 200 participantes e assinalou, oficialmente, o encerramento das celebrações em Portugal do Ano Internacional da Estatística. Neste âmbito destaca-se a sessão comemorativa, durante a qual foi entregue, pela primeira vez, o prémio Carreira SPE, com o objetivo de reconhecer individualidades que desempenham um papel relevante no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal. Foram distinguidos com o referido prémio os Professores Bento Murteira, Dinis Pestana e Ivette Gomes. Foi ainda prestada homenagem à memória do Professor Daniel Muller, falecido a 17 de outubro de 2013.

O programa científico do XXI Congresso da SPE incluiu o já tradicional e muito esperado minicurso que este ano versou sobre Análise de Valores Extremos: uma introdução, ministrado por Ivette Gomes (Universidade de Lisboa), com a colaboração de Isabel Fraga Alves (Universidade de Lisboa) e Cláudia Neves (Universidade de Aveiro); quatro conferências plenárias, com os oradores convidados Adrien Bowman (Universidade de Glasgow), Maria Antónia Turkman (Universidade de Lisboa), Carlos Braumann (Universidade de Évora) e Esther Ruiz (Universidade Carlos III de Madrid); nove sessões temáticas, cada uma com quatro comunicações, organizadas pelos colegas Alexandra Ramos (Universidade do Porto), Alfredo Egídio (Universidade de Lisboa), Conceição Amado (Universidade de Lisboa), Fátima Ferreira (Universidade do Minho), Lisete Sousa (Universidade de Lisboa), Nuno Sepúlveda (London School of Hygiene and Tropical Medicine, UK), Raquel Menezes (Universidade

do Minho) e Valeska Andreozzi (Universidade de Lisboa); sessenta e duas comunicações orais abrangendo as diversas áreas da Estatística e ainda duas sessões de pósteres envolvendo quarenta e sete trabalhos. Agradecemos a todos o empenho que muito contribuiu para o sucesso científico do congresso.

A consecução das tarefas inerentes à realização deste congresso deveu-se às respetivas Comissão Organizadora (CO) e Comissão Científica (CC). Os editores deste livro desejam exprimir aqui o seu agradecimento a todos os elementos da CO e da CC, bem como aos autores dos artigos candidatos à publicação nestas atas e ao corpo de revisores de todo o material submetido. O nosso agradecimento é extensivo ao INE por mais uma vez ter aceite encarregar-se da impressão deste tipo de documento no quadro da frutuosa colaboração que mantém com a SPE. Sem o envolvimento de todos eles não teria sido possível dar à luz mais um volume das atas da SPE a disseminar parte da produção científica da comunidade estatística portuguesa. Salienta-se que, pela primeira vez, teremos a publicação on-line das atas no sítio da SPE de forma a facilitar a acessibilidade dos trabalhos revistos e aceites.

Por fim, queremos expressar o nosso reconhecimento a todos os congressistas pelos trabalhos apresentados na certeza de que é a divulgação do que de melhor se faz em Estatística que promove, na sociedade, a importância de *a Ciência da Incerteza*.

Aveiro, dezembro de 2014
*Os Editores*

## Agradecimentos

Aos seguintes colegas, listados por ordem alfabética do primeiro nome, pelo generoso trabalho de revisão dos artigos submetidos a estas Atas e que em muito valorizou o conteúdo desta publicação:

Adelaide Freitas
Adelaide Figueiredo
Alexandra Ramos
Ana Cristina Braga
Ana Ferreira
Ana Luísa Papoila
Ana Maria Abreu
Ana Rita Gaio
Anabela Flores
Antónia Turkman
António Pedro Duarte Silva
Bruno de Sousa
Cláudia Neves
Conceição Lopes
Conceição Rocha
Conceição Serra
Cristina Martins
Cristina Simões Rocha
Denisa Mendonça
Dinis Pestana
Esmeralda Gonçalves
Fátima Pina
Fernanda Figueiredo
Fernando Rosado
Giovani Silva
Graça Temido
Helena Ferreira
Inês Sousa
Isabel Barão
Isabel Rodrigues
Isabel Silva
Ivette Gomes
João Nicolau
Jorge Cadima
Júlia Teles
Lisete Sousa
Luís Machado
Luzia Gonçalves
Magda Monteiro
Manuel Morais
Manuela Souto
Marco Costa
Maria João Polidoro
Marília Antunes
Nélia Silva
Nuno Sepúlveda
Patrícia Bermudéz
Paulo Infante
Paulo Soares
Paulo Teles
Pedro Silva
Regina Bispo
Rosário Oliveira
Sónia Gouveia
Susana Faria
Vanda Lourenço

## Agradecimentos

Agradecemos às seguintes entidades o valioso apoio concedido à realização do XXI Congresso Anual da SPE

- **Aveiforca**
- **BH Commercial & Home Fitness - Health & Beauty**
- **CIDMA**
- **Fundação para a Ciência e a Tecnologia**
- **Hotel Meliá Ria**
- **Instituto Nacional de Estatística**
- **Livraria Escolar Editora**
- **Pedro Pereira, lda**
- **SAS Portugal**
- **Stand Vicente, lda**
- **Universidade de Aveiro**
- **Universidade do Porto**

Um agradecimento especial é devido à Direção da Sociedade Portuguesa de Estatística e em particular à sua secretária Elena Codreanu e aos colegas das Comissões Científica e Organizadora do Congresso, listados por ordem alfabética do primeiro nome.

## Comissão Científica

- **Carlos Daniel Paulino** (Universidade de Lisboa)
- **Denisa Mendonça** (ICBAS)
- **Esmeralda Gonçalves** (Universidade de Coimbra)
- **Helena Ferreira** (Universidade da Beira Interior)
- **Isabel Pereira** (Universidade de Aveiro)

## Comissão Organizadora

- **Isabel Pereira** (CIDMA & Universidade de Aveiro)
- **Adelaide Freitas** (CIDMA & Universidade de Aveiro)
- **Cláudia Neves** (CEAUL & Universidade de Aveiro)
- **Eugénio Rocha** (CIDMA & Universidade de Aveiro)
- **Manuel Scotto** (CIDMA & Universidade de Aveiro)
- **Maria Eduarda Silva** (CIDMA & Universidade do Porto)
- **Nélia Silva** (CIDMA & Universidade de Aveiro)

# Índice

# Transectos lineares em ungulados de montanha: um estudo de simulação

Anabela Afonso
CIMA-UE e ECT, Universidade de Évora, Portugal, *aafonso@uevora.pt*

Russell Alpizar-Jara
CIMA-UE e ECT, Universidade de Évora, Portugal, *alpizar@uevora.pt*

Jesús M. Pérez
Departamento de Biología Animal, Biología Vegetal y Ecología, Universidade de Jaén, Espanha, *jperez@ujaen.es*

**Palavras–chave**: Densidade, função de deteção, monotonia

**Resumo**: Desde os finais de 1980 que a amostragem por transectos lineares tem sido utilizada para estimar a densidade e abundância de ungulados de montanha. No entanto, a aplicação desta técnica nestas populações pode apresentar algumas particularidades como sejam: a topografia do terreno é bastante irregular, os indivíduos destas populações tendem formar grupos com dimensões bastante distintas, e não se observa relação linear entre o tamanho dos grupos e a distância de afastamento a que se encontram do observador. Estas características fazem com que a função de deteção nem sempre seja uma função monótona decrescente como seria expectável na amostragem por transectos lineares convencional. Para facilitar o processo de estimação da função de deteção é usual truncar os dados eliminando da análise os animais detetados a uma grande distância do observador. Neste trabalho pretendemos avaliar o efeito da truncatura nas estimativas da densidade populacional quando a função de deteção não é estritamente monótona mas se assume como sendo. Numa análise de dados de ungulados de montanha verificámos que, contrariamente ao que seria de esperar, ao truncar os dados as estimativas da densidade tendiam a aumentar.

# 1 Introdução

A amostragem por transectos lineares (TL) foi inicialmente concebida para regiões de estudo planas ou pouco acidentadas. Nesta técnica a probabilidade de deteção é estimada a partir de uma função ajustada às distâncias perpendiculares observadas entre os animais e o observador, a qual é designada por função de deteção [1].
Devido à sua simplicidade de aplicação, no final da década de 1980 a amostragem por TL começou a ser utilizada também em terrenos montanhosos [3]. Contudo, nas populações de ungulados de montanha existe uma tendência para formar agrupamentos de dimensões bastante variáveis, i.e., de um modo geral observam-se vários grupos com um pequeno número de animais e alguns grupos de elevada dimensão. Estes grupos tanto são detetados perto como afastados da linha percorrida pelo observador, e por vezes não se observa associação linear entre as dimensões dos grupos e as distâncias a que se encontram em relação ao observador. É muito vulgar a observação de grupos a distâncias atípicas o que dificulta o processo de estimação da função de deteção e, consequentemente, da densidade ou abundância de animais na região de estudo.
Na estimação da função de deteção podemos optar por várias estratégias de análise dos dados: *i*) truncar ou não truncar os dados de distâncias, *ii*) agrupar ou não agrupar as distâncias em classes, ou *iii*) combinar as estratégias anteriores. Neste trabalho estamos interessados em avaliar o efeito de truncar grupos de indivíduos localizados a distâncias extremas. Para tal simulamos populações agrupadas, das quais serão retiradas amostras com várias dimensões. Esta simulação está baseada numa aplicação com dados reais de várias populações de ungulados de montanha utilizados em [4].

# 2 Transectos lineares

Na amostragem por TL um ou vários observadores percorrem uma linha de comprimento, $L$, e registam a distância perpendicular, $x$, a que cada indivíduo (ou grupo de indivíduos) se encontra da linha.

Sempre que é avistado um grupo de indivíduos, o observador regista também o tamanho $s$ do grupo. A densidade $D$ de indivíduos na área de estudo $A$ é estimada por

$$\hat{D} = \hat{D}_s \hat{E}(s), \tag{1}$$

sendo $\hat{D}_s = n/(a\hat{P}_a)$, $n$ o número de grupos detetados na área amostrada $a$ (sendo $a \leq A$), $\hat{E}(s)$ uma estimativa para o tamanho médio dos grupos e $\hat{P}_a$ a probabilidade de deteção estimada a partir função de deteção $g(x) = P(detetar|distância\ x)$ ajustada às distâncias observadas aos grupos na área amostrada $a$. Detalhes sobre esta metodologia e os pressupostos básicos podem ser consultados em [1].

É habitual considerar-se que a função $g(x)$ é decrescente com o aumento das distâncias, $x$, uma vez que é natural que a capacidade de deteção do observador diminua com o aumento da distância. Usualmente, a estimação de $g(x)$ consiste em ajustar um modelo de regressão não linear aos dados das distâncias observadas, em concreto alguns modelos cuja literatura considera possuírem boas propriedades [1]. No entanto, também têm sido propostas alternativas não paramétricas [2]. Para facilitar o processo de estimação de $g(x)$, é vulgar truncar parte das maiores distâncias observadas. Habitualmente, truncam-se entre 5 e 10% das maiores distâncias observadas, ou à distância $w$ tal que $g(w) = 0{,}15$. Quando não se truncam as distâncias observadas, então utilizam-se todos os dados disponíveis na amostra e assume-se que $w$ corresponde à distância máxima observada, i.e. $w = max(x_i)$ [1].

# 3 Dados de montanha

No trabalho [4] são analisadas 15 amostras de dados relativos a 3 espécies de ungulados de montanha observadas em 6 serras localizadas em Espanha, França e Itália. Para estimar os parâmetros de interesse, os autores utilizaram o programa Distance [6]. Na análise destes dados a metodologia mostrou ser consistente com resultados

publicados por outros autores, tanto nas estimativas da densidade populacional, como da sua precisão, perante as várias formas de agrupamento em classes de distâncias. Verificou-se também que a precisão das estimativas não foi afetada pelo nível de truncatura. No entanto, verificou-se que as estimativas da densidade populacional eram mais elevadas quando se consideraram percentagens de truncatura maiores, ao contrário do que seria de esperar segundo os vários trabalhos de simulação e as considerações metodológicas prevalecentes na literatura. Um padrão característico observado no histograma de dados de montanha é o de uma distribuição bimodal das distâncias perpendiculares a que se observam os vários grupos de animais, como se mostra na Figura 1(a). Este padrão deve-se muitas vezes à irregularidade topográfica do terreno. Os dados destas amostras referem-se a grupos de animais, que variam entre 1 e 10 indivíduos (Figura 1(c)) e não há evidências de associações lineares entre as distâncias e as dimensões dos grupos detetados (Figura 1(d)). $E(s)$ é o valor esperado da dimensão dos grupos de indivíduos. Na literatura são propostas várias abordagens de estimação ([1], p. 119) sendo a mais simples e mais usada a média amostral do tamanho dos grupos.

# 4 Simulação

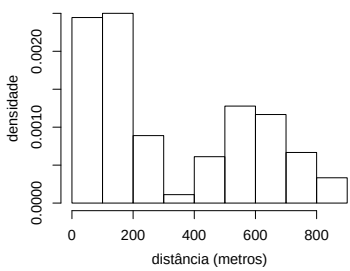
O estudo de simulação realizado teve como principal objetivo compreender quais as consequências de truncar as maiores distâncias, quando a função de deteção não é monótona decrescente em populações que ocorrem em grupos. Para tal, efetuaram-se 1000 replicações tendo por base a informação contida nos dados cuja análise gráfica se apresenta na Figura 1:

1. amostras com 40, 60 e 80 grupos detetados;
2. função de deteção do tipo *half*-normal ($\theta > 0$) com um termo de ajustamento de coseno($\alpha > 0$)

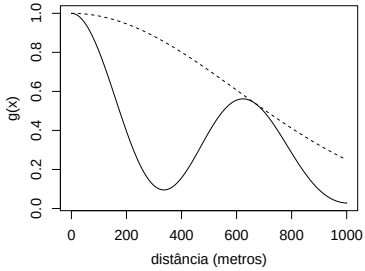
$$g(x) = \frac{1}{1+\alpha} e^{\left(-\frac{x^2}{2\theta^2}\right)} \left(1 + \alpha cos\left(3\pi \frac{x}{w}\right)\right), 0 \leq x \leq w, \quad (2)$$

com $w = 1000$ metros, $\theta = 600$, $\alpha = 0{,}8$ (Figura 1(b));

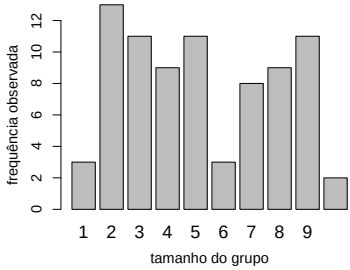
3. dimensão dos grupos segue uma distribuição uniforme{1,...,10} (Figura 1(c));
4. não há associação entre os tamanhos dos grupos e as distâncias (Figura 1(d)).



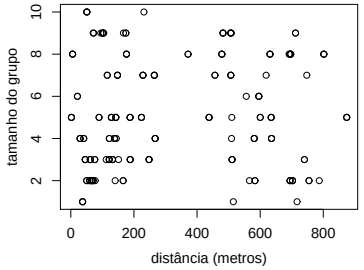

(a) Distribuição das distâncias perpendiculares observadas.




(b) Função de deteção $g(x)$: eq. 2 (linha sólida) *vs.* *half*-normal (linha tracejada).




(c) Distribuição do tamanho do grupo.




(d) Distância *vs.* tamanho do grupo.

Figura 1: Características observadas nos dados de ungulados de montanha e reproduzidas na simulação.

Com base nas características anteriores, para uma área de estudo de $A = 200$ ha, temos os seguintes parâmetros que definem as abun-

dâncias ($N_s = D_s A$) e densidades populacionais de grupos: para $n = 40$, $E(\hat{N}_s) = 105$ e $E(\hat{D}_s) = 0{,}53$; para $n = 60$, $E(\hat{N}_s) = 158$ e $E(\hat{D}_s) = 0{,}79$; e para $n = 80$, $E(\hat{N}_s) = 210$ e $E(\hat{D}_s) = 1{,}05$.

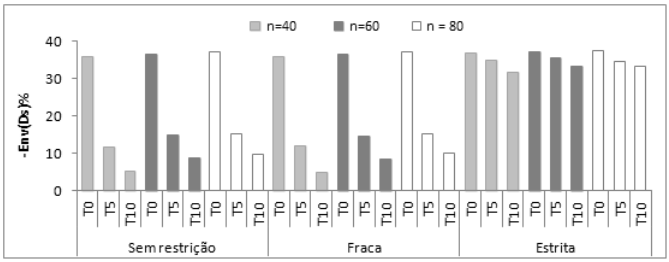
# 5 Resultados

Para cada um dos cenários anteriores, recorrendo ao programa Distante [6], obtiveram-se estimativas para $D_s$ e respetivo coeficiente de variação (CV), considerando três níveis de truncatura à direita (0%, 5% e 10%) e três níveis de restrição para o ajustamento da função de deteção (opções: monotonia estrita, fraca e sem restrição). Apresentamos também o erro quadrático médio ($EQM$) como uma medida para avaliar a perfomance dos estimator da densidade.

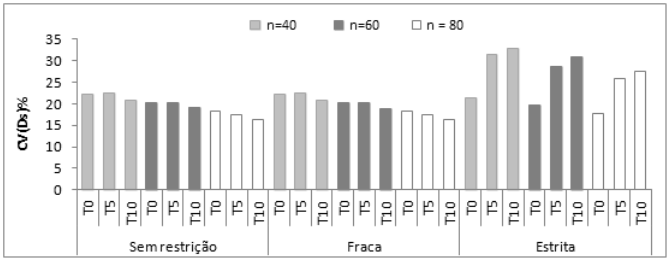
Os resultados destas simulações revelam que o estimador de $D_s$ é sempre enviesado negativamente, diminuindo o enviesamento com o aumento da percentagem de truncatura das maiores distâncias (Figura 2(a)). O enviesamento relativo é severo (entre -30% e -40%) quando o modelo ajustado assume monotonia estrita, ou quando não se truncam os dados independentemente da restrição imposta a $g(x)$.

O CV do estimador diminui com o aumento da dimensão da amostra (Figura 2(b)), e também com o aumento da percentagem de truncatura, exceto quando é assumida a monotonia estrita. Nesta última situação, ao aumentar a truncatura aumenta também o CV podendo atingir valores entre os 20% e 35%.

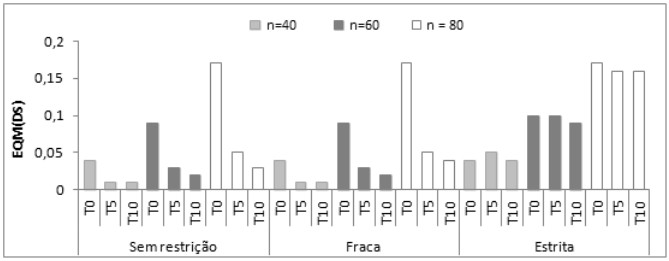
Os melhores cenários em termos globais, considerando o erro quadrático médio (EQM), obtêm-se quando $g(x)$ é ajustada sem restrições de monotonia (ou com apenas restrição fraca), mas para níveis de truncatura de 5% e 10%. Caso contrário, o estimador tem uma performance muito precária (Figura 2(c)). Isto deve-se essencialmente à drástica redução do seu enviesamento (superior a 20%) com o aumento da truncatura.




(a) Enviesamento percentual da densidade de grupos estimada (multiplicado por -1).




(b) Coeficiente de variação da densidade de grupos estimada.




(c) Erro quadrático médio (EQM) da densidade de grupos estimada.

Figura 2: Medidas de qualidade para as amostras com 40, 60 e 80 grupos, por percentagem de truncatura (T0 - 0%; T5 - 5%, T10 - 10%) e tipo de monotonia (sem restrição, fraca e estrita).

# 6 Conclusão e trabalho futuro

Do ponto de vista prático, na análise dos dados provenientes de um esquema de amostragem por TL, o utilizador é confrontado com uma série de decisões relacionadas com estratégias para realizar essa análise. Por exemplo, se é necessário truncar ou não os dados, que percentagem truncar, se as distâncias devem ou não ser agrupadas em classes, e nesse caso em quantas classes, etc... O efeito que estas decisões tem nas estimativas da densidade deve sempre ser avaliado, sem descurar o cumprimento dos pressupostos associados à metodologia.

A truncatura muitas vezes reduz o enviesamento no estimador da densidade e/ou melhora a sua precisão para além de tornar os dados mais fáceis de modelar [1]. No nosso trabalho mostrámos que com dados atípicos, como os que são algumas vezes obtidos em amostragem de ungulados de montanha, este resultado nem sempre é válido no que se refere à variabilidade do estimador quando se impõe que a função de deteção seja estritamente monótona, mesmo após a truncatura de 10% dos dados. Neste tipo de dados a função de deteção poderá ser bimodal, evidenciando uma mistura de distribuições. Este comportamento deve-se tipicamente à topografia do terreno e não necessariamente a um plano de amostragem inadequado. Em terrenos montanhosos o declive do terreno é muito irregular e existem zonas com fraca visibilidade; quando se projetam as distâncias num plano esta irregularidade tem reflexos na distribuição das distâncias perpendiculares projetadas. Ao forçar que a função deteção seja estritamente monótona a função ajustada terá um parâmetro de dispersão com um valor muito elevado de forma a conseguir adaptar-se a este comportamento. Truncar 10% das maiores distâncias pode não ser suficiente para eliminar o efeito mistura. Por conseguinte, a restrição de monotonia estrita nunca deve ser considerada nestes casos.

Futuramente pretendemos estudar o comportamento dos estimadores: *i*) com outras funções de deteção, incluindo funções de deteção bivariadas, $g(x,s)$; *ii*) com outras distribuições do tamanho do grupo

relacionadas ou não com a distância; *iii*) modelando o tamanho do grupo como covariável.

# Agradecimentos

Os autores agradecem ao Dr. Paulino Fandos pela cedência dos dados recolhidos. As atividades de investigação foram parcialmente suportadas pelo Plan Andaluz de Investigación, Junta de Andalucía (RNM - 118). Alpizar-Jara e Afonso são membros do CIMA-UE, centro de investigação financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal, pelo projeto PEst-OE/MAT/UI0117/2014.

# Referências

[1] Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., Thomas, L. (2001) *Introduction to Distance Sampling. Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.

[2] Chen, S.X. (1996). A kernel estimate for the density of a biological population by using line transect sampling. *Applied Statistics* 45, 135–50.

[3] Escós, J., Alados, C.L. (1988). Estimating mountain ungulate density in Sierras de Cazorla y Segura. *Mammalia* 52, 425–428.

[4] Pérez, J.M., Sarasa, M., Moço, G., Granados, J.F., Crampe, J.P., Serrano, E., Maurino, L., Meneguz, P.M., Afonso, A., Alpizar-Jara, R. (no prelo). The effect of data analysis strategies in density estimation of mountain ungulates using Distance Sampling. *Italian Journal of Zoology*.

[5] Southwell, C., Weaver, K. (1993). Esvaluation of analytical procedures for density estimation from line-transect data: data grouping, data truncation and the unit analysis. *Wildlife Research* 20, 433-444.

[6] Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R.B., Marques,T.A., Burnham, K.P. (2010). Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47, 5–14.

# Da utilização de cadeias de Markov multivariadas enquanto regressores num problema de previsão

Bruno Damásio
Universidade de Lisboa, ISEG, CEMAPRE, *bdamasio@iseg.utl.pt*
João Nicolau
Universidade de Lisboa, ISEG, CEMAPRE, *nicolau@iseg.utl.pt*

**Palavras–chave**: Cadeia de Markov multivariada, cadeia de Markov de ordem superior

**Resumo**: Este artigo propõe um novo conceito: a utilização de cadeias de Markov multivariadas (CMM) enquanto regressores. A nossa abordagem baseia-se na observação de que se podem tratar possíveis variáveis discretas, cujos valores são desconhecidos no período de previsão, como uma CMM com o intuito de se reduzirem os erros de previsão de uma determinada variável dependente. Portanto, usufrui-se da informação sobre as interações entre os estados passados das categorias da CMM para prever os regressores discretos e, assim, melhorar a previsão da dita variável dependente.

## 1 Introdução

Considere-se o seguinte modelo simples de alterações de regime: $y_t = \beta x_t + \delta z_t + u_t$ onde $z_t$ é uma variável binária latente que evolui ao longo do tempo de acordo com uma cadeia de Markov homogénea de primeira ordem[1], no sentido em que $P\left(z_t = i_0 | \, z_{t-1} = i_1\right) = P\left(z_{t+h} = i_0 | \, z_{t+h-1} = i_1\right), i_0, i_1 = 0{,}1$ para todo o $h{>}0$. Este tipo de

[1] Numa cadeia de Markov de primeira ordem para se prever o futuro apenas se necessita do presente. Numa cadeia de Markov de ordem superior na $\sigma$-álgebra gerada pela informação disponível até $t$, $\mathcal{F}_t$, está também contida explicitamente informação relativa ao passado do processo, como, aliás, se verá mais à frente.

modelos, e consequentes desenvolvimentos, foram extensivamente estudados (Hamilton [3]). Se $z_t$ fosse observável então estimar-se-iam sem grandes problemas os parâmetros $\beta$ e $\delta$ recorrendo a métodos de estimação standard. No entanto, prever $y_t$ pode levantar algumas dificuldades porque sendo $z_t$ uma variável aleatória, não é observável no período de previsão. Nestas circunstâncias, é necessária uma estrutura probabilística para $z_t$, por exemplo uma cadeia de Markov como nos modelos de alteração de regime (para simplificar assumimos que $x_t$ é um termo dinâmico, como um $AR(1)$, ou uma simples tendência determinística). Neste artigo analisa-se um problema de previsão quando $y_t$ depende de $s > 1$ variáveis discretas cujas dependências são governadas por uma cadeia de Markov multivariada. Esta abordagem é nova na literatura e o modelo mais próximo do nosso é, porventura, o modelo de alterações de regime supracitado. No entanto, ao contrário dos modelos de alteração de regime, que apenas podem incorporar cadeias de Markov univariadas com poucos estados devido à complexidade dos procedimentos de estimação, o nosso modelo pode incorporar diversas variáveis $z_t$ com múltiplos estados cada uma, graças à especificação MTD-probit, como veremos adiante. Este artigo considera um problema de previsão de uma série temporal $y_t$ que depende de variáveis quantitativas ($x_t$) e de $s$ variáveis discretas ($S_{1,t},\ldots,S_{s,t}$) onde cada $S_{j,t}$ ($j = 1,\ldots,s$) pode assumir valores no conjunto finito $\{1,2,\ldots,m\}$. Assume-se que $S_{j,t}$ depende dos valores passados de $S_{1,t-1},\ldots,S_{j,t-1},\ldots,S_{s,t-1}$, e que esta estrutura de dependências pode ser bem modelada por uma CMM de primeira ordem. Para levar a cabo um modelo de regressão que relacione $y_t$ com as variáveis discretas convertem-se as categorias de $S_{j,t}$ num conjunto de variáveis binárias

$$z_{jk,t} = \mathcal{I}_{\{S_{j,t}=k\}}, \tag{1}$$

onde $\mathcal{I}_{\{\cdot\}}$ é a função indicatriz, $\mathcal{I}_{\{S_{j,t}=k\}} = 1$ se $S_{j,t} = k$ e 0 caso contrário, para $k = 1,\ldots,m-1$. Assuma-se, sem qualquer perca de generalidade, uma especificação linear do tipo

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\delta} + u_t, \tag{2}$$

onde $\boldsymbol{x}'_t$ é um vector de elementos predeterminados e $\mathcal{F}_{t-1}$ ou $\mathcal{F}_t$ mensuráveis ($\mathcal{F}_t$ representa a $\sigma$-algebra gerada por todos os eventos até ao instante $t$); $\boldsymbol{z}'_t$ é um vector de variáveis binárias $z_{kj,t}$, respeitantes à CMM, definidas em (1); $\{u_t\}$ é um processo ruido branco gaussiano independente de $\boldsymbol{x}'_t$ e de $\boldsymbol{z}'_t$. Para prever $y_{t+h}$ emprega-se o melhor previsor em termos de erro quadrático médio

$$E\left(y_{t+h}|\,\mathcal{F}_t\right) = E\left(\boldsymbol{x}'_{t+h}|\,\mathcal{F}_t\right)\boldsymbol{\beta} + E\left(\boldsymbol{z}'_{t+h}|\,\mathcal{F}_t\right)\boldsymbol{\delta}. \tag{3}$$

Para ilustrar, considerem-se apenas duas variáveis categóricas ($s = 2$) onde cada uma delas assume valores no conjunto finito $\{1,2,3\}$, i.e. $m = 3$. Expandindo os vectores $\boldsymbol{z}'_t$ e $\boldsymbol{\delta}$ decorre que:

$$y_{t+h} = \boldsymbol{x}'_{t+h}\boldsymbol{\beta} + \delta_{11}\mathcal{I}_{\{S_{1,t}=1\}} + \delta_{12}\mathcal{I}_{\{S_{1,t}=2\}} + \delta_{21}\mathcal{I}_{\{S_{2,t}=1\}} + \delta_{22}\mathcal{I}_{\{S_{2,t}=2\}} + u_t \tag{4}$$

onde $S_{j,t}$ representa a $j$-ésima série categórica da CMM. Tendo em conta o facto de os valores de $S_{j,t+h}$ serem desconhecidos no período de previsão, i.e. para $h \geqslant 1$, exploramos as possíveis dependências entre $S_{j,t+h}$ e os os valores passados de $S_{1,t}$ e de $S_{2,t}$, através da abordagem CMM, objetivando prever $S_{j,t+h}$ e, consequentemente, $y_{t+h}$. Se tanto $S_{1,t}$ como $S_{2,t}$ fossem variáveis quantitativas discretas a equação seria mais simples

$$y_{t+h} = \boldsymbol{x}'_{t+h}\boldsymbol{\beta} + \delta_1 S_{1,t+h} + \delta_2 S_{2,t+h} + u_t. \tag{5}$$

Das equações (4) ou (5), é evidente que para prever $y_{t+h}$ se tem que avaliar $P\left(S_{j,t+h} = k|\,\mathcal{F}_t\right)$, para $k = 1,2,\ldots,m$. Para operacionalizar estas expressões assumem-se as seguintes hipóteses:

*Hipótese 1: CMM de primeira ordem*

$$P\left(S_{j,t} = k|\,\mathcal{F}_{t-1}\right) = P\left(S_{j,t} = k|\,S_{1,t-1} = i_1, \cdots, S_{s,t-1} = i_s\right). \tag{6}$$

Ou seja, $S_{j,t}$ dado $\{S_{1,t-1}, \cdots, S_{s,t-1}\}$ é indepentente das restantes variáveis em $\mathcal{F}_{t-1}$.

*Hipótese 2: CMM homogénea*

$$P\left(S_{j,t}=k|\,S_{1,t-1},\cdots,S_{s,t-1}\right)=P\left(S_{j,t+h}=k|\,S_{1,t+h-1},\cdots,S_{s,t+h-1}\right). \tag{7}$$

*Hipótese 3: Termos contemporâneos negligenciáveis.* $S_{j,t}$ é independente de $\{S_{1,t},\cdots,S_{j-1,t},S_{j+1,t},\cdots,S_{s,t}\}$ dado $\{S_{1,t-1},\cdots,S_{s,t-1}\}$, ou seja

$$\begin{aligned}P\left(S_{j,t}=k|\,S_{1,t}=i_1,\cdots,S_{j-1,t}=i_{j-1},S_{j+1,t}=i_{j+1},\cdots,S_{s,t}=i_s,S_{1,t-1},\cdots,S_{s,t-1}\right)\\ =P\left(S_{j,t}=k|\,S_{1,t-1},\cdots,S_{s,t-1}\right)\end{aligned} \tag{8}$$

Para se conseguir prever $y_{t+h}$, tem de se calcular $E\left(\boldsymbol{x}'_{t+h}\middle|\,\mathcal{F}_t\right)$ e $E\left(\boldsymbol{z}'_{t+h}\middle|\,\mathcal{F}_t\right)$. Por hipótese a primeira expressão é conhecida, donde debrucemo-nos apenas sobre a última expressão. Um elemento genérico de $E\left(\boldsymbol{z}'_{t+h}\middle|\,\mathcal{F}_t\right)$ é $E\left(\boldsymbol{z}_{kj,t+h}|\,\mathcal{F}_t\right)$ que, da hipótese 1, pode ser escrito como:

$$\begin{aligned}E\left(\boldsymbol{z}_{kj,t+h}|\,\mathcal{F}_t\right)&=P\left(\boldsymbol{z}_{kj,t+h}=1|\,\mathcal{F}_t\right)=P\left(S_{j,t+h}=k|\,\mathcal{F}_t\right)\\&=P\left(S_{j,t+h}=k|\,S_{1,t}=i_1,\cdots,S_{s,t}=i_s\right).\end{aligned} \tag{9}$$

A expressão (9) será estimada recorrendo à teoria de estimação das CMM que, em última instância, conduz a $E\left(\boldsymbol{z}'_{t+h}\middle|\,\mathcal{F}_t\right)$ and $E\left(y_{t+h}|\,\mathcal{F}_t\right)$. Na próxima secção abordar-se-ão de forma sucinta os principais aspectos em torno da estimação de CMM.

# 2 Cadeias de Markov multivariadas enquanto regressores: estimação do modelo

Nesta secção discute-se a estratégia de estimação dos parâmetros do modelo: os parâmetros $\boldsymbol{\psi}=(\boldsymbol{\beta},\boldsymbol{\delta})$ e os parâmetros da cadeia de Markov $\boldsymbol{\eta}$. Seja $\boldsymbol{\theta}=(\boldsymbol{\psi},\boldsymbol{\eta})$ o vector de ambos os parâmetros e $B$ e $D$ os espaços-parâmetro respectivamente de $\boldsymbol{\psi}=(\boldsymbol{\beta},\boldsymbol{\delta})$ e de $\boldsymbol{\eta}$. Por

construção, tanto $\boldsymbol{\psi}$ como $\boldsymbol{\eta}$ são de variação livre (ver o conceito em [2]), i.e. $\boldsymbol{\psi}$ e $\boldsymbol{\eta}$ não são sujeitos a restrições cruzadas. Por outras palavras, para qualquer valor admissível de $\boldsymbol{\psi}$ em $B$, $\boldsymbol{\eta}$ pode assumir qualquer valor em $D$ (e vice-versa). Nestas circunstâncias, a distribuição condicional de $y_t|\,\mathbf{S}_t,\mathcal{F}_{t-1}$, depende apenas de $\boldsymbol{\psi}$ e a distribuição condicional de $\mathbf{S}_t|\,\mathcal{F}_{t-1}$ depende apenas de $\boldsymbol{\eta}$. Desta feita, a densidade conjunta pode ser factorizada sequencialmente tal que:

$$\begin{aligned} f\,(y_0,y_1,\ldots,y_n;S_{j0},S_{j1},\ldots,S_{jn};\boldsymbol{\theta}) &= \prod_{t=1}^{n} f\,(y_t,\mathbf{S}_t|\,\mathcal{F}_{t-1};\boldsymbol{\theta}) \\ &= \prod_{t=1}^{n} f\,(y_t|\,\mathbf{S}_t,\mathcal{F}_{t-1};\psi) \prod_{t=1}^{n} P\,(\mathbf{S}_t|\,\mathcal{F}_{t-1};\boldsymbol{\eta})\,. \end{aligned}$$

A expressão $P\,(\mathbf{S}_t|\,\mathcal{F}_{t-1};\boldsymbol{\eta}) = P\,(S_{1,t},\ldots,S_{s,t}|\,\mathcal{F}_{t-1};\boldsymbol{\eta})$ pode ser escrita como:

$$\begin{aligned} P\,(S_{1,t},\ldots,S_{s,t}|\,\mathcal{F}_{t-1};\boldsymbol{\eta}) &= P\,(S_{1,t},\ldots,S_{s,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}) \quad (10) \\ &= \prod_{j=1}^{s} P\,(S_{j,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}) \quad (11) \\ &= \prod_{j=1}^{s} P\,(S_{j,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_j)\,, \quad (12) \end{aligned}$$

onde (10) e (11) decorrem respectivamente das hipóteses 1 e 3. Na equação (12) decompomos o vector $\boldsymbol{\eta}$ em $\boldsymbol{\eta} = (\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_s)$, onde $\boldsymbol{\eta}_j$ são os parâmetros associados à distribuição condicional $S_{jt}|\,S_{1,t-1},\ldots,S_{s,t-1}$. Tal como indagámos há pouco, $\boldsymbol{\eta}$, o vector dos parâmetros $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_s$, e $\boldsymbol{\psi}$ são de variação livre, como veremos de seguida. Rearranjando

os termos decorre que

$$
\begin{aligned}
&f\,(y_0,y_1,\ldots,y_n;S_{j0},S_{j1},\ldots,S_{jn};\boldsymbol{\theta}) \\
&\quad=\prod_{t=1}^{n} f\,(\,y_t|\,\mathbf{S}_t,\mathcal{F}_{t-1};\boldsymbol{\psi})\prod_{t=1}^{n}\prod_{j=1}^{s} P\,(\,S_{j,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_j) \\
&\quad=\prod_{t=1}^{n} f\,(\,y_t|\,\mathbf{S}_t,\mathcal{F}_{t-1};\boldsymbol{\psi})\prod_{t=1}^{n} P\,(\,S_{1,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_1) \\
&\qquad\qquad\cdots\prod_{t=1}^{n} P\,(\,S_{s,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_s)\,, \qquad (13)
\end{aligned}
$$

onde o logaritmo da verosimilhança é

$$
\begin{aligned}
\log f\,(y_0,y_1,\ldots,y_n;S_{j0},S_{j1},\ldots,S_{jn};\boldsymbol{\theta}) &= \sum_{t=1}^{n}\log f\,(\,y_t|\,\mathbf{S}_t,\mathcal{F}_{t-1};\boldsymbol{\psi}) + \\
&\sum_{t=1}^{n}\log P\,(\,S_{1,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_1) + \cdots + \\
&\sum_{t=1}^{n}\log P\,(\,S_{s,t}|\,S_{1,t-1},\ldots,S_{s,t-1};\boldsymbol{\eta}_s)\,. \qquad (14)
\end{aligned}
$$

Esta decomposição permite mostrar que os parâmetros podem ser estimados separadamente, maximizando, para o efeito, as diversas expressões na equação anterior, sem qualquer perda de consistência nem de eficiência. Portanto, $\boldsymbol{\psi}=(\boldsymbol{\beta},\boldsymbol{\delta})$ é estimado, por exemplo, através da máxima verosimilhança da equação (2), e os $\boldsymbol{\eta}_j$ $(j=1,\ldots,s)$ são estimados um a um considerando cada uma das distribuições condicionais $S_{j,t}|\,S_{1,t-1},\ldots,S_{s,t-1}$, tal como veremos com mais detalhe na próxima secção (ver, por exemplo a equação (18)).

# 3 Cadeias de Markov multivariadas: estimação

Na secção anterior mostrou-se que as expressões

$$\sum_{t=1}^{n} \log P\left(S_{j,t} | S_{1,t-1}, \ldots, S_{s,t-1}; \boldsymbol{\eta}_j\right), (j = 1, \ldots, s), \tag{15}$$

podem ser mazimizadas de forma independente dos restantes termos da função de verosimilhança (FV). É sabido que é impraticável modelar as probabilidades

$$P_j\left(i_0 | i_1, \ldots, i_s\right) \equiv P\left(S_{j,t} = i_0 | S_{1,t-1} = i_1, \ldots, S_{s,t-1} = i_s\right), \tag{16}$$

quando $s$ e $m$ são elevados e a dimensão da amostra é baixa, ou mesmo moderada, já que o número de parâmetros do modelo é igual a $m^s\left(m-1\right)$. Para resolver este problema Raftery [5] considerou uma hipótese simplificadora para modelar cadeias de Markov de ordem superior (CMOS). Recentemente Nicolau [4] propôs uma especificação alternativa denominada modelo MTD-Probit:

$$P_j^{\Phi}\left(i_0 | i_1, \ldots, i_s\right) \equiv \frac{\Phi\left(\eta_{j0} + \eta_{j1} P_{j1}\left(i_0 | i_1\right) + \cdots + \eta_{js} P_{js}\left(i_0 | i_s\right)\right)}{\sum_{k=1}^{m} \Phi\left(\eta_{j0} + \eta_{j1} P_{j1}\left(k | i_1\right) + \cdots + \eta_{js} P_{js}\left(k | i_s\right)\right)}, \tag{17}$$

onde $\eta_{ji} \in \mathbb{R}\ (j = 1, \ldots, s; i = 0, \ldots, m)$, $\Phi$ é a função de distribuição normal padrão e $P_{jk}\left(i_0 | i\right) \equiv P\left(S_{j,t} = i_0 | S_{k,t-1} = i\right)$. Quando $S_{j,t}$ é a variável dependente, e $n_{i_1 i_2 \ldots i_{i_s} i_0}$ a frequência dos casos em que $S_{1,t-1} = i_1, \ldots, S_{s,t-1} = i_s, S_{j,t} = i_0$, a FV é

$$\log L = \sum_{i_1 i_2 \ldots i_s i_0} n_{i_1 i_2 \cdots s i_0} \log\left(P_j^{\Phi}\left(i_0 | i_1, \ldots, i_s\right)\right) \tag{18}$$

e o estimador da máxima verosimilhança é: $\hat{\eta}_j = \arg\max_{\eta_{j1}, \ldots, \eta_{js}} \log L$. Os parâmetros $P_{jk}\left(i_0 | i_1\right),\ k = 1, \ldots, s$ podem ser estimados consistentemente com $\hat{P}_{jk}\left(i_0 | i_1\right) = \frac{n_{i_1 i_0}}{\sum_{i_0=1}^{n} n_{i_1 i_0}}$ onde $n_{i_1 i_0}$ é o número de transições de $S_{k,t-1} = i_1$ para $S_{j,t} = i_0$. Este procedimento tem

a vantagem de simplificar bastante a estimação sem afectar a consistência do estimador da máxima verosimilhança $\hat{\eta}_j$ na medida em que $\hat{P}_{jk}$ é um estimador consistente de $P_{jk}$.

# 4 Modelo de previsão multi-passos

A secção anterior discutiu o problema da estimação das probabilidades $P\left(S_{j,t} = i_0 \middle| S_{1,t-1} = i_1, \ldots, S_{s,t-1} = i_s\right)$. Esta secção introduz a questão da previsão a $h$-passos com a CMM. Dado ter-se, por hipótese, uma CMM homogénea (Hipótese 1), a expressão da previsão a 1-passo é trivial. A obtenção das expressões a $h$-passos não é assim tão imediata. Para o efeito, utilizando a versão discreta das equações de Chapman-Kolmogorov, a regra da probabilidade total e as hipóteses 1 a 3, considera-se

$$\begin{aligned} &P\left(S_{j,t+h} = i_0 \middle| S_{1,t} = i_1, \ldots, S_{s,t} = i_s\right) \\ &= \frac{\Phi\left(\eta_{j0} + \eta_{j1} P\left(S_{j,t+h} = i_0 \middle| S_{1,t} = i_1\right) + \cdots + \eta_{js} P\left(S_{j,t+h} = i_0 \middle| S_{s,t} = i_s\right)\right)}{\sum_{k=1}^{m} \Phi\left(\eta_{j0} + \eta_{j1} P\left(S_{j,t+h} = i_0 \middle| S_{1,t} = i_1\right) + \cdots + \eta_{js} P\left(S_{j,t+h} = i_0 \middle| S_{s,t} = i_s\right)\right)}, \end{aligned} \tag{19}$$

que não é mais do que uma extensão natural da equação (4) e que necessita do cálculo prévio de $P\left(S_{j,t+h} = i_0 \middle| S_{k,t} = i_k\right)$. Mostra-se facilmente que:

$$\begin{aligned} &P\left(S_{j,t+h} = i_0 \middle| S_{k,t} = i_k\right) = \\ &\sum_{\alpha=1}^{m} P\left(S_{j,t+h} = i_0 \middle| S_{k,t+h-1} = \alpha\right) P\left(S_{k,t+h-1} = \alpha \middle| S_{k,t} = i_k\right). \end{aligned} \tag{20}$$

Esta expressão é igual ao elemento $(i_0, i_k)$ da matriz $P^{(jk)}\left(P^{(kk)}\right)^{h-1}$ onde $P^{(jk)}$ é a matriz cujos elementos são $P\left(S_{j,t} = i_0 \middle| S_{k,t-1} = i_k\right)$. Podemos agora estabelecer o algoritmo subjacente à previsão de $y_{t+h}$

1. Estimar o modelo $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\delta} + u_t$ e os respectivos parâmetros $\beta$ e $\boldsymbol{\delta}$ através do OLS ou de qualquer outro método.

2. Obter $\hat{\eta}_j = \arg\max_{\eta_{j1},\ldots,\eta_{js}} \log L$ onde o logaritmo da verosimilhança se refere à equação (18).

3. A partir de $\hat{\eta}_j$ calcular $P\left(S_{j,t+1} = k|S_{1,t}, \cdots, S_{s,t}\right)$ e deduzir as expressões $P\left(S_{j,t+h} = k|S_{1,t}, \cdots, S_{s,t}\right)$

4. Por fim, obter a previsão $y_{t+h}$ calculando

$$E\left(y_{t+h}|\,\mathcal{F}_t\right) = E\left(\boldsymbol{x}'_{t+h}\middle|\,\mathcal{F}_t\right)\boldsymbol{\beta} + E\left(\boldsymbol{z}'_{t+h}\middle|\,\mathcal{F}_t\right)\boldsymbol{\delta}.$$

# 5 Estudo de simulação de Monte Carlo

## 5.1 Procedimento

Nesta secção avaliamos o potencial preditivo das CMM, onde cada uma das cadeias desempenha o papel de regressor, através de um problema de simulação de Monte Carlo. Consideramos um processo simples com duas categorias e três estados ($s = 2$ e $m = 3$). A CMM é simulada de acordo com o seguinte algoritmo:

1. Inicializar o processo $\{(S_{1,t}, S_{2,t})\}$ atribuindo valores arbitrários a $S_{1,0}$ e a $S_{2,0}$.

2. Definir duas matrizes de probabilidades de transição $m^s \times m$ com elementos:

$$\begin{array}{l} P\left(S_{1,t} = i_o|\, S_{1,t-1} = i_1, S_{2,t-1} = i_2\right) \\ P\left(S_{2,t} = i_o|\, S_{1,t-1} = i_1, S_{2,t-1} = i_2\right) \end{array} \tag{21}$$

   (ver a definição do processo gerador de dados mais à frente)

3. Dadas as condições iniciais $S_{1,0}$ e $S_{2,0}$ (passo 1), simular o processo multivariado $\{(S_{1,t}, S_{2,t})\}, t = 1,..,T$:

   (a) simular $U_1$, uniformemente distribuido em [0, 1];

(b) atribuir valores a $S_{1,t}$ de acordo com a regra:

$$S_{1,t} = \begin{cases} 1 & \text{if} \quad 0 \leq U_1 < p_1^{[1]} \\ 2 & \text{if} \quad p_1 \leq U_1 < p_1^{[1]} + p_2^{[1]} \\ 3 & \text{if} \quad p_1 + p_2 \leq U_1 < 1 \end{cases}$$

onde $p_i^{[1]} \equiv P\left(S_{1,t} = i | S_{1,t-1} = i_1, S_{2,t-1} = i_2\right), i = 1,2,3;$

(c) repetir o processo para $S_{2,t}$ (com $U_2 \sim U\,(0,1)$ independente de $U_1$) .

4. Repetir os passos 1-3 até $t = T$.

A partir deste algoritmo constroem-se as 4 variáveis binárias tal como em (1): $z_{jk,t} = \mathcal{I}_{\{S_{j,t}=k\}}, k = 1, \cdots, m-1$.
Consideram-se ainda

- $\boldsymbol{z}_t' \equiv \left[\begin{array}{cccc} z_{11} & z_{12} & z_{21} & z_{22} \end{array}\right]$, $\boldsymbol{\delta} = \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \end{array}\right]'$,
- $\boldsymbol{x}_t' = \left[\begin{array}{cc} 1 & x_t \end{array}\right]$, $x_t$ e $u_t$ são i.i.d. $N\,(0,1)$, $\beta = \left[\begin{array}{cc} 1 & 1 \end{array}\right]'$.

Por conveniência, assumimos que $S_{1,t}$ e $S_{2,t}$ têm as mesmas probabilidades de transição:

$$P\left(S_{2,t} = i_0 | S_{1,t-1} = i_1, S_{2,t-1} = i_2\right) = P\left(S_{1,t} = i_0 | S_{1,t-1} = i_1, S_{2,t-1} = i_2\right).$$

Posteriormente comparamos os erros de previsão a $h$-passos da variável dependente produzidos por quatro diferentes hipóteses para os valores das variáveis binárias em $t + h$:
1) São conhecidos, ou seja $\hat{z}_{jk,t+h}^{(1)} = z_{jk,t+h}$;
2) São previstos de acordo com a metodologia proposta,

$$\hat{z}_{jk,t+h}^{(2)} = \hat{P}\left(S_{j,t+h} = k | S_{1,t} = i_1, S_{2,t} = i_2\right) \tag{22}$$

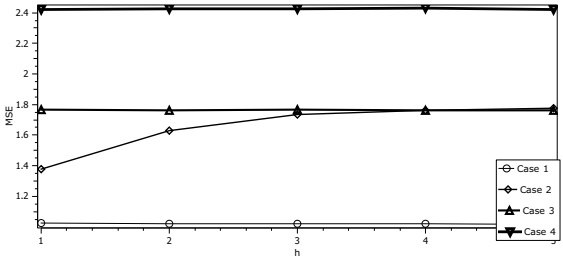
onde $\hat{P}\left(S_{j,t+h} = k | S_{1,t} = i_1, S_{2,t} = i_2\right)$ se obtem através da expressão (20);
3) São previstos utilizando as respectivas médias marginais, $\hat{z}_{jk,t+h}^{(3)} = T^{-1} \sum_{t=1}^{T} z_{jk,t}$.

4) As variáveis binárias são omitidas, i.e. $\hat{z}^{(4)}_{jk,t+h} \equiv 0$.
As previsões *out-of-sample* são geradas através do método *recursive (expanding windows) forecast*. Uma amostra de $t = 1$ a $T = 1000$ é utilizada para estimar os modelos e, posteriormente, são produzidas previsões a $h$-passos *out-of-sample*, começando em $T = 1000$. O limite superior da amostra aumenta sucessivamente em um período, os modelos são re-estimados e são produzidas novas previsões a $h$-passos começando, agora, em $T + 1$. O procedimento é repetido 1000 vezes, i.e. consideram-se 1000 previsões *out-of-sample*. O horizonte de previsão é definido como sendo $h =$ 1,2,3,4,5. Por fim, avalia-se a qualidade das previsões utilizando as estatísticas $EQM_{lh} = N^{-1} \sum_{t=T}^{T+999} \hat{e}^2_{l,t+h}$, onde $N = 1000$ é o número de réplicas consideradas e $e_{l,t+h}$ é o erro de previsão produzido pelo modelo $l$ ($l = 1,2,3,4$) no $h$-ésimo passo de previsão, i.e. $e_{l,t+h} \equiv y_{t+h} - \hat{y}^{(l)}_{t+h}$, onde $\hat{y}^{(l)}_{t+h} \equiv \boldsymbol{x}'_{t+h}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{z}}^{(l)'}_{t+h}\hat{\boldsymbol{\delta}}$ e $\hat{\boldsymbol{z}}^{(l)'}_{t+h} \equiv \left[ \begin{array}{cccc} z^{(l)}_{11,t+h} & z^{(l)}_{12,t+h} & z^{(l)}_{21,t+h} & z^{(l)}_{22,t+h} \end{array} \right]$, for $l = 1,2,3,4$.

## 5.2 Estudo de simulação de Monte Carlo: discussão dos resultados

A figura 1 expõe os resultados dos erros de previsão: $EQM_{lh}$ para $l = 1,2,3,4$ e $h = 1,2,3,4$.
Como era expectável o caso 1 conduziu aos melhores resultados, na medida em que as previsões de $y_{t+h}$ se basearam nos valores conhecidos de $z_{jk,t+h}$, e o caso 3 originou os piores resultados, já que as variáveis binárias $z_{jk,t+h}$ foram pura e simplesmente ignoradas. Quando ao caso 2, onde foi aplicada a metodologia proposta e, portanto, onde se exploraram as probabilidades de intra e inter-transiçao de estados entre os regressores categóricos produz claramente melhores resultados do que os do caso 3, onde as previsões foram obtidas recorrendo às estimativas das probabilidades marginais $P(S_{j,t+h} = k)$. Para se confirmarem as vantagens da metodolo-

Figura 1: Resultados dos erros de previsão $EQM_{lh}$




gia proposta em relação às probabilidades marginais levámos a cabo os testes Diebold-Mariano (DM) [1], que nos permitiram avaliar a significância estatística da diferença entre os EQM dos diferentes casos. Os resultados atestam a supremacia da metodologia proposta em relação às probabilidades marginais para $h = 1$ e para $h = 2$ (p-value zero) e possivelmente também para $h = 3$ (p-value 0.08). À medida que $h$ aumenta as vantagens do modelo proposto vão-se dissipando tal como seria expectável, na medida em que, dada a hipótese da estacionaridade e fraca dependência, se verifica a convergência das probabilidades condicionadas para as probabilidades estacionárias, i.e. $P\left(S_{j,t+h} = i_0 | S_{1,t} = i_1, S_{2,t} = i_2\right) \rightarrow P\left(S_{j,t} = i_0\right)$ quando $h \rightarrow \infty$.

# Referências

[1] Diebold, F.X., Mariano, R.S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics* 20, 134–144.

[2] Engle, R.F., Hendry, D.F., Richard, J.-F. (1983). Exogeneity. *Econometrica* 277–304.

[3] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 357–384.

[4] Nicolau, J. (2014). A new model for multivariate Markov chains. *Scandinavian Journal of Statistics* 41, 1124–1135.

[5] Raftery, A.E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society: Series B* 47, 528–539.

# Modelos de cura aplicados ao cancro da mama

Carina Alves
NGDEstatística, SESARAM, *anacarina.alves@gmail.com*
Ana Maria Abreu
CCCEE e CCM, Universidade da Madeira, *abreu@uma.pt*

**Palavras–chave**: Análise de sobrevivência, distribuição de Chen, estimador de Kaplan-Meier, modelo de cura, regressão de Cox

**Resumo**: A constante evolução da Medicina, associada às medidas de deteção precoce de certas doenças, têm permitido que a cura seja cada vez mais uma realidade. O cancro da mama é, em certas condições, um dos exemplos. Por este motivo, analisaram-se dados relativos a 833 mulheres diagnosticadas com cancro da mama, entre 1998 e 2005, na Região Autónoma da Madeira, aplicando métodos de Análise de Sobrevivência clássica e modelos de cura, com o objetivo de conhecer melhor a realidade da doença nesta região. Verificou-se que o risco de morte por cancro da mama é maior em mulheres na faixa etária dos 50 aos 59 anos. Dos tratamentos a que os doentes foram submetidos, a realização de cirurgia está associada a um melhor prognóstico. Os modelos de cura foram aplicados às mulheres nos estádios III e IV, obtendo-se proporções estimadas de indivíduos curados de 0.332 e de 0.074, respetivamente.

## 1 Introdução

A Análise de Sobrevivência permite estudar a distribuição do tempo de vida de indivíduos desde um instante inicial bem definido (que neste trabalho corresponde à data do diagnóstico do cancro da mama), até à ocorrência do acontecimento de interesse (aqui definido como a morte pela doença). Este acontecimento nem sempre é observado no período do estudo (dando origem a observações censuradas) quer

por o período de *follow-up* não ser suficientemente longo, quer devido à existência de indivíduos para os quais esse acontecimento nunca ocorrerá. Neste último caso, torna-se adequado a utilização de modelos de cura, os quais pressupõem que uma parte dos indivíduos fica curada. Os mais comuns, e que serão aqui utilizados, são os modelos de mistura, ([1], [5]).
Em Portugal, em 2007, a taxa de incidência padronizada de cancro da mama feminino (por 100000 mulheres) foi de 82.4, sendo a taxa de mortalidade padronizada de 19.6 por 100000, de acordo com a Direção-Geral da Saúde, [8]. Como atualmente, em determinadas circunstâncias, o cancro da mama é uma doença para a qual existe cura (por exemplo, no sentido em que os indivíduos com este diagnóstico que estejam curados têm um padrão de mortalidade semelhante ao da população em geral, ([6], [7]), a aplicação dos modelos de cura complementa a informação que se obtém através da Análise de Sobrevivência clássica. Importa notar que o conceito de cura que aqui será considerado refere-se a um indivíduo sobreviver durante, pelo menos, 5 anos.
Assim, na secção 3.1 serão utilizados métodos que não contemplam a existência de indivíduos curados com o intuito de identificar covariáveis relevantes para a sobrevivência dos indivíduos. Em seguida, na secção 3.2, os dados relativos às mulheres no estádio IV serão analisados recorrendo ao modelo de cura sem covariáveis, com o objetivo de estimar a proporção de indivíduos curados neste estádio. Ainda na mesma secção, os dados relativos às mulheres nos estádios III e IV serão analisados recorrendo ao modelo de cura com covariáveis, de modo a tentar identificar covariáveis relevantes para a sobrevivência dos indivíduos não curados e para a proporção de indivíduos curados.

# 2 Metodologia

Realizou-se um estudo prospetivo utilizando dados fornecidos pelo Registo Oncológico Regional - Sul. A base de dados tem 833 re-

gistos, os quais correspondem a mulheres diagnosticadas com cancro da mama (critério de inclusão), exceto as do tipo inflamatório (critério de exclusão), entre 1998 e 2005, na Região Autónoma da Madeira. O *follow-up* foi feito até fevereiro de 2012, garantindo um período mínimo de 5 anos para todos os indivíduos. O tempo máximo de *follow-up* foi de cerca de 14 anos, sendo a mediana igual a 7.4 anos. As covariáveis consideradas foram: "Grupo Etário", "Estádio", "Cirurgia" e "Quimioterapia", as quais estão definidas na Tabela 1.

As probabilidades de sobrevivência foram estimadas através do estimador de Kaplan-Meier (KM) da função de sobrevivência (f.s.) e, para determinar a influência das covariáveis no tempo de vida, foi usado o modelo de regressão de Cox [4], que é um modelo de riscos proporcionais. Este modelo pode ser escrito à custa da função de risco, no instante $t$, na forma que se segue,

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) = h_0(t)\exp(\beta_1 z_1 + \cdots + \beta_p z_p), \qquad (1)$$

onde $\mathbf{z} = (z_1, \ldots, z_p)'$ representa o vetor de covariáveis associado a cada indivíduo, $\beta_1, \ldots, \beta_p$ os correspondentes coeficientes de regressão e $h_0(t)$ a função de risco subjacente.

Posteriormente aplicou-se o modelo de cura sem covariáveis, [5], o qual se pode escrever da seguinte forma,

$$S(t) = p + (1-p)S_d(t), \qquad (2)$$

onde $S(t)$ designa a f.s. populacional da variável aleatória (v.a.) $T$ (que representa o tempo de vida do indivíduo), $S_d(t)$ a f.s. correspondente aos indivíduos não curados e $p$ a proporção de indivíduos curados na população. Para a distribuição do tempo de vida dos indivíduos não curados foi utilizada a distribuição de Chen [3], ou seja, que tem a f.s. caracterizada por (3),

$$S_d(t) = \exp[\lambda(1 - \exp(t^\alpha))], \qquad (3)$$

onde $\lambda$ é o parâmetro de escala e $\alpha$ o parâmetro de forma.

Por último, considerou-se o modelo de cura semi-paramétrico com

covariáveis [1], que é dado por

$$S(t|\mathbf{x},\mathbf{z}) = p(\mathbf{z}) + (1 - p(\mathbf{z}))S_d(t|\mathbf{x}),$$

onde $\mathbf{x}$ e $\mathbf{z}$ são vetores de covariáveis. Para os indivíduos não curados foi usado o modelo de Cox e para a proporção de cura o modelo de regressão logístico.
A utilização dos modelos de cura pressupõe que o *follow-up* seja longo de modo a que o acontecimento de interesse possa ser observado para a maioria dos indivíduos não curados. Tal facto pode ser visível através da estabilização da estimativa de KM da f.s. e complementado analiticamente através da estimativa da mediana do tempo de vida dos indivíduos não curados, [10], visto que o tempo de *follow-up* deve ser superior ao valor da mediana. Note-se, contudo, que é condição suficiente o valor da mediana obtido através da estimativa de KM verificar o pressuposto anterior, mas não necessária dado que a mediana dos não curados será, necessariamente, inferior.
Os dados foram analisados, utilizando o programa de análise estatística *PASW Statistics for Windows*, versão 18, bem como o *software* R [9], em particular, o *package smcure* [2].

# 3 Resultados

## 3.1 Análise de sobrevivência clássica

A informação relevante da base de dados está organizada na Tabela 1, com o objetivo de evidenciar a eventual relação da variável "Estádio" com as restantes. Na Tabela 1, os valores entre parêntesis referem-se às percentagens dentro de cada categoria de cada variável e a idade é indicada em anos. Das 833 mulheres deste estudo, foi observada a morte pela doença em 253 (30.4%). Houve 463 (55.6%) observações censuradas correspondentes a mulheres que estavam vivas no final do *follow-up* e 117 (14.0%) censuradas devido a morte por outra causa ou por causa desconhecida.

Tabela 1: Características das mulheres diagnosticadas com cancro da mama entre 1998 e 2005. Percentagem dentro das variáveis.

| | Estádio | | | | |
|---|---|---|---|---|---|
| | 0 ou I (%) | II (%) | III (%) | IV (%) | Desc. (%) |
| **Gp Etário** | | | | | |
| <40 (n=54) | 14 (25.9) | 19 (35.2) | 6 ( 11.1) | 2 (3.7) | 13 (24.1) |
| 40 a 49 (n=156) | 39 (25.0) | 50 (32.1) | 16 (10.3) | 6 (3.8) | 45 (28.8) |
| 50 a 59 (n=199) | 29 (14.6) | 72 (36.2) | 18 (9.0) | 13 (6.5) | 67 (33.7) |
| >=60 (n=424) | 34 (8.0) | 112 (26.4) | 38 (9.0) | 31 (7.3) | 209 (49.3) |
| **Cirurgia** | | | | | |
| Fez (n=745) | 116 (15.6) | 253 (34.0) | 70 (9.4) | 17 (2.3) | 289 (38.8) |
| Não fez (n=88) | 0 (0.0) | 0 (0.0) | 8 (9.1) | 35 (39.8) | 45 (51.1) |
| **Quimio** | | | | | |
| Fez (n=456) | 51 (11.2) | 188 (41.2) | 68 (14.9) | 30 (6.6) | 119 (26.1) |
| Não fez (n=377) | 65 (17.2) | 65 (17.2) | 10 (2.7) | 22 (5.8) | 215 (57.0) |

Desc. = Desconhecido; Gp Etário = Grupo Etário; Quimio = Quimioterapia

A dificuldade em identificar a presença ou ausência de metástases à distância, à data do diagnóstico, originou um grande número de casos em que o estádio é desconhecido. No entanto, suspeita-se que a maior parte dos casos no estádio desconhecido sejam na realidade do estádio II, devido a ser neste último onde é mais frequente ocorrer o diagnóstico e devido à semelhança das estimativas de KM das correspondentes f.s..

Na Figura 1, observa-se que o prognóstico é tanto melhor quanto menos grave for o estádio. As mulheres diagnosticadas com os estádios 0 ou I, a partir dos 8 anos após o diagnóstico, mantêm uma probabilidade de sobrevivência estimada acima dos 85%. Por seu lado, as mulheres diagnosticadas no último estádio, o mais grave, têm uma probabilidade muito reduzida (6.9%) de estarem vivas ao fim de 12.8 anos, e é nos 2 primeiros anos após o diagnóstico que se verifica um decrescimento acentuado da probabilidade de sobrevivência.

A aplicação do modelo de regressão de Cox deu origem aos resultados

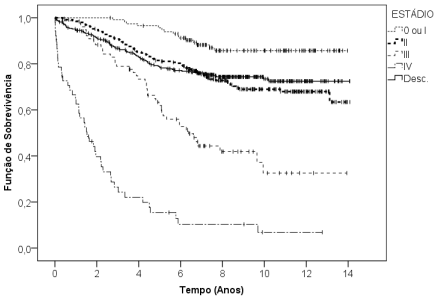



Figura 1: Estimativa de Kaplan-Meier das f.s. por estádio.

que constam na Tabela 2. Estima-se que as mulheres no grupo etário dos 50 a 59 anos têm um risco de morte que é 37.8% superior ao risco de morte no grupo etário $< 40$ anos. As mulheres nos estádios II, III, IV ou desconhecido têm um acréscimo no risco de morte de 126.6%, 353.7%, 818.3% ou 98.5%, respetivamente, em relação às mulheres no estádio 0 ou I. Quem fez cirurgia tem apenas 30% do risco de morte de quem não fez cirurgia e quem fez quimioterapia tem um risco de morte que é 36.4% superior ao risco de quem não fez.

Tabela 2: Modelo de Cox.

| Covariáveis | $\hat{\beta}$ | Exp($\hat{\beta}$) | Int. confiança 95% | $p^*$ |
|---|---|---|---|---|
| G. Etário 50 a 59 anos | .321 | 1.378 | (1.048, 1.813) | .022 |
| Estádio II | .818 | 2.266 | (1.292, 3.972) | .004 |
| Estádio III | 1.512 | 4.537 | (2.471, 8.330) | <.001 |
| Estádio IV | 2.217 | 9.183 | (4.672, 18.052) | <.001 |
| Estádio desconhecido | .686 | 1.985 | (1.132, 3.481) | .017 |
| Cirurgia | -1.198 | 0.302 | (.204, .445) | <.001 |
| Quimioterapia | .311 | 1.364 | (1.010, 1.842) | .043 |

*teste de Wald.

## 3.2 Estimação do modelo de cura

Foi aplicado o modelo de cura sem covariáveis apenas ao grupo das mulheres no estádio IV, não só por o tempo de *follow-up* ser suficiente, [10], (a mediana obtida através da estimativa de KM é 1,5 anos) mas também por a distribuição de Chen ser adequada para situações deste tipo. De facto, quando o parâmetro de forma, $\alpha$, é inferior a 1, a função de risco começa por ter um valor elevado e vai decrescendo ao longo do tempo até ao instante $\left(\frac{1}{\alpha}-1\right)^{\frac{1}{\alpha}}$, momento em que a função inverte o sentido de crescimento. Esta forma da função de risco é assim compatível com a estimativa de KM da f.s. (ver Figura 1), em que inicialmente existe um decrescimento rápido, sendo a estabilização justificada pela quase inexistência de indivíduos não curados. Como nos estádios II e III o decrescimento inicial não é tão rápido, esta não é uma boa distribuição para modelar estes dados. Assim sendo, tendo em conta as equações (2) e (3), o modelo de cura que se obteve para o estádio IV foi

$$\hat{S}(t) = 0.074 + 0.926\exp[0.3339661(1-\exp(t^{0.4543725}))],$$

a partir do qual se estima que a mediana do tempo de vida das mulheres não curadas seja de 1.29 anos. Neste modelo observa-se ainda que a estimativa da proporção de indivíduos curados é de 0.074.
A inexistência ou exiguidade de casos nos estádios II e III, respetivamente, em que a cirurgia não ocorreu levou a que esta covariável não fosse considerada no modelo de cura. Através do *package smcure* tentámos ajustar um modelo de cura para estes estádios. No entanto, no estádio II a estimativa da mediana do tempo de vida foi de 6.6 anos, ultrapassando o valor mínimo do *follow-up*, pelo que não foi considerado. No estádio III a mediana já foi de 1.28, viabilizando a utilização deste modelo. Contudo, na proporção de indivíduos curados não houve covariáveis significativas (apesar de termos mantido a quimioterapia devido à estrutura do algoritmo do *package*) e na f.s. dos indivíduos não curados restou apenas a quimioterapia. Devido à instabilidade da convergência do algoritmo, não se obtém um valor

único para a significância do teste variando, no caso da quimioterapia, em valores ora inferiores a 0.05 ora inferiores a 0.10. O valor da estimativa do parâmetro para a quimioterapia foi -1.092353, resultando numa proporção de indivíduos curados de 0.332. No estádio IV não houve covariáveis significativas, não constituindo assim uma vantagem em relação ao modelo anterior.

# 4 Conclusão

Há uma grande diferença na probabilidade estimada de sobrevivência das mulheres consoante o estádio em que a doença é detetada (e.g. aos 10 anos: estádio II – 80% vs estádio IV – 10%). O elevado número de casos em que o estádio é desconhecido constitui uma limitação deste estudo. Através do modelo de Cox verifica-se que o grupo etário dos 50 aos 59 anos é o único que tem um risco de morte acrescido (38%) em relação ao grupo com idade inferior a 40 anos. Além disso, a cirurgia tem um efeito benéfico (há um decrécimo de 70% no risco de morte) e a quimioterapia é um fator de pior prognóstico. Note-se que esta última afirmação deve ser interpretada com algum cuidado uma vez que este tratamento está também relacionado com a severidade da doença. Com base no modelo de cura com a distribuição de Chen, obtém-se uma estimativa da proporção de indivíduos curados no estádio IV de 0.074. O modelo de cura com a covariável quimioterapia permite estimar a proporção de indivíduos curados no estádio III em 0.332, embora a instabilidade do algoritmo em relação à significância seja algo a merecer mais estudo.

# Agradecimentos

Investigação parcialmente financiada pela FCT – Fundação para a Ciência e a Tecnologia, projeto PEst-OE/MAT/UI0219/2011 – Projeto Estratégico do CCM (Centro de Ciências Matemáticas).

# Referências

[1] Abreu, A.M., Rocha, C.S. (2013). A Parametric Cure Model with Covariates. Em: Lita da Silva, J., Caeiro, J., Natário, I., Braumann, C.A. (eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications* 37–45, Springer–Verlag, Berlin Heidelberg.

[2] Cai, C., Zou, Y., Peng, Y., Zhang, J. (2013). smcure: Semiparametric mixture cure model. R package version 2.0.

[3] Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters* 49, 2, 155–161.

[4] Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society.* B, 34, 187–220.

[5] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 39, 1–38.

[6] Haybittle, J.L. (1965). A two–parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association* 60, 16–26.

[7] Pocock, S.J., Gore, S.M., Kerr, G.R. (1982). Long term survival analysis: the curability of breast cancer. *Statistics in Medicine* 1, 93–104.

[8] Programa Nacional para as Doenças Oncológicas e Direção de Serviços de Informação (2013). Portugal. Doenças Oncológicas em Números – 2013. Ministério da Saúde.

[9] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-00051-07-0, URL http://www.r-project.org/.

[10] Yu, B., Tiwari, R.C., Cronin, K.A., Feuer, E.J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine* 23, 1733–1747.

# Uma abordagem não paramétrica à previsão da dose individualizada de *atracurium*

Conceição Rocha
Faculdade de Ciências, Universidade do Porto, *mnrocha@fc.up.pt*
Maria Eduarda Silva
Faculdade de Economia, Universidade do Porto & CIDMA, *mesilva@fep.up.pt*
Teresa Mendonça
Faculdade de Ciências, Universidade do Porto & CIDMA, *tmendo@fc.up.pt*

**Palavras–chave**: Previsão, aplicação às ciências médicas

**Resumo**: Numa intervenção cirúrgica a definição da dose de fármaco deve atender à intensidade do estímulo cirúrgico e à resposta do paciente. A variabilidade farmacocinética e farmacodinâmica entre pacientes impede que se aplique a todos os pacientes um regime único de concentração de fármaco ou de combinação de fármacos. O objetivo deste trabalho é obter uma previsão robusta em tempo real da dose individualizada de *atracurium* a administrar em estado estacionário ($u_{ss}$) que conduza e fixe o nível do bloqueio neuromuscular, NMB, no valor desejado. Para atingir este fim, a análise espetral de Walsh-Fourier e a regressão robusta são conjugadas com sucesso para conduzirem a um conjunto de preditores informativo a usar na análise preditiva individualizada com base no princípio de aprendizagem. As previsões assim obtidas apresentam erros absolutos relativos inferiores a 10% em mais de 99% dos casos.

# 1 Introdução

O facto de, habitualmente, serem administrados três tipos de fármacos na anestesia geral reforça a necessidade de estratégias, de previsão da dose individualizada, adequadas a este tipo de fármacos e, em particular, à sua utilização na anestesia geral. A adequação da dose de cada fármaco às características do paciente para, simultaneamente, obter o efeito fisiológico desejado e evitar efeitos secundários requer, portanto, estratégias rápidas e precisas da avaliação da dose de fármaco adequada ao paciente e à intervenção clínica.
A prática clínica seguida para a administração do agente NMB *atracurium* consiste em administrar um *bolus* inicial de 500 $\mu g\ kg^{-1}$ após o que, mas não antes de decorridos 10 minutos, uma dose de manutenção é infundida de modo contínuo. A duração, relativamente curta, da ação do *atracurium* leva a que, na busca da dose adequada a infundir de forma constante, ocorram variações consideráveis no nível de NMB antes de este atingir um valor relativamente estável.
De acordo com estudos anteriores [8, 6] é razoável assumir que a resposta do paciente ao *bolus* inicial inclui informação valiosa que deve ser tida em conta na caracterização da resposta individual. Em particular, a resposta ao *bolus*, doravante representado por $r(t)$, na fase de indução (durante os primeiros 10 minutos) é caracterizada, principalmente, por mudanças de nível e apresenta uma grande variabilidade interindividual como se ilustra na Figura 1. Isto sugere a utilização de medidas extraídas de $r(t)$, $t \in [0{,}10]$, como possíveis preditores para a conceção do chamado perfil de dose de manutenção a fim de obter o efeito pretendido e predefinido. Os atributos aqui considerados, e descritos em pormenor na secção 2, destinam-se a descrever a tendência individual de $r(t)$. Tendo caracterizado o indivíduo, serão considerados dois algoritmos na análise preditiva levada a cabo na secção 3. Note-se que a disponibilidade do modelo populacional [7] permite o cálculo da $u_{ss}$ uma vez que em ambiente clínico isso não acontece. Finalmente, na secção 4 apresentam-se alguns comentários finais.

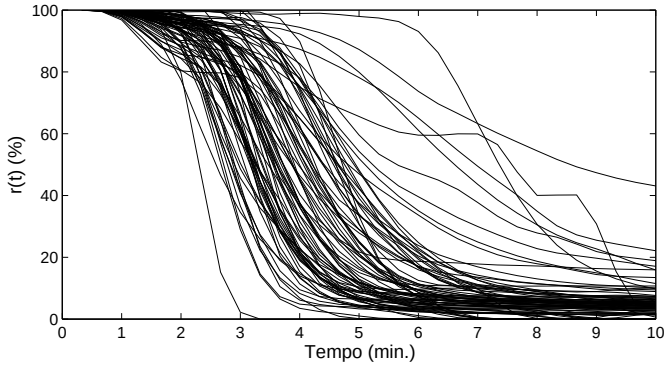



Figura 1: Resposta individual à administração inicial do *bolus* de 500 $\mu g\ kg^{-1}$ de *atracurium*: 84 casos do nível de NMB, $r(t)$, induzido pela administração de *atracurium*.

# 2 Caracterização da resposta individual

O nível NMB nos primeiros 10 minutos é principalmente caracterizado por mudanças na tendência e apresenta grande variabilidade interindividual entre os pacientes. Este facto sugere que a caracterização individual seja realizada em duas etapas: detetar primeiro os pontos de mudança e, em seguida, caracterizar a tendência entre os pontos de mudança através do declive. Nesta secção, a análise espetral de Walsh-Fourier (WFA) [4] é usada para detetar pontos de mudança na resposta NMB. De seguida, estima-se o declive entre os pontos de mudança através da regressão robusta. Finalmente, a forma global da resposta de NMB é caracterizada pela área sob a curva durante os primeiros 10 minutos.

No presente estudo, os periodogramas Walsh dos dados [9] mostram picos correspondentes, aproximadamente, aos períodos médios de 2.8 e 14.2 minutos; indicam outros, menos frequentes, que correspondem, aproximadamente, a um período médio de 6.0, 7.1, 8.5 minutos. Tendo em consideração que: a frequência do registo do

nível de NMB, $r(t)$, é de 20 em 20 segundos; os três períodos médios de 6.0, 7.1, 8.5 minutos são consecutivos; existe ruído sobreposto no sinal e existe correlação entre as observações do sinal, os resultados sugerem $v_i = r(t_i)$, $t_i \in \{2.7, 3.0, 7.3, 7.7\}$,$i = 1, \ldots, 4$, como variáveis adequadas para preditores.

A Figura 2(a) ilustra as quatro variáveis selecionadas como representativas dos pontos de mudança das séries. Estas variáveis dividem as séries temporais em três zonas e o passo seguinte é associar a cada uma destas três zonas um declive que será estimado por um método robusto [2, 3], devido à presença de ruído e ao reduzido número de observações.

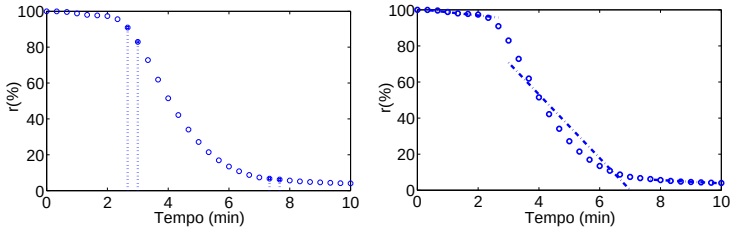



Figura 2: Representação geométrica das variáveis preditoras $v_i$,$i = 1, \ldots, 4$, em (a) e das variáveis preditoras $v_i$,$i = 5,6,7$, em (b), para um caso real.

Portanto, as tendências do nível NMB, $RT_1$, $RT_2$, $RT_3$, nos intervalos de tempo [0, 2.7], [3.0, 7.3], [7.7, 10.0], são avaliadas por meio de regressão robusta. Assim, são adicionadas mais três variáveis $v_5 = RT_1$, $v_6 = RT_2$, $v_7 = RT_3$ para caracterizar o perfil do nível de NMB durante os primeiros 10 minutos. A Figura 2(b) ilustra os resultados da estimação desses declives para um caso clínico real. Para completar a caracterização da resposta individual nos primeiros 10 minutos após a administração do *bolus*, uma última variável ou preditor, $v_8$, é incluída neste conjunto de variáveis. Esta variável, designada por FAUC, representa a razão entre a área sob a curva durante os primeiros 10 minutos (área preenchida na Figura 3(a)) e a área total (área do retângulo). Este conjunto de oito variáveis,

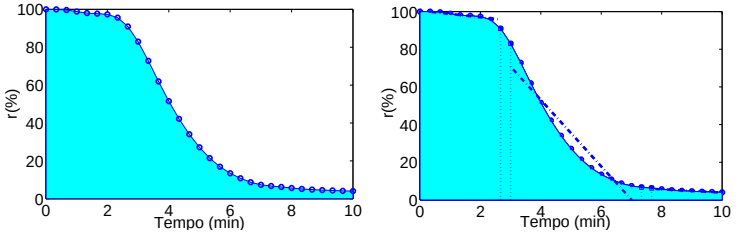



Figura 3: Representação geométrica das variáveis preditoras: $v_8$ em (a), e $v_i$,$i = 1,\ldots,8$, em (b), para um caso real.

$\mathcal{P} = \{v_i\}_{i=1,\ldots,8}$, representadas geometricamente na Figura 3(b), caracteriza a resposta de cada paciente à administração do mesmo *bolus* de *atracurium* e constitui o conjunto de atributos a ser utilizado para prever a dose de *atracurium* em estado estacionário $u_{ss}$.

# 3 Abordagem não paramétrica - Classificação supervisionada

Para conduzir a análise preditiva individualizada com base no princípio de aprendizagem, são consideradas duas famílias conhecidas de métodos de classificação supervisionada: aprendizagem baseada em casos, *Instance Based* [5] (algoritmo *k-vizinhos mais próximos* - *k*NN), e árvores de decisão, *Decision Trees* [1] (algoritmo baseado na lógica, *classificação e árvores de regressão* - CART). A matriz de dados a utilizar como conjunto de treino neste estudo para a obtenção dos classificadores é um conjunto de dados simulados a partir do modelo estocástico para o nível de NMB induzido por administração de *atracurium* e descrito em [7]. Tal como já foi referido, o acesso à dose $u_{ss}$ só é possível para os dados obtidos por simulação. Assim sendo, simulam-se 5000 casos do nível de NMB, $r(t)$, usando o modelo populacional [7] bem como a correspondente dose no estado

estacionário $u_{ss}$, que induz o valor de referência $r(t) = 10\%$. Este conjunto de dados será de agora em diante designado por $\mathcal{NB}_{\mathcal{R}5000}$. Para cada um dos 5000 casos são calculados, não só os preditores $\mathcal{P} = \{v_i\}_{i=1,\ldots,8}$ como, também, o nível de NMB no estado estacionário $r_{ss}$ induzido pela previsão da dose $u_{ss}$. O facto de, clinicamente, o nível de NMB no estado estacionário poder assumir valores compreendidos entre 5% e 15%, sem que tal seja encarado como um problema, irá, também, ser utilizado para avaliar a eficácia da metodologia proposta.

Como a eficácia do classificador $k$NN está relacionada com o número de vizinhos ($k$) utilizados na aplicação do método $k$NN, é feita uma seleção prévia com base na precisão da previsão quando são considerados os valores de $k$ de 1 a 10 para os classificadores. Nesta aplicação, a maior precisão ocorre para $k = 2$.

A previsão de dose para $\mathcal{NB}_{\mathcal{R}5000}$ apresenta um erro absoluto percentual máximo de 17% para o algoritmo $CART$ e de 15% para o $k$NN. As duas previsões levam a erros absolutos percentuais, $|\hat{u}_{ss} - u_{ss}|/u_{ss} \times 100$, inferiores a 10% em mais de 99.7% dos casos. Para a precisão da previsão e para a validação cruzada 'deixa-um-fora' obtiveram-se, respetivamente, os valores da raiz quadrada do erro quadrático médio ($RMSE$) de 0.245 e 0.403 para o classificador $CART$, e de 0.120 e 0.225 para o classificador $k$NN. Estes resultados encontram-se resumidos na tabela 1 e indicam que o classificador $k$NN é mais preciso do que o $CART$.

As doses previstas $\hat{u}_{ss}$ para todos os 5000 casos conduziram a valores para o nível de NMB dentro do intervalo clinicamente requerido $[5\%, 15\%]$.

# 4 Comentários

O sistema de bloqueio neuromuscular induzido pelo *atracurium* é um sistema dinâmico. Sistemas dinâmicos aparecem frequentemente em biomedicina e, algumas vezes, não são identificáveis em tempo real

| Classificador | | $CART$ | $k$NN |
|---|---|---|---|
| $RMSE$ | Precisão da previsão | 0.245 | 0.120 |
| | 'deixa-um-fora' | 0.403 | 0.225 |
| Máximo do erro absoluto percentual, EAP | | 17% | 15% |
| % de casos com EAP inferior a 10% | | 99.72% | 99.96% |

Tabela 1: Algumas medidas dos erros para os dois classificadores por regressão, $CART$ e $k$NN.

ou são mal identificados, como neste caso.
Da aplicação dos dois métodos de classificação supervisionada resultam dois classificadores para prever a variável $u_{ss}$. Ambos os classificadores apresentam desempenhos comparáveis com erros absolutos percentuais inferiores a 10% em quase todos os casos.
É de referir que a aplicação destes modelos a um conjunto de 46 pacientes submetidos a anestesia geral conduziu a previsões da dose $u_{ss}$ próximas do valor da dose média que lhes foi administrada para manter o nível do NMB próximo de 10%. Tendo em conta que, neste conjunto de 46 pacientes, as doses necessárias para estabilizar o nível do NMB próximo de 10% variam entre 3.9 e 20.8 $\mu g\, kg^{-1}\, min^{-1}$ e que o clínico inicia a administração do fármaco com uma dose próxima de 5 $\mu g\, kg^{-1}\, min^{-1}$, a previsão obtida pelo método proposto neste trabalho constitui, de modo geral, uma melhor aproximação à dose individualizada.
Este mesmo estudo pode facilmente estender-se a outros protocolos de administração do fármaco, a outros fármacos anestésicos ou mesmo a outros sistemas fisiológicos.

# Agradecimentos

Conceição Rocha agradece à FCT/ESF, bolsa SFRH/BD/61781 /2009. Trabalho desenvolvido com suporte parcial do *FEDER* através do *COMPETE*—Programa Operacional Factores de Competitividade e de fundos Portugueses através do CIDMA e da FCT com os projetos PEst-C/MAT/UI4106/2011 com o número COMPETE FCOMP-01-0124-FEDER-022690 e GALENO - Modeling and Control for personalized drug administration, PTDC/SAU-BEB/103667/2008.

# Referências

[1] Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and Regression Trees*. Edições Wadsworth and Brooks. Monterey, CA.

[2] Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Publicações Sage.

[3] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.

[4] Kohn, R. (1980). On the Spectral Decomposition of Stationary Time Series using Walsh Functions. *Advances in Applied Probability* 12, 183–199.

[5] Kotsiantis, S.B. (2007). Supervised machine learning: a review of classification rechniques. *Informatica* 31, 249–268.

[6] Rocha, C., Mendonça, T., Silva, M.E. (2009). Online individualized dose estimation. *Proceedings of the 6th IEEE International Symposium on Intelligent Signal Processing (WISP2009)*, Budapest, Hungary, 26–28.

[7] Rocha, C., Mendonça, T., Silva, M.E. (2013). Modelling neuromuscular blockade: a stochastic approach based on clinical data. *Mathematical and Computer Modelling of Dynamical Systems* 19, 540–556.

[8] Silva, M.E., Mendonça, T., Silva, I., Magalhães, H. (2005). Statistical analysis of neuromuscular blockade response: contributions to an automatic controller calibration. *Computational Statistics & Data Analysis* 49, 955–968.

[9] Stoffer, D.S. (1991). Walsh-Fourier analysis and its statistical applications. *Journal American Statistical Association* 86, 461–479.

# Aplicação do método dos excessos de nível a valores extremos de precipitação na ilha da Madeira

Délia Gouveia-Reis
Universidade da Madeira, Centro de Estatística e Aplicações da Universidade de Lisboa, Centro de Investigação de Montanha, *delia@uma.pt*

Luiz Guerreiro Lopes
Universidade da Madeira, Centro de Investigação de Montanha, Instituto de Ciências Agrárias e Ambientais Mediterrânicas, *lopes@uma.pt*

Sandra Mendonça
Universidade da Madeira, Centro de Estatística e Aplicações da Universidade de Lisboa, *smfm@uma.pt*

**Palavras–chave**: Estatística de extremos, *peaks over threshold*, precipitação intensa

**Resumo**: A metodologia POT (*peaks over threshold*) requer a análise das observações que excedem um certo limiar. Neste estudo, são encontradas estimativas para os limiares correspondentes aos valores diários de precipitação na Ilha da Madeira referentes ao período de 1950 a 1980. Apresentam-se também as estimativas resultantes desses valores para os parâmetros da função de distribuição generalizada de Pareto.

## 1 Introdução

A Ilha da Madeira, situada no Atlântico Norte Oriental, apesar de ser uma ilha vulcânica com uma área de apenas 737 km$^2$, apresenta diferentes regiões relativamente aos extremos de precipitação [4], em consequência da sua complexa orografia e do pronunciado desnível

das suas vertentes. Os extremos de precipitação constituem um factor importante na ocorrência de cheias rápidas e fluxos de lamas e de detritos, que têm marcado o passado da ilha. O evento natural extremo mais significativo, ocorrido no dia 9 de Outubro de 1803, provocou a morte a mais de 800 pessoas [5]. Dos eventos ocorridos num passado mais recente, foi o de 20 de Fevereiro de 2010 o mais devastador, com 45 mortes oficiais e danos significativos em muitas infraestruturas [2]. Assim, a Ilha da Madeira, dadas as suas características orográficas e o seu historial de eventos naturais extremos induzidos por chuvas intensas, constitui um laboratório natural para o estudo de extremos de precipitação e sua modelação estatística. O estudo de valores extremos por meio da abordagem de Gumbel ou método dos blocos apresenta a desvantagem de considerar apenas um máximo por bloco, o que pode limitar esse estudo [1]. Tal limitação é atenuada pela abordagem POT (*peaks over threshold*) ou método dos excessos de nível, uma vez que todas as observações que excedem um certo nível ou limiar (*threshold*), $u$, são integradas na análise. A escolha do limiar é um problema controverso e em aberto, envolvendo uma conciliação entre a redução do viés dos estimadores ou da sua variância, em consequência dos valores de $u$ serem, respectivamente, mais ou menos elevados [3]. Aqui, com base nos dados disponíveis, procura-se escolher esse limiar para os valores diários de precipitação registados na Ilha da Madeira durante um período de 31 anos. Neste estudo, apresentam-se também as estimativas dos parâmetros de escala e de forma da função de distribuição generalizada de Pareto, obtidas por meio do método da máxima verosimilhança.

## 2 Metodologia

Seja $X$ uma variável aleatória com uma dada função de distribuição. Um acontecimento extremo de $X$ pode ser definido como um valor de $X$ que ultrapassa um certo limiar $u$. Sob determinadas condições, e para $u$ suficientemente grande, a função de distribuição de $X - u$, dado $X > u$, é dada aproximadamente por $H(y) =$

$1-\left(1+\frac{\gamma y}{\sigma}\right)^{-1/\gamma}$, onde $y>0$, $1+\frac{\gamma y}{\sigma}>0$, $\sigma>0$ e $\gamma\neq 0$. Para $\gamma=0$, $H(y)=1-e^{-\frac{y}{\sigma}}$. A função $H$ é denominada função de distribuição de Pareto generalizada, sendo os valores $\sigma$ e $\gamma$ os parâmetros de escala e de forma, respectivamente. Assim, a escolha do limiar $u$ é um ponto fulcral na definição de eventos extremos. Um método gráfico para tal escolha assenta na interpretação do gráfico de vida residual média. Considere-se a amostra aleatória $x_1, x_2,..., x_n$ de observações independentes e identicamente distribuídas. O gráfico de vida residual média é constituído pelos pontos do conjunto $\left\{\left(u, \frac{1}{k}\sum_{i=1}^{k}(x_{(i)}-u)\right) : u<x_{max}\right\}$, onde as excedências $x_{(1)},..., x_{(k)}$ são os $k$ valores que excedem $u$, $x_{max}$ é o maior valor da amostra aleatória, sendo os valores $x_{(i)}-u$, com $i=1,...,k$, denominados excessos. Se a função de distribuição de Pareto generalizada for um modelo adequado para os excessos em relação a um dado limiar $u_0$, com $u_0<u$, então o gráfico de vida residual média deverá ser aproximadamente linear em $u$. Outro método gráfico complementar para a escolha do limiar $u$ resulta do ajuste da distribuição de Pareto generalizada para uma variedade de limiares. Se a distribuição de Pareto generalizada for um modelo adequado para os excessos relativos a $u_0<u$, então também o será para os excessos relativos a $u$, sendo idênticos os parâmetros de forma das duas distribuições. Por outro lado, se denotarmos por $\sigma_{u_0}$ o valor do parâmetro de escala da distribuição associada ao limiar $u_0$, então $\sigma=\sigma_{u_0}+\gamma(u-u_0)$, variando $\sigma$ com $u$. No entanto, se tomarmos $\sigma^*=\sigma-\gamma u$, então, $\sigma^*$ é constante relativamente a $u$. Assim, se a distribuição de Pareto generalizada for um modelo adequado para os excessos relativos a este valor, as estimativas de $\sigma^*$ e $\gamma$ deverão ser aproximadamente constantes para valores superiores a $u_0$ [1].

# 3 Resultados

Os dados analisados neste estudo correspondem a séries de valores diários de precipitação na Ilha da Madeira referentes ao período

de 1950 a 1980, fornecidos pelo Departamento de Hidráulica e Tecnologias Energéticas do Laboratório Regional de Engenharia Civil (LREC). Estes dados são provenientes de 12 estações meteorológicas e postos udométricos, representados na Figura 1 e indicados na Tabela 1, então mantidos pela antiga Junta Geral do Distrito Autónomo do Funchal.

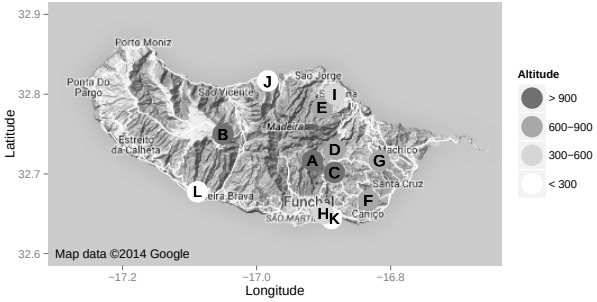



Figura 1: Localização geográfica e faixa de altitude das estações consideradas.

Para cada uma das estações, a análise do gráfico de vida residual média foi complementada com a análise dos gráficos das estimativas dos parâmetros do modelo em função de $u$, sendo o método da máxima verosimilhança aplicado para a obtenção das estimativas $\sigma^*$ e $\gamma$. Tomando como exemplo a estação Funchal (K), observa-se que um bom candidato para o limiar $u$ é o valor 35 mm, tanto no correspondente gráfico de vida residual média (Figura 2, esquerda), como nos gráficos de $\widehat{\sigma}^*$ e de $\widehat{\gamma}$ em função do limiar (Figura 2, direita). Na Tabela 1, são apresentados os valores para o limiar $u$ e o respectivo número de excedências $k$, resultantes da interpretação dos gráficos correspondentes a cada estação. As estimativas dos parâmetros de escala e de forma da função de distribuição de Pareto generalizada obtidas por meio do método da máxima verosimilhança, considerando cada um dos limiares escolhidos, são também apresentadas na Tabela 1.
A estação localizada à maior altitude, Areeiro (A), apresenta o maior

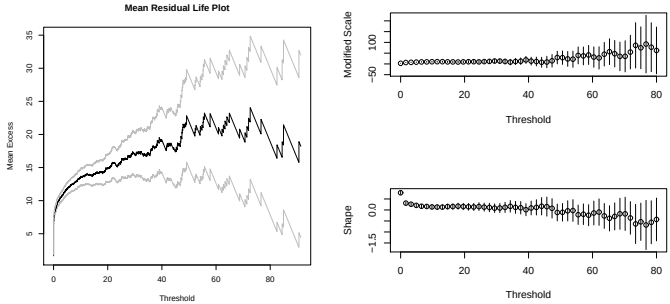



Figura 2: Gráficos de vida residual média (esq.) e de $\widehat{\sigma}^*$ (*Modified Scale*) e $\widehat{\gamma}$ (*Shape*) (dir.) em função do limiar para valores diários de precipitação, em mm, da estação do Funchal (K).

Tabela 1: Limiar, número de excedências e estimativas dos parâmetros de escala e de forma.

| Nome (Id.) | Altitude | $u$ | $k$ | $\widehat{\sigma}$ | $\widehat{\gamma}$ |
|---|---|---|---|---|---|
| Areeiro (A) | 1610 m | 120 | 110 | 48.67 | 0.09 |
| Bica da Cana (B) | 1560 m | 100 | 109 | 34.01 | −0.07 |
| Montado do Pereiro (C) | 1260 m | 110 | 58 | 44.75 | −0.20 |
| Ribeiro Frio (D) | 874 m | 100 | 97 | 49.84 | −0.23 |
| Queimadas (E) | 860 m | 80 | 111 | 34.99 | −0.26 |
| Camacha (F) | 680 m | 70 | 92 | 35.59 | −0.20 |
| Santo da Serra (G) | 660 m | 80 | 87 | 41.41 | −0.12 |
| Sanatório (H) | 380 m | 40 | 124 | 16.95 | 0.19 |
| Santana (I) | 380 m | 60 | 91 | 27.62 | 0.14 |
| Ponta Delgada (J) | 136 m | 50 | 86 | 24.35 | 0.18 |
| Funchal (K) | 58 m | 35 | 91 | 14.78 | 0.14 |
| Lugar de Baixo (L) | 15 m | 30 | 106 | 11.74 | 0.05 |

valor para $u$. Este valor, tal como os limiares correspondentes a todas as demais estações com altitude superior a 870 m, é maior ou igual a 100 mm. No entanto, maiores altitudes não correspondem necessariamente a limiares mais elevados. A estação Sanatório (H), por exemplo, que se encontra localizada na vertente sul a uma altitude de 380 m, apresenta um valor de $u$ inferior aos obtidos para as estações Santana (I) e Ponta Delgada (J), ambas localizadas na vertente norte a 380 m e 136 m, respectivamente. Observa-se também que os limiares correspondentes às estações localizadas mais perto do mar são inferiores aos valores apresentados pelas estações localizadas mais no interior da ilha. Por exemplo, a estação Camacha (F) apresenta um limiar menor do que as estações Queimadas (E) e Santo da Serra (G), localizadas a uma maior distância da linha da costa. Observa-se, igualmente, que o valor da estação Santana (I), localizada na vertente norte, é superior ao da estação Ponta Delgada (J), localizada na mesma vertente, porém mais próximo à costa. Da mesma forma, os valores dos limiares das estações Funchal (K) e Lugar de Baixo (L), localizadas na vertente sul, são inferiores ao da estação Sanatório (H), situada mais no interior da ilha. Estas três estações e as estações Santana (I) e Ponta Delgada (J), localizadas também junto ao mar mas na vertente norte, apresentam estimativas positivas $\widehat{\gamma}$. Com a excepção da estação Areeiro (A), todas as estações localizadas acima dos 600 m apresentam estimativas negativas para o parâmetro de forma. O valor esperado do modelo em estudo é dado pela expressão $\sigma/(1-\gamma)$, quando $\gamma < 1$, pelo que um aumento no valor de $\sigma$ traduz-se, mantendo-se o valor de $\gamma$, num aumento do valor esperado. Mais, quando $\gamma < 0$, este valor esperado é inferior a $\sigma$ e, quando $0 < \sigma < 1$, o valor esperado é superior a $\sigma$. Para cada uma das estações, o ajustamento da distribuição generalizada de Pareto foi analisado por meio da comparação da respectiva função densidade de probabilidade com o histograma das excedências relativas a cada possível limiar. Além destas duas representações gráficas, o ajustamento do modelo foi também examinado por meio de gráficos de probabilidade, de quantis e de nível de retorno. Um exemplo encontra-se na Figura 3, que apresenta os gráficos do estudo

da qualidade do ajuste do modelo obtido para os valores diários de precipitação da estação Bica da Cana (B).

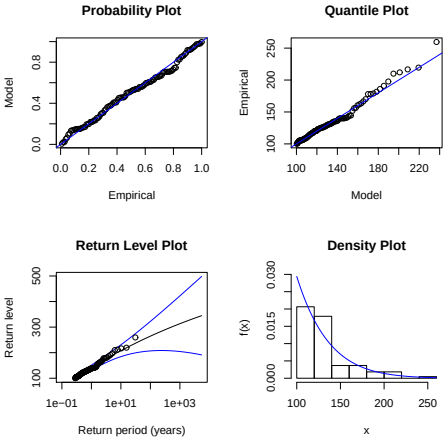



Figura 3: Avaliação gráfica da qualidade do ajuste do modelo obtido para os valores diários de precipitação da estação da Bica da Cana (B).

# 4 Considerações finais

Neste estudo exploratório, foram seleccionados limiares para séries de valores diários de precipitação na Ilha da Madeira, considerados como observações independentes e identicamente distribuídas. Para cada limiar, foi determinado o número de excedências e, por meio do método da máxima verosimilhança, obtidas estimativas para os parâmetros de escala e de forma da função de distribuição generalizada de Pareto. Os valores obtidos nesta análise indiciam que a caracterização dos extremos de precipitação por meio da abordagem POT, nesta ilha caracterizada por uma complexa orografia, deverá

ter em conta factores como a vertente e a altitude onde estão localizadas as estações e a sua proximidade ao mar.

# Agradecimentos

À Fundação para a Ciência e a Tecnologia, pelo apoio financeiro concedido através da bolsa de doutoramento SFRH/BD/39226/2007 e dos projectos PEst-OE/MAT/UI0006/2011 e PEst-OE/MAT/UI0006/2014, financiados por fundos nacionais do MCTES. Ao Departamento de Hidráulica e Tecnologias Energéticas do LREC e, em particular, ao Doutor Carlos Magro, pela disponibilização dos dados de precipitação utilizados. À Universidade da Madeira, por propiciar as condições adequadas à realização deste estudo. Ao Revisor, pelos seus comentários e sugestões, que contribuiram para melhorar o texto deste artigo.

# Referências

[1] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

[2] Fragoso, M., Trigo, R.M., Pinto, J.G., Lopes, S., Lopes, A., Ulbrich, S., Magro, C. (2012). The 20 February 2010 Madeira flash-floods: Synoptic analysis and extreme rainfall assessment. *Natural Hazards Earth System Sciences* 12, 715–730.

[3] Gomes, M.I., Fraga Alves, M., Neves, C. (2013). *Análise de Valores Extremos: Uma Introdução*. Edições SPE, Lisboa.

[4] Gorricha, J., Lobo, V., Costa, A.C. (2012). Spatial characterization of extreme precipitation in Madeira Island using geostatistical procedures and a 3D SOM. In Rückermann, C.-P., Resch, B. (eds.): *Proceedings of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 98–104, IARIA, Wilmington, DE.

[5] Quintal, R. (1999). Aluviões da Madeira; séculos XIX e XX. *Territorium* 6, 31–48.

# Propriedade de Taylor em processos autorregressivos

Esmeralda Gonçalves
CMUC, Departamento de Matemática, Universidade de Coimbra, Portugal, *esmerald@mat.uc.pt*

Cristina M. Martins
Departamento de Matemática, Universidade de Coimbra, Portugal, *cmtm@mat.uc.pt*

Nazaré Mendes-Lopes
CMUC, Departamento de Matemática, Universidade de Coimbra, Portugal, *nazare@mat.uc.pt*

**Palavras–chave**: Modelos AR, curtose, propriedade de Taylor

**Resumo**: Este artigo tem como principal objetivo analisar a presença da propriedade de Taylor em modelos lineares. Assim, desenvolve-se um estudo teórico sobre a sua ocorrência numa classe de modelos autorregressivos não negativos, obtendo-se uma condição necessária e suficiente para que tais modelos possuam a referida propriedade. Estes resultados são analisados em modelos com características de curtose significativamente diferentes, permitindo tirar conclusões sobre a relação entre a presença da propriedade de Taylor e a curtose do modelo.

## 1 Introdução

O efeito de Taylor é uma característica presente em séries temporais de natureza diversa. Este facto estilizado foi, pela primeira vez, detetado por Taylor (1986) ao analisar os retornos de algumas séries financeiras. Taylor constatou que as autocorrelações das observações em módulo eram positivas e sistematicamente superiores às correspondentes autocorrelações das observações ao quadrado, tam-

bém positivas, isto é, dispondo de $T$ observações, $X_1, X_2, \ldots, X_T$, de um processo $X = (X_t, t \in \mathbb{Z})$, observou que $\widehat{\rho}_{|X|}(h) > \widehat{\rho}_{X^2}(h)$, $h = 1,2,\ldots$, onde $\widehat{\rho}$ designa a função de autocorrelação empírica.

A propriedade teórica correspondente denomina-se propriedade de Taylor e, supondo que as funções $\rho_{|X|}$ e $\rho_{X^2}$ são positivas, é expressa pela condição

$$\rho_{|X|}(h) > \rho_{X^2}(h), \ h \in \mathbb{Z},$$

onde $\rho$ representa a função de autocorrelação. A presença desta propriedade numa determinada classe de modelos para séries temporais poderá contribuir para a seleção de um modelo mais adequado para descrever a dinâmica da série de interesse. No entanto, este estudo exige o conhecimento de momentos de $X$ de ordem superior a 2 o que o torna teoricamente elaborado.

A referida propriedade tem sido sobretudo analisada em modelos com características claras de leptocurtose, como os condicionalmente heteroscedásticos, conforme pode ser encontrado nos trabalhos de He e Teräsvirta (1999), Gonçalves, Leite e Mendes-Lopes (2009), Haas (2009) e Leite (2013), dedicados ao seu estudo na classe geral dos modelos GTARCH, tendo os resultados obtidos revelado uma relação forte entre a sua ocorrência e valores elevados da curtose do modelo. Esta leitura é sobretudo coerente com a leptocurtose assinalada em séries temporais, em particular de natureza financeira como as que Taylor analisou, em que o referido facto estilizado está, em geral, claramente presente.

O facto dos estudos desenvolvidos terem vindo a mostrar que a presença da propriedade de Taylor pode estar mais ligada ao peso das caudas do que ao carácter heteroscedástico dos modelos, motivou a investigação desta propriedade em modelos não condicionalmente heteroscedásticos. Tanto quanto é do nosso conhecimento, não há qualquer informação sobre a presença da propriedade de Taylor nos modelos lineares o que, tendo em conta a importância destes na modelação geral de dados temporais, nos parece necessário ultrapassar.

O estudo que desenvolvemos neste trabalho concentra-se em modelos lineares não negativos. A modelação de séries temporais não

negativas tem vindo a ser amplamente considerada na literatura nomeadamente pela necessidade de introduzir modelos adequados para descrever a volatilidade de séries temporais. Uma alternativa às representações condicionalmente heteroscedásticas consiste em definir a variância condicional diretamente como um processo ARMA quase certamente não negativo (Hong-zhi, 1992; Tsay e Chan, 2007). Em Tsay e Chan (2007) são desenvolvidas caracterizações para que um processo ARMA seja quase certamente não negativo e é ainda referido o interesse destes processos no contexto da modelação estocástica de séries financeiras; este último aspeto reforça o interesse de estudar a propriedade de Taylor neste tipo de modelos de séries temporais.

Com o objetivo de dar uma contribuição para este estudo, vamos neste trabalho avaliar a eventual ocorrência da propriedade de Taylor em processos autorregressivos não negativos de ordem 1. A relação entre a validade desta propriedade e a curtose do modelo será analisada. Serão desenvolvidas aplicações deste estudo, e consequente discussão, a modelos autorregressivos com processos de erro apresentando curtoses significativamente diferentes, sendo notória a importância deste parâmetro na presença no modelo da dita propriedade de Taylor.

## 2 Modelo AR(1) e propriedade de Taylor

Seja $X = (X_t, t \in \mathbb{Z})$ um processo estocástico real tal que

$$X_t = \phi X_{t-1} + \varepsilon_t \tag{1}$$

com $0 < \phi < 1$ e $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ um processo estocástico não negativo, de componentes identicamente distribuídas com momentos até à ordem 4 e tal que $E\left(\varepsilon_t^i | \underline{\varepsilon}_{t-1}\right) = m_i$, $i = 1, 2, 3$, $t \in \mathbb{Z}$, designando por $\underline{\varepsilon}_{t-1}$ a $\sigma$-álgebra gerada pelo passado de $\varepsilon_t$, i.e., $\sigma\left(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots\right)$. Assim, $E\left(\varepsilon_t^i\right) = E\left(E\left(\varepsilon_t^i | \underline{\varepsilon}_{t-1}\right)\right) = m_i$, $i = 1, 2, 3$. De modo natural, denotamos $E\left(\varepsilon_t^4\right)$ por $m_4$, $t \in \mathbb{Z}$. Observamos

que os resultados enunciados neste trabalho apenas requerem a existência do momento simples de ordem 4 de $\varepsilon_t$, não sendo necessária qualquer hipótese sobre o valor esperado condicionado desta ordem. O lema seguinte resume, nas condições enunciadas, as expressões das funções de autocorrelação de $X$ e $X^2$ para $h \in \mathbb{N}$, apresentando-se a prova desta última.

**Lema 2.1** *Tem-se*
*a)* $\rho_X(h) = \phi^h$, $h \in \mathbb{N}$.
*b)* $\rho_{X^2}(h) = \phi^{2h} + 2m_1 \dfrac{\phi^h - \phi^{2h}}{1-\phi} \dfrac{Cov\left(X_t, X_t^2\right)}{V\left(X_t^2\right)}$, $h \in \mathbb{N}$.

*Dem.*: De (1) obtemos $X_t = \phi^h X_{t-h} + \sum_{k=0}^{h-1} \phi^k \varepsilon_{t-k}$, $h \geq 1$, pelo que $X_t^2 = \phi^{2h} X_{t-h}^2 + \sum_{k=0}^{h-1} \phi^k \varepsilon_{t-k} \sum_{j=0}^{h-1} \phi^j \varepsilon_{t-j} + 2\phi^h X_{t-h} \sum_{k=0}^{h-1} \phi^k \varepsilon_{t-k}$. Então

$$Cov\left(X_t^2, X_{t-h}^2\right) = \phi^{2h} V(X_t^2) + \sum_{k=0}^{h-1} \phi^k \sum_{j=0}^{h-1} \phi^j Cov\left(\varepsilon_{t-k}\varepsilon_{t-j}, X_{t-h}^2\right) + 2\phi^h \sum_{k=0}^{h-1} \phi^k Cov\left(X_{t-h}\varepsilon_{t-k}, X_{t-h}^2\right). \quad (2)$$

Comecemos por calcular $Cov\left(\varepsilon_{t-k}\varepsilon_{t-j}, X_{t-h}^2\right)$ para $k \neq j$, $k,j \in \{0, \ldots, h-1\}$. Sem perda de generalidade, consideremos $j < k$. Atendendo às propriedades da esperança condicionada, temos

$$\begin{aligned} Cov\left(\varepsilon_{t-k}\varepsilon_{t-j}, X_{t-h}^2\right) &= E\left(\varepsilon_{t-k}\varepsilon_{t-j} X_{t-h}^2\right) - E\left(\varepsilon_{t-k}\varepsilon_{t-j}\right) E\left(X_t^2\right) \\ &= E\left(\varepsilon_{t-k} X_{t-h}^2 E\left(\varepsilon_{t-j} | \underline{\varepsilon}_{t-j-1}\right)\right) - E\left(\varepsilon_{t-k} E\left(\varepsilon_{t-j} | \underline{\varepsilon}_{t-j-1}\right)\right) E\left(X_t^2\right) \\ &= m_1 E\left(E\left(\varepsilon_{t-k} X_{t-h}^2 | \underline{\varepsilon}_{t-k-1}\right)\right) - m_1^2 E\left(X_t^2\right) \\ &= m_1 E\left(X_{t-h}^2 E\left(\varepsilon_{t-k} | \underline{\varepsilon}_{t-k-1}\right)\right) - m_1^2 E\left(X_t^2\right) = 0. \end{aligned}$$

Se $j = k$, o valor de $Cov\left(\varepsilon_{t-k}\varepsilon_{t-j}, X_{t-h}^2\right)$ é dado por

$$\begin{aligned} Cov\left(\varepsilon_{t-k}^2, X_{t-h}^2\right) &= E\left(X_{t-h}^2 E\left(\varepsilon_{t-k}^2 | \underline{\varepsilon}_{t-k-1}\right)\right) - E\left(\varepsilon_{t-k}^2\right) E\left(X_{t-h}^2\right) \\ &= m_2 E\left(X_{t-h}^2\right) - m_2 E\left(X_{t-h}^2\right) = 0. \end{aligned}$$

Quanto a $Cov\left(X_{t-h}\varepsilon_{t-k},X_{t-h}^2\right)$, temos

$$\begin{aligned} Cov\left(X_{t-h}\varepsilon_{t-k},X_{t-h}^2\right) &= E\left(X_{t-h}^3\varepsilon_{t-k}\right) - E\left(X_{t-h}\varepsilon_{t-k}\right)E\left(X_{t-h}^2\right) \\ &= E\left(X_{t-h}^3\right)m_1 - E\left(X_{t-h}\right)m_1 E\left(X_{t-h}^2\right) \\ &= m_1\left[E\left(X_t^3\right) - E\left(X_t\right)E\left(X_t^2\right)\right]. \end{aligned}$$

Retomando a igualdade (2), e como $\rho_{X^2}(h) = \frac{Cov\left(X_t^2,X_{t-h}^2\right)}{V\left(X_t^2\right)}$, vem

$$\rho_{X^2}(h) = \phi^{2h} + 2\phi^h\, m_1\, \frac{E\left(X_t^3\right)-E(X_t)E\left(X_t^2\right)}{V\left(X_t^2\right)}\, \sum_{k=0}^{h-1}\phi^k,$$

obtendo-se a expressão enunciada. ■

Tal como $\rho_X$, a função $\rho_{X^2}$ presente no lema anterior é positiva, atendendo em particular ao facto de $Cov\left(X_t, X_t^2\right) \geq 0$, uma vez que $X_t$ é não negativo. Decorre então, de modo imediato, o seguinte resultado:

**Teorema 2.2** *O processo $X$ definido em (1) verifica a propriedade de Taylor se e somente se*

$$\frac{Cov\left(X_t, X_t^2\right)}{V\left(X_t^2\right)} < \frac{1-\phi}{2m_1}. \tag{3}$$

Note-se que os momentos de $X$ presentes nesta condição podem ser obtidos, de modo recursivo, a partir de

$$E\left(X_t^n\right) = \sum_{i=0}^{n}\binom{n}{i}\phi^{n-i}m_i E\left(X_t^{n-i}\right),\ n = 1,2,3,4. \tag{4}$$

Observamos ainda que a condição (3) pode exprimir-se em termos dos momentos simples e centrados do processo de erro, $m_i$, $i = 1,2,3$, e $\mu_{k,\varepsilon}$, $k = 2,3,4$, respetivamente. De facto, tendo em conta os resultados apresentados no Apêndice A, obtém-se

$$Cov\left(X_t, X_t^2\right) = \frac{-2\phi(1+2\phi)m_1^3+\left(-1+2\phi+5\phi^2\right)m_1m_2+\left(1-\phi^2\right)m_3}{(1-\phi^2)(1-\phi^3)}$$

$$V\left(X_t^2\right) = \frac{\left(1-\phi^2\right)\mu_{4,\varepsilon}+6\phi^2(\mu_{2,\varepsilon})^2}{(1-\phi^2)(1-\phi^4)} + \frac{4\mu_{3,\varepsilon}}{1-\phi^3}\frac{m_1}{1-\phi} + \frac{4\mu_{2,\varepsilon}}{1-\phi^2}\frac{m_1^2}{(1-\phi)^2} - \frac{(\mu_{2,\varepsilon})^2}{(1-\phi^2)^2}.$$

No corolário seguinte apresenta-se uma condição suficiente para que os modelos AR aqui considerados verifiquem a propriedade de Taylor expressa em termos da respetiva curtose. Esta condição vem ao encontro dos estudos sobre a propriedade de Taylor anteriormente referidos, nos quais parece emergir uma relação forte entre a presença desta propriedade e valores elevados da curtose do processo $X$. Denotemos o momento centrado de ordem $k$ de $X_t$ por $\mu_{k,X}$ e seja $K_X$ o coeficiente de curtose de $X_t$, isto é, $K_X = \frac{\mu_{4,X}}{(\mu_{2,X})^2}$.

**Corolário 2.3** *Seja $X$ o processo definido em (1) com $\mu_{3,\varepsilon} = 0$. Se $K_X > 1 + 4\frac{(1+\phi)}{(1-\phi)}\frac{m_1^2}{V(\varepsilon_t)}$, então $X$ verifica a propriedade de Taylor.*

*Dem.*: Comecemos por considerar as igualdades seguintes, relacionando momentos simples e centrados de $X_t$.

$$\begin{aligned} E\left(X_t^3\right) &= E\left((X_t - E(X_t)) + E(X_t)\right)^3 \\ &= \mu_{3,X} + 3\mu_{2,X}E(X_t) + \left(E(X_t)\right)^3 \end{aligned}$$

e, de modo análogo,

$$E\left(X_t^4\right) = \mu_{4,X} + 4\mu_{3,X}E(X_t) + 6\mu_{2,X}\left(E(X_t)\right)^2 + \left(E(X_t)\right)^4.$$

Daqui deduzimos

$$\begin{aligned} &Cov\left(X_t, X_t^2\right) = \mu_{3,X} + 2V(X_t)E\left(X_t\right) \\ &V\left(X_t^2\right) = \mu_{4,X} + 4\mu_{3,X}E(X_t) + 4V(X_t)\left(E(X_t)\right)^2 - \left(V(X_t)\right)^2. \end{aligned}$$

Se $\mu_{3,X} = 0$, teremos então

$$\frac{Cov\left(X_t, X_t^2\right)}{V(X_t^2)} = \frac{2E(X_t)}{V(X_t)\left(K_X + 4\frac{(E(X_t))^2}{V(X_t)} - 1\right)} < \frac{2E(X_t)}{V(X_t)(K_X - 1)}.$$

Notemos que a condição $\mu_{3,X} = 0$ é equivalente a $\mu_{3,\varepsilon} = 0$ (pois $\mu_{3,X} = \frac{\mu_{3,\varepsilon}}{1-\phi^3}$), sendo em particular verificada se a lei de $\varepsilon_t$ é simétrica relativamente a $E(\varepsilon_t)$. Sob aquela condição, a propriedade de Taylor

verificar-se-á se $\frac{2E(X_t)}{V(X_t)\{K_X-1\}} < \frac{1-\phi}{2m_1}$, ou de modo equivalente, se

$$K_X > 1 + 4\frac{E(X_t)}{V(X_t)}\frac{m_1}{1-\phi}.$$

Mas, tendo em conta que $\frac{E(X_t)}{V(X_t)} = \frac{m_1(1+\phi)}{V(\varepsilon_t)}$, concluímos que a propriedade de Taylor está presente se $K_X > 1 + 4\,\frac{(1+\phi)}{1-\phi}\,\frac{m_1^2}{V(\varepsilon_t)}$. ■

# 3 Propriedade de Taylor e curtose de $\varepsilon$

Nesta secção avaliamos a presença da propriedade de Taylor no modelo (1) considerando processos de erro com distribuições não negativas e com pesos nas caudas significativamente diferentes.
Conforme está provado no Apêndice A, as curtoses dos processos $X$ e $\varepsilon$ estão ligadas pela relação

$$K_X(\phi) = \frac{6\phi^2}{1+\phi^2} + \frac{1-\phi^2}{1+\phi^2}K_\varepsilon, \tag{5}$$

deduzindo-se que o processo $X$ é leptocúrtico se o seu processo gerador, $\varepsilon$, o for também.
Da igualdade anterior é ainda fácil concluir que $K_X$ é uma função crescente de $\phi$ se $K_\varepsilon < 3$, sendo decrescente se $K_\varepsilon > 3$. Além disso, quando $\phi$ tende para 1, $K_X$ tende para 3, independentemente de $K_\varepsilon$.
Neste estudo, teremos em conta a condição necessária e suficiente de presença da propriedade de Taylor que se pode escrever na forma

$$T_L(\phi) = 2m_1\frac{Cov\left(X_t, X_t^2\right)}{V\left(X_t^2\right)} + \phi - 1 < 0, \tag{6}$$

onde $L$ designa a lei marginal do processo de erro, $\varepsilon$. No Apêndice B, apresentam-se as expressões de $T_L(\phi)$ para as leis $L$ consideradas neste trabalho.

## 3.1 Erros com distribuição platicúrtica

**i) Lei uniforme** - Seja $X$ o processo definido em (1) com $\varepsilon_t$ seguindo a lei uniforme no intervalo $]0,\alpha[$, $U(]0,\alpha[)$, $\alpha > 0$, para a qual se tem

$K_\varepsilon = 1.8$. Facilmente se verifica que a condição (6) é verdadeira para qualquer $\phi \in ]0,1[$, pelo que, neste caso, a propriedade de Taylor está sempre presente no modelo AR(1).

**ii) Lei triangular** - Suponhamos agora que $\varepsilon_t$ segue a lei triangular no intervalo $]0,\alpha[$, $Tr(]0,\alpha[)$, $\alpha > 0$. Tem-se $K_\varepsilon = 2.4$. Analogamente ao caso anterior, a condição (6) é verdadeira para qualquer $\phi \in ]0,1[$, pelo que também neste caso a propriedade de Taylor está sempre presente.
Nos dois casos considerados, verifica-se que as funções $K_X(\phi)$ e $T_L(\phi)$, definidas em (5) e (6), não dependem do parâmetro $\alpha$. Na Figura 1 apresentam-se os gráficos destas funções.

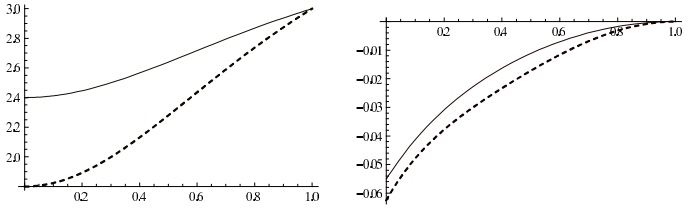



Figura 1: Gráficos de $K_X(\phi)$ (esq.) e de $T_L(\phi)$ (dir.) nos casos $U(]0,\alpha[)$ (tracejado) e $Tr(]0,\alpha[)$ (cheio)

A presença da propriedade de Taylor é muito fraca (valores de $T_L(\phi)$ próximos de zero), sendo um pouco mais significativa no caso de menor curtose.

**iii) Uma lei de densidade assimétrica negativa** - Consideremos agora o caso em que $\varepsilon_t$ segue a lei $L_\alpha$ de função densidade $f(x) = \frac{2x}{\alpha^2}\mathbb{I}_{]0,\alpha[}(x)$, $\alpha > 0$. A curtose desta lei é 2.4. Tem-se $T_{L_\alpha}(\phi) > 0$, pelo que a propriedade de Taylor não está presente nestes modelos AR. Na Figura 2 apresenta-se o gráfico de $T_{L_\alpha}$.
Embora a propriedade de Taylor não esteja presente, nota-se que os valores de $T_{L_\alpha}(\phi)$ são, em valor absoluto, da ordem de grandeza dos

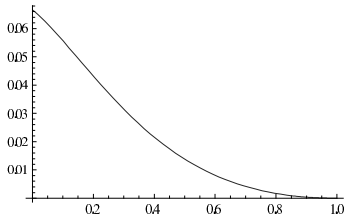



Figura 2: Gráfico de $T_{L_\alpha}(\phi)$

de $T_{U(]0,\alpha[)}(\phi)$ e de $T_{Tr(]0,\alpha[)}(\phi)$. Assim, podemos inferir que não há, nestes casos que correspondem a processos geradores platicúrticos, diferença significativa entre as funções $\rho_{|X|}$ e $\rho_{X^2}$.

## 3.2 Erros com distribuição leptocúrtica

**i) Lei gama** - Consideremos o processo $X$ definido em (1) com $\varepsilon_t$ seguindo a lei gama de parâmetros $\theta$ e $\alpha$, $\gamma\,(\theta,\alpha)$, $\alpha > 0$, $\theta > 0$, isto é, de densidade $f(x) = \frac{\alpha^\theta}{\Gamma(\theta)}\, e^{-\alpha x} x^{\theta-1} \mathbb{I}_{]0,+\infty[}\,(x)$, cuja curtose é dada por $K_\varepsilon = 3 + \frac{6}{\theta}$.
Verifica-se que $T_{\gamma(\theta,\alpha)}(\phi)$ depende também de $\theta$ (mas não de $\alpha$).
Os gráficos apresentados na Figura 3 permitem comparar as funções $T_{\gamma(\theta,\alpha)}\,(\phi)$, para $\theta = 0.1, 1, 5, 10, 100$ à luz da evolução em $\theta$ das funções $K_X\,(\phi) = \frac{6\phi^2}{1+\phi^2} + \frac{1-\phi^2}{1+\phi^2}\left(3 + \frac{6}{\theta}\right)$.
A propriedade de Taylor está sempre presente, sendo de notar que, quando $\theta \to \infty$, $T_{\gamma(\theta,\alpha)}\,(\phi) \to 0$, isto é, as funções de autocorrelação em estudo estão, à medida que $\theta$ aumenta, cada vez mais próximas. Este resultado é compatível com a evolução da curtose pois o processo é tanto mais leptocúrtico quanto menor for o parâmetro de forma, $\theta$.

**ii) Lei de Pareto** - Suponhamos agora que $\varepsilon_t$ segue a lei de Pareto de parâmetros $\alpha$ e $\theta$, $Par\,(\alpha,\theta)$, com densidade $f(x) = \frac{\theta\alpha^\theta}{x^{\theta+1}} \mathbb{I}_{]\alpha,+\infty[}\,(x)$,

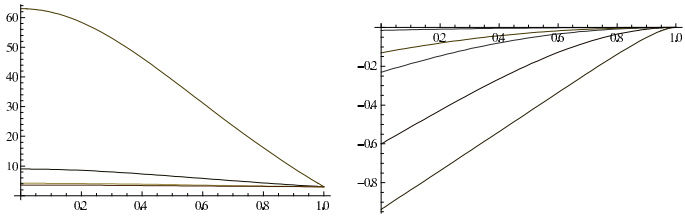



Figura 3: Gráficos, no caso $\gamma(\theta,\alpha)$, de $K_X(\phi)$ (esq.), $\theta = 0.1, 1, 5, 10$ (de cima para baixo) e de $T_{\gamma(\theta,\alpha)}(\phi)$ (dir.), $\theta = 0.1, 1, 5, 10, 100$ (de baixo para cima)

$\alpha > 0$, $\theta > 4$, e de curtose $K_\varepsilon = \frac{3(\theta-2)(3\theta^2+\theta+2)}{\theta(\theta^2-7\theta+12)}$.
Analogamente ao caso anterior, verifica-se que a função $T_{Par(\alpha,\theta)}(\phi)$ também depende do parâmetro de forma, $\theta$. Na Figura 4 encontram-se os gráficos das funções $T_{Par(\alpha,\theta)}(\phi)$, para $\theta = 4.1, 5, 9, 12, 100$, bem como os das funções $K_X(\phi)$, para $\theta = 4.1, 5, 10$.

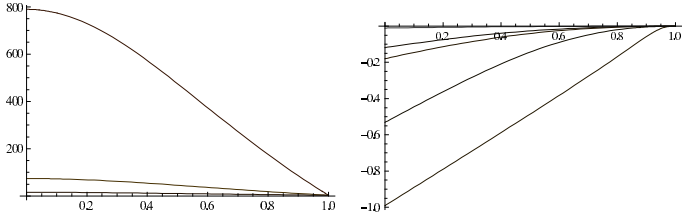



Figura 4: Gráficos, no caso $Par(\theta,\alpha)$, de $K_X(\phi)$ (esq.), $\theta = 4.1, 5, 10$ (de cima para baixo) e de $T_{Par(\alpha,\theta)}(\phi)$ (dir.), $\theta = 4.1, 5, 9, 12, 100$ (de baixo para cima)

As conclusões são análogas às referidas para a lei gama.

**iii) Comparação dos casos gama e Pareto** - Na Figura 5 estão representados os gráficos das funções $K_X(\phi)$, relativas aos dois modelos, e os das funções $T_{\gamma(\theta,\alpha)}(\phi)$ e $T_{Par(\alpha,\theta)}(\phi)$, no domínio $\{(\theta,\phi) \in ]4,10[ \times ]0,1[\}$. Esta figura evidencia o facto de a propriedade de Taylor se manifestar de modo mais forte no caso Pareto, i.e., de se acentuar a sua presença com o aumento da curtose do modelo.

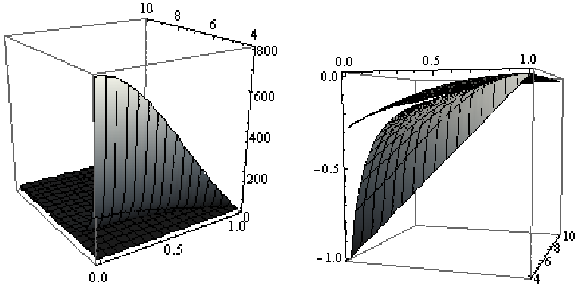



Figura 5: Gráficos, nos casos $\gamma(\theta,\alpha)$ e $Par(\theta,\alpha)$, de $K_X(\phi)$ (esq., Pareto "por cima") e de $T_L(\phi)$ (dir., Pareto "por baixo")

# 4 Conclusão

O estudo desenvolvido mostra que os modelos lineares podem também reproduzir o efeito de Taylor. É de salientar que nesta classe de modelos foi possível ir teoricamente bastante mais longe do que noutras classes já analisadas. De facto, estabelecemos uma condição necessária e suficiente que garante a presença, nestes modelos, da propriedade de Taylor em toda a sua generalidade, isto é, para todo o horizonte $h$ das funções $\rho_{|X|}$ e $\rho_{X^2}$. Neste contexto geral, continua a ser notória a forte ligação da propriedade de Taylor com a curtose do processo. Efetivamente, nos casos platicúrticos estudados, a propriedade de Taylor, quando se manifesta, é de modo muito suave,

acentuando-se ligeiramente à medida que a curtose se afasta do valor de referência, 3. Pode mesmo afirmar-se que, nestes casos, a verificação ou não da propriedade é muito pouco significativa, isto é, as funções $\rho_{|X|}$ e $\rho_{X^2}$ são quase indistinguíveis. Em contrapartida, nos processos leptocúrticos, a presença desta propriedade é constante e significativa, sendo tanto mais acentuada quanto maior é a curtose do processo. Concluímos, mais uma vez, que a propriedade de Taylor é fortemente dependente do maior ou menor peso das caudas da lei subjacente ao processo em análise.

## Apêndice A - Relação entre os momentos de $X$ e de $\varepsilon$

A partir de (4), obtemos $E(X_t) = \frac{m_1}{1-\phi}$ e $E\left(X_t^2\right) = \frac{2\phi m_1^2+(1-\phi)m_2}{(1-\phi)(1-\phi^2)}$, bem como expressões para $E(X_t^3)$ e $E(X_t^4)$.

Temos $\mu_{2,X} = E[(\phi X_{t-1} - \phi E(X_t) + \varepsilon_t - m_1)^2] = \phi^2\mu_{2,X} + \mu_{2,\varepsilon}$, pelo que $\mu_{2,X} = \frac{\mu_{2,\varepsilon}}{1-\phi^2}$. Analogamente, conclui-se que $\mu_{3,X} = \frac{\mu_{3,\varepsilon}}{1-\phi^3}$ e que $\mu_{4,X} = \frac{6\phi^2\mu_{2,\varepsilon}\mu_{2,X}+\mu_{4,\varepsilon}}{1-\phi^4}$.

Destas relações vem $K_X = \frac{\mu_{4,X}}{(\mu_{2,X})^2} = \frac{6\phi^2\mu_{2,\varepsilon}\frac{\mu_{2,\varepsilon}}{1-\phi^2}+\mu_{4,\varepsilon}}{(\mu_{2,\varepsilon})^2}\frac{1-\phi^2}{1+\phi^2}$, obtendo-se $K_X = \frac{6\phi^2}{1+\phi^2} + \frac{1-\phi^2}{1+\phi^2}K_\varepsilon$. Daqui, e como $\phi \in ]0,1[$, deduz-se a equivalência $K_X > 3 \Leftrightarrow K_\varepsilon > 3$, isto é, o processo $X$ é leptocúrtico se e só se o seu processo de erro, $\varepsilon$, o é.

## Apêndice B - Expressões de $T_L(\phi)$ para as leis $L$ estudadas

$$T_{U(]0,\alpha[)}(\phi) = -\frac{(1-\phi)^2(1+4\phi^2)}{16+14\phi+19\phi^2+11\phi^3}$$

$$T_{Tr(]0,\alpha[)}(\phi) = -\frac{(1-\phi)^2(7+13\phi^2)}{127+113\phi+133\phi^2+107\phi^3}$$

$$T_{L_\alpha}(\phi) = \frac{(1-\phi)^2(9+25\phi+12\phi^2+19\phi^3+3\phi^4)}{135+288\phi+493\phi^2+473\phi^3+352\phi^4+179\phi^5}$$

$$T_{\gamma(\theta,\alpha)}(\phi) = -\frac{(1-\phi)^2 N_{\gamma(\theta,\alpha)}(\phi)}{D_{\gamma(\theta,\alpha)}(\phi)}, \text{ com}$$

$$N_{\gamma(\theta,\alpha)}(\phi) = 3(\theta+1) + (5\theta+3)\phi + 6\theta\phi^2 + (5\theta-3)\phi^3 + 3(\theta-1)\phi^4$$

$$D_{\gamma(\theta,\alpha)}(\phi) = 2\theta^2 + 5\theta + 3 + 4\theta(\theta+1)\phi + (6\theta^2+\theta-3)\phi^2 + (6\theta^2-\theta-3)\phi^3 + 4\theta(\theta-1)\phi^4 + (2\theta^2-5\theta+3)\phi^5$$

$T_{Par(\theta,\alpha)}(\phi) = -\frac{(1-\phi)^2 N_{Par(\theta,\alpha)}(\phi)}{D_{Par(\theta,\alpha)}(\phi)}$, com

$$N_{Par(\theta,\alpha)}(\phi) = \theta^4 - 3\theta^3 + 5\theta - 3 + (2\theta^4 - 8\theta^3 + 2\theta^2 + 13\theta - 3)\phi + \theta(2\theta^3 - 9\theta^2 - 3\theta + 28)\phi^2 + (2\theta^4 - 11\theta^3 - \theta^2 + 31\theta + 3)\phi^3 + (\theta^4 - 6\theta^3 - 3\theta^2 + 23\theta + 3)\phi^4$$

$$D_{Par(\theta,\alpha)}(\phi) = (\theta-3)(\theta-1)^4 + 2\theta(\theta-1)^2(\theta^2-6\theta+8)\phi + (3\theta^5 - 27\theta^4 + 77\theta^3 - 77\theta^2 + 15\theta + 3)\phi^2 + (3\theta^5 - 27\theta^4 + 76\theta^3 - 70\theta^2 + 3\theta + 3)\phi^3 + 2\theta(\theta^4 - 10\theta^3 + 31\theta^2 - 26\theta - 8)\phi^4 + (\theta^5 - 11\theta^4 + 37\theta^3 - 23\theta^2 - 31\theta - 3)\phi^5.$$

# Referências


[1] Gonçalves, E., Leite, J., Mendes-Lopes, N. (2009). A mathematical approach to detect the Taylor property in TARCH processes. *Statistics& Probability Letters* 79, 602–610.

[2] Leite, J. (2013). *Processos Threshold GARCH com potência: estrutura probabilista e aplicações a cartas de controlo*. Tese de Doutoramento, Universidade de Coimbra.

[3] Gonçalves, E., Martins, C.M., Mendes-Lopes, N. (2012). Propriedade de Taylor em modelos bilineares não negativos. *Livro de resumos do XX Congresso Anual da SPE*, 605–608.

[4] Haas, M. (2009). Persistence in volatility, conditional kurtosis, and the Taylor property in absolute value GARCH processes. *Statistics& Probability Letters* 79, 1674–1683.

[5] He, C., Teräsvirta, M. (1999). Properties of moments of a family of GARCH processes. *Journal of Econometrics* 92, 173–192.

[6] Hong-zhi, A. (1992). Non-negative autorregressive models. *Journal of Time Series Analysis* 13, 283–295.

[7] Taylor, S. (1986). *Modelling Financial Time Series*. Wiley.

[8] Tsay, H., Chan, K.S. (2007). A Note on Non-Negative Arma Processes. *Journal of Time Series Analysis* 28, 350–360.

# Propriedade de Taylor no modelo TGARCH(1,1)

Esmeralda Gonçalves
CMUC, Departamento de Matemática, Universidade de Coimbra, *esmerald@mat.uc.pt*

Joana Leite
CMUC, Instituto Politécnico de Coimbra - ISCAC, *jleite@iscac.pt*

Nazaré Mendes-Lopes
CMUC, Departamento de Matemática, Universidade de Coimbra, *nazare@mat.uc.pt*

**Palavras–chave**: Propriedade de Taylor, Processos TGARCH, Autocorrelações

**Resumo**: Neste trabalho é estudada a propriedade de Taylor, que relaciona as autocorrelações do quadrado e do módulo de uma série temporal, no modelo TGARCH(1,1). Depois de apresentadas as expressões das autocorrelações, estabelece-se a existência de um conjunto de parametrizações do modelo verificando tal propriedade. A região de verificação da propriedade é claramente explicitada considerando o modelo TARCH(1) e adotando algumas distribuições particulares para o processo gerador; o estudo mostra também que valores elevados do coeficiente de curtose deste processo contribuem para a presença desta propriedade.

## 1 Factos estilizados e o efeito de Taylor

"*...[In] literature on prices, returns, and volatility* (...) *a number of "stylized facts" have been accumulated, these being empirical "facts" that have been observed to occur for many (possibly all) assets in most (possibly all) markets, most time periods and most data frequencies.*" Granger (2005).

"*Wide-ranging applications to financial data have discovered important stylized facts and illustrated both strengths and weaknesses of the models.*" Engle (2002).

A presença sistemática de algumas características marcantes em séries temporais de determinada natureza tem sido destacada por diversos autores, como ilustram as citações anteriores. Por exemplo, nas séries de retornos financeiros associadas aos preços de certos bens, isto é, em séries $X = (X_t, t \in \mathbb{Z})$ tais que $X_t = \left(\log \frac{p_t}{p_{t-1}}\right) \times 100$, com $p_t$ o preço do bem no instante $t$, observam-se usualmente os seguintes factos estilizados:
- aglomerados de volatilidade (*volatility clustering*),
- distribuição marginal com caudas pesadas,
- reacção assimétrica ao sinal dos valores passados (*leverage effects*),
- ausência de autocorrelação na série dos retornos, $X$,
- correlação positiva nas séries $|X|$ e $X^2$.

Para além dos factos estilizados já referidos, a análise mais pormenorizada das séries financeiras tem revelado aspectos curiosos na estrutura das autocorrelações. Concretamente, em 1986, Taylor observa, em 40 séries de retornos, um facto sistematicamente presente. Considerando a autocorrelação de ordem $n$ do processo $|X|^k$, $\rho_n(k) = corr\left(|X_t|^k, |X_{t-n}|^k\right)$, com $k > 0$, ele verifica que para $n = 1, \ldots, 30$, tem sempre a relação seguinte entre as autocorrelações empíricas de $|X|$ e de $|X|^2$: $\widehat{\rho}_n(1) > \widehat{\rho}_n(2)$.
Ding, Granger e Engle (1993), Granger e Ding (1995), Granger, Spear e Ding (2000) e Taylor (2007) alargam aquele estudo confirmando a presença de tal característica, que Granger e Ding (1995) denominam efeito de Taylor e que se traduz pela verificação de $\widehat{\rho}_n(1) > \widehat{\rho}_n(k)$ para todo $n$ e todo $k \neq 1$. Granger (2005) junta este efeito à lista de factos estilizados que caracterizam a dinâmica das séries de retornos.
O efeito de Taylor parece estar fortemente relacionado com dois dos factos estilizados anteriormente referidos, nomeadamente os aglomerados de volatilidade e as caudas pesadas. As figuras seguintes,

relativas aos retornos das cotações de fecho das acções do Banco Espírito Santo (BES) no período de 02/01/2004 a 31/12/2013, esclarecem esta ligação.

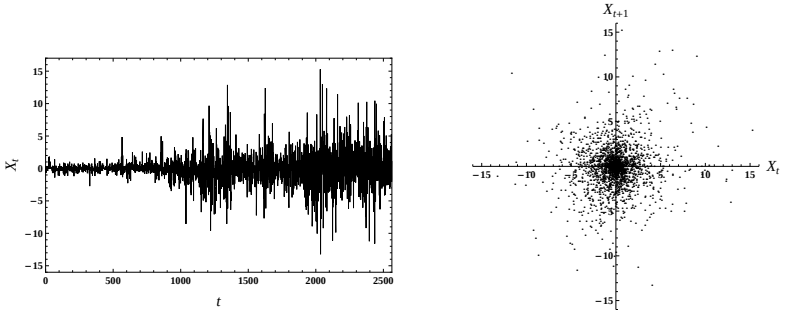



Figura 1: Retornos das cotações de fecho das acções do BES

Na Figura 1 está presente a trajectória da série observada e a representação da série no plano $(X_t, X_{t+1})$.

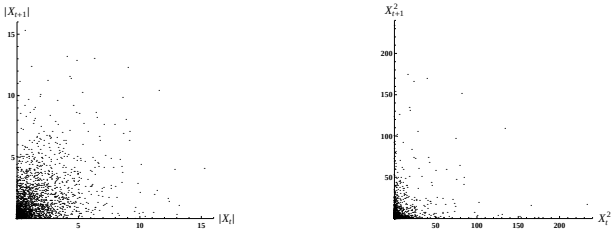



Figura 2: Transformação dos retornos das cotações de fecho das acções do BES

Na Figura 2 é possível observar os diagramas de dispersão de transformações dos retornos em instantes consecutivos, concretamente, do lado esquerdo, dos retornos em valor absoluto e, do lado direito, do quadrado dos retornos. Da confrontação destes diagramas percebe-

-se o impacto do aglomerados de volatilidade, pois, como um valor elevado de $|X_t|$ tende a ser seguido por um valor também elevado de $|X_{t+1}|$, os pontos ($|X_t|$,$|X_{t+1}|$) alinham-se mais pela bissectriz dos eixos representados do que os pontos ($X_t^2$,$X_{t+1}^2$), que se situam mais perto dos eixos. A explicação do comportamento do quadrado dos retornos parece estar nas caudas pesadas da distribuição dos retornos, que faz com que haja o aparecimento mais frequente de grandes valores de $|X_t|$, e também na transformação quadrado, que provoca, então, um afastamento entre $X_t^2$ e $X_{t+1}^2$, com consequente perda da ligação linear.

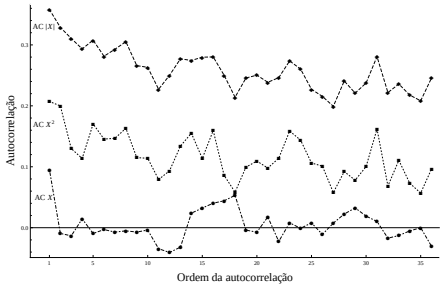



Figura 3: Autocorrelações dos retornos, do quadrado dos retornos e do módulo dos retornos das cotações de fecho das acções do BES

A Figura 3, que apresenta as autocorrelações dos retornos, do quadrado dos retornos e dos retornos em valor absoluto, para as ordens de 1 a 36, vem, então, confirmar que há bem mais dependência linear entre os retornos transformados do que entre os próprios retornos, com especial incidência na transformação módulo.

Assim, a procura de um modelo para descrever a dinâmica destas séries de retornos financeiros deverá ser feita numa classe de processos capazes de reproduzir o efeito de Taylor. A propriedade de Taylor é a relação teórica que traduz o referido efeito de Taylor, detectado em múltiplas séries temporais, não só de natureza financeira mas também física. Dizemos, então, que a propriedade de Taylor está

presente numa série temporal $X = (X_t, t \in \mathbb{Z})$ quando, para qualquer $n \in \mathbb{N}$, a autocorrelação de ordem $n$ da série em valor absoluto, $\rho_n(1) = corr(|X_t|, |X_{t-n}|)$, é maior do que a autocorrelação de ordem $n$ do quadrado da série, $\rho_n(2) = corr\left(X_t^2, X_{t-n}^2\right)$.

A análise da propriedade de Taylor tem vindo a ser desenvolvida na classe dos modelos autorregressivos condicionalmente heteroscedásticos (GARCH e TARCH) por He e Teräsvirta (1999), Gonçalves, Leite e Mendes-Lopes (2009), Haas (2009), Malmstem e Teräsvirta (2010) e também em modelos autorregressivos de volatilidade estocástica (ARSV) por Mora-Galán, Pérez e Ruiz (2004), Veiga (2009), Pérez, Ruiz e Veiga (2010), Malmstem e Teräsvirta (2010).
Na secção seguinte apresentamos o modelo TGARCH(1,1) com potência $\delta$ e as expressões das autocorrelações da série em valor absoluto e do quadrado da série. Estabelecemos depois, na secção 3, a existência de um conjunto de parametrizações do modelo que verificam a propriedade de Taylor, sendo a região de verificação claramente explicitada no caso de modelos TARCH(1) com distribuições particulares para o processo gerador. O estudo feito mostra uma ligação forte entre valores elevados do coeficiente de curtose do processo gerador e a presença desta propriedade. Terminamos, na secção 4, com algumas notas finais, apontando caminho para o desenvolvimento do estudo da propriedade de Taylor.

# 2 Modelo TGARCH(1,1) com potência

O processo estocástico $X = (X_t,\, t \in \mathbb{Z})$ segue o modelo TGARCH(1,1) com potência $\delta$ se

$$\begin{cases} X_t = Z_t \sigma_t \\ \sigma_t^{\delta} = \alpha_0 + \alpha_1 \left(X_{t-1}^{+}\right)^{\delta} + \beta_1 \left(X_{t-1}^{-}\right)^{\delta} + \gamma_1 \sigma_{t-1}^{\delta} \end{cases}$$

onde $X_t^{+} = \max\{X_t, 0\}$, $X_t^{-} = \max\{-X_t, 0\}$, $\delta > 0$, $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, $\gamma_1 > 0$ e $Z = (Z_t, t \in \mathbb{Z})$ é uma família de variáveis aleatórias

reais independentes e identicamente distribuídas, com $Z_t$ independente de $\underline{X}_{t-1}$.

Esta especificação de $\sigma_t$ permite considerar modelos assimétricos e é também flexível, ao não fixar *a priori* o valor da potência $\delta$, pois $\delta$ pode ser um qualquer número real positivo. Notemos ainda que tal especificação inclui, por exemplo, modelos TGARCH ($\delta = 1$), GARCH ($\delta = 2$ e $\alpha_1 = \beta_1$) e power-GARCH ($\alpha_1 = \beta_1$). Em Gonçalves, Leite e Mendes-Lopes (2012) é feito o estudo da estacionaridade forte e à ordem $\delta$ para o modelo TGARCH($p$,$q$) com potência $\delta \neq 0$ e ordens $p$ e $q$ quaisquer.

Relativamente à propriedade de Taylor, diversos estudos de simulação sugerem que o modelo GARCH não é tão capaz de a fazer surgir como o modelo TGARCH. Além disso, as expressões de $\rho_n(1)$ e $\rho_n(2)$ estão apenas determinadas para o modelo TGARCH. Notemos ainda que parece natural começar com este modelo já que, quando $\alpha_1 = \beta_1$ e $\gamma_1 = 0$, se tem $|X_t| = |Z_t|(\alpha_0 + \alpha_1 |X_{t-1}|)$ , ou seja, a menos de uma perturbação homotética pelo módulo de um processo de tipo ruído branco, $|Z_t|$ , há uma relação linear flagrante entre $|X_t|$ e $|X_{t-1}|$.

Apresentamos, em seguida, as expressões das autocorrelações para o modelo TGARCH(1,1), deduzidas por He e Teräsvirta (1999) e Leite (2013). Consideremos $\vartheta_i = E\left[\left(\alpha_1 Z_0^+ + \beta_1 Z_0^- + \gamma_1\right)^i\right]$, para $i = 1,2,3,4$.

Se o modelo TGARCH(1,1), com processo gerador de distribuição marginal simétrica, centrada e reduzida, é tal que $\vartheta_2 < 1$, então fica assegurada a existência de $\rho_n(1)$, para qualquer $n$, tendo-se

$$\rho_1(1) = \frac{\alpha_1 + \beta_1}{2} E(|Z_0|) + \gamma_1 + \frac{\gamma_1\left(1 - \vartheta_1^2\right)\left(\left[E(|Z_0|)\right]^2 - 1\right)}{\left(1 - \vartheta_1^2\right) - \left[E(|Z_0|)\right]^2 (1 - \vartheta_2)}$$

e, para $n > 1$, $\rho_n(1) = \vartheta_1^{n-1}\rho_1(1)$.

De modo análogo, se o modelo TGARCH(1,1), com processo gerador de distribuição marginal simétrica, centrada e reduzida tal que

$E\left(Z_0^4\right) < +\infty$, é tal que $\vartheta_4 < 1$, então fica ainda assegurada a existência de $\rho_n(2)$, para qualquer $n$, tendo-se

$$\rho_1(2) = \frac{\alpha_1^2 + \beta_1^2}{2} + (\alpha_1 + \beta_1)\gamma_1 \frac{E\left(|Z_0|^3\right)}{E\left(Z_0^4\right)} + \frac{\gamma_1^2}{E\left(Z_0^4\right)} + (1 - \vartheta_4)\frac{\Upsilon}{\Delta}$$

e, para $n > 1$,

$$\rho_n(2) = \vartheta_2^{n-1}\rho_1(2) + (1 - \vartheta_4)\frac{\Phi^*}{\Delta}\sum_{i=1}^{n-1}\vartheta_2^{n-i-1}\vartheta_1^i,$$

onde

$$\begin{aligned}
\Upsilon &= \left[(\alpha_1 + \beta_1)E\left(|Z_0|^3\right) + 2\gamma_1\right](1 + 2\vartheta_1 + 2\vartheta_2 + \vartheta_1\vartheta_2)(1 - \vartheta_1)(1 - \vartheta_2) + \\
&\quad + \left(\frac{\alpha_1^2 + \beta_1^2}{2} + (\alpha_1 + \beta_1)\gamma_1\frac{E\left(|Z_0|^3\right)}{E\left(Z_0^4\right)} + \frac{\gamma_1^2}{E\left(Z_0^4\right)} - 1\right)(1 + \vartheta_1)^2(1 - \vartheta_3) + \\
&\quad + \left(1 - \vartheta_1^2\right)(1 - \vartheta_2)(1 - \vartheta_3), \\
\Phi^* &= -2\vartheta_1(1 + \vartheta_1)(1 - \vartheta_3) + \\
&\quad + \left[(\alpha_1 + \beta_1)E\left(|Z_0|^3\right) + 2\gamma_1\right](1 + 2\vartheta_1 + 2\vartheta_2 + \vartheta_1\vartheta_2)(1 - \vartheta_1), \\
\Delta &= E\left(Z_0^4\right)\Psi(1 - \vartheta_1)(1 - \vartheta_2) - (1 + \vartheta_1)^2(1 - \vartheta_3)(1 - \vartheta_4), \\
\Psi &= 1 + 3\vartheta_1 + 5\vartheta_2 + 3\vartheta_3 + 3\vartheta_1\vartheta_2 + 5\vartheta_1\vartheta_3 + 3\vartheta_2\vartheta_3 + \vartheta_1\vartheta_2\vartheta_3.
\end{aligned}$$

# 3 Propriedade de Taylor no modelo TGARCH(1,1)

O resultado seguinte estabelece a presença da propriedade de Taylor na classe dos modelos TGARCH(1,1).

**Teorema 3.1** *O modelo TGARCH*(1,1)*, com processo gerador de distribuição marginal simétrica, centrada e reduzida tal que* $\left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}} < E(|Z_0|) < 1$ *e* $\vartheta_4 < 1$*, admite um conjunto de parametrizações para as quais a propriedade de Taylor é verificada.*

A demonstração (Leite, 2013) passa fundamentalmente por serem estabelecidos, para todo $n$, os seguintes comportamentos assintóticos:

$$\rho_n(1) = corr(|X_t|, |X_{t-n}|) \text{ tende para } \left(E(|Z_0|)\left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}}\right)^n$$

e

$$\rho_n(2) = corr\left(X_t^2, X_{t-n}^2\right) \text{ tende para } \left(\left[E\left(Z_0^4\right)\right]^{-\frac{1}{2}}\right)^n,$$

quando $(\alpha_1,\beta_1,\gamma_1)$ tende para $\left(\left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}}, \left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}}, 0\right)$.

Embora a região de parametrizações indicada na prova do teorema anterior esteja restrita a uma vizinhança da parametrização particular do modelo, é possível alargar esta região. De facto, na demonstração do teorema consideramos apenas que $(\alpha_1,\beta_1,\gamma_1)$ tende para $(\mathfrak{a},\mathfrak{a},0)$, onde $\mathfrak{a} = \left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}}$, tendo-se $\vartheta_4$ a tender para 1; no entanto, quando se faz $(\alpha_1,\beta_1,\gamma_1)$ tender para $(\mathfrak{b},m\mathfrak{b},0)$, onde $\mathfrak{b} = \left[E\left(Z_0^4\right)\right]^{-\frac{1}{4}}\left(\frac{2}{m^4+1}\right)^{\frac{1}{4}}$, com $m > 0$, também decorre que $\vartheta_4$ tende para 1. Assim, se atendermos às expressões das autocorrelações e admitirmos verificadas as hipóteses do teorema, podemos concluir que existem parametrizações do modelo pertencentes a vizinhanças cujo centro pode estar situado sobre qualquer ponto $(\mathfrak{b},m\mathfrak{b},0)$ tal que $0.2814 < m < 3.5546$, para as quais a propriedade de Taylor está presente. Na Figura 4 apresenta-se a representação da região de existência das autocorrelações ($\vartheta_4 < 1$), sendo nela assinalada a faixa de parametrizações, encostada à fronteira de existência das autocorrelações, onde está, então, garantida a verificação da propriedade de Taylor.

A extensão da conclusão para $\gamma_1$ não tão próximo de 0 usa o corolário 3 de Haas (2009).

Com o objectivo de ilustrar o teorema anterior, consideramos o modelo TARCH(1) ($\gamma_1 = 0$) com diferentes distribuições para o processo gerador, que apresentam coeficientes de curtose, $\kappa$, iguais a 2.4, 3 e 6; exibimos, na Figura 5, as regiões de verificação da relação $\rho_1(1) > \rho_1(2)$ .

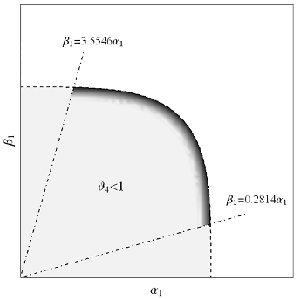



Figura 4: Região de existência das autocorrelações (cinzento claro) e região de verificação da propriedade de Taylor (cinzento escuro).

As distribuições consideradas foram: (i) distribuição *triangular* de densidade $f(x) = \frac{\sqrt{6}-|x|}{6}\mathbb{1}_{]-\sqrt{6},\sqrt{6}[}(x)$, $x \in \mathbb{R}$, com coeficiente de curtose $\kappa = \frac{12}{5}$, portanto, uma distribuição platicúrtica; (ii) distribuição *Gaussiana* padrão, de densidade $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$, para a qual $\kappa = 3$, mesocúrtica; (iii) distribuição de *Laplace* de densidade $f(x) = \frac{\sqrt{2}}{2}e^{-\sqrt{2}|x|}$, $x \in \mathbb{R}$, tendo $\kappa = 6$, logo, leptocúrtica; (iv) distribuição baseada na distribuição de *Student* com 6 graus de liberdade de densidade $f(x) = \frac{1}{\sqrt{4\pi}}\frac{\Gamma\left(\frac{7}{2}\right)}{\Gamma(3)}\left(1+\frac{x^2}{4}\right)^{-7/2}$, $x \in \mathbb{R}$, para a qual $\kappa = 6$, portanto, também leptocúrtica e com o mesmo valor do coeficiente de curtose da distribuição de Laplace indicada.

Estes exemplos sugerem claramente que a subclasse dos modelos TARCH(1) capaz de exibir a propriedade de Taylor aumenta significativamente com a curtose do processo gerador.

# 4 Notas finais

Neste trabalho estabelecemos a presença da propriedade de Taylor no modelo TGARCH(1,1). A obtenção de condições expressas em função dos parâmetros do modelo e dos momentos do processo gera-

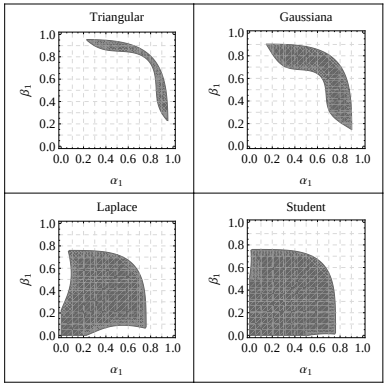



Figura 5: Modelo TARCH(1): lei do processo gerador e região de verificação da relação $\rho_1(1) > \rho_1(2)$.

dor que permitam assegurar a presença da propriedade de Taylor é um dos desafios imediatos assim como explorar a propriedade no modelo TGARCH de ordens gerais $(p,q)$. O estudo apresentado mostra que valores elevados do coeficiente de curtose das variáveis do processo gerador parecem favorecer o aparecimento da propriedade de Taylor. A avaliação da importância do coeficiente de curtose, nomeadamente a análise da sua ordem de grandeza, para o aparecimento da propriedade de Taylor é uma questão em aberto que será também objecto de estudo futuro.

# Referências

[1] Ding, Z., Granger, C.W.J., Engle, R.F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1, 83–106.

[2] Engle, R.F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics* 17, 425–446.

[3] Gonçalves, E., Leite, J., Mendes-Lopes, N. (2009). A mathematical approach to detect the Taylor property in TGARCH processes. *Statistics & Probability Letters* 79, 602–610.

[4] Gonçalves, E., Leite, J., Mendes-Lopes, N. (2012). On the probabilistic structure of power threshold generalized ARCH stochastic processes. *Statistics & Probability Letters* 82, 1597–1609.

[5] Granger, C.W.J. (2005). Past and future of empirical finance: some personal comments. *Journal of Econometrics* 129, 35–40.

[6] Granger, C.W.J., Ding, Z. (1995). Some properties of absolute return: an alternative measure of risk. *Annales d'Économie et de Statistique* 40, 67–95.

[7] Granger, C.W.J., Spear, S., Ding, Z. (2000). *Stylized facts on the temporal distributional properties of absolute returns: an update.* Proceedings of the IWSF, 97–120.

[8] Haas, M. (2009). Persistence in volatility, conditional kurtosis, and the Taylor property in absolute value GARCH processes. *Statistics & Probability Letters* 79, 1674–1683.

[9] He, C., Teräsvirta, T. (1999). Properties of moments of a family of GARCH processes. *Journal of Econometrics* 92, 173–192.

[10] Leite, J. (2013). *Processos Threshold GARCH com potência: estrutura probabilista e aplicação a cartas de controlo.* Tese de Doutoramento, Universidade de Coimbra.

[11] Malmsten, H., Teräsvirta, T. (2010). Stylized facts of financial time series and three popular models of volatility. *Eur. J. Pure App. Math.*, 3, 443-477.

[12] Mora-Galán, A., Pérez, A., Ruiz, E. (2004). Stochastic volatility models and the Taylor effect. *Stat. Econ. Working Papers* ws046315, Univ. Carlos III.

[13] Pérez, A., Ruiz, E., Veiga, H. (2010). A note on the properties of power-transformed returns in long-memory stochastic volatility models with leverage effect. *Computational Statistics & Data Analysis* 53, 3593–3600.

[14] Taylor, S.J. (1986). *Modelling Financial Time Series*, Wiley.

[15] Taylor, S.J. (2007). *Asset Price Dynamics, Volatility, and Prediction.* Princeton.

[16] Veiga, H. (2009). Financial Stylized Facts and the Taylor-Effect in Stochastic Volatility Models. *Economics Bulletin* 29, 265–276.

# Modelos GARCH de valores inteiros associados a leis infinitamente divisíveis

Esmeralda Gonçalves
CMUC, Departamento de Matemática, Universidade de Coimbra, *esmerald@mat.uc.pt*

Nazaré Mendes-Lopes
CMUC, Departamento de Matemática, Universidade de Coimbra, *nazare@mat.uc.pt*

Filipa Silva
CMUC, Departamento de Matemática, Universidade de Coimbra, *mat0504@mat.uc.pt*

**Palavras–chave**: Séries temporais de valores inteiros, leis de probabilidade infinitamente divisíveis discretas, modelos GARCH, distribuições Poisson Compostas

**Resumo**: Este trabalho tem como objectivo propor um modelo de valores inteiros com distribuição condicional marginal pertencente à classe geral das leis discretas infinitamente divisíveis. Para isso, introduz-se uma vasta classe de séries de contagem que inclui, em particular, os modelos Poisson INGARCH , binomial negativo INGARCH e Poisson generalizado INGARCH ([5], [9] e [10], respectivamente), introduzidos recentemente na literatura. Estabelece-se a existência de solução fortemente estacionária e er- gódica numa subclasse que inclui os modelos Poisson INGARCH e Poisson generalizado INGARCH.

## 1 Introdução

A necessidade de oferecer respostas adequadas à modelação estocástica de séries temporais de valores inteiros não negativos, tem sido realçada por vários autores em particular porque tal tipo de séries

surge com frequência e de forma natural em diversos contextos e áreas científicas tais como a medicina, a economia, o turismo ou a meteorologia. Neste sentido, têm surgido na literatura vários modelos para séries temporais de valores inteiros não negativos, entre os quais se destacam os modelos INGARCH (INteger-valued GARCH), propostos em 2006 por Ferland, Latour e Oraichi [5], claramente inspirados nos GARCH clássicos de Bollerslev [2] mas com distribuição condicional de Poisson.

Com o objetivo de alargar estes estudos, particularmente no que diz respeito à família de distribuições condicionais, vamos no presente trabalho introduzir e analisar uma nova classe de modelos de valores inteiros com evolução análoga para a média condicional mas em que a família de leis condicionais associada é a das leis infinitamente divisíveis com suporte em $\mathbb{N}_0$. A equivalência, no conjunto das leis discretas com suporte $\mathbb{N}_0$ , entre leis infinitamente divisíveis e leis de Poisson compostas, [7], leva-nos a designá-lo por modelo GARCH de valor inteiro Poisson composto. Esta família é introduzida explicitando a função característica da distribuição condicional de $X_t$ que, pelo facto de ser de Poisson composta, admite a forma geral $\Phi(u) = e^{\lambda[\varphi(u)-1]}$, $u \in \mathbb{R}$, onde $\varphi$ é uma função característica e $\lambda$ um valor real positivo ([7]). Para além do modelo de Poisson ([5]) e dos modelos GP-INGARCH ([10]), em que a lei condicional é de Poisson generalizada, esta nova classe inclui outras famílias de modelos que destacaremos neste trabalho. Será estabelecida, para a classe geral introduzida, uma condição necessária e suficiente de estaciona- riedade em média que provar-se-á ser também necessária e suficiente para garantir a estacionariedade forte e fraca, bem como a ergo- dicidade, numa subclasse geral que inclui, entre outros, os modelos estudados em [5] e [10].

## 2 Definição do modelo

Seja $X = (X_t, t \in \mathbb{Z})$ um processo estocástico com valores em $\mathbb{N}_0$ e designemos por $\underline{X}_{t-1}$ a $\sigma-$álgebra gerada por $\{X_{t-i}, i \geq 1\}$.

**Definição 2.1** *Diz-se que $X$ verifica um modelo GARCH de valor inteiro Poisson Composto de ordens $p$ e $q \in \mathbb{N}$, abreviadamente, um CP-INGARCH($p$,$q$) (Compound Poisson INteger-valued GARCH), se $\forall t \in \mathbb{Z}$, a função característica de $X_t|\underline{X}_{t-1}$, $\Phi_{X_t|\underline{X}_{t-1}}$, é dada por*

$$\begin{cases} \Phi_{X_t|\underline{X}_{t-1}}(u) = e^{i\frac{\lambda_t}{\varphi_t'(0)}[\varphi_t(u)-1]}, \quad u \in \mathbb{R} \\ E(X_t|\underline{X}_{t-1}) = \lambda_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{k=1}^{q} \beta_k \lambda_{t-k} \end{cases} \tag{1}$$

*com $\alpha_0 > 0$, $\alpha_j \geq 0$ ($j = 1, \ldots, p$), $\beta_k \geq 0$ ($k = 1, \ldots, q$), onde $(\varphi_t, t \in \mathbb{Z})$ é uma família de funções características sobre $\mathbb{R}$, $\underline{X}_{t-1}$-mensuráveis, associada a uma família de leis discretas de suporte contido em $\mathbb{N}_0$ e média finita e $i$ designa a unidade imaginária.*

Quando $\beta_1 = \cdots = \beta_q = 0$, o modelo designa-se $CP\text{-}INARCH(p)$. Notemos que, como $\varphi_t$ é a função característica de uma lei discreta de suporte $\mathbb{N}_0$ e média finita, a derivada de $\varphi_t$ em $u = 0$, $\varphi_t'(0)$, existe e não se anula.

**Observação 2.2 (Geração de um modelo $CP\text{-}INGARCH(p,q)$)** *Como a distribuição condicional de $X_t$ é uma lei de Poisson composta discreta de suporte $\mathbb{N}_0$, então, $\forall t \in \mathbb{Z}$ e condicionalmente a $\underline{X}_{t-1}$, é válida a seguinte igualdade em distribuição*

$$X_t \stackrel{d}{=} \sum_{j=1}^{N_t} X_{t,j}, \tag{2}$$

*onde $N_t$ segue uma lei de Poisson de parâmetro $\lambda_t^* = i\ \lambda_t/\varphi_t'(0)$ e $X_{t,1}, \ldots, X_{t,N_t}$ são variáveis aleatórias (v.a.) discretas com média finita, suporte contido em $\mathbb{N}_0$, independentes entre si e independentes de $N_t$, com função característica $\varphi_t$. Notemos que estas funções características sendo $\underline{X}_{t-1}$-mensuráveis podem ser funções aleatórias.*

**Exemplo 2.3** ***1.*** *Ferland et al. [5] introduz o, já referido, modelo $INGARCH(p,q)$ que corresponde ao presente considerando $\varphi_t(u) = \varphi(u) = e^{iu}$, isto é, a função característica de uma lei de Dirac concentrada em $\{1\}$.*

***2.*** *Em [9] estudou-se o modelo NB-INGARCH(p,q) inspirado em [5] mas em que a distribuição condicional de $X_t$ é a binomial negativa de parâmetros $(r, p_t)$ com $p_t = \frac{1}{1+\lambda_t}$ e $r \in \mathbb{N}$. Com o modelo geral* (1) *recuperamos, a menos de um factor de escala, o anterior considerando $\varphi_t(u) = \frac{\ln(1-(1-p_t)e^{iu})}{\ln p_t}$ e $p_t = e^{-\lambda_t^*/r}$.*

***3.*** *Zhu [10] apresenta o modelo GP-INGARCH(p,q), em analogia com os anteriores, em que a lei de $X_t|\underline{X}_{t-1}$ é a lei de Poisson generalizada de parâmetros $(\lambda_t^*, \kappa)$ com $\lambda_t^* = (1-\kappa)\lambda_t$ e $0 < \kappa < 1$. Este modelo resulta do modelo geral aqui apresentado considerando $\varphi_t = \varphi$ a função característica de uma lei de Borel de parâmetro $\kappa$ ([3], [8]).*

*Para além destes modelos recentemente estudados na literatura, uma vasta classe de processos está incluída no modelo CP-INGARCH. Os próximos exemplos mostram como obter este tipo de processos e ainda um caso particular onde $(\varphi_t)$ é uma família de funções características deterministas dependentes de t (caso 5).*

***4.*** *Sejam $(X_{t,j}, t \in \mathbb{Z})$ v.a. independentes seguindo qualquer lei discreta de média finita, independente de t e suporte $\mathbb{N}_0$, e $N_t$ uma v.a. independente de $X_{t,j}$ e seguindo uma lei de Poisson de parâmetro $\frac{\lambda_t}{E(X_{t,j})}$. O processo $X_t = \sum_{j=1}^{N_t} X_{t,j}$ verifica o modelo* (1).

***5.*** *Se $(X_{t,j}, t \in \mathbb{Z})$ são v.a. independentes seguindo a lei binomial de parâmetros $r \in \mathbb{N}$ e $\frac{1}{t^2+1}$, isto é, $\varphi_t(u) = (\frac{e^{iu}+t^2}{t^2+1})^r$, $u \in \mathbb{R}$, $t \in \mathbb{Z}$, então o processo $X_t = \sum_{j=1}^{N_t} X_{t,j}$ verifica um modelo CP-INGARCH considerando $N_t$ independente de $X_{t,j}$ e a seguir a lei de Poisson de parâmetro $\frac{\lambda_t(t^2+1)}{r}$.*

Apresentam-se de seguida, na Figura 1, trajectórias do processo $X$, bem como os resumos descritivos da sua lei marginal, para $p = q = 1$.

**Observação 2.4 (Representação** *CP-INARCH*$(\infty)$**)**
*Considerem-se os polinómios*

$$A(L) = \alpha_1 L + \cdots + \alpha_p L^p \quad e \quad B(L) = 1 - \beta_1 L - \cdots - \beta_q L^q,$$

*onde L representa o operador atraso. Para garantir a existência da inversa de $B(L)$ suponha-se que as raízes de $B(z) = 0$ estão fora*

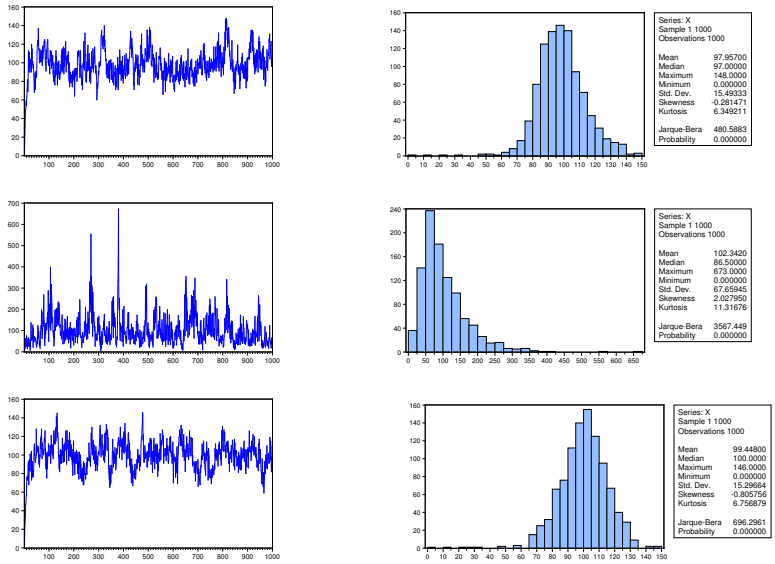



Figura 1: Lei condicional de Poisson (em cima), binomial negativa com $r = 5$ (no centro) e com $\varphi_t$ função característica de uma distribuição binomial$(15, \frac{1}{t^2+1})$ (em baixo): $\alpha_0 = 10$, $\alpha_1 = 0.4$ e $\beta_1 = 0.5$.

*do círculo unitário, o que é equivalente à hipótese* $\boldsymbol{H1}: \sum_{j=1}^{q} \beta_j < 1$. *Assim, sob esta hipótese, podemos reescrever a esperança condicional do modelo* (1) *na forma*

$$B(L)\lambda_t = \alpha_0 + A(L)X_t \Leftrightarrow \lambda_t = \alpha_0 B^{-1}(1) + H(L)X_t$$

*com* $H(L) = B^{-1}(L)A(L) = \sum_{j=1}^{\infty} \psi_j L^j$, *onde* $\psi_j$ *é o coeficiente de* $z^j$ *no desenvolvimento de MacLaurin da função* $A(z)/B(z)$, *e portanto temos uma representação* $CP\text{-}INARCH(\infty)$ *do modelo em estudo, nomeadamente*

$$\lambda_t = \alpha_0 B^{-1}(1) + \sum_{j=1}^{\infty} \psi_j X_{t-j}.$$

# 3 Propriedades de estacionariedade

No estudo de modelos de ajustamento a séries temporais é importante verificar propriedades de estabilidade ao longo do tempo. O estudo da estacionariedade de tais modelos é uma questão básica na sua análise probabilística e será agora desenvolvido.

## 3.1 Estacionariedade em média

O teorema seguinte estabelece uma condição necessária e suficiente de estacionariedade em média do modelo *CP-INGARCH(p,q)* geral.

**Teorema 3.1** *Seja* $(X_t, t \in \mathbb{Z})$ *um processo seguindo o modelo* (1). *O processo é estacionário em média sse* $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$.

*Dem.*: Sendo $X_t$ uma função mensurável positiva, pode formalmente escrever-se

$$\mu_t = E(X_t) = E(E(X_t|\underline{X}_{t-1})) = E(\lambda_t)$$
$$\Leftrightarrow \mu_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \mu_{t-j} + \sum_{k=1}^{q} \beta_k \mu_{t-k},$$

tendo em conta que os somatórios envolvidos existem, podendo não ser finitos. Observe-se que esta equação de diferenças não homogénea possui uma solução estável, finita e independente de $t$ sse os zeros $z_1, \ldots, z_{\max(p,q)}$ da equação $1 - \sum_{j=1}^{p} \alpha_j z^j - \sum_{k=1}^{q} \beta_k z^k = 0$ estão no exterior do disco unitário, ou seja, sse $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$. ■

**Observação 3.2** *Como consequência do teorema anterior, os processos* $X$ *e* $\lambda = (\lambda_t, t \in \mathbb{Z})$ *são simultaneamente estacionários em média se* $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$ *e tem-se*

$$E(X_t) = E(\lambda_t) = \mu = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j - \sum_{k=1}^{q} \beta_k}.$$

Uma aplicação importante do teorema 3.1 é considerar o seu contrarecíproco: se a soma das constantes do modelo for superior ou igual a 1 então podemos garantir que o processo não será estacionário em média e, neste caso, também não será nem forte nem fracamente estacionário. O gráfico da Figura 2 mostra claramente a não estacionariedade de uma série em que a soma dos parâmetros é 1.

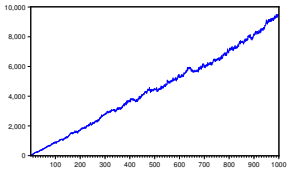



Figura 2: Lei condicional de Poisson: $\alpha_0 = 10$, $\alpha_1 = 0.4$ e $\beta_1 = 0.6$.

## 3.2 Estacionariedade forte e fraca

Nesta secção vamos estudar a existência de soluções forte e fracamente estacionárias na subclasse dos modelos $CP\text{-}INGARCH(p,q)$ para os quais a função característica $\varphi_t$ é determinista independente de $t$ como acontece, em particular, quando a lei condicional é a lei de Poisson ou de Poisson generalizada. Veremos ainda que é possível estabelecer a ergodicidade, propriedade que tal como a estacionariedade, é bastante importante no estudo e análise de uma série temporal já que o carácter ergódico de um processo estocástico permite garantir que as suas propriedades estatísticas sejam deduzidas de uma única e suficientemente longa trajectória do processo.

### 3.2.1 Construção de uma sucessão $\{(X_t^{(n)}, t \in \mathbb{Z}), n \in \mathbb{Z}\}$ de processos estacionários em média e propriedades

Considere-se uma função característica $\varphi$ relativa a uma lei discreta e o correspondente modelo (1) verificando a hipótese **H1** e seja $(\psi_j, j \in \mathbb{N})$ a sucessão de coeficientes associada à representação $CP\text{-}INARCH(\infty)$ do modelo. Seja $(U_t, t \in \mathbb{Z})$ uma sucessão de v.a. inteiras não negativas independentes seguindo a lei de Poisson composta de função característica

$$\Phi_{U_t}(u) = exp\left\{\psi_0 \frac{i}{\varphi'(0)}\left[\varphi(u) - 1\right]\right\},$$

com $\psi_0 = \alpha_0 B^{-1}(1)$. Para cada $t \in \mathbb{Z}$ e $k \in \mathbb{N}$, seja $\mathcal{Z}_{t,k} = \{Z_{t,k,j}\}_{j\in\mathbb{N}}$ a sucessão de v.a. inteiras não negativas independentes possuindo a lei de Poisson composta de função característica

$$\Phi_{Z_{t,k,j}}(u) = exp\left\{\psi_k \tfrac{i}{\varphi'(0)}\left[\varphi(u) - 1\right]\right\}.$$

Note-se que $E(U_t) = \psi_0$, $E(Z_{t,k,j}) = \psi_k$ e que $Z_{t,k,j}$ são identicamente distribuídas para cada par $(t,j) \in \mathbb{Z} \times \mathbb{N}$. Assuma-se ainda que todas as variáveis $U_s$, $Z_{t,k,j}$ com $s$, $t \in \mathbb{Z}$, $k$, $j \in \mathbb{N}$, são independentes entre si. Baseada nestas v.a. define-se a sucessão $X_t^{(n)}$ da seguinte forma:

$$X_t^{(n)} = \begin{cases} 0, & n < 0 \\ U_t, & n = 0 \\ U_t + \sum_{k=1}^{n} \sum_{j=1}^{X_{t-k}^{(n-k)}} Z_{t-k,k,j}, & n > 0 \end{cases},$$

onde se convenciona que $\sum_{j=1}^{0} Z_{t-k,k,j} = 0$.

No que se segue usamos a notação $\mu_n = E(X_t^{(n)})$, $\forall n \in \mathbb{Z}$, uma vez que por indução facilmente se verifica a sua independência relativamente a $t$. Usando o facto de $\mu_k = 0$ quando $k < 0$, deduz-se para $n > 0$, que a sucessão $\{\mu_n\}$ verifica uma equação de diferenças finitas de grau $\max(p,q)$ e coeficientes constantes, nomeadamente,

$$\mu_n = \psi_0 + \sum_{k=1}^{\infty} \psi_k \mu_{n-k} = B^{-1}(L)\left[A(L)\mu_n + \alpha_0\right] \Leftrightarrow K(L)\mu_n = \alpha_0.$$

O polinómio característico desta equação, $K(z) = B(z) - A(z)$, possui todos os zeros fora do disco unitário se $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$. Assim, sob esta condição, $\{(X_t^{(n)}, t \in \mathbb{Z}),\ n \in \mathbb{Z}\}$ é uma sucessão de processos estacionários em média, de onde se deduz que

$$\lim_{n\to\infty} \mu_n = \frac{\psi_0}{1 - \sum_{k=1}^{\infty} \psi_k} = \frac{\alpha_0 B^{-1}(1)}{1 - H(1)} = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j - \sum_{k=1}^{q} \beta_k} = \mu.$$

A condição anterior, para além de estabelecer a estacionariedade em média de $X_t^{(n)}$, é fundamental para provar a existência de uma solução estacionária e ergódica do modelo (1), que não é mais do que o limite quase certo e em $L^2$ da sucessão $\{X_t^{(n)}\}$.

**Proposição 3.3** *Seja* $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$.

***a)*** *A sucessão* $\{(X_t^{(n)}, t \in \mathbb{Z}),\ n \in \mathbb{Z}\}$ *converge quase certamente (q.c.) e em* $L^1$ *para um processo* $X^* = (X_t^*, t \in \mathbb{Z})$.

***b)*** *Se* $\varphi$ *é derivável no ponto zero pelo menos até à ordem 2, então* $\{(X_t^{(n)}, t \in \mathbb{Z}),\ n \in \mathbb{Z}\}$ *converge também em* $L^2$ *para* $(X_t^*, t \in \mathbb{Z})$.

*Dem.*: ***a)*** Como facilmente se estabelece usando indução, a sucessão $\{(X_t^{(n)}, t \in \mathbb{Z}), n \in \mathbb{Z}\}$ é não decrescente. Usando esta monotonia e a hipótese sobre os coeficientes do modelo prova-se que a sucessão $\{(X_t^{(n)}, t \in \mathbb{Z}), n \in \mathbb{Z}\}$ converge q.c. para um processo, $(X_t^*, t \in \mathbb{Z})$, que é finito q.c., aplicando o teorema de Borel-Cantelli como na proposição 2 de [5]. Assim, pelo teorema de Beppo Lévy tem-se que o primeiro momento de $X_t^*$ é finito já que, da estacionariedade em média de $X_t^{(n)}$, se conclui

$$\mu = \lim_{n\to\infty} \mu_n = \lim_{n\to\infty} E\left(X_t^{(n)}\right) = E\left(X_t^*\right),$$

e consequentemente, a convergência de $\{(X_t^{(n)}, t \in \mathbb{Z}), n \in \mathbb{Z}\}$ em $L^1$.
***b)*** Para a convergência em $L^2$ seguimos o raciocínio apresentado nas proposições 4 e 5 de [5] notando que

$$V\left(Z_{t-k,k,j}\right) = -\Phi''_{Z_{t-k,k,j}}(0) - \psi_k^2 = -i\frac{\varphi''(0)}{\varphi'(0)}\psi_k < \infty. \qquad \blacksquare$$

### 3.2.2 O processo $X^* = (X_t^*, t \in \mathbb{Z})$ é uma solução forte e fracamente estacionária e ergódica do modelo

Como consequência dos resultados da secção anterior, conclui-se o teorema seguinte.

**Teorema 3.4** *Seja $X$ um processo seguindo o modelo* (1) *com $\varphi_t$ determinista independente de $t$ e verificando a hipótese* ***H1****.*
***a)*** *$\{(X_t^{(n)}, t \in \mathbb{Z}), n \in \mathbb{Z}\}$ é uma sucessão de processos fortemente estacionários e ergódicos;*
***b)*** *Existe um processo fortemente estacionário e ergódico, solução do modelo CP-INGARCH(p,q) se $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$.*

*Dem.*: ***a)*** A prova da estacionariedade forte segue o procedimento apresentado na proposição 3 de [5], já que as sucessões $(U_t, t \in \mathbb{Z})$ e $(\mathcal{Z}_{t,k}, t \in \mathbb{Z}, k \in \mathbb{N})$ definidas na secção 3.2.1 são de v.a. independentes e identicamente distribuídas (i.i.d.). Para além disso, $(X_t^{(n)})$ é uma sucessão de processos ergódicos porque é função mensurável da sucessão $\{(U_t, \mathcal{Z}_{t,j}), t \in \mathbb{Z}, j \in \mathbb{N}\}$ de v.a. i.i.d. e, por isso, ergódica ([4], p. 332).

***b)*** O limite q.c. da sucessão $(X_t^{(n)})$ é solução do modelo, já que

$$\Phi_{X_t^*|\underline{X}_{t-1}^*}(u) \overset{(1)}{=} \lim_{n\to+\infty} \Phi_n(u) \overset{(2)}{=} e^{i\frac{\lambda_t}{\varphi'(0)}[\varphi(u)-1]}, \quad u \in \mathbb{R},$$

com $\Phi_n$ a função característica da sucessão $r_t^{(n)}|\underline{X}_{t-1}^*$, onde

$$r_t^{(n)} = U_t + \sum_{k=1}^{n} \sum_{j=1}^{X_{t-k}^*} Z_{t-k,k,j}.$$

De facto, a igualdade (1) resulta do teorema de Paul Lévy uma vez que, analogamente à secção 2.6 de [5], se prova que para um $t$ fixo, a sucessão $r_t^{(n)} - X_t^{(n)}$ converge em média para zero e portanto $r_t^{(n)}$ converge em probabilidade para $X_t^*$. Da independência das variáveis envolvidas na definição de $r_t^{(n)}$ e condicionalmente a $\underline{X}_{t-1}^*$, deduz-se

$$\Phi_n(u) = \exp\left\{\left(\psi_0 + \sum_{k=1}^{n} \psi_k X_{t-k}^*\right) \frac{i}{\varphi'(0)} [\varphi(u) - 1]\right\},$$

e portanto quando $n \to \infty$, conclui-se a igualdade (2).
Basta-nos agora provar que esta solução é fortemente estacionária e ergódica. Recorde-se que quando $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$, $(X_t^{(n)})$ é uma sucessão de processos fortemente estacionários que converge q.c. para $(X_t^*)$. Assim, considerando sem perda de generalidade os índices $\{1,\ldots,k\}$ tem-se, $\forall h \in \mathbb{Z}$ e para $n \to \infty$,

$$(X_1^{(n)},\ldots,X_k^{(n)}) \longrightarrow (X_1^*,\ldots,X_k^*), \quad (X_{1+h}^{(n)},\ldots,X_{k+h}^{(n)}) \longrightarrow (X_{1+h}^*,\ldots,X_{k+h}^*),$$

q.c., e consequentemente em lei. Usando a estacionariedade forte do processo $(X_t^{(n)})$ e a unicidade do limite conclui-se que o processo $(X_t^*)$ é fortemente estacionário. Para provar a ergodicidade de $(X_t^*)$ basta ter em conta que este processo é o limite quase certo de uma sucessão de funções mensuráveis de um processo ergódico, isto é, uma função mensurável do processo ergódico $(U_t, \mathcal{Z}_{t,j}, j \in \mathbb{N})$ ([1], Teorema 36.4). ■

**Observação 3.5** *Se $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$, $(X_t^*)$ é fracamente estacionário já que, pela alínea b) da proposição* 3.3*, é um processo de $L^2$ fortemente estacionário.*

# 4 Conclusões

Neste trabalho introduzimos uma classe geral de modelos INGARCH que inclui como casos particulares algumas contribuições recentes na modelação de séries temporais de valores inteiros ([5], [9], [10]). Esta generalização é obtida considerando que a distribuição de $X_t$ dado o seu passado pertence à família das leis infinitamente divisíveis discretas e definindo o modelo através da correspondente função característica.

A existência de uma solução forte e fracamente estacionária e ergódica na subclasse dos modelos CP-INGARCH gerados por uma função característica determinista e não dependente de $t$ é um primeiro passo, seguramente importante, para a análise destas propriedades na família geral de tais processos. Estabelecidas tais propriedades fundamentais na referida subfamília, que inclui alguns dos modelos particulares presentes na literatura, novos estudos probabilísticos, como os ligados à existência e cálculo dos momentos serão desenvolvidos com vista à utilização desta família abrangente de processos de contagem na modelação de séries temporais de valores inteiros.

# Agradecimentos

As autoras agradecem aos *referees* a leitura cuidada e os comentários feitos. Este trabalho teve o apoio do Centro de Matemática da Universidade de Coimbra (pelo European Regional Development Fund, programa COMPETE, e pelo Governo Português através da Fundação para a Ciência e Tecnologia (FCT), projecto PEst-C/MAT/UI0324/2011). O trabalho do terceiro autor é apoiado pela Bolsa de Investigação SFRH/BD/85336/2012 da FCT.

# Referências

[1] Billingsley, P. (1995). *Probability and Measure*, New York, Wiley, 3rd edition.

[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Econometrics* 31, 307–327.

[3] Consul, P.C., Famoye, F. (2006) *Lagrangian Probability Distributions*, Birkhäuser, Boston.

[4] Durrett, R. (2010). *Probability: Theory and Examples*, Cambridge University Press, 4th ed..

[5] Ferland, R., Latour, A., Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis* 27, 923–942.

[6] Gonçalves, E., Mendes Lopes, N., Silva, F. (2013). Infinitely divisible distributions in integer valued GARCH models, DMUC, Preprint 13-38.

[7] Steutel, F.W., van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line*, Marcel-Dekker, Inc, New York.

[8] Weiß, C.H. (2008). Thinning operations for modelling time series of counts- a survey. *Advances in Statistical Analysis* 92, 319–341.

[9] Zhu, F. (2011). A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis* 32, 54–67.

[10] Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications* 389 1, 58–71.

# Metodologias estatísticas para estudo da interacção genótipo×ambiente em clones de videira

Elsa Gonçalves
Secção de Matemática/DCEB e Centro de Botânica Aplicada à Agricultura, Instituto Superior de Agronomia, Universidade de Lisboa, *elsagoncalves@isa.ulisboa.pt*

Antero Martins
Centro de Botânica Aplicada à Agricultura, Instituto Superior de Agronomia, Universidade de Lisboa, *anteromart@isa.ulisboa.pt*

**Palavras–chave**: Modelos mistos, correlações genéticas, interacção genótipo×ambiente, *biplot*

**Resumo**: Na videira as metodologias de estudo da interacção genótipo ×ambiente (G×E) encontram-se ainda insuficientemente desenvolvidas. Neste trabalho propõe-se uma metodologia de estudo da interacção adaptada às fases iniciais da selecção baseada no ajustamento de um modelo linear misto que estuda a interacção G×E através da correlação genética entre ambientes, seguido de uma abordagem descritiva de síntese dos resultados referentes aos preditores dos efeitos genotípicos recorrendo à análise em componentes principais. A aplicação da metodologia leva à identificação de um grupo de genótipos da casta Arinto simultaneamente superiores em rendimento e estáveis na gama de ambientes estudada.

## 1 Introdução

Os materiais de propagação hoje largamente usados na viticultura - clones - são geneticamente homogéneos, por isso, podem exibir sensibilidade à interacção genótipo×ambiente (G×E) particularmente elevada. Isto significa que os resultados da avaliação de um genó-

tipo em determinados ambientes onde foi seleccionado poderão não se observar noutros ambientes, nos quais o genótipo irá ser posteriormente cultivado pelo viticultor.

No melhoramento de plantas em geral, as principais técnicas aplicáveis ao estudo da interacção recorrem à análise de regressão dos rendimentos do genótipo sobre os índices ambientais [4], a métodos não paramétricos baseados na ordenação dos genótipos em diferentes ambientes [10], a técnicas baseadas na análise em componentes principais [8] e a modelos mistos [17]. As técnicas originais baseadas na análise de regressão evoluíram para modelos mistos nos quais é incorporada uma covariável ambiental (média dos genótipos num determinado ambiente) [6], embora a sua utilização seja fortemente criticada pelo facto de a covariável ser estimada a partir dos próprios dados. Os modelos baseados na análise em componentes principais são os chamados modelos GGE (*Genotype main effects and Genotype×Environment interaction effects*) e AMMI (*Additive main effects and multiplicative interaction*). Presentemente, estes estão entre os métodos estatísticos mais usados para analisar dados de rendimento de ensaios agrícolas [18, 5]. De forma resumida, o que os distingue é a transformação aplicada aos dados antes de se fazer a decomposição em valores singulares. Nos primeiros a decomposição em valores singulares é aplicada a $(\bar{y}_{ij} - \bar{y}_{.j})$, e nos segundos a $(\bar{y}_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})$, onde: $\bar{y}_{ij}$ é o rendimento médio do genótipo $i$ no ambiente $j$; $\bar{y}_{.j}$ é o rendimento médio de todos os genótipos no ambiente $j$; $\bar{y}_{i.}$ é o rendimento médio do genótipo $i$ em todos os ambientes e $\bar{y}_{..}$ é a média geral. Estes modelos consideram os efeitos genotípicos e os efeitos da interacção G×E como fixos, são adequados para conjuntos de dados equilibrados, não consideram a heterogeneidade de variâncias entre ambientes, etc.. Na tentativa de os tornar mais flexíveis e adaptados aos problemas concretos do melhoramento de plantas, ambas as abordagens têm sido alvo de sucessivos desenvolvimentos. No caso dos modelos GGE, têm sido propostas várias transformações alternativas [19]: $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SE_j}$, $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SE_j/\sqrt{r}}$, $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SD_j/\sqrt{h_j^2}}$, onde $SE_j$ é o erro padrão residual no ambiente $j$, $r$ é o

número de repetições no ambiente $j$, $SD_j$ é o desvio padrão fenotípico ao nível das médias dos genótipos no ambiente $j$, dado por $\sqrt{\hat{\sigma}_g^2 + \hat{\sigma}_E^2/r}$, $h_j^2$ é a heritabilidade em sentido lato no ambiente $j$, definida como a fracção da variância fenotípica que é atribuível a causas genotípicas, isto é, $\frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_E^2/r}$, sendo $\hat{\sigma}_g^2$ e $\hat{\sigma}_E^2$ as estimativas das variâncias genotípica e do erro, respectivamente. No fundo, todas estas transformações têm como objectivo aproximar os valores fenotípicos (observados) dos valores genotípicos pela redução da parte não genética do valor fenotípico. No caso dos modelos AMMI, estes têm evoluído para modelos mistos, designados por modelos FA (*Factor analytic multiplicative mixed model*) [12, 13, 16, 17], em que se propõe o uso de modelos multiplicativos para os efeitos genotípicos em cada ambiente (que se admitem aleatórios) baseados na técnica multivariada de análise factorial [9]. Sendo $\mathbf{u}$ o vector $mt \times 1$ dos $m$ efeitos genotípicos do rendimento nos $t$ ambientes, vem

$$\mathbf{u} = (\boldsymbol{\lambda}_1 \otimes \mathbf{I}_m)\mathbf{f}_1 + \cdots + (\boldsymbol{\lambda}_k \otimes \mathbf{I}_m)\mathbf{f}_k + \boldsymbol{\delta},$$

em que $\mathbf{f}_r$ $(r = 1, \ldots, k < t)$ representa os vectores $m \times 1$ dos k factores aleatórios, $\boldsymbol{\lambda}_r$ representa os vectores $t \times 1$ de *loadings* dos factores nos ambientes e $\boldsymbol{\delta}$ é o vector $mt \times 1$ dos resíduos deste modelo. Frequentemente a sistematização dos resultados dos efeitos genotípicos é feita num gráfico dos *loadings* do factor 1 contra os *loadings* do factor 2. A sua utilização é defendida dada a facilidade de convergência do algoritmo de estimação dos parâmetros quando o número de ambientes em estudo é elevado. Actualmente são muito usados no melhoramento genético animal e vegetal [2].

De entre as técnicas atrás referidas, quando aplicadas à videira, quase todas são mais apropriadas para o estudo da interacção em fases finais de selecção, quando um número de genótipos reduzido (entre 30 a 40) é avaliado num maior número de ambientes, com o objectivo de seleccionar um ou poucos genótipos (selecção clonal). As mais adequadas para o estudo da interacção numa fase inicial de selecção passam necessariamente pelo ajustamento de modelos mistos. Nessa fase trabalha-se com uma amostra representativa da

variabilidade da casta (compreendendo, em geral, entre 100 a 300 genótipos) num número reduzido de ambientes (em geral, entre 3 e 6). Nesta fase o objectivo é seleccionar um grupo de genótipos superiores (entre 20 a 40) para utilização imediata na viticultura e/ou entrada num novo ciclo de experimentação conducente à selecção clonal.
O presente trabalho propõe uma metodologia de estudo da interacção genótipo×ambiente adaptada às fases iniciais de selecção. Concretamente, propõe-se uma metodologia que permita a selecção de um grupo de genótipos geneticamente superiores em rendimento e estáveis na gama de ambientes estudada.

## 2 A metodologia proposta

Matricialmente, o modelo linear misto pode ser genericamente descrito como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}, \tag{1}$$

em que $\mathbf{Y}$ é o vector $n\times 1$ das observações (rendimentos observados), $\mathbf{X}$ é a matriz de delineamento $n\times p$ dos efeitos fixos, $\boldsymbol{\beta}$ é o vector $p\times 1$ de efeitos fixos, $\mathbf{Z}$ é a matriz de delineamento $n \times q$ dos efeitos aleatórios, $\mathbf{u}$ é o vector $q \times 1$ de efeitos aleatórios e $\boldsymbol{\varepsilon}$ é o vector $n\times 1$ de erros aleatórios.

Os vectores $\mathbf{u}$ e $\boldsymbol{\varepsilon}$ admitem-se independentes, com distribuição normal multivariada de vector de valores médios nulo e matrizes de covariâncias $\mathbf{G}$ e $\mathbf{R}$, respectivamente, isto é,

$$cov\left[\mathbf{u},\boldsymbol{\varepsilon}\right] = \mathbf{0}, \qquad \mathbf{u} \cap \mathcal{N}_q\left(\mathbf{0},\mathbf{G}\right), \qquad \boldsymbol{\varepsilon} \cap \mathcal{N}_n\left(\mathbf{0},\mathbf{R}\right).$$

A distribuição de $\mathbf{Y}$ vem assim normal multivariada, com vector de valores médios $\mathbf{X}\boldsymbol{\beta}$ e matriz de covariâncias $\mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$,

$$\mathbf{Y} \cap \mathcal{N}_n\left(\mathbf{X}\boldsymbol{\beta},\mathbf{V}\right).$$

No estudo deste tipo de modelo, os grandes objectivos são estimar as componentes de covariância e obter o melhor preditor empírico linear não enviesado de $\mathbf{u}$ [7, 15],

$$\tilde{\mathbf{u}}_{EBLUP} = \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}(\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{EBLUE}),$$

sendo $\hat{\boldsymbol{\beta}}_{EBLUE} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{Y}$ o melhor estimador empírico não enviesado de $\boldsymbol{\beta}$ , $(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}$ a inversa generalizada de $(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})$ e $\hat{\mathbf{G}}$ e $\hat{\mathbf{V}}$ as matrizes de covariâncias estimadas. O método de máxima verosimilhança restrita, REML [11], é actualmente o mais recomendado e utilizado para estimar componentes de covariância em grandes conjuntos de dados com estrutura complexa [15] e a inferência relativa às componentes de covariância é, em geral, baseada em testes de razão de verosimilhanças restritas. No contexto biológico em estudo, as variantes de modelos lineares mistos que poderão surgir estão focadas na composição do vector $\mathbf{u}$ e na estrutura das matrizes de covariâncias $\mathbf{G}$ e $\mathbf{R}$. Um modelo clássico para estudo da interacção G×E do rendimento, no contexto de um delineamento experimental em blocos completos casualizados, pode ser descrito como

$$y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \tau\,(\gamma)_{jk} + e_{ijk},$$

para $i = 1, \ldots, m$, $j = 1, \ldots, t$, $k = 1, \ldots, r$, em que $y_{ijk}$ representa o valor fenotípico do rendimento do clone $i$ no bloco $k$ do ambiente $j$, $\mu$ representa o rendimento médio populacional, $\alpha_i$ representa o efeito genotípico do clone $i$, $\gamma_j$ representa o efeito do ambiente $j$, $(\alpha\gamma)_{ij}$ representa o efeito da interacção do clone $i$ com o ambiente $j$, $\tau\,(\gamma)_{jk}$ representa o efeito do bloco $k$ no ambiente $j$, $e_{ijk}$ representa o erro aleatório associado à observação $y_{ijk}$. Os efeitos genotípicos e da interacção admitem-se aleatórios, os restantes como fixos. Utilizando a notação matricial descrita em (1), neste caso, a componente $\mathbf{Zu}$ toma a forma

$$\mathbf{Zu} = \begin{bmatrix} \mathbf{Z}_g & \mathbf{Z}_{ge} \end{bmatrix} \begin{bmatrix} \mathbf{u}_g \\ \mathbf{u}_{ge} \end{bmatrix},$$

em que $\mathbf{Z}_g$ é a matriz de delineamento $n \times m$ dos efeitos genotípicos do rendimento, $\mathbf{Z}_{ge}$ é a matriz $n \times mt$ de delineamento dos efeitos da interacção G×E, $\mathbf{u}_g$ é o vector $m \times 1$ dos efeitos genotípicos do rendimento, $\mathbf{u}_{ge}$ é o vector $mt \times 1$ dos efeitos da interacção G×E. Os vectores $\mathbf{u}_g$ e $\mathbf{u}_{ge}$ admitem-se independentes e os seus elementos admitem-se variáveis aleatórias independentes e identicamente distribuídas. A matriz de covariâncias do vector $\mathbf{u}$ é, neste contexto, definida como

$$\mathbf{G} = \mathbf{G}_g \oplus \mathbf{G}_{ge},$$

em que $\mathbf{G}_g = \sigma^2_g \mathbf{I}_m$ e $\mathbf{G}_{ge} = \sigma^2_{ge} \mathbf{I}_{mt}$, sendo $\sigma^2_g$ e $\sigma^2_{ge}$ as variâncias genotípica e da interacção G×E , respectivamente, e $\mathbf{I}_m$ e $\mathbf{I}_{mt}$ matrizes identidade $m \times m$ e $mt \times mt$, respectivamente, e $\oplus$ representa a soma directa de matrizes.
Os erros aleatórios admitem-se variáveis aleatórias independentes e identicamente distribuídas. A matriz de covariâncias do vector $\boldsymbol{\varepsilon}$ é, neste contexto, definida como

$$\mathbf{R} = \sigma^2_{\varepsilon} \mathbf{I}_n,$$

em que $\sigma^2_{\varepsilon}$ é a variância dos erros aleatórios e $\mathbf{I}_n$ a matriz identidade $n \times n$. Este é, de facto, um dos modelos lineares mistos mais utilizados para estudo da interacção G×E, mas muito questionável no actual contexto biológico (ensaios de grande dimensão, sujeitos a uma forte variação ambiental, relativos a uma planta perene, etc.).
Um modelo mais realista admite, então, variâncias genéticas distintas para cada ambiente e efeitos genotípicos entre ambientes correlacionados. Nesta abordagem, fortemente alicerçada na teoria da genética quantitativa [3], a interacção G×E é estudada a partir da correlação genética entre ambientes. O modelo proposto para estudo da interacção G×E do rendimento em fases iniciais de selecção da videira, no contexto de um delineamento experimental em blocos completos casualizados, é, então, descrito como

$$y_{ijk} = \mu + \alpha\,(\gamma)_{ji} + \gamma_j + \tau\,(\gamma)_{jk} + e_{ijk},$$

para $i=1,\ldots,m$, $j=1,\ldots,t$, $k=1,\ldots,r$, em que $y_{ijk}$ representa o valor fenotípico do rendimento do clone $i$ no bloco $k$ do ambiente $j$, $\mu$ representa o rendimento médio populacional, $\alpha(\gamma)_{ji}$ representa o efeito genotípico do clone $i$ no ambiente $j$, $\gamma_j$ representa o efeito do ambiente $j$, $\tau(\gamma)_{jk}$ representa o efeito do bloco $k$ no ambiente $j$, $e_{ijk}$ representa o erro aleatório associado à observação $y_{ijk}$. Os efeitos genotípicos em cada ambiente assumem-se aleatórios, os restantes como fixos. Utilizando a notação matricial descrita em (1), a componente $\mathbf{Zu}$ é agora definida como

$$\mathbf{Zu}=\begin{bmatrix} \mathbf{Z}_{g_1} & \cdots & \mathbf{Z}_{g_t} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{g_1} \\ \vdots \\ \mathbf{u}_{g_t} \end{bmatrix} = \sum_{j=1}^{t} \mathbf{Z}_{g_j} \mathbf{u}_{g_j},$$

em que $\mathbf{Z}_{g_j}$ é a matriz de delineamento $n\times m$ dos efeitos genotípicos do rendimento no ambiente $j$ e $\mathbf{u}_{g_j}$ o vector $m\times 1$ dos efeitos genotípicos do rendimento no ambiente $j$. A matriz de covariâncias do vector $\mathbf{u}$ é, neste contexto, definida como

$$\mathbf{G}=\begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} & \sigma_{g_{13}} & \cdots & \sigma_{g_{1t}} \\ \sigma_{g_{21}} & \sigma^2_{g_2} & \sigma_{g_{23}} & \cdots & \sigma_{g_{2t}} \\ \sigma_{g_{31}} & \sigma_{g_{32}} & \sigma^3_{g_3} & \vdots & \sigma_{g_{3t}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{g_{t1}} & \sigma_{g_{t2}} & \sigma_{g3} & \cdots & \sigma^2_{g_t} \end{bmatrix} \otimes \mathbf{I}_m,$$

sendo $\sigma^2_{g_j}$ (para $j=1,\ldots,t$), a variância genotípica do rendimento no ambiente $j$, $\sigma_{g_{jj'}}$ $(\forall j \neq j')$ a covariância genética entre o ambiente $j$ e o ambiente $j'$, $\mathbf{I}_m$ a matriz identidade $m\times m$ e $\otimes$ o produto de Kronecker de matrizes. A correlação genética entre o ambiente $j$ e o ambiente $j'$ é dada por:

$$r_{g_{jj'}}=\frac{\sigma_{g_{jj'}}}{\sigma_{g_j}\sigma_{g_{j'}}}.$$

Quanto menor for a correlação genética maior é a sensibilidade ambiental exibida pelos genótipos, ou seja, maior é a interacção G×E.

Na maioria das situações faz sentido o mesmo tipo de pressupostos para os erros aleatórios, sendo a matriz de covariâncias do vector $\boldsymbol{\varepsilon}$, neste contexto, definida como

$$\mathbf{R}=\begin{bmatrix} \sigma^2_{\varepsilon_1} & \sigma_{\varepsilon_{12}} & \sigma_{\varepsilon_{13}} & \cdots & \sigma_{\varepsilon_{1t}} \\ \sigma_{\varepsilon_{21}} & \sigma^2_{\varepsilon_2} & \sigma_{\varepsilon_{23}} & \cdots & \sigma_{\varepsilon_{2t}} \\ \sigma_{\varepsilon_{31}} & \sigma_{\varepsilon_{32}} & \sigma^3_{\varepsilon_3} & \vdots & \sigma_{\varepsilon_{3t}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_{t1}} & \sigma_{\varepsilon_{t2}} & \sigma_{\varepsilon_{t3}} & \cdots & \sigma^2_{\varepsilon_t} \end{bmatrix} \otimes \mathbf{I}_{mr}.$$

Trata-se, portanto, de um modelo complexo, pelo que o algoritmo de estimação dos parâmetros poderá não convergir quando o número de ambientes é elevado. Deverá ser comparado com o modelo que assuma $\sigma_{g_{jj'}}=0$ e $\sigma_{\varepsilon_{jj'}}=0$ $(\forall j \neq j')$.
Com esta análise obtêm-se os melhores preditores empíricos lineares não enviesados (EBLUPs) dos efeitos genotípicos do rendimento de todos os clones para cada ambiente. Note-se que todas as decisões de selecção são baseadas nestes EBLUPs dos efeitos genotípicos. Contudo, se esta informação não for sintetizada, será difícil de interpretar. Torna-se, portanto, necessário complementar esta análise com uma abordagem descritiva que permita a compreensão da relação entre genótipos e ambientes. A alternativa proposta é fazer uma análise em componentes principais, não com as transformações dos dados descritas na secção 1, mas sim com os EBLUPs dos efeitos genotípicos do rendimento obtidos com o ajustamento do modelo linear misto proposto. Ou seja, a decomposição em valores singulares será feita com base na matriz

$$\mathbf{H} = \left[ \begin{array}{c|c|c|c} \tilde{\mathbf{u}}_{EBLUP_1} & \tilde{\mathbf{u}}_{EBLUP_2} & \cdots & \tilde{\mathbf{u}}_{EBLUP_t} \end{array} \right],$$

sendo $\tilde{\mathbf{u}}_{EBLUP_j}$ (para $j=1,\ldots,t$) o vector $m \times 1$ dos melhores preditores empíricos lineares não enviesados dos efeitos genotípicos do rendimento no ambiente $j$.
Adoptando este procedimento e, se as duas primeiras componentes

principais conseguirem explicar uma elevada percentagem da variabilidade total desses dados, então a imagem reproduzida num *biplot* será uma ferramenta muito útil para compreender a relação entre genótipos e ambientes, e assim conseguir seleccionar de uma forma expedita um grupo de clones simultaneamente superiores em rendimento e estáveis na gama de ambientes estudada.

# 3 Uma aplicação

Os dados utilizados referem-se ao rendimento (kg/planta) da casta Arinto, uma das castas brancas com mais forte identidade e mais cultivadas do país. Esses dados são provenientes de ensaios iniciais de selecção, contendo amostras de genótipos representativas da diversidade da casta em distintas regiões de cultura e perfazendo um elevado efectivo de clones. Foram estudados 169 genótipos em dois locais com delineamento experimental em blocos completos casualizados, com 4 repetições. Em cada local o rendimento foi avaliado em vários anos, tendo-se considerado cada combinação local/ano como um ambiente distinto. Estudaram-se no total 6 ambientes: A1 (Alenquer/1988); A2 (Alenquer/1989); S1 (Setúbal/1995); S2 (Setúbal/1998); S3 (Setúbal /1999); S4 (Setúbal /2000).

As análises foram efectuadas com o *Software R*. Para o ajustamento do modelo linear misto proposto recorreu-se ao *package* $ASReml-R$ [1] (método de estimação REML, usando o algoritmo de informação média). Para a análise em componentes principais e construção do *biplot* foram usados comandos *prcomp* e *biplot*, respectivamente.

Com o ajustamento do modelo misto proposto obtiveram-se as estimativas das componentes de variância genotípica do rendimento nos 6 ambientes ($\hat{\sigma}^2_{gA1} = 0{,}040$; $\hat{\sigma}^2_{gA2} = 0{,}455$; $\hat{\sigma}^2_{gS1} = 0{,}034$; $\hat{\sigma}^2_{gS2} = 0{,}198$; $\hat{\sigma}^2_{gS3} = 0{,}180$; $\hat{\sigma}^2_{gS4} = 0{,}128$, em todos os casos valor-p$<0{,}001$), rejeitando-se a hipótese de inexistência de variância genotípica do rendimento em todos os ambientes, para qualquer nível de signifi-

cância usual. O valor do critério de informação de Akaike [14] obtido com o ajustamento deste modelo ($AIC = -744{,}3278$) foi menor que o valor obtido com o ajustamento do modelo que assume independência entre ambientes ($AIC = -108{,}9314$).

Com base nas covariâncias genéticas estimadas, construiu-se a matriz das correlações genéticas estimadas entre os diferentes ambientes (Tabela 1).

| | **A1** | **A2** | **S1** | **S2** | **S3** | **S4** |
|---|---|---|---|---|---|---|
| **A1** | 1 | 0,471 | 0,719 | 0,754 | 0,852 | 0,802 |
| **A2** | 0,471 | 1 | 0,502 | 0,338 | 0,384 | 0,381 |
| **S1** | 0,719 | 0,502 | 1 | 0,735 | 0,718 | 0,705 |
| **S2** | 0,754 | 0,338 | 0,735 | 1 | 0,950 | 0,882 |
| **S3** | 0,852 | 0,384 | 0,718 | 0,950 | 1 | 0,927 |
| **S4** | 0,802 | 0,381 | 0,705 | 0,882 | 0,927 | 1 |

Tabela 1: Matriz das correlações genéticas estimadas entre ambientes A1, A2, S1, S2, S3 e S4.

A presença de correlações genéticas baixas a moderadas entre alguns ambientes evidencia a existência de interacção G×E. De entre os ambientes estudados, o ambiente A2 revelou ser o mais distinto de todos os outros. Nos ambientes S2, S3 e S4 os genótipos reagiram aproximadamente da mesma maneira. As correlações genéticas entre anos no mesmo local não foram necessariamente as mais elevadas, evidenciando a grande influência do ano na definição de ambiente.
Para cada ambiente obtiveram-se os EBLUPS dos efeitos genotípicos do rendimento dos 169 genótipos. A interpretação desta informação foi feita com base numa análise em componentes principais. A relação entre o comportamento dos genótipos nos diversos ambientes pode ser visualizada no *biplot* da Figura 1, graças à elevadíssima percentagem de variabilidade explicada pelas duas primeiras componentes principais (98,58%). De acordo com o já observado na Tabela 1, também se observa na Figura que o comportamento comparado

dos genótipos no ambiente A2 foi claramente distinto do observado nos outros ambientes e que os efeitos genotípicos do rendimento dos clones nos ambientes S2, S3 e S4 estão altamente correlaccionados. Também se verifica que a variabilidade genotípica do rendimento é menor nos ambientes A1 e S1 comparativamente à observada nos restantes ambientes. Da observação da Figura consegue-se distinguir clones menos sensíveis à interacção G×E, uns com rendimento acima da média em todos os ambientes (clones AR2404, AR3605, AR0310, AR8201, AR0664, AR1501, AR8007, AR9005, AR0230, AR2410, AR0498, etc.), outros com rendimento abaixo da média em todos os ambientes (clones AR6118, AR6502, AR1627, etc). Também é possível identificar a existência de clones mais sensíveis à interacção G×E: clones que revelam um bom rendimento nos ambientes S2, S3, S4, mas baixo rendimento em A2 (clones AR8807, AR8801, AR6116, AR6707, etc.), e clones que exibem bom redimento no ambiente A2, mas baixo rendimento em S2, S3, S4 (por exemplo, clones AR6619, AR3602 e AR8505). Assim, se o objectivo for seleccionar clones que apresentem rendimento acima da média em todos os ambientes, o mais indicado será seleccionar os clones que se encontram dentro do círculo desenhado no *biplot*.

Em suma, o ajustamento do modelo linear misto proposto e a síntese dos resultados referentes aos preditores dos efeitos genotípicos do rendimento recorrendo à análise em componentes principais permitiu caracterizar a interacção G×E em clones de videira. Com esta metodologia conseguiu-se identificar um grupo de clones (no mínimo 30 clones) que são simultaneamente bons em rendimento e estáveis na gama de ambientes estudada. Esse grupo poderá ser disponibilizado aos viticultores para a plantação de novas vinhas e, paralelamente, entrar num novo ciclo de experimentação.

# 4 Considerações finais

Algumas considerações são devidas sobre as vantagens da abordagem proposta face a outras referidas na secção 1. A sua comparação

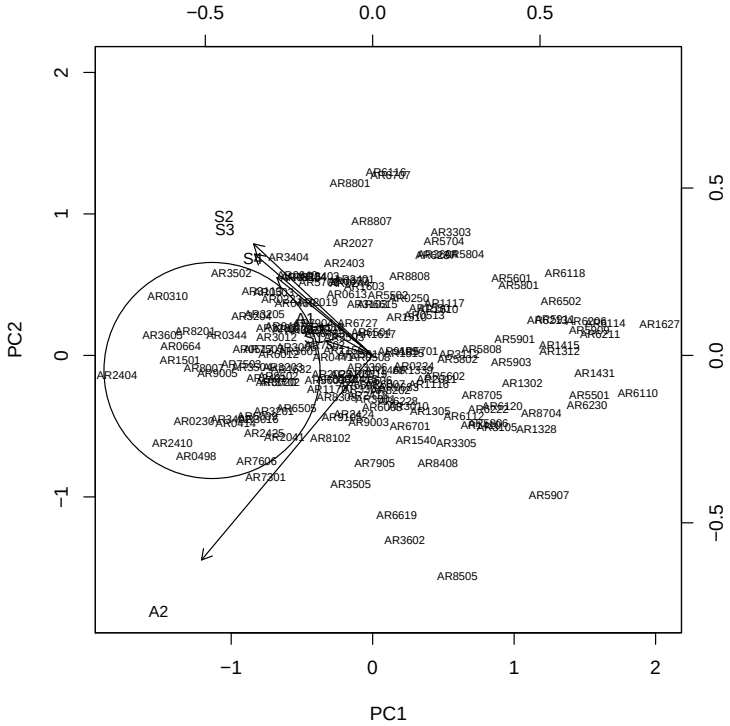



Figura 1: *Biplot* baseado nos EBLUPs dos efeitos genotípicos do rendimento de 169 clones de Arinto em 6 ambientes (A1, A2, S1, S2, S3 e S4). PC1 e PC2 são a primeira e segunda componentes principais, respectivamente (PC1=72,34%, PC2=26,24%, Soma = 98,58%). A circunferência identifica um grupo de clones que são simultaneamente superiores em rendimento e estáveis na gama de ambientes estudada.

com os modelos GGE e AMMI faz pouco sentido, já que estes admitem efeitos genotípicos fixos (para além de outros pressupostos questionáveis no actual contexto biológico). Ainda assim, para o caso da casta Arinto os resultados da análise em componentes principais obtidos com as transformações mais recentes propostas nos modelos GGE foram os seguintes: para $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SE_j}$, $PC1 = 55{,}84\%$, $PC2 = 21{,}68\%$ (Soma = 77,52%); para $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SE_j/\sqrt{r}}$, $PC1 = 55{,}84\%$, $PC2 = 21{,}68\%$ (Soma = 77,52%); para $\frac{(\bar{y}_{ij}-\bar{y}_{.j})}{SD_j/\sqrt{h_j^2}}$, $PC1 = 57{,}20\%$, $PC2 = 17{,}95\%$ (Soma = 75,15%). Em todos os casos a percentagem da variabilidade explicada em duas componentes foi menor que a obtida com a metodologia proposta.

Por outro lado, a comparação da abordagem proposta com o modelo misto FA [17] faz mais sentido, pois este acaba por ser uma simplificação do modelo misto proposto. Recorde-se que o ajustamento do modelo FA é defendido devido à sua menor complexidade e à facilidade de convergência do algoritmo de estimação dos parâmetros quando o número de ambientes em estudo é elevado. No actual contexto biológico o número de ambientes em estudo é reduzido, pelo que adoptando o modelo misto proposto (estrutura das matrizes de covariâncias não estruturada) será a abordagem que nos fornecerá uma visão mais completa das correlações genéticas entre ambientes. No entanto, dada a complexidade deste modelo, a questão da sua aplicabilidade é pertinente. Como tal, fez-se ainda um estudo complementar com dados de rendimento de ensaios iniciais de selecção da videira distribuídos pelas várias regiões vitícolas de Portugal (ensaios em que a heritabilidade foi superior a 0,35), envolvendo 25 castas, sendo que o número de ambientes incluídos nos modelos ajustados variou entre 3 e 8. De entre os 25 modelos ajustados, o algoritmo de estimação dos parâmetros não convergiu apenas em 2 casos, ambos com 3 ambientes em estudo (o mesmo aconteceu com o ajustamento do modelo misto FA). Portanto, estes resultados suportam a aplicabilidade da metodologia proposta para a avaliação

da interacção G×E em ensaios iniciais de selecção da videira.

# Agradecimentos

Aos colegas da "Rede Nacional de Selecção da Videira" pela sua contribuição em todo o processo de selecção das castas. Ao PRODER, medida 2.2.3.1 (PA18999), e à Fundação para a Ciência e Tecnologia (PEst-OE/AGR/UI0240/2011 e PEst-OE/AGR/UI0240/2014 ) pelo apoio financeiro.

# Referências

[1] Butler, D., Cullis, B.R., Gilmour, A.R, Gogel, B.J. (2007). ASReml-R reference manual.AsReml-R estimates variance componentes under a general linear mixed model by residual maximum likelihood (REML). NSW Department of Primary Industries, Quensland Government. Queensland.

[2] Cullis, B.R., Jefferson, P., Thompson, R., Smith, A.B. (2014). Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. *Theoretical and Applied Genetics* 127, 2193–2210.

[3] Falconer, D.S., Mackay, T.F.C. (1996). *An Introduction to Quantitative Genetics*. 4th ed. Prentice Hall, London.

[4] Finlay, K., Wilkinson, G. (1963). The analysis of adaptation in a plant breeding program. *Australian Journal of Agricultural Research* 14, 742–754.

[5] Gauch, H.G., Piepho, H.P., Annicchiarico, P. (2008). Statistical analysis of yield trials by AMMI and GGE: further considerations.*Crop Science* 48, 866–889.

[6] Gogel, B.J., Cullis, B.R., Verbyla, A.P. (1995). REML estimation of multiplicative effects in multi-environment trials. *Biometrics* 51, 744–749.

[7] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447.

[8] Kempton, R. (1984). The use of biplots in interpreting cultivar by environment interactions. *Journal of Agricultural Science* 103, 123–135.

[9] Mardia, K.V., Kent, J.T., Bibby, J.M. (1988). *Multivariate Analysis*. London: Academic.

[10] Nassar, R., Huhn, M. (1987). Studies on estimation of phenotypic stability: tests of significance for non parametric measures of phenotypic stability. *Biometrics* 43, 45–53.

[11] Patterson, H.D., Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554.

[12] Piepho, H.P. (1997). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53, 761–767.

[13] Piepho, H.P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics* 97, 195–201.

[14] Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). Akaike information criterion statistics. Dordrecht: D. Reidel.

[15] Searle, S.R., Casella, G., McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons, New York.

[16] Smith, A.B., Cullis, B.R., Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147.

[17] Smith, A.B., Cullis, B.R., Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science* 143, 449–462.

[18] Yan, W., Kang, M.S., Ma, B., Woods, S., Cornelius, P. (2007). GGE Biplot vs. AMMI analysis of genotype-by-environment Data. *Crop Science* 47, 643–655.

[19] Yan, W., Holland, J. (2010). A heritability-ajusted GGE biplot for test environment evaluation. *Euphytica* 171, 355–369.

# Análise bayesiana semiparamétrica de resposta binária com covariável contínua sujeita a omissão não aleatória

Frederico Z. Poleto
IME, Universidade de São Paulo, Brasil, *fpoleto@ime.usp.br*
Carlos Daniel Paulino
IST e CEAUL, Universidade de Lisboa, Portugal, *dpaulino@math.ist.utl.pt*
Julio M. Singer
IME, Universidade de São Paulo, Brasil, *jmsinger@ime.usp.br*
Geert Molenberghs
Hasselt University, Bélgica, *geert.molenberghs@uhasselt.be*

**Palavras–chave**: Mistura por processo Dirichlet, dados incompletos, MNAR, regressão binária, análise bayesiana não paramétrica

**Resumo**: A omissão em variáveis explicativas requer um modelo para estas, mesmo que o interesse recaia apenas no modelo condicional para as respostas dadas as covariáveis. Uma especificação incorreta dos modelos para as covariáveis ou para o mecanismo de omissão pode levar a inferências enviesadas para os parâmetros de interesse. A literatura conhecida segue uma de duas vias: uso para as covariáveis de distribuições flexíveis, não paramétricas ou semiparamétricas, juntamente com uma suposição MAR, ou de distribuições paramétricas aliadas a um mecanismo de omissão mais geral de tipo MNAR. Considera-se aqui uma análise de variáveis respostas combinando um mecanismo MNAR com um modelo não paramétrico baseado numa mistura por processo Dirichlet para covariáveis contínuas sujeitas a omissão. A via descrita é ilustrada com dados simulados e também por análise de um conjunto de dados reais.

# 1 Introdução

Em muitos estudos surgem dados omissos para algumas das variáveis explicativas ($\mathbf{X}$) de tal modo que não se afigura conveniente excluir tais variáveis ou unidades amostrais da análise. Ainda que o interesse possa estar na distribuição condicional das variáveis respostas ($\mathbf{Y}$) dado $\mathbf{X}$, necessita-se de especificar também um modelo para a distribuição marginal das variáveis sofrendo omissão ou, pelo menos, para a distribuição condicional delas dadas as covariáveis que são sempre observadas.

Em casos onde pelo menos uma covariável é contínua, pode-se não ter *a priori* qualquer informação para um modelo paramétrico plausível. Suposições incorretas para o mecanismo de omissão ou para a distribuição das covariáveis podem gerar inferências enviesadas para a distribuição condicional das respostas dadas as covariáveis. Para fazer face à complexidade analítica adota-se pragmaticamente uma metodologia bayesiana para a análise de um modelo global composto de distribuições paramétricas condicionais para $\mathbf{Y}$ dado $\mathbf{X}$ e de um modelo flexível para $\mathbf{X}$ concretizado numa mistura não paramétrica por um processo Dirichlet (Ishwaran e James [2]), aliado a um mecanismo de omissão que se permite ser não aleatório. Para simplicidade restringir-se-á o modelo semiparamétrico ao caso de uma única covariável contínua sujeita a omissão.

O resto do artigo é desenhado da seguinte forma. Na Secção 2 introduz-se o modelo não paramétrico para uma variável contínua, o qual vai constituir uma componente dos modelos semiparamétricos abordados a seguir, em especial na Secção 3. O modelo semiparamétrico desta secção vai ser comparado com modelos paramétricos alternativos no quadro de um estudo de simulação na Secção 4. Um conjunto de dados reais é analisado na Secção 5 através de um modelo semiparamétrico que se propõe incorporar particularmente alguns juízos apriorísticos sobre o mecanismo de omissão que tornam a sua modelação diferente da referida nas secções anteriores. O artigo termina com uma referência a breves conclusões.

# 2 Modelo não paramétrico para dados completos contínuos

Seja $X_i$, $i = 1, \ldots, n$, uma amostra aleatória de tamanho $n$ de uma função de distribuição $F$. Num quadro paramétrico supõe-se uma forma conhecida para $F$, indexada por um parâmetro dimensionalmente finito especificado a priori mas geralmente desconhecido. Para permitir uma maior flexibilidade na modelação e robustez contra uma incorreta especificação de $F$, consideram-se modelos não paramétricos.

Um modo de evitar a especificação da forma de $F$ é empregar medidas de probabilidade aleatórias (RPM), que são distribuições de probabilidade sobre o espaço de medidas de probabilidade, tal como o chamado processo Dirichlet (DP) simbolizado por $F \sim \mathrm{DP}(\alpha, F_0)$. Este processo significa que, para qualquer partição mensurável do espaço amostral $A_1, \ldots, A_M$, o vetor probabilístico de componentes $F(A_j), j = 1, \ldots, M$ segue uma distribuição Dirichlet com vetor paramétrico $[\alpha F_0(A_1), \ldots, \alpha F_0(A_M)]$, onde $\alpha$ é um parâmetro de precisão e $F_0$ é uma distribuição de referência sobre o espaço amostral. Com esta parametrização, $F_0$ é a esperança *a priori* da distribuição $F$ e à medida que $\alpha$ aumenta há uma maior concentração de $F$ em torno de $F_0$. Todavia, sabe-se que o DP gera (talvez inesperadamente) uma distribuição discreta quase certamente, o que pode não ser apropriado para muitas aplicações.

Um modo de gerar uma RPM compatível com distribuições absolutamente contínuas é supor que $X_i$ segue uma distribuição absolutamente contínua dado um valor de um parâmetro específico $\theta_i$ e que, por sua vez, $\theta_i$, $i = 1, \ldots, n$, constitui uma amostra aleatória de um DP, i.e., $X_i|\theta_i \overset{\text{ind.}}{\sim} F_{\theta_i}$, $i = 1, \ldots, n$, $(\theta_1, \ldots, \theta_n)|G \overset{\text{i.i.d.}}{\sim} G$, $G|(\alpha, G_0) \sim \mathrm{DP}(\alpha, G_0)$. Admitindo que os parâmetros $\{\theta_i\}$ seguem uma distribuição *a priori* do tipo DP centrada em $G_0$, em vez da abordagem comum de supor que eles seguem diretamente uma distribuição paramétrica $G_0$, acrescenta uma flexibilidade desejável ao modelo. O termo mistura por processo Dirichlet (DPM) advém da formulação

hierárquica que implica que a distribuição marginal para $X_i$ é uma mistura, i.e., $f(x_i) = \int f(x_i|\theta_i)dG(\theta_i)$, $G|(\alpha,G_0) \sim \text{DP}(\alpha,G_0)$.
A definição construtiva do DP (Sethuraman [8]) mostra que $G|(\alpha,G_0)$ $\sim \text{DP}(\alpha,G_0)$ pode ser representado por $G(A) = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}(A)$ para qualquer subconjunto mensurável $A$ do espaço de valores de $\{\theta_j\}$, onde $p_1 = V_1$, $p_j = V_j \prod_{k=1}^{j-1}(1 - V_k)$,$j > 1$, $V_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1,\alpha)$, $j = 1,2,\ldots$, $\delta_{\theta_j}(A)$ é a medida de Dirac (i.e., igual a um se $\theta_j \in A$ e a zero, no caso contrário), e $\theta_j \overset{\text{i.i.d.}}{\sim} G_0$, $j = 1,2,\ldots$. Esta construção dos pesos aleatórios $\{p_j\}$ é rotulada de procedimento quebra-vara (*stick-breaking*). Tal representação do DP permite delinear algoritmos eficientes para ajustar modelos DPM reescrevendo a distribuição marginal da mistura como $f(x_i) = \sum_{j=1}^{\infty} p_j f(x_i|\theta_j)$, $i = 1,\ldots,n$.
Na prática, contudo, por razões de simplicidade é comum truncar a mistura a $M$ componentes (veja, e.g., Ishwaran e James [2]), o que equivale a aproximar $\text{DP}(\alpha,G_0)$ por um DP truncado (TDP), denotado por $\text{TDP}(\alpha,G_0,M)$. Neste caso, para obter os pesos $p_1,\ldots,p_M$, geram-se as variáveis $V_j \sim \text{Beta}(1,\alpha)$, $j = 1,\ldots,M-1$, e fixa-se $V_M = 1$. A opção por um TDP permite implementar o modelo em software disponível, como BUGS, JAGS e R; respetivamente, *vide, e.g.*, Lunn *et al.* [4], Plummer [5] e R Core Team [7]).
A escolha de $M$ é a questão-chave da via de recorrer a distribuições *a priori* TDP. Recorrendo a resultados de Antoniak [1] sobre o DP e à distribuição *a posteriori* do número de valores distintos dos $\{\theta_i\}$ e de $\alpha$ pode-se definir valores razoáveis para o ponto $M$ de truncatura (*vide* Poleto, Paulino, Singer e Molenberghs [6]).

# 3 Um modelo semiparamétrico para respostas binárias com uma covariável contínua sujeita a omissão não aleatória

Seja $Y_i$ uma resposta binária sempre observada, $X_i$ uma covariável contínua com valores potencialmente omissos e $R_i$ uma variável in-

dicadora assumindo o valor 1 se $X_i$ é observado e 0, se $X_i$ é omisso, $i = 1,\ldots,n$. Embora o interesse se concentre na distribuição condicional de $Y_i$ dado $X_i$, é necessário considerar um modelo para $X_i$ pois não se deseja desprezar a porção da amostra em que $X_i$ é omisso. Como se admite que o mecanismo gerador de dados omissos pode depender dos próprios valores não observados, necessita-se de modelar $R_i$. Adotando a denominada fatorização em modelo de seleção, considera-se o modelo

$$R_i|(Y_i,X_i,\delta_0,\delta_1,\delta_2,\delta_3) \overset{\text{ind.}}{\sim} \text{Bern}(\theta_i),\ \text{logito}(\theta_i) = \delta_0 + \delta_1 X_i + \delta_2 Y_i + \delta_3 X_i Y_i, \quad (1)$$

$$Y_i|(X_i,\beta_0,\beta_1) \overset{\text{ind.}}{\sim} \text{Bern}(\pi_i),\ \text{logito}(\pi_i) = \beta_0 + \beta_1 X_i, \quad (2)$$

$$X_i|(\mu_i,V) \overset{\text{ind.}}{\sim} N(\mu_i,V), \quad (3)$$

em que $\text{Bern}(\theta_i)$ denota a distribuição Bernoulli com probabilidade de sucesso $\theta_i$, $i = 1,\ldots,n$, juntamente com as distribuições *a priori* mutuamente independentes $\delta_j|(\mu_{\delta_j},\sigma_{\delta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\delta_j},\sigma_{\delta_j})$, $j = 0,1,2,3$, $\beta_j|(\mu_{\beta_j},\sigma_{\beta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\beta_j},\sigma_{\beta_j})$, $j = 0,1$, $(\mu_1,\ldots,\mu_n)|G \overset{\text{i.i.d.}}{\sim} G$, $G|\alpha,G_0,M \sim \text{TDP}(\alpha,G_0,M)$, $V|T \sim \text{Unif}[0,T]$, $\alpha|(\lambda_1,\lambda_2) \sim Ga(\lambda_1,\lambda_2)$, $G_0\ |(\mu_0,\tau) = N(\mu_0,\tau)$, $\mu_0|(a,A) \sim N(a,A)$. Recorde-se que o acrónimo TDP designa um processo Dirichlet truncado em que o ponto de truncatura $M$ foi baseado no argumento avançado por Antoniak [1] e Ishwaran e James [2]. Uma justificativa para o uso da distribuição *a priori* uniforme para a variância $V$, em vez da comummente empregada distribuição gama, é apresentada por Ishwaran e James [2] e Poleto *et al.* [6].

O modelo é considerado semiparamétrico porque emprega uma estrutura dita não paramétrica (na realidade, massivamente paramétrica) para o modelo marginal de $X_i$ e estruturas paramétricas convencionais para as distribuições condicionais de $Y_i$ dado $X_i$ e $R_i$ dado $Y_i$ e $X_i$.

O mecanismo (1) é do tipo omissão não ao acaso (MNAR de *missing not at random*) porque considera que a probabilidade de ocorrerem

covariáveis omissas pode depender dos seus próprios valores não observados. Por outro lado, se se incluir a suposição de omissão ao acaso (MAR de *missing at random*) $\delta_1 = \delta_3 = 0$, o mecanismo torna-se ignorável do ponto de vista de inferências bayesianas para $\beta_0$ e $\beta_1$ devido à suposta independência apriorística entre ($\delta_0$,$\delta_2$) e os outros parâmetros (Little e Rubin [3]). Uma subclasse do modelo MAR é o mecanismo de omissão completamente ao acaso (MCAR de *missing completely at random*) que pode ser formulado fixando $\delta_1 = \delta_2 = \delta_3 = 0$.

Neste cenário com omissão em variáveis explicativas, é importante notar que a chamada análise de casos completos (CCA de *complete case analysis*), na qual se desprezam as unidades com dados omissos, gera usualmente inferências não enviesadas para $\beta_0$ e $\beta_1$, não só sob o mecanismo MCAR mas também sob quaisquer outros mecanismos que não dependem de $Y_i$ tais como na versão reduzida do mecanismo de omissão não ao acaso, $\text{MNAR}_{\text{red}} : \delta_2 = \delta_3 = 0$. A CCA em dados gerados sob o mecanismo $\text{MNAR}_{\text{red}}$ resulta em inferências enviesadas para a distribuição marginal de $X_i$, mas não para a distribuição condicional de $Y_i$ dado $X_i$. Note-se que a CCA não requer a especificação dum modelo marginal para $X_i$ se o interesse jaz apenas na distribuição condicional de $Y_i$ dado $X_i$.

# 4 Alguns resultados de um estudo de simulação

Considera-se as seguintes distribuições para a variável explicativa: $X^N \sim N(12,3^2)$, $X^L \sim$ Log-normal($2.45,0.246^2$), $X^C = 0.8X^{C1} + 0.2X^{C2}$, $X^{C1} \sim$ Unif[8,12], $X^{C2} \sim$ Log-normal($2.79,0.642^2$), em que Log-normal($\mu$,$\sigma^2$) denota uma distribuição log-normal e $\mu$ e $\sigma$ são respetivamente a média e o desvio padrão da variável subjacente na escala logarítmica. A média e o desvio padrão de $X^L$ e $X^C$ coincidem com os correspondentes parâmetros de $X^N$, embora as densidades sejam muito diferentes conforme ilustrado na Figura 1.

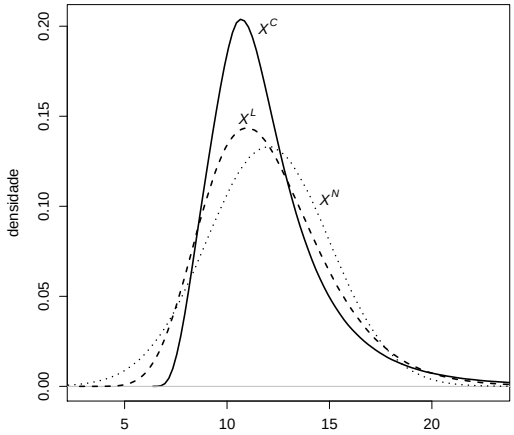



Figura 1: Densidades de distribuições normal ($X^N$), log-normal ($X^L$) e combinação linear ($X^C$) de uma uniforme e uma log-normal, com a mesma média e desvio padrão.

Com o propósito de avaliar o impacte de resultados obtidos sob diferentes suposições distribucionais para a covariável, simulou-se uma amostra de $X$ de tamanho $n = 10000$ de cada uma de três distribuições $X^N$, $X^L$ e $X^C$; de seguida, para cada valor simulado sob cada uma das distribuições da covariável gerou-se $Y$ de (2) com $\beta_0 = 6$ e $\beta_1 = -0.5$; finalmente, gerou-se $R$ de (1) com $\delta_0 = -3$, $\delta_1 = 0.5$ e $\delta_2 = \delta_3 = 0$. Para cada um dos três conjuntos de dados simulados (com $X^N$, $X^L$ e $X^C$), ajustou-se o modelo semiparamétrico da secção anterior bem como os modelos paramétricos normal e log-normal. Para estes últimos o modelo não paramétrico com o 1º nível (3) é substituído por $X_i|\mu_0,\tau \overset{\text{i.i.d.}}{\sim} N(\mu_0,\tau)$ e $X_i|\mu_0,\tau \overset{\text{i.i.d.}}{\sim}$ Log-normal$(\mu_0,\tau)$, $i = 1,\ldots,n$, dotado de distribuições

Tabela 1: Valores esperados (VE), desvios padrões (DP) e intervalos de credibilidade equicaudais (IC) a 95% da distribuição *a posteriori*.

| Análise de Casos Disponíveis | | | | | | | |
|---|---|---|---|---|---|---|---|
| Distr. | Covariável | $\beta_0$ | | | $\beta_1$ | | |
| Gerado | Suposto | VE | DP | IC 95% | VE | DP | IC 95% |
| $X^N$ | Normal | 6.22 | 0.15 | [5.93; 6.51] | −0.515 | 0.012 | [−0.538; −0.491] |
| | Log-norm. | 6.37 | 0.14 | [6.09; 6.66] | −0.525 | 0.012 | [−0.549; −0.502] |
| | Não-Param. | 6.21 | 0.15 | [5.92; 6.51] | −0.514 | 0.012 | [−0.538; −0.491] |
| $X^L$ | Normal | 5.06 | 0.13 | [4.82; 5.32] | −0.428 | 0.011 | [−0.449; −0.407] |
| | Log-norm. | 6.01 | 0.14 | [5.73; 6.29] | −0.501 | 0.012 | [−0.525; −0.478] |
| | Não-Param. | 6.00 | 0.14 | [5.71; 6.28] | −0.500 | 0.012 | [−0.524; −0.477] |
| $X^C$ | Normal | 4.72 | 0.12 | [4.49; 4.96] | −0.395 | 0.010 | [−0.416; −0.375] |
| | Log-norm. | 5.08 | 0.13 | [4.83; 5.34] | −0.425 | 0.011 | [−0.447; −0.404] |
| | Não-Param. | 5.78 | 0.15 | [5.49; 6.08] | −0.481 | 0.013 | [−0.505; −0.456] |
| Análise de Casos Completos | | | | | | | |
| $X^N$ | | 6.22 | 0.15 | [5.93; 6.52] | −0.515 | 0.012 | [−0.539; −0.492] |
| $X^L$ | | 6.02 | 0.14 | [5.74; 6.30] | −0.502 | 0.012 | [−0.525; −0.479] |
| $X^C$ | | 5.83 | 0.15 | [5.54; 6.12] | −0.484 | 0.012 | [−0.509; −0.460] |
| Valores Verdadeiros | | $\beta_0 = 6.00$ | | | $\beta_1 = -0.500$ | | |

*a priori* vagas. Para todos os modelos as distribuições *a priori* vagas para $\delta_j$ e $\beta_j$ envolveram os hiperparâmetros $\mu_{\delta_j} = \mu_{\beta_j} = 0$ e $\sigma_{\delta_j} = \sigma_{\beta_j} = 10^3$, $j$ = 0,1. Analogamente, $M = 10$, $T = s_x^2$ (variância empírica dos valores de $X$), $\tau = 16s_x^2$, $\lambda_1 = \lambda_2 = 2$, $a = 0$ e $A = 10^3$. Além disso, supôs-se sempre a estrutura correta para o mecanismo de omissão, i.e., $\delta_2 = \delta_3 = 0$, de modo que as únicas componentes que variavam no estudo eram a distribuição usada para gerar a covariável e a distribuição admitida para esta na análise.

De acordo com a Tabela 1, as amostras obtidas das distribuições *a posteriori* dos parâmetros $\beta_0$ e $\beta_1$ indicam que o modelo não paramétrico para a covariável gera resultados muito próximos dos obtidos com o correspondente modelo paramétrico verdadeiro sob qualquer das distribuições normal e log-normal. Nesses casos, os intervalos de credibilidade contêm os verdadeiros valores de $\beta_0$ e $\beta_1$; isto não ocorre nas análises sob modelos paramétricos incorretos para $X^N$ e $X^L$. Por outro lado, no caso de $X^C$, só os intervalos de credibilidade

da análise sob o modelo não paramétrico para a covariável continham os verdadeiros valores de $\beta_0$ e $\beta_1$. A CCA produz resultados para os parâmetros do modelo logístico muito próximos dos obtidos com todos os dados disponíveis, o que poderia ser antecipado para este modelo MNAR identificável.

# 5 Análise de dados de embolia pulmonar

Wicki *et al.* [9] analisaram dados de 1090 pacientes que foram consecutivamente admitidos no banco de urgência do Hospital Universitário de Genebra por suspeita de embolia pulmonar, i.e., bloqueio da artéria pulmonar ou de alguma das suas ramificações. O objetivo do seu estudo era desenvolver um sistema de pontuação que indicasse a probabilidade de ocorrência desta doença cardiovascular baseado em testes de diagnóstico e outra informação facilmente obtida. Por simplicidade, considera-se aqui só algumas das variáveis explicativas incluídas no modelo final apresentado por estes autores.

O indicador da presença de embolia pulmonar (variável resposta), bem como quatro variáveis explicativas (idade, embolia pulmonar prévia ou trombose venosa profunda, cirurgia recente e frequência cardíaca), foram observadas para todos os pacientes, enquanto duas variáveis indicando presença de certas caraterísticas (atelectasia laminar e elevação do hemidiafragma) apresentaram valores faltantes para um único paciente que, por esta razão, foi removido do conjunto de dados. Por outro lado, a pressão parcial do dióxido de carbono($PaCO_2$), obtida por gasometria arterial, surgiu omissa para 103 (9%) pacientes.

Análises preliminares permitiram mostrar que os dados observados para $PaCO_2$ parecem ser melhor acomodados pela distribuição preditiva *a posteriori* do modelo não paramétrico do que pelas correspondentes densidades dos modelos normal, log-normal e gama (vide Figura 2 para o modelo gama, que forneceu um melhor ajuste do que o normal e o log-normal). Por outro lado, elas não mostraram evidência de associação entre $PaCO_2$ e as outras variáveis explicativas.

Tendo isto em mente considerou-se um modelo não paramétrico marginal em vez de condicional para $PaCO_2$, do género daquele descrito na Secção 3.

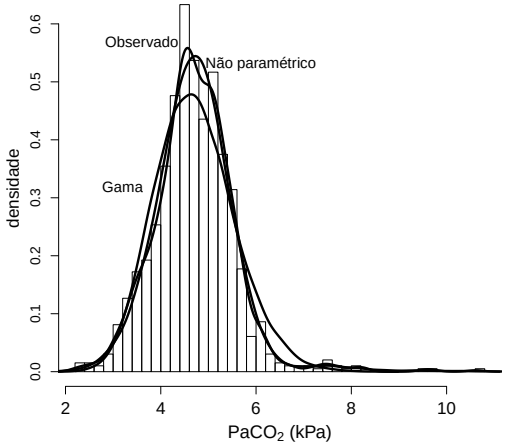



Figura 2: Histograma de dados observados para $X = PaCO_2$ e estimativas da densidade pelo método de núcleo gaussiano baseado em dados e valores amostrados da distribuição preditiva *a posteriori* decorrente do ajuste dos modelos não paramétrico e Gama.

Wicki *et al.* [9] mencionam que $PaCO_2$ estava em falta para alguns pacientes porque a gasometria arterial não foi realizada ou foi executada enquanto os pacientes estavam respirando oxigénio. Perrier (comunicação pessoal) afirma que isso ocorreu para pacientes que estavam muito pouco doentes ou tão doentes que necessitavam da administração de oxigénio. Contudo, não foi registrado em qual dos dois casos os pacientes com dados de $PaCO_2$ faltantes seriam classificados. Os comentários de Perrier sugerem que é razoável supor que

Tabela 2: Valores esperados (VE) e desvios padrões (DP) *a posteriori* e intervalos de credibilidade equicaudais (IC) a 95%.

| Parâ- | Modelo não paramétrico | | | Análise Casos Completos | | |
|---|---|---|---|---|---|---|
| metros | VE | DP | IC 95% | VE | DP | IC 95% |
| $\beta_0$ | −2.476 | 0.642 | [−3.727; −1.192] | −2.585 | 0.672 | [−3.901; −1.273] |
| $\beta_1$ | 1.375 | 0.268 | [ 0.852; 1.903] | 1.512 | 0.293 | [ 0.943; 2.086] |
| $\beta_2$ | 1.080 | 0.177 | [ 0.735; 1.429] | 1.087 | 0.187 | [ 0.724; 1.453] |
| $\beta_3$ | 0.706 | 0.187 | [ 0.339; 1.070] | 0.732 | 0.200 | [ 0.343; 1.126] |
| $\beta_4$ | 0.590 | 0.189 | [ 0.221; 0.962] | 0.591 | 0.201 | [ 0.198; 0.990] |
| $\beta_5$ | 0.268 | 0.046 | [ 0.179; 0.359] | 0.288 | 0.048 | [ 0.195; 0.384] |
| $\beta_6$ | 1.158 | 0.331 | [ 0.508; 1.809] | 1.221 | 0.352 | [ 0.538; 1.914] |
| $\beta_7$ | −0.405 | 0.101 | [−0.609; −0.209] | −0.429 | 0.102 | [−0.631; −0.231] |
| $\delta_0$ | 2.624 | 0.223 | [ 2.200; 3.077] | | | |
| $\delta_1$ | 0.482 | 0.210 | [ 0.082; 0.914] | | | |
| $\delta_2$ | −0.112 | 0.250 | [−0.587; 0.399] | | | |

a probabilidade de se observar $PaCO_2$ ($\theta_i$) pode (i) ser máxima para pacientes com probabilidade de embolia pulmonar ($\pi_i$) próxima da prevalência de embolia pulmonar ($\pi$) e (ii) diminuir, à medida que a probabilidade de embolia pulmonar fica mais distante da prevalência. Tendo isto em vista, propõe-se um modelo de omissão com uma regressão segmentada que permite que $\theta_i$ decaia com velocidade diferente à medida que $\pi_i \longrightarrow 0$ e $\pi_i \longrightarrow 1$ e, assim, espera-se que $\delta_1 > 0$ e $\delta_2 < 0$. Este modelo juntamente com um modelo condicional para a resposta indicando embolia pulmonar dadas as covariáveis são indicados como segue: $R_i|(\delta_0,\delta_1,\delta_2,\mathrm{LN}_i,\mathrm{LP}_i) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\theta_i)$, $\mathrm{logito}(\theta_i) = \delta_0 + \delta_1\mathrm{LN}_i + \delta_2\mathrm{LP}_i$, $Y_i|(\beta_0,\{X_{ji},\beta_j,j = 1,\ldots,7\}) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\pi_i)$, $\mathrm{logito}(\pi_i) = \beta_0 + \sum_{j=1}^{7}\beta_j X_{ji}$, para $i = 1,\ldots,n$, onde $Y_i$ é a variável indicadora de embolia pulmonar, as variáveis explicativas $X_{1i},\ldots,X_{7i}$ são, respetivamente (i) um indicador de recente cirurgia, (ii) um indicador de anterior embolia pulmonar ou de profunda trombose venosa , (iii) um indicador de atelectasia pulmonar em radiografia peitoral (AL-RX), (iv) um indicador de elevação dum hemidiafragma em radiografia peitoral (EH-RX), (v) idade, em décadas, (vi) frequência cardíaca em centenas de batimentos por minuto

(bpm) e (vii) pressão parcial de dióxido de carbono ($PaCO_2$) em kPa, $R_i$ é o indicador de observação de $PaCO_2$ ($X_{7i}$), $LN_i = LC_i$, se $LC_i < 0$, e $LN_i = 0$, caso contrário, $LP_i = LC_i$, se $LC_i > 0$, e $LP_i = 0$, caso contrário e $LC_i = \text{logito}(\pi_i) - \text{logito}(\hat{\pi})$, em que $\hat{\pi}$ é a prevalência de embolia pulmonar no hospital, supostamente conhecida e dada pela proporção de embolia pulmonar na amostra (27%). A componente DPM para $PaCO_2$ do modelo global é idêntica à referida no modelo da Secção 3.

A Tabela 2 exibe alguns resultados *a posteriori* das análises de casos disponíveis mediante o modelo TDP para $X_{7i}$ e de casos completos para propósitos comparativos.

As análises de todos os dados disponíveis baseados no modelo global referido acima acabam por ser mais adequadas do que as análises de casos completos porque, ao incorporarem suposições sobre o modo de ocorrência dos dados omissos, devem proporcionar resultados menos enviesados sobre a associação entre embolia pulmonar e $PaCO_2$, e gerar resultados mais precisos para as outras associações.

# 6 Conclusões

Este artigo centra-se na modelação de respostas binárias quando há uma covariável contínua sujeita a omissão informativa (MNAR). Mostra-se que uma abordagem bayesiana com um modelo semiparamétrico baseado numa mistura por processo Dirichlet para a distribuição marginal daquela covariável é uma alternativa viável para evitar eventuais vieses em inferências de interesse introduzidos pela adoção de uma distribuição paramétrica incorreta.

Algumas extensões podem ser tomadas em consideração como a possibilidade de se ter duas ou mais covariáveis contínuas sujeitas a omissão. Nesse caso, o modelo para o mecanismo de omissão pode assentar num produto de distribuições bernoullianas univariadas e o modelo para as covariáveis pode ser baseado em DPM unidimensionais. Havendo possibilidade de omissão também na variável resposta, tal pode ser manuseado através de modelação simultânea dos

respetivos indicadores de observação/omissão.
Com ou sem as extensões referidas, os maiores desafios na aplicação dos modelos contemplados são provavelmente os casos em que suposições para o mecanismo de omissão originam modelos inidentificáveis e em que o tamanho amostral é grande em oposição a escassa informação *a priori*. Nessas circunstâncias, torna-se lento o processo computacional de gerar as distribuições *a posteriori* de interesse via MCMC, nomeadamente pelo elevado número de componentes do TDP e pela autocorrelação das respectivas cadeias. Nesses casos especiais, algumas das análises como as realizadas por Poleto *et al.* [6] podem levar alguns dias para serem produzidas, até mesmo para modelos paramétricos. Por outro lado, análises de modelos identificáveis como as realizadas neste artigo podem ser produzidas em algumas horas.

# Agradecimentos

Expressa-se com gratidão os apoios financeiros concedidos a este trabalho de investigação: Frederico Z. Poleto e Julio M. Singer, pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil, e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil; Carlos Daniel Paulino, pela Fundação para a Ciência e Tecnologia (FCT) através da unidade CEAUL-FCUL, Portugal e projetos Pest-OE/MAT/UI0006 de 2011 e 2014; Geert Molenberghs, por IAP research network P6/03 do Governo Belga (Belgian Science Policy). Os autores agradecem ao Dr. Arnaud Perrier e ao Dr. Henri Bounameaux do Hospital Universitário de Genebra por fornecerem o conjunto de dados.

# Referências

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2, 1152–1174.

[2] Ishwaran, H., James, L.F. (2002). Approximate Dirichlet process computing finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* 11, 508–532.

[3] Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. John Wiley & Sons, New York.

[4] Lunn, D.J., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: evolution, critique and future directions (with discussion). *Statistics in Medicine* 28, 3049–3082.

[5] Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22.

[6] Poleto, F., Paulino, C.D., Singer, J., Molenberghs, G. (2014). Semiparametric Bayesian analysis of binary responses with a continuous covariate subject to non-random missingness. *Statistical Modeling* (no prelo).

[7] R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

[8] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* 4, 639–650.

[9] Wicki, J., Perneger, T.V., Junod, A.F., Bounameaux, H., Perrier, A. (2001). Assessing clinical probability of pulmonary embolism in the emergency ward. *Archives of Internal Medicine* 161, 92–97.

# Dependência extremal: risco de contágio de valores extremos

Helena Ferreira
Departamento de Matemática, Universidade da Beira Interior, Covilhã, Portugal, *helena.ferreira@ubi.pt*
Marta Ferreira
Departamento de Matemática e Aplicações, Centro de Matemática da Universidade do Minho, Braga, Portugal, *msferreira@math.uminho.pt*

**Palavras–chave**: Teoria de valores extremos multivariada, coeficientes de dependência na cauda de subvetores, inferência

**Resumo**: O fenómeno da globalização, juntamente com um relaxamento na supervisão dos mercados financeiros, tornou-os mais vulneráveis e mais dependentes entre si. A ocorrência de grandes perdas em mercados fortes acaba por se reflectir ao nível das principais bolsas mundiais, e vice-versa. A necessidade de medir esta interdependência extremal conduziu ao aparecimento de diversos coeficientes no seio da teoria multivariada de valores extremos. Neste trabalho apresentam-se coeficientes para a dependência extremal entre dois vetores aleatórios, estendendo medidas existentes na literatura. A estimação será também abordada e uma ilustração do conceito será feita com dados reais.

## 1 Introdução

Devido à globalização e a uma falta de supervisão, vem-se assistindo a um aumento da dependência nos mercados financeiros. Não admira, portanto, que áreas como a gestão de riscos em finanças esteja altamente interessada na quantificação dessa dependência. Por exemplo, a medida de risco sobejamente conhecida por VaR (*Value-at-Risk*), depende fortemente da estrutura de dependência

entre eventos extremos, tornando importante a modelação e a análise da dependência extremal. Na verdade, a teoria de valores extremos multivariada oferece-se como a ferramenta natural para abordar esta questão. A medida mais usada neste contexto é o coeficiente de dependência de cauda, usualmente denotado TDC (*tail dependence coefficient*), definido do seguinte modo ([4]):

$$\lambda := \lim_{t \downarrow 0} P(F_2(X_2) > 1 - t | F_1(X_1) > 1 - t), \tag{1}$$

onde $F_1$ e $F_2$ são funções de distribuição (f.d.'s) das variáveis aleatórias (v.a.'s) $X_1$ e $X_2$, respetivamente, consideradas contínuas. Diz-se que o par $(X_1,X_2)$ apresenta dependência de cauda sempre que $\lambda > 0$ e independência quando $\lambda = 0$. Formulações multivariadas de coeficientes de dependência de cauda podem ser usadas para descrever o grau de dependência na cauda de um ortante de uma distribuição multivariada (e.g. [6], [5] e [7]). Muitas destas medidas, recentemente cada vez mais utilizadas em tempos tão exigentes, consideram que os eventos extremos ocorrem em todas as componentes do vetor. Isto implica uma maior complexidade de trabalho e de compreensão das mesmas, quando em comparação com o caso bivariado. Não surpreendentemente, as aplicações raramente vão além da dimensão três. Relaxando um pouco a dimensão no sentido de considerar a ocorrência de pelo menos um evento extremo em sub-vetores (blocos) de um vetor aleatório, pode ser suficiente para avaliar a dependência de cauda. Por exemplo, até que ponto a queda de um mercado na Europa pode influenciar a ocorrência de pelo menos um *crash* de um mercado norte-americano, e vice-versa? Em Ferreira e Ferreira ([2], 2012) foi considerada uma função de dependência de cauda que permite avaliar a probabilidade de ocorrer simultaneamente algum valor extremo em cada um dos blocos.

Neste trabalho apresentamos uma função para a dependência de cauda de um vetor aleatório, baseada na probabilidade de ocorrerem valores extremos para o máximo de um bloco, dado que o máximo de outro bloco assume um valor extremo também. No vetor unitário, esta função origina o aqui designado BTDC (*block tail dependence*

*coefficient*). O condicionamento nos dois sentidos permite diferenciar o nível de "gravidade", consoante a propagação parte do primeiro ou do segundo blocos. Relações com outros coeficientes conhecidos da literatura serão também exploradas. O estabelecimento de propriedades conduzirá a um estimador simples fortemente consistente. No final, uma aplicação a dados financeiros ilustrará o conceito.

# 2 Dependência extremal em blocos

Seja $\mathbf{X} = (X_1, \ldots, X_d)$ um vetor aleatório com f.d. $F$ e f.d.'s marginais $F_i$ contínuas, $i = 1, \ldots, n$. Para $I \subset \{1, \ldots, d\}$, defina-se $M(I) = \bigvee_{i \in I} F_i(X_i)$ e $\mathbf{X}_I$ o sub-vetor de $\mathbf{X}$ cujas v.a.'s estão indexadas em $I$. Considere-se $C_F$ a função cópula de $F$, i.e.,

$$F(x_1, \ldots, x_d) = C_F(F_1(x_1), \ldots, F_d(x_d)), \;\; (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Vamos estudar a dependência entre eventos extremos respeitantes a dois sub-vetores (blocos), $\mathbf{X}_{I_1}$ e $\mathbf{X}_{I_2}$, onde $I_1$ e $I_2$ são subconjuntos disjuntos de $\{1, \ldots, d\}$.
A partir do TDC, a referência [7] introduz o conceito de função de dependência de cauda, definida por:

$$\Lambda(x,y) := \lim_{t \downarrow 0} P(F_2(X_2) > 1 - x/t | F_1(X_1) > 1 - y/t).$$

De notar que $\Lambda(1,1) = \lambda$. Começamos por estender, na Definição 2.1, a função de dependência de cauda para dois blocos de vetores e, a partir desta, introduzimos um novo coeficiente para a dependência na cauda.

**Definição 2.1** *Sejam $I_1$ e $I_2$ dois subconjuntos não vazios de $\{1, \ldots, d\}$. A função de dependência de cauda de $\mathbf{X}_{I_1}$ dado $\mathbf{X}_{I_2}$ é definida, para $(x,y) \in (0,\infty)^2$,*

$$\Lambda^{(I_1|I_2)}(x,y) := \lim_{t \to \infty} P\Big(M(I_1) > 1 - \frac{x}{t} \Big| M(I_2) > 1 - \frac{y}{t}\Big),$$

*desde que o limite exista.*

Denominaremos o coeficiente $\Lambda^{(I_1|I_2)}(1,1)$ por BTDC (*block tail dependence coefficient*), o qual nos dá informação acerca da probabilidade de ocorrer algum valor extremo no bloco $\{F_i(X_i), i \in I_1\}$, dado que algum valor extremo ocorre no bloco $\{F_i(X_i), i \in I_2\}$. No caso de $I_1 = \{1\}$ e $I_2 = \{2\}$, $\Lambda^{(I_1|I_2)}(1,1)$ corresponde ao TDC dado em (1). As medidas $\Lambda^{(I_1|I_2)}(1,1)$ e $\Lambda^{(I_2|I_1)}(1,1)$ permitem avaliar até que ponto a excedência de um valor de risco no bloco $I_2$ se propaga ao bloco $I_1$, no primeiro caso, e vice-versa, no segundo caso. Assim sendo, diferencia simultaneamente o nível de "gravidade", consoante a propagação parte do primeiro ou do segundo blocos. De notar que o coeficiente $\Lambda^{(I_1,I_2)}(1,1)$ analisado em Ferreira e Ferreira ([2], 2012), corresponde à probabilidade de ocorrer simultaneamente algum valor extremo em cada um dos blocos, não permitindo a referida diferenciação.

No que segue, fixamos a notação a ser usada no resto do artigo: para $(x,y) \in (0,\infty)^2$, $\emptyset \neq I_1,I_2 \subseteq \{1,\ldots,d\}$ e $i \in \{1,\ldots,d\}$, seja

$$\alpha_i^{(I_1,I_2)}(u,v) = u\mathbf{1}_{I_1}(i) + v\mathbf{1}_{I_2}(i) + \mathbf{1}_{\overline{I_1 \cup I_2}}(i)$$

onde $\mathbf{1}(\cdot)$ é a função indicatriz, e para $G$ uma distribuição multivariada de valor extremo (MEV) e $C_G$ a função cópula respetiva, seja

$$l^{(I_1,I_2)}(x^{-1},y^{-1}) = -\log C_G(\alpha_1^{(I_1,I_2)}\big(e^{-x},e^{-y}\big),\ldots,\alpha_d^{(I_1,I_2)}\big(e^{-x},e^{-y}\big)).$$

Usaremos ainda a notação $\wedge = \min$ e $\vee = \max$. O coeficiente extremal é uma medida introduzida em [9] e mais tarde abordada em [8]. O coeficiente extremal de $\mathbf{X}_{I_1 \cup I_2}$, denotado $\epsilon_{I_1 \cup I_2}$, define-se por

$$C_G(\alpha_1^{(I_1,I_2)}\big(e^{-x},e^{-x}\big),\ldots,\alpha_d^{(I_1,I_2)}\big(e^{-x},e^{-x}\big)) = \exp(-x)^{\epsilon_{I_1 \cup I_2}},$$

pelo que, se tem

$$l^{(I_1,I_2)}(x^{-1},x^{-1}) = x\epsilon_{I_1 \cup I_2}.$$

Se $F$ pertencer ao domínio de atração de uma distribuição MEV $G$, então é possível deduzir

$$\Lambda^{(I_1|I_2)}(x,y) = \quad 1 + \frac{-\log(\exp(-x))^{\epsilon_{I_1}}}{-\log(\exp(-y))^{\epsilon_{I_2}}} - \frac{l^{(I_1,I_2)}(x^{-1},y^{-1})}{-\log(\exp(-y))^{\epsilon_{I_2}}}.$$

Este resultado pode ser visto em [2] (Proposição 2.2).

**Exemplo 2.2** *Considerando o modelo simétrico logístico, tem-se*

$$l^{(I_1,I_2)}(x^{-1},y^{-1}) = \Big(\sum_{j=1}^{d}(-\log\alpha_j^{(I_1,I_2)}(\exp(-x),\exp(-y)))^{1/\theta}\Big)^{\theta},$$

*com* $\theta \in (0,1]$, $x,y > 0$. *Logo,*

$$\begin{aligned}\Lambda^{(I_1|I_2)}(x,y) = \quad & 1 + \frac{\left(\sum_{j\in I_1} x^{1/\theta}\right)^{\theta}}{\left(\sum_{j\in I_2} y^{1/\theta}\right)^{\theta}} - \frac{\left(\sum_{j\in I_1} x^{1/\theta} + \sum_{j\in I_2} y^{1/\theta}\right)^{\theta}}{\left(\sum_{j\in I_2} y^{1/\theta}\right)^{\theta}} \\ = \quad & 1 + \frac{|I_1|^{\theta}x}{|I_2|^{\theta}y} - \frac{\left(|I_1|x^{1/\theta}+|I_2|y^{1/\theta}\right)^{\theta}}{|I_2|^{\theta}y}.\end{aligned}$$

**Proposição 2.3** *Se $F$ pertence ao domínio de atração de uma distribuição MEV, então*

$$0 \le \Lambda^{(I_1|I_2)}(x,y) \le 1 \wedge \frac{x\epsilon_{I_1}}{y\epsilon_{I_2}}.$$

*Dem.*: A desigualdade da esquerda é imediata pela definição de $\Lambda^{(I_1|I_2)}(x,y)$. Para a desigualdade da direita, veja-se que

$$\begin{aligned} & F(a_1^{(I_1,I_2)}(x^{-1},y^{-1}),\ldots,a_d^{(I_1,I_2)}(x^{-1},y^{-1})) \\ \le \quad & F(a_1^{(I_1,\emptyset)}(x^{-1},x^{-1}),\ldots,a_d^{(I_1,\emptyset)}(x^{-1},x^{-1})) \\ & \wedge F(a_1^{(\emptyset,I_2)}(y^{-1},y^{-1}),\ldots,a_d^{(\emptyset,I_2)}(y^{-1},y^{-1})),\end{aligned}$$

donde

$$\begin{aligned} & l^{(I_1,\emptyset)}(x^{-1},x^{-1}) + l^{(\emptyset,I_2)}(y^{-1},y^{-1}) - l^{(I_1,I_2)}(x^{-1},y^{-1}) \\ \leq \; & l^{(I_1,\emptyset)}(x^{-1},x^{-1}) + l^{(\emptyset,I_2)}(y^{-1},y^{-1}) \\ & -\big(l^{(I_1,\emptyset)}(x^{-1},x^{-1}) \vee l^{(\emptyset,I_2)}(y^{-1},y^{-1})\big) \\ = \; & x\epsilon_{I_1} \wedge y\epsilon_{I_2}. \end{aligned}$$

■

A proposição seguinte encontra-se em [2] (Proposição 3.1) e vai-nos permitir deduzir estimadores para as funções e coeficientes de dependência de cauda aqui apresentados, no caso de estarmos perante modelos MEV.

**Proposição 2.4** *Se $F$ é uma distribuição MEV com função cópula $C_F$, então, para $l(x_1,\ldots,x_d) = -\log C_F(\exp(-x_1^{-1}),\ldots,\exp(-x_d^{-1}))$, tem-se*

$$l(x_1,\ldots,x_d) = \frac{E(F_1(X_1)^{x_1} \vee \ldots \vee F_d(X_d)^{x_d})}{1 - E(F_1(X_1)^{x_1} \vee \ldots \vee F_d(X_d)^{x_d})}.$$

**Corolário 2.5** *Sob as condições da Proposição 2.4, tem-se*

$$l^{(I_1,I_2)}(x^{-1},y^{-1}) = \frac{E(M(I_1)^{1/x} \vee M(I_2)^{1/y})}{1 - E(M(I_1)^{1/x} \vee M(I_2)^{1/y})},$$

*e*

$$\Lambda^{(I_1|I_2)}(x,y) = 1 + \frac{\frac{E(M(I_1)^{1/x})}{1-E(M(I_1)^{1/x})}}{\frac{E(M(I_2)^{1/y})}{1-E(M(I_2)^{1/y})}} - \frac{\frac{E(M(I_1)^{1/x} \vee M(I_2)^{1/y})}{1-E(M(I_1)^{1/x} \vee M(I_2)^{1/y})}}{\frac{E(M(I_2)^{1/y})}{1-E(M(I_2)^{1/y})}}.$$

Propõe-se assim o estimador

$$\widetilde{\Lambda}^{(I_1|I_2)}(x,y) = 1 + \frac{x\widetilde{\epsilon}_{I_1}}{y\widetilde{\epsilon}_{I_2}} - \frac{\widetilde{l}^{(I_1,I_2)}(x^{-1},y^{-1})}{y\widetilde{\epsilon}_{I_2}}, \tag{2}$$

onde

$$x\widetilde{\epsilon}_{I_1} = \frac{\overline{M(I_1)^{1/x}}}{1-\overline{M(I_1)^{1/x}}}, \quad y\widetilde{\epsilon}_{I_2} = \frac{\overline{M(I_2)^{1/y}}}{1-\overline{M(I_2)^{1/y}}}$$

$$\text{e} \quad \widetilde{l}^{(I_1,I_2)}(x^{-1},y^{-1}) = \frac{\overline{M(I_1)^{1/x} \vee M(I_2)^{1/y}}}{1-\overline{M(I_1)^{1/x} \vee M(I_2)^{1/y}}},$$

tendo-se

$$\overline{M(I_l)^{1/z}} = \frac{1}{n}\sum_{i=1}^{n} \bigvee_{j\in I_l} F_j(X_j^{(i)})^{1/z}, l = 1,2, z \in \mathbb{R}$$

e

$$\overline{M(I_1)^{1/x} \vee M(I_2)^{1/y}} = \frac{1}{n}\sum_{i=1}^{n} \Big( \bigvee_{j\in I_1} F_j(X_j^{(i)})^{1/x} \vee \bigvee_{j\in I_2} F_j(X_j^{(i)})^{1/y} \Big).$$

com $\widetilde{l}^{(I_1,I_2)}(1,1) = \widetilde{\epsilon}_{I_1\cup I_2}$.
Se as margens forem desconhecidas, podemos substituir $F_j$, $j = 1,\ldots,d$, por um estimador $\widetilde{F}_j$. Por questões de maior precisão nas estimativas, é usual considerar variantes da f.d. empírica, como

$$\widehat{F}_j(u) = \frac{1}{n+1}\sum_{k=1}^{n} \mathbf{1}_{\{X_j^{(k)} \le u\}}. \tag{3}$$

Outras sugestões e mais detalhes sobre este tópico podem ser vistos em [1].

No caso de margens conhecidas, a normalidade assintótica de $x\widetilde{\epsilon}_{I_1}$, $y\widetilde{\epsilon}_{I_2}$ e $\widetilde{l}^{(I_1,I_2)}(x^{-1},y^{-1})$, facilmente se deduz do Teorema Limite Central, pois todos se baseiam numa média amostral (ver [2]; Proposição 3.4 e Corolário 3.5). De notar, contudo, que o estimador (2), pelo fato de consistir em quocientes de estimadores assintoticamente normais, não podemos concluir deste modo que também o seja. Este

tópico será objeto de trabalho futuro com base em outra(s) abordagem(ns). Por outro lado, a consistência forte é facilmente deduzida da Lei Forte dos Grandes Números (veja-se Ferreira e Ferreira [2], 2012; Proposição 3.6), resultado este que se estende para o caso de margens desconhecidas (Ferreira e Ferreira [2], 2012; Proposição 3.8). Também é possível estabelecer a normalidade assintótica de $x\widetilde{\epsilon}_{I_1}$, $y\widetilde{\epsilon}_{I_2}$ e $\widetilde{l}^{(I_1,I_2)}(x^{-1},y^{-1})$ para margens desconhecidas, embora o mesmo problema acima descrito subsista para o referido estimador.

## 2.1 Uma aplicação a dados financeiros

Nesta secção abordaremos a questão levantada aquando da Introdução: até que ponto a queda de um mercado na Europa pode influenciar a ocorrência de pelo menos um *crash* de um mercado norte-americano, e vice-versa? A nossa análise baseia-se na aplicação dos coeficientes, $\Lambda^{(I_1|I_2)}(1,1)$ e $\Lambda^{(I_2|I_1)}(1,1)$, aos log-retornos negativos dos valores de fecho dos índices CAC 40, FTSE100, SMI e XDAX ao nível da Europa, e dos índices Dow Jones, Nasdaq e SP500 nos EUA, de janeiro de 1993 a Março de 2004. De maneira a construirmos uma amostra de máximos e, assim, assemelhá-los a um modelo MEV, vamos considerar os máximos mensais, obtendo-se uma amostra de dimensão 84. A nuvem de pontos na Figura 1 parece evidenciar a presença de alguma dependência. De modo a quantificar essa dependência, usaremos os estimadores $\widetilde{\Lambda}^{(I_1|I_2)}(1,1)$ e $\widetilde{\Lambda}^{(I_2|I_1)}(1,1)$, facilmente deduzidos de (2), considerando que as margens são desconhecidas, ou seja, estimando as distribuições marginais com base em (3). Na Tabela 1 encontram-se os resultados obtidos, considerando os blocos Europa e EUA, denotados, respetivamente, $I_1$ e $I_2$. A dependência extremal entre os mercados dos blocos considerados é dada por cada um dos coeficientes $\widetilde{\epsilon}_{I_1}$, $\widetilde{\epsilon}_{I_2}$ e $\widetilde{\epsilon}_{I_1\cup I_2}$. Veja-se que $\widetilde{\Lambda}^{(I_1|I_2)}(1,1) > \widetilde{\Lambda}^{(I_2|I_1)}(1,1)$ o que indica que, para os mercados considerados, o efeito de propagação do fenómeno de queda de mercados bolsistas tende a ser mais grave quando tem início nos EUA.

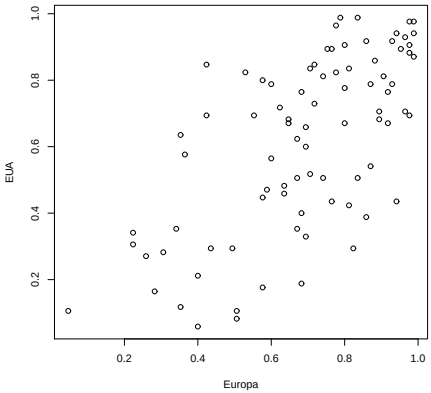



Figura 1: Nuvem de pontos relativa aos máximos mensais na Europa versus EUA.

| | |
|---|---|
| $\widetilde{\epsilon}_{I_1}$ | 2.24398 |
| $\widetilde{\epsilon}_{I_2}$ | 1.59071 |
| $\widetilde{\epsilon}_{I_1 \cup I_2}$ | 2.82637 |
| $\widetilde{\Lambda}^{(I_1 \mid I_2)}(1,1)$ | 0.63388 |
| $\widetilde{\Lambda}^{(I_2 \mid I_1)}(1,1)$ | 0.44935 |

Tabela 1: Estimativas obtidas para os blocos, $I_1$ (Europa) e $I_2$ (EUA).

# Agradecimentos

Marta Ferreira é financiada por Fundos FEDER através do Programa Operacional Factores de Competitividade - COMPETE e por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projecto PEst-OE/MAT/UI0013/-2014. Helena Ferreira é parcialmente financiada pelo "Centro de Matemática" da

Universidade da Beira Interior e pelo projeto PEst-OE/MAT/UI0212/2014 através da Fundação para a Ciência e a Tecnologia (FCT) co-financiado por FEDER/COMPETE.

# Referências

[1] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. (2004). *Statistics of Extremes: Theory and Application*. Wiley, New York.

[2] Ferreira, H., Ferreira, M. (2012). On extremal dependence of block vectors. *Kybernetika* 48, 988–1006.

[3] Frahm, G., Junker, M., Schmidt R. (2005). Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: Mathematics & Economics* 37, 80–100.

[4] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

[5] Li, H. (2009). Orthant tail dependence of multivariate extreme value distributions. *Journal of Multivariate Analysis* 100, 243–256.

[6] Li, H., Sun, Y. (2009). Tail dependence for heavy-tailed scale mixtures of multivariate distributions. *Journal of Applied Probability* 46, 925–937.

[7] Schmidt, R., Stadtmüller, U. (2006). Nonparametric estimation of tail dependence. *Scandinavian Journal of Statistics* 33, 307–335.

[8] Smith, R.L. (1990). Max-stable processes and spatial extremes. Preprint, Univ. North Carolina, USA.

[9] Tiago de Oliveira, J. (1962/63). Structure theory of bivariate extremes: extensions. *Estudos de Matemática, Estatistica e Economicos* 7, 165–195.

# Estimação do índice de valores extremos em ambiente R - as abordagens paramétrica e semi-paramétrica

Helena Penalva
Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal e CEAUL, *helena.penalva@esce.ips.pt*

Sandra Nunes
Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal, CEAUL e CMA/FCT/UNL, *sandra.nunes@esce.ips.pt*

Manuela Neves
Instituto Superior de Agronomia e CEAUL, Universidade de Lisboa, *manela@isa.ulisboa.pt*

**Palavras–chave**: Estimação paramétrica e semi-paramétrica, índice de valores extremos, *software* R, teoria de valores extremos

**Resumo**: Este trabalho tem como objectivos principais: ilustrar a utilização do *software* R numa análise de dados de valores extremos e estimação do índice de valores extremos, $\xi$, utilizando as abordagens paramétrica e semi-paramétrica; fazer uma discussão dos resultados obtidos por diferentes procedimentos em cada uma daquelas abordagens. A ilustração foi feita num conjunto de dados de níveis médios diários de caudais de um rio, registados durante 50 anos.

## 1 Introdução

A Teoria de Valores Extremos (EVT, do inglês "*Extreme Value Theory*") estuda acontecimentos que poderão ser mais extremos do que aqueles que alguma vez já foram observados. Tem vindo a afirmar-se como uma das áreas da Estatística de grande relevo em várias ciências aplicadas onde se observa a ocorrência de valores extremos que é necessário modelar. Nesta modelação é necessário estimar pa-

râmetros de entre os quais tem primordial relevância o parâmetro de forma, $\xi$, que descreve o comportamento da cauda direita, $1-F$, do modelo subjacente aos dados. A sua estimação precisa é muito importante e de enorme influência na estimação de outros parâmetros de valores extremos, tais como *quantis elevados* ou *período de retorno de quantis elevados*. Pretende-se utilizar e construir, quando não exista, procedimentos em R para estimar $\xi$.

# 2 Resultados preliminares

Os estudos em teoria assintótica de valores extremos tiveram início com Fréchet [8], mas ficaram definitivamente estabelecidos com Gnedenko [9] ao formular que a distribuição limite do máximo de variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (i.i.d.), convenientemente normalizado, se existir, é do tipo,

$$EV_\xi(x) = \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0, \; se \; \xi \neq 0 \; ; \\ \exp[-\exp(-x)], & x \in \mathbb{R}, \; se \; \xi = 0. \end{cases} \quad (1)$$

O parâmetro de forma, $\xi$, é usualmente designado por *índice de valores extremos* e a sua estimação é da maior importância em EVT. A função de distribuição (f.d.) $EV_\xi$ incorpora as três leis limite: a lei de Gumbel: $\Lambda(x) = \exp(-\exp(-x)) = EV_0(x)$, $x \in \mathbb{R}$, $\xi = 0$, com cauda direita do tipo exponencial; a lei de Fréchet: $\Phi_\xi(x) = \exp(-x^{-1/\xi}) = EV_\xi(\frac{x-1}{\xi})$, $x > 0$, $\xi > 0$ com cauda direita pesada, do tipo exponencial negativo; a lei de Weibull: $\Psi_\xi(x) = \exp(-(-x)^{1/\xi}) = EV_\xi(\frac{-x-1}{\xi})$, $x < 0$, $\xi < 0$ com cauda direita curta. A f.d. $EV_\xi$ pode também incluir parâmetros de localização, $\lambda \in \mathbb{R}$, e de escala, $\delta > 0$, denotando-s por $EV_\xi(x;\lambda,\delta) = EV_\xi((x-\lambda)/\delta)$. Em vez se pensar no máximo de uma amostra, Balkema and de Haan [1] e Pickands [19] propuseram considerar todos os valores que excedem um nível elevado $u$. Mostraram que, nas condições de existência de (1), a f.d. condicional dos excessos do nível $u$, quando $u \to \infty$, i.e. de $Y = (X-u)|X > u$ é bem aproximada pela Generalizada de Pareto (GP), cuja f.d. é definida como,

$H_\xi(y) = 1 - (1 + \xi y/\widetilde{\delta})^{-1/\xi}$, $\{y : y > 0 \ e \ (1 + \xi y/\widetilde{\delta}) > 0\}$, onde o parâmetro de forma, $\xi$, é o mesmo em ambas as distribuições, mas relativamente ao parâmetro de escala tem-se $\widetilde{\delta} = \delta + \xi(u - \lambda)$.

# 3 Estimação

A estimação em EVT começou por ser realizada numa abordagem paramétrica, baseada na propriedade de max-estabilidade. Nesta abordagem, por limitações de espaço, faremos apenas referência à metodologia de Gumbel conhecida como *Modelo dos Máximos Anuais* (MMA) e à metodologia dos excessos acima de um limiar, conhecida como POT (do inglês *'Peaks over Threshold'*). Em cada uma das metodologias a estimação dos parâmetros foi feita usando o método de máxima verosimilhança (ML) do inglês *'Maximum Likelihood'* e o método dos momentos ponderados de probabilidade (PWM) do inglês *'Probability Weighted Moments'*.

Na década de 70 os procedimentos de estimação começaram a ser efectuados considerando uma abordagem semi-paramétrica, na qual não se adopta um modelo limite, mas se admite apenas que a função $F$ está no domínio de atracção da $EV_\xi$, para um valor $\xi$ adequado. A estimação de $\xi$ é, nestes procedimentos, baseada nas $k$ maiores estatísticas ordinais, em que se admite $k \to \infty$ e $k/n \to 0$ quando $n \to \infty$. Nesta abordagem vários estimadores têm sido propostos. Para ilustrar a utilização do *software* R vamos referir apenas os estimadores clássicos de Hill [15], de Pickands [19] e dos Momentos, Dekkers *et al.* [6], e ainda um estimador mais recente, de viés reduzido e com variância mínima, Caeiro *et al.* [3].

Sendo $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ as estatísticas ordinais associadas à amostra $(X_1, X_2, \ldots, X_n)$, estes estimadores são assim definidos:

$$\text{Hill } (\xi > 0) \qquad \widehat{\xi}^{H}_{k,n} := \frac{1}{k} \sum_{i=1}^{k} \ln(X_{n-i+1:n}) - \ln(X_{n-k:n}) \tag{2}$$

$$\text{Pickands } (\xi \in \mathbb{R}) \quad \widehat{\xi}^{P}_{k,n} := \frac{1}{\ln 2} \ln\left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}}\right) \tag{3}$$

$$\text{Momentos } (\xi \in \mathbb{R}) \qquad \widehat{\xi}^{M}_{k,n} := M^{(1)}_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{(M^{(1)}_{k,n})^2}{M^{(2)}_{k,n}}\right)^{-1} \quad (4)$$

onde $M^{(r)}_{k,n} := \dfrac{1}{k}\sum_{i=1}^{k}[\ln(X_{n-i+1:n}) - \ln(X_{n-k:n})]^r$.

A classe de estimadores de viés reduzido e com variância mínima, MVRB (do inglês *minimum-variance reduced-bias*), introduzida por Caeiro, *et al.* [3] revelou um desempenho superior ao dos estimadores clássicos, em contexto de caudas pesadas ($\xi > 0$). Esta classe de estimadores tem a forma funcional

$$\overline{H}(k) \equiv \overline{H}_{\hat{\beta},\hat{\rho}}(k) := H(k)\big(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})\big), \quad (5)$$

com $H(k) \equiv \widehat{\xi}^{H}_{k,n}$ o estimador de Hill, e $\hat{\beta}$ e $\hat{\rho}$ estimadores consistentes dos parâmetros de escala e forma de segunda ordem, $\beta$ e $\rho$, ver Caeiro *et al.* [3] e Gomes *et al.* [14], para mais detalhes. Sobre estimação de viés reduzido podem referir-se os trabalhos de Gomes *et al.* [12] e Beirlant *et al.* [2], entre outros.

Para a estimação de $\rho$ consideraremos um membro particular da classe dos estimadores de $\rho$ introduzidos em Fraga Alves, *et al.* [7]. Esta classe, parametrizada num parâmetro de controlo $\tau \in \mathbb{R}$ que aqui tomaremos como $\tau = 0$, ver Gomes *et al.* [14], é definida como: $\hat{\rho}(k) \equiv \hat{\rho}_0(k) := \min\left(0, \frac{3(T^{(0)}_n(k)-1)}{T^{(0)}_n(k)-3}\right)$, sendo, $T^{(0)}_n(k)$ assim definido, $T^{(0)}_n(k) := \left[\ln\left(M^{(1)}_{k,n}\right) - \frac{1}{2}\ln\left(M^{(2)}_{k,n}/2\right)\right] / \left[\frac{1}{2}\ln\left(M^{(2)}_{k,n}/2\right) - \frac{1}{3}\ln\left(M^{(3)}_{k,n}/6\right)\right]$, com $M^{(j)}_{k,n}(k)$, $j = 1, 2, 3$, definido acima. Para a estimação do parâmetro de escala de segunda-ordem, $\beta$, vamos considerar $\hat{\beta}_{\hat{\rho}}(k) := \left(\frac{k}{n}\right)^{\hat{\rho}} [d_{\hat{\rho}}(k)\ D_0(k) - D_{\hat{\rho}}(k)] \,/\, [d_{\hat{\rho}}(k)\ D_{\hat{\rho}}(k) - D_{2\hat{\rho}}(k)]$, com $\hat{\rho} = \hat{\rho}_0(k)$, e para $\alpha \leq 0$, $d_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\alpha}$ e $D_\alpha(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\alpha} U_i$, com $U_i := i\big[\ln(X_{n-i+1:n}/X_{n-i:n})\big]$, $1 \leq i \leq k$.

De modo a não termos um aumento de variância, no estimator $\overline{H}(k)$ devem considerar-se os estimadores $\hat{\rho}_0(k)$ e $\hat{\beta}_{\hat{\rho}}(k)$, calculados em $k = k_1$, com $k_1 = \lfloor n^{1-\epsilon} \rfloor$, $\epsilon = 0.001$, ver Gomes e Martins [11], Gomes *et al.* [12] e Caeiro *et al.* [5], para mais detalhes. Estimadores

alternativos para $\beta$ podem ser vistos em Caeiro e Gomes [4] e Gomes *et al.* [13].

# 4 Aplicação do *software* R na modelação de dados e na estimação de $\xi$

O software R, linguagem de código aberto para computação estatística e tratamento de dados, possui vários *packages* estatísticos e permite a implementação pelos utilizadores de *packages* adicionais. Para a modelação e estimação em EVT, quer na abordagem paramétrica quer na semi-paramétrica, podemos referir alguns *packages* que o *software* R possui: `evd; ismev; evir; fExtremes; POT; evdbayes; copula; SpatialExtremes.`

Na ilustração que iremos apresentar serão considerados dados de níveis médios diários do caudal ($m^3/s$) do rio Paiva, medidos na estação hidrométrica de Fragas da Torre, entre 1946/47 a 1995/96. Os 50 anos observados correspondem exatamente ao período entre 1 de Outubro de 1946 e 30 de Setembro de 1996. O estudo do fluxo deste rio é extremamente importante uma vez que é uma das alternativas ao rio Douro como fonte de abastecimento de água para a região sul do Porto.

Gomes [10] utilizou os dados relativos aos meses entre Novembro e Fevereiro. Contudo, como o interesse é analisar os valores extremos, verificámos que na maioria dos anos, os meses de Março e Abril apresentaram valores muito elevados de caudal, pelo que decidimos considerar aqueles meses na amostra a estudar. Ficamos então com uma amostra de 6 meses (em cada ano) de dados diários durante 50 anos, num total de 9050 dados. O teste *Augmented Dickey-Fuller* foi aplicado à amostra, tendo-se obtido um *p-value* inferior a 0.01 pelo que iremos admitir a estacionaridade dos dados, hipótese esta também admitida em Gomes [10].

A Figura 1 apresenta o cronograma dos dados em estudo (direita) e o histograma associado (esquerda).

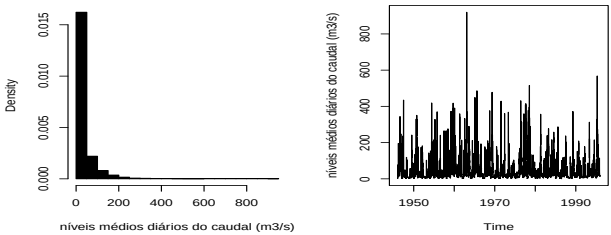


Figura 1: Histograma (esquerda) e cronograma (direita).

As principais medidas descritivas, das quais referimos a Skewness= 4.12 e Kurtosis= 27.60, obtidas com recurso ao *package* `fBasics`, conjuntamente com o histograma, indicam uma cauda muito mais pesada que a do modelo de Gauss.

**Abordagem paramétrica**

Para utilizarmos a metodologia MMA, é necessário obter os máximos de cada bloco, o que facilmente se consegue com *package* `evir` e o comando `gev(dados, block=181)`, onde *block=181*, indica o tamanho de cada bloco. O gráfico da autocorrelação parcial, Figura 2, mostra que os máximos são observações fracamente correlacionadas. Portanto faz sentido ajustar uma distribuição $EV$ aos máximos anuais. Obtivemos as estimativas ML para a distribuição $EV_\xi$, os intervalos de confiança de Wald e os intervalos determinados com recurso à log-verosimilhança de perfil para todos os parâmetros, recorrendo ao *package* `evd` com os seguintes comandos no R:

```
fgev(gev(dados, block=181)$data);
confint(fgev(gev(dados, block=181)$data)),level=0.95);
confint(profile(fvev(gev(dados, block=181)$data)),
         level=0.95).
```

O método PWM também foi aplicado para determinar as estimativas

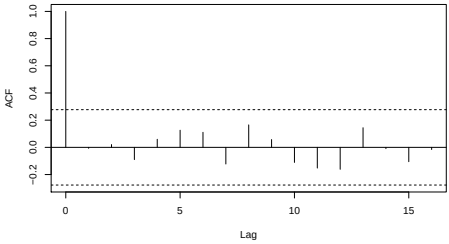


Figura 2: Função de autocorrelação parcial.

dos parâmetros. Neste caso utilizou-se o *package* `fExtremes` e o comando

```
gevFit(dados, block=181, type="pwm").
```

A aplicação da metodologia POT baseia-se na modelação da distribuição de excessos acima de um nível elevado, na qual se restringe a nossa atenção às observações que excedem esse nível. O desafio desta análise é exatamente a escolha deste limiar. Esta escolha pode ser feita a partir do gráfico da vida residual média, e outros dois gráficos, um referente ao parâmetro de escala e outro ao parâmetro de forma, onde se ajusta a distribuição $GP$ a um conjunto de limiares.

Os gráficos foram construídos recorrendo ao *package* `POT` com os comandos : `mrlplot(dados)` e `tcplot(dados, c(100,500))`.

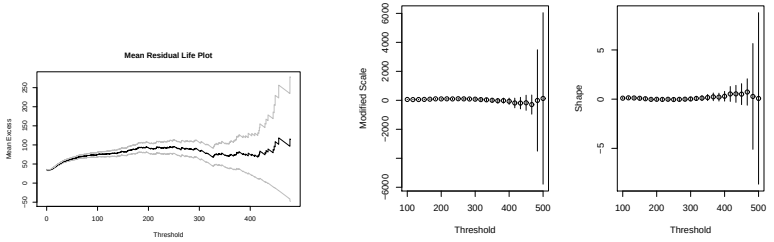



Figura 3: Vida residual média (esquerda) e ajustamento da distribuição $GP$ a um conjunto de limiares (centro e direita)

A Figura 3. sugere que se considere um limiar perto de 300. As estimativas ML e as estimativas PWM para os parâmetros da distribuição $GP$ podem ser obtidas utilizando o *package* `POT` e executando os seguintes comandos:

```
fitgpd(dados, 300, est="mle")
fitgpd(dados, 300, est="pwmb")
```

Os valores das estimativas, e respectivos desvios padrão (d.p.), fornecidos pela aplicação dos métodos de estimação referentes à abordagem paramétrica encontram-se resumidos na Tabela 1.

Tabela 1. Estimativas dos parâmetros e intervalo de confiança

| Métodos Estimação | $\hat{\xi}$ | $\hat{\lambda}$ | $\hat{\delta}$ |
|---|---|---|---|
| MMA - ML (d.p.) | -0.02 (0.09) | 227.35 (19.34) | 122.73 (13.87) |
| I.C. de perfil | $(-0.16, 0.21)$ | (190.53, 267.22) | (99.55, 155.46) |
| MMA - PWM | -0.08 | 231.25 | 129.78 |
| | $\hat{\xi}$ | | $\hat{\tilde{\delta}}$ |
| POT - ML (d.p.) | 0.02 (0.09) | | 83.24 (12.98) |
| POT - PWMB (d.p.) | -0.19 (0.16) | | 100.98 (19.39) |

Como o intervalo de confiança para $\xi$ inclui o zero, não se rejeita a hipótese nula, $\xi = 0$. Assim a distribuição Gumbel é uma possível

candidata para modelar os máximos dos dados referentes aos níveis médios diários do rio.

### Abordagem semi-paramétrica

Na estimação semi-paramétrica são consideradas as $k$ maiores estatísticas ordinais associadas à totalidade das observações, não se definindo nenhum modelo paramétrico, como dissémos. Neste trabalho foram utilizados o estimador de Hill [15], o estimador de Pickands [19], o estimador dos momentos Dekkers *et al.* [6] e o estimador de viés reduzido com variância mínima (MVRB), Caeiro *et al.* [3]. Estes estimadores foram programados em linguagem R. Por questões de espaço não é possível incluir neste trabalho os comandos do R contruídos.
A Figura 4. apresenta as trajectórias das estimativas provenientes daqueles estimadores para $1 \leq k \leq 5000$. Verifica-se a dificuldade, já mencionada atrás, da escolha do nível $k$. Este assunto, apesar de já abordado por vários autores dos quais citamos Neves *et al.* [16] continua a ser assunto de investigação.

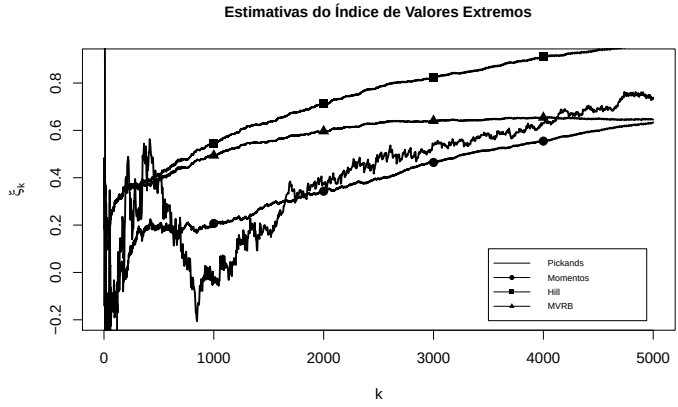



Figura 4: Estimativas de $\xi$, com os estimadores de Hill, Pickands, Momentos e MVRB.

**Notas:** Chama-se a atenção para maior estabilidade das estimativas MVRB, como é conhecido da literatura. É no entanto de salientar que, mesmo sem efectuatrmos a escolha de $k$ correspondente à zona de maior estabilidade, parece haver concordância no valor da estimativa apontada pelos métodos dos momentos e MVRB (e acompanhada pelas estimativas de Pickands, apesar da volatilidade deste estimador). No entanto parece ser apontada como estimativa para $\xi$ um valor bem superior ao que as abordagens paramétricas obtiveram.

# 5 Comentários e trabalho futuro

Neste trabalho procurou-se continuar a explorar os procedimentos de modelação e inferência de dados de valores extremos com o R, iniciados em trabalhos anteriores, Penalva *et al.* [17] e Penalva *et al.* [18]. Alguns estimadores foram por nós programados em R, havendo outros em fase de programação. As estimativas de $\xi$ apresentaram valores bastante diferentes consoante as abordagens, o que motiva trabalho futuro de pesquisa destas disparidades.

# Agradecimentos

Investigação parcialmente suportada por fundos nacionais através da FCT-Fundação para a Ciência e a Tecnologia, projectos PEst-OE/MAT/UI006/2011,2014 (CEAUL) e PEst-OE/MAT/UI0297/2011,
2014(CMA/FCT/UNL).

# Referências

[1] Balkema, A.A., de Haan, L. (1974). Residual life time at great age. *Annals of Probability* 2, 792–804.

[2] Beirlant, J., Caeiro, F., Gomes, M.I. (2012). An overview and open research topics in the field of statistics of univariate extremes. *REVSTAT - Statistical Journal* 10, 1–31.

[3] Caeiro, F., Gomes, M.I., Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator. *REVSTAT - Statistical Journal* 3, 111–136.

[4] Caeiro, F., Gomes, M.I. (2006). A new class of estimators of a "scale" second order parameter. *Extremes* 9, 193–211.

[5] Caeiro, F., Gomes, M.I., Henriques-Rodrigues, L. (2009). Reduced-bias tail index estimators under a third order framework. *Communications in Statistics - Theory and Methods* 38, 1019–1040.

[6] Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution.*Annals of Statistics* 17, 1833–1855.

[7] Fraga Alves, M.I., Gomes M.I., de Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica* 60, 194–213.

[8] Frechet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique (Cracovie)* 6, 93–116.

[9] Gnedenko, B.V. (1943). Sur la distribution limite d'une série aléatoire. *Annals of Mathematics* 44, 423–453.

[10] Gomes, M.I. (1993). on the estimation parameters of rare events in environmental times series. *Statistics for the Environment*, 226–241.

[11] Gomes, M.I., Martins, M.J. (2002). "Asymptotically unbiased"estimators of the tail index based on external estimation of the second order parameter. *Extremes* 5, 5–31.

[12] Gomes, M.I., de Haan, L., Henriques-Rodrigues, L. (2008). Tail Index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society: Series B* 70, 31–52.

[13] Gomes, M.I., Henriques-Rodrigues, L., Pereira, H., Pestana, D. (2010). Tail index and second order parameters' semi-parametric estimation based on the log-excesses. *Journal of Statistical Computation and Simulation* 80, 653–666.

[14] Gomes, M.I., Martins, M.J. and Neves, M.M. (2013) Generalized Jackknife-based estimators for univariate extreme-value modeling. *Communications in Statistics - Theory and Methods* 42, 1227–1245.

[15] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3, 1163–1174.

[16] Neves, M.M., Gomes, M.I., Figueiredo, F., Prata Gomes, D. (2015). Modeling extreme events: sample fraction adaptive choice in parameter estimation. *Journal of Statistical Theory and Practice* 9, 184–199.

[17] Penalva, H., Neves, M.M., Nunes, S. (2013). Topics in Data Analysis Using R in Extreme value Theory. *Advances in Methodology and Statistics / Metodoloski zvezki* 10, 17–29.

[18] Penalva, H., Nunes, S., Neves, M.M. Statistical Modeling and Inference in Extremes: Applications with R. *Biometrie und Medizinische Informatik Greifswalder Seminarberichte, Statistical and Biometrical Challenges, Shaker Verlag*, 281–309.

[19] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics* 3, 119–131.

# Diagnóstico em regressão binária

Isabel Natário
CEAUL, Dep. Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516, Caparica, *icn@fct.unl.pt*
Sílvia Shrubsall
CESUR, Instituto Superior Técnico - Universidade de Lisboa, Lisboa, *silviashrubsall@gmail.com*

**Palavras–chave**: Regressão binária, função de ligação, diagnóstico

**Resumo**: O modelo de regressão binária pressupõe uma variável aleatória binária de interesse, cujo valor esperado, coincidente com a probabilidade de sucesso, é modelado através de uma função de covariáveis que influenciam essa resposta, função dita de ligação. O diagnóstico do modelo empregue é essencial, mas frequentemente omitido ou apenas aflorado. Neste trabalho revêm-se e sistematizam-se algumas técnicas disponíveis para abordar esta questão, indicando-se algumas bibliotecas existentes do programa R-project para o efeito. Ilustram-se estas metodologias através de um exemplo com dados reais de acidentes rodoviários com vítimas na cidade de Lisboa.

## 1 Introdução

As aplicações dos modelos de regressão binária são variadas, ocupando um lugar de destaque em estudos de biologia e medicina [12]. Estes modelos pressupõem que há uma variável aleatória binária de interesse, $Y$, cujo valor esperado, coincidente com sua probabilidade de sucesso $p = P(Y = 1)$, é modelado através de uma função de covariáveis $x_1, \ldots x_k$ que possivelmente influenciam essa resposta - combinadas num preditor, $\eta$. Usualmente considera-se que este preditor é linear nas covariáveis, *i.e.*, $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. A estimação é feita por máxima verosimilhança [17].
A função que relaciona $\mathrm{E}[Y|x_1, \ldots, x_k] = p$ com o preditor designa-se

por função de ligação do modelo,

$$g(\mathrm{E}[Y|x_1,\ldots,x_p]) = g(p) = \eta. \quad (1)$$

Esta função não é única, havendo motivos relacionados com a observação na prática da forma dos preditores, por exemplo, ou motivos de simplicidade que presidem à sua escolha. Diferentes preditores resultam em diferentes modelos, pelo que se coloca a questão de qual será o mais adequado. Efetivamente verifica-se que a má especificação da função de ligação pode conduzir a enviesamentos consideráveis na estimação dos parâmetros de regressão e nas estimativas da probabilidade de sucesso da resposta [4].
Adicionalmente, a análise de diagnóstico do modelo escolhido deve também questionar se os outros pressupostos do modelo são verificados pelos dados - avaliando a qualidade do ajuste. Para tal deve-se levar a cabo uma análise cuidada dos resíduos do modelo ajustado, das observações ditas influentes e discrepantes, sem esquecer que se trabalha num contexto de regressão binária e não de outro tipo de regressão, como por exemplo a regressão linear simples, em que a variável resposta se assume normalmente distribuída.
Este trabalho tem por objectivo rever e sistematizar algumas técnicas disponíveis para abordar esta questão, indicando referências a bibliotecas (pacotes) do programa R-project disponíveis para tal. Na secção 2 detalham-se os pontos relacionados com a função de ligação e na secção 3 com os outros aspetos da modelação. Estes pontos são ilustrados num exemplo com dados reais de acidentes rodoviários com vítimas em Lisboa, em que se procuram fatores de risco para a gravidade de um acidente, secção 4. Na secção 5 conclui-se.

# 2 Função de ligação

O valor esperado de uma variável resposta binária $Y$, com distribuição Bernoulli($p$), é $p$, um valor em ]0,1[. Igualar tal quantidade a um preditor linear, que à partida pode tomar qualquer valor em $\mathbb{R}$, não é adequado. Adicionalmente, os exemplos práticos revelam que a variação desta probabilidade com os valores das covariáveis tende

a apresentar um gráfico semelhante ao de uma função distribuição, sugerindo uma para a função de ligação (1).
A interpretação do modelo binário na chamada formulação latente ajuda a entender melhor esta escolha. Pense-se em $Y$ como a variável resposta, que é observada. Esta relaciona-se com uma outra variável de valores reais, $Y^*$, chamada a variável resposta latente, de acordo com o seu posicionamento com um certo valor de corte $\theta$: $Y = 0$ se $Y^* < \theta$ e $Y = 1$ se $Y^* > \theta$. A interpretação deste par ($Y$,$Y^*$) depende do contexto do problema, por exemplo pode representar a resposta binária a um certo tratamento ($Y$) e a dose de um certo medicamento ($Y^*$).
Temos então que $p = P(Y = 1) = P(Y^* > \theta)$. De forma a identificar o modelo fixa-se $\theta = 0$ e padroniza-se $Y^*$ para ter um desvio padrão unitário. Pode-se então expressar a dependência da variável resposta nas covariáveis através de um modelo linear para a variável latente, $Y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + U = \eta + \text{erro}$, onde se assume que $U$ se distribui de acordo com uma função distribuição $F(u)$ (não necessariamente normal). Desta forma,

$$p = P(Y = 1) = P(Y^* > 0) = P(U > -\eta) = 1 - F(-\eta).$$

Daqui se conclui que $\eta = -F^{-1}(1-p) = g(p)$, no caso de a distribuição de $U$ não ser simétrica em torno de 0, e que $\eta = F^{-1}(p) = g(p)$, no caso de a distribuição de $U$ ser simétrica em torno de 0, já que $F(u) = 1 - F(-u)$. Em ambos os casos a função de ligação é apenas uma função dos quantis da distribuição do erro, permitindo várias alternativas paramétricas para a função de ligação. A mais utilizada é a baseada na função de distribuição logística, $F(u) = \frac{e^u}{1+e^u}$ (modelo de regressão logística),

$$p = \frac{e^\eta}{1+e^\eta} \quad \Leftrightarrow \quad \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \eta \quad \Leftrightarrow \quad g(p) = \eta.$$

A segunda mais usada a função distribuição de uma variável aleatória normal padrão, $\Phi(u)$ (modelo de regressão probit):

$$p = \Phi(\eta) \quad \Leftrightarrow \quad \text{probit}(p) = \Phi^{-1}(p) = \eta \quad \Leftrightarrow \quad g(p) = \eta.$$

Outra função de ligação comum é a chamada complementar log-log, baseada na função distribuição uma variável log-Weibull padrão de $-U$, $F(-u) = 1 - e^{-e^u}$ (modelo de regressão log-log):

$$p = 1 - e^{-e^\eta} \;\Leftrightarrow\; \text{cloglog}(p) = \ln\left(-\ln(1-p)\right) = \eta \;\Leftrightarrow\; g(p) = \eta.$$

Outras opções estão ainda disponíveis. Na verdade a oferta pode-se sistematizar em famílias de funções de ligação, $\mathfrak{F} = \{F(\cdot,\psi), \psi \in \Psi,\ F(\cdot,\psi)$ função distribuição $, F(\cdot,\cdot)$ conhecido$\}$, que incluem as funções de ligação atrás descritas como seus elementos particulares para algum valor de $\psi$ e, com base em algum dos seus membros, fazer $p = F(\eta,\psi) = F_\psi(\eta)$, onde $\psi$ é o chamado parâmetro de ligação, que pode ser estimado conjuntamente com os coeficientes de regressão. Exemplos destas famílias mais comuns são descritos em [22, 1, 3].
Apesar da variedade, a escolha mais frequente para a função de ligação é a logística, principalmente porque o modelo resultante tem associado um tratamento matemático simples e flexível, com coeficientes de regressão que se prestam a uma interpretação simples. E na verdade, as funções de ligação logística e probit são sensivelmente funções lineares uma da outra para valores de $p$ entre $]0.1; 0.9[$, pelo que resultam em estimativas semelhantes na maioria dos casos. Contudo, o mesmo não acontece com as restantes funções de ligação. As consequências da má especificação da função de ligação são diversas, mas destacam-se o aumento substancial do enviesamento das probabilidades estimadas, agravado por se usar uma função de ligação simétrica quando deveria ser enviesada ou com uma cauda mais pesada, ou por probabilidades de sucesso extremas, aumento do erro quadrático médio das estimativas (pior para a má especificação por enviesamento do que por caudas pesadas), entre outras [4, 13, 14].
É então importante avaliar se a escolha da função de ligação é adequada. Para tal existem alguns métodos desenvolvidos envolvendo testes para os quais a possibilidade de incluir a função de ligação pressuposta numa classe paramétrica de funções de ligação é essencial. Destacam-se seguidamente alguns métodos disponíveis para tal, pela relevância da sua frequência nas aplicações:

- O teste de Pregibon [22], desenvolvido para uma família de funções de ligação de Tukey generalizadas, que inclui como membros as mais comuns funções de ligação. Este teste examina a adequação da função de ligação pressuposta, digamos $g_0(p)$, linearizando a verdadeira função de ligação, digamos $g_*(p)$, em torno de $g_0(p)$, ambas elementos da mesma família de funções de ligação cujos elementos genéricos dependem de alguns parâmetros adicionais. Esta linearização permite decompor $g_0(p)$ na soma da verdadeira função de ligação $g_*(p) = \eta$ com outra parcela linear nos parâmetros específicos da família das funções de ligação, que passam assim a ser considerados como coeficientes extra da regressão. O teste à sua significância revela a discrepâncias entre as funções de ligação pressuposta e real. Este teste está implementado nos pacotes do R-project `LDdiag`, [18], o pacote `rms` [8] e no o pacote `binomTools` [9].

- O teste de diagnóstico de Hosmer-Lemeshow modificado [10, 12] verifica se as probabilidades estimadas a partir do modelo são consistentes com a resposta binária observada. Para tal ordena as probabilidades estimadas para cada observação, divide-as em (10) grupos de sensivelmente o mesmo número de probabilidades, calcula as probabilidades médias dentro de cada grupo e, multiplicando-as pelo número de observações do grupo, obtém o número esperado de sucessos nesse grupo. Esse número é então comparado com o número efectivo de sucessos observado no grupo através de um teste de $\chi^2$ de Pearson. Este teste está implementado nos pacotes do R-project `LDdiag`, [18], `rms` [8], `binomTools` [9] ou `ResourceSelection` [16].

- Existem outros testes menos conhecidos, como por exemplo o teste de Cheng e Wu [2] que se baseia num resultado de equivalência de "informação" e uma técnica de redução de dimensão, construindo um teste que tem a vantagem de não necessitar de especificar uma família particular de funções de ligação alternativas. Este teste não se encontra disponível no R.

# 3 Outros Aspectos da Modelação

A análise de diagnóstico em modelos de regressão binária deve-se ainda ocupar de eventuais problemas na especificação do preditor linear (seleção das covariáveis), com observações discrepantes ou mesmo com os pressupostos distribucionais. Os resíduos, a estatística desvio e a taxa de erro são quantidades que têm um papel central nesta pesquisa, mas por vezes são mal usados.

## 3.1 Resíduos

Para um modelo de regressão binária, tal como para um modelo de regressão linear, os resíduos $r$ podem ser definidos como a diferença entre os valores observados e os valores esperados [5],

$$r = Y - p = Y - g^{-1}(\eta).$$

Sendo os dados discretos também o são os resíduos, i.e, para cada valor de $g^{-1}(\eta)$, o resíduo só poderá valer um de dois valores, $-g^{-1}(\eta)$ ou $(1-g^{-1}(\eta))$, dependendo de se $Y = 0$ ou $Y = 1$. Assim, o gráfico destes resíduos contra os valores preditos pelo modelo frequentemente não é muito útil. Seguindo as ideias de [15] e de [21] para o modelo de regressão logística, generalizadas para dados discretos em [5] e [6], sugere-se o uso de resíduos compartimentados (*binned*) ou suavizados, para maior facilidade de interpretação, que devem ser simetricamente próximos de zero, caso cada compartimento inclua em si um número suficiente de resíduos.

Suponha-se que se pretende colocar em gráfico os resíduos $r$ versus os valores preditos (também poderia ser contra os valores de uma qualquer covariável). Podem-se compartimentar os resíduos e os valores preditos ordenando os valores destes últimos e arrumando-os ordenadamente num certo número compartimentos com sensivelmente o mesmo número de pontos. Para cada compartimento calculam-se então as médias dos valores preditos e dos correspondentes resíduos, fazendo-se o gráfico de umas contra as outras.

É ainda possível acrescentar ao gráfico os limites de $\pm$ 2 desvios padrão, dados por $\pm 2\sqrt{p(1-p)/m}$, em que $m$ representa o número

de pontos em cada compartimento, dentro das quais se esperam que caiam cerca de 95% dos resíduos compartimentados, permitindo a identificação de pontos discrepantes.
O pacote do R `arm` [7], para a ajuste de modelos de regressão e multinível/hierárquicos, incorpora uma função para fazer estes gráficos. Verifica-se ainda que os resultados associados a modelos de regressão binária apresentam grande sensibilidade aos ditos pontos influentes, sendo a sua identificação objeto de preocupação e de alguns desenvolvimentos recentes [19, 24].

## 3.2 Taxa de erro

A taxa de erro determina-se como a proporção de casos para os quais a observação $Y$ vale 0 e o correspondente valor predito $g^{-1}(\hat{\eta})$ é maior que 0.5 ou para os quais a observação $Y$ vale 1 e o correspondente valor predito $g^{-1}(\hat{\eta})$ é menor que 0.5. A taxa de erro é sempre menor que 0.5, frequentemente muito menor. Este erro pode-se comparar com a taxa de erro do modelo nulo para o qual todas as observações têm igual probabilidade $p$, significando que o preditor é constante e que, por isso, $p$ se estima como a proporção de valores observados não nulos. Neste caso a taxa de erro vale $\min(p,1-p)$ [5].

## 3.3 Desvio

Para testar a função de ligação e o preditor linear usados num modelo ou a significância de uma dada covariável, pode-se empregar o desvio (*deviance*), que é uma estatística análoga ao desvio padrão dos resíduos [5], definida, a menos de uma constante que não interessa para efeitos de comparação, como -2 vezes o logaritmo da função de verosimilhança calculado nas estimativas obtidas para o modelo proposto. Menores desvios correspondem melhores ajustes.
Para dados agrupados (em que para grupos de observações correspondem os mesmos valores de covariáveis), a distribuição do desvio pode ser aproximadamente qui-quadrado, mas para os dados individuais que aqui se tratam (observações em nível individual) esta aproximação não é adequada, não podendo ser usada para testar a

qualidade de ajustamento. Contudo, se o número de combinações possíveis dos valores das covariáveis não for muito elevado relativamente ao tamanho amostral total, pode ser possível agrupar os dados por essas combinações e fazer um teste à qualidade de ajustamento recorrendo à aproximação qui-quadrado do desvio [23].
Alternativamente pode-se recorrer ao teste de Hosmer e Lemeshow [10, 12] atrás descrito, que pode ser usado com qualquer tipo de dados, usando as probabilidades preditivas para criar os referidos grupos.

# 4 Aplicação

O número de acidentes *per capita* é em Lisboa dos maiores, quando comparado com as outras congéneres Europeias. Cerca de 11% destes acidentes são classificados como fatais, envolvendo pelo menos um morto, ou sérios, envolvendo pelo menos uma pessoa hospitalizada. É importante identificar-se corretamente os fatores que contribuem para a gravidade de um acidente e estimar o seu impacto.
Neste contexto, analisaram-se os dados dos acidentes com vítimas ocorridos na cidade de Lisboa, entre 2004 e 2007, obtidos da base de dados nacional gerida pela Autoridade Nacional de Segurança Rodoviária (ANSR), baseada nos relatórios preenchidos pelas forças policiais - 7565 acidentes com toda a informação correspondente completa. Aqui há informação sobre a variável resposta binária $Y$, representando a gravidade dos acidentes (6798 acidentes fatais e sérios, correspondendo a $Y = 1$, e 773 acidentes ligeiros, correspondendo a $Y = 0$), e informação complementar sobre o tipo de acidente (colisão, atropelamento, despiste), sobre aspetos circunstanciais (condições atmosféricas, hora, intensidade luminosa, dia da semana), características da via, características dos intervenientes do acidente (vítimas e condutores) e tipo de veículos envolvidos.
Com base numa primeira análise destes dados em [20], consideraram-se no modelo as covariáveis apresentadas na Tabela 1. A este conjunto de covariáveis, vários modelos foram ajustados (não considerando aqui interações, por simplicidade), destacando-se aqui os mo-

| Variável | Descrição | Valores | Variável | Descrição | Valores |
|---|---|---|---|---|---|
| $x_1$ | Tempo atmosférico | 1=Bom<br>2=Outro | $x_5$ | Semáforos | 1=Funcionando<br>2=Não funcionando<br>3=Inexistentes |
| $x_2$ | Hora do acidente | 1 = [0 : 00,6.59]<br>2 = [7 : 00,10.59]<br>3 = [11 : 00,15.59]<br>4 = [16 : 00,20.59]<br>5 = [21 : 00,23.59] | $x_6$ | Idade condutores | 1=Pelo menos 1 jovem mas nenhum idoso<br>2=Pelo menos 1 idoso mas nenhum jovem<br>3=Outro |
| $x_3$ | Luz do dia | 1=Lusco-fusco<br>2=Dia<br>3=Noite | $x_7$ | Peões | 1=Sim<br>2=Não |
| $x_4$ | Tipo de acidente | 1=Atropelamento<br>2=Colisão<br>3=Despiste | $x_8$ | Sexo condutores | 1=Só feminino<br>2=Só masculino<br>3=Masculino e feminino |

Tabela 1: Covariáveis incluidas na modelação da gravidade dos acidentes rodoviários.

delos logit, probit e complemetar log-log, por serem os mais usuais:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 \tag{2}$$

$$\text{probit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 \tag{3}$$

$$\text{cloglog}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 \tag{4}$$

Na Tabela 2 apresentam-se as estimativas dos seus coeficientes e a medida de qualidade dos modelos AIC (*Akaike Information Criterion*) que se revela indicadora da semelhante qualidade dos ajustes.

Para estes modelos efetuou-se o teste de Pregibon para testar se a função de ligação utilizada seria adequada. Para tal utilizou-se a função `pregibon` do pacote do R `LDdiag`. Os modelos Logit e Probit passaram o teste, com valores-p de 0.052 e 0.251, respetivamente, mas o modelo Cloglog não passou (valor-p= 0.017). Várias outras funções de ligação foram ainda consideradas sendo, em geral, rejeitadas por este teste. As funções de ligação logit e probit são essencialmente semelhantes para probabilidades pouco extremas, não estranhando o mesmo resultado do teste.

| | Logit | | Probit | | Cloglog | |
|---|---|---|---|---|---|---|
| Variável | Estimativa | Ep | Estimativa | Ep | Estimativa | Ep |
| Ordenada na origem | -2.66* | 0.44 | -1.52* | 0.24 | -2.69* | 0.41 |
| $x_1$(2) | -0.34* | 0.13 | -0.18* | 0.06 | -0.32* | 0.12 |
| $x_2$.(2) | -0.34. | 0.19 | -0.18. | 0.10 | -0.32. | 0.17 |
| $x_2$.(3) | -0.19 | 0.18 | -0.10 | 0.09 | -0.18 | 0.17 |
| $x_2$.(4) | -0.30* | 0.14 | -0.16* | 0.08 | -0.27* | 0.13 |
| $x_2$.(5) | -0.52* | 0.16 | -0.28* | 0.08 | -0.48* | 0.15 |
| $x_3$.(2) | -0.14 | 0.27 | -0.07 | 0.14 | -0.13 | 0.26 |
| $x_3$.(3) | 0.42 | 0.27 | 0.22 | 0.14 | 0.40 | 0.25 |
| $x_4$.(2) | 0.17 | 0.32 | 0.09 | 0.17 | 0.15 | 0.29 |
| $x_4$.(3) | 0.64. | 0.33 | 0.33. | 0.18 | 0.60* | 0.30 |
| $x_5$.(2) | -0.23 | 0.38 | -0.10 | 0.19 | -0.23 | 0.36 |
| $x_5$.(3) | -0.23* | 0.08 | -0.12* | 0.04 | -0.26* | 0.08 |
| $x_6$.(2) | -0.41* | 0.17 | -0.20* | 0.09 | -0.39* | 0.16 |
| $x_6$.(3) | -0.16. | 0.09 | -0.08. | 0.05 | -0.15. | 0.09 |
| $x_7$.(2) | 1.03* | 0.31 | 0.53* | 0.17 | 0.96* | 0.28 |
| $x_8$.(2) | 0.54* | 0.13 | 0.29* | 0.07 | 0.51* | 0.12 |
| $x_8$.(3) | 0.46* | 0.17 | 0.24* | 0.09 | 0.43* | 0.16 |
| AIC | 4832.2 | | 4832.8 | | 4831.9 | |

Tabela 2: Estimativas dos coeficientes dos modelos (2), (3) e (4) e correspondentes erros padrão.

Assim, a função de ligação logit é adequada neste contexto, pelo que a partir deste ponto nos focaremos no respectivo modelo (2). De forma a se avaliar outros aspetos do ajustamento, foram-se considerar para este modelo os resíduos correspondentes. Na Figura 1 encontram-se dispostos em gráfico, do lado esquerdo, os resíduos correspondentes versus os valores preditos pelo modelo e do lado direito os resíduos compartimentados. Estes últimos poderiam ter sido calculados e depois dispostos em gráfico utilizando a função `binnedplot` do pacote `arm` do R. Apenas um dos resíduos compartimentados cai fora dos limites e não parece haver nenhum padrão preocupante nestes, pelo que tal vem corroborar num bom ajuste.

Este bom ajuste é confirmado quando se efetua o teste da qualidade de ajustamento de Hosmer-Lemeshow, realizado no R com recurso à função `hoslem.test` do pacote `ResourceSelection` do R, com as parametrizações usuais (valor-p de 0.837). Poderia também ter sido realizado com base na função `HLtest.Rsq` do pacote `binomTools` do R. É ainda confirmado quando se agrupam os dados pelos níveis das covariáveis e se faz uso da distribuição por amostragem aproximada qui-quadrado do desvio (valor-p de 0.986) e quando se calcula a taxa

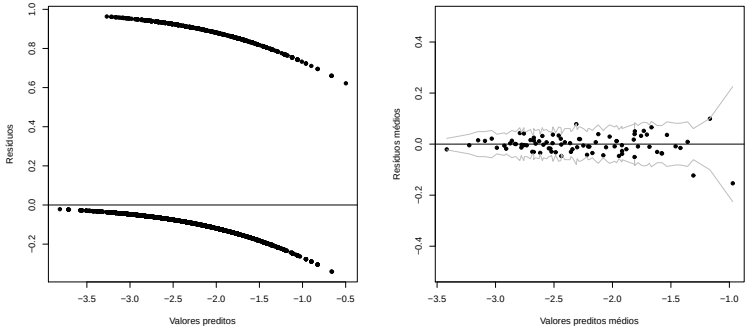



Figura 1: Resíduos (esquerda) e resíduos compartimentados (direita) do modelo (2).

de erro (0.102). O agrupamento das observações pode ser feito, por exemplo, recorrendo à função `group` do pacote `binomTools`.

# 5 Conclusões

Qualquer modelo de regressão binária deve ser questionado quanto à qualidade do seu ajuste, e tal envolve não apenas uma breve análise de resíduos, mas também a avaliação de se a função de ligação usada é adequada, já que tal pode levar a enviesamentos graves nas estimativas e nos valores preditos pelo modelo, bem como aumentos no erro quadrático médio. Os resíduos devem ser avaliados tendo sempre em mente de que são resíduos de dados discretos e, por isso, não devem ser tratados como resíduos de um modelo de regressão linear. Não há razão para se evitar a avaliação correcta de um modelo de regressão binária, havendo soluções acessíveis, algumas das quais apresentadas neste trabalho, evitando assim os problemas inerentes à incorreção.

# Agradecimentos

A ANSR e o LNEC forneceram parte da informação contida nos dados. Este trabalho foi parcialmente financiado por fundos nacionais através da Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT, pelos projeto PEst-OE/MAT/UI0006/2014 e pelo projetos de investigação Spatial Analysis of Child Road Accidents (SACRA), PTDC/TRA/66161/2006. Os autores agradecem ainda os comentários dos revisores anónimos que permitiram melhorar a qualidade do trabalho.

# Referências

[1] Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* 68, 357–364.

[2] Cheng, K. F., Wu, J. W. (1994). Testing goodness of fit for a parametric family of link functions. *Journal of the American Statistical Society*, 89 657–664.

[3] Czado, C. (1997). On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and Inference*, 61 125–139.

[4] Czado, C., Santner, T.J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* 33, 213–231.

[5] Gelman, A., Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.

[6] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis - 3rd edition*. CRC Press, Boca-Raton.

[7] Gelman, A., Su, Y-S, Yajima, M., Hill, J., Pittau, M.G., Kerman, J., Zheng, T., Dorie, V. (2007). `arm` package manual. R-project.

[8] Harrell Jr, F.E. (2002). `rms` package manual. R-project.

[9] Haubo, R., Christensen, B., Hansen, M.K. (2011). `binomTools` package manual. R-project.

[10] Hosmer, D.W., Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* 9, 1043–1068.

[11] Hosmer, D.W., Jovanovic, B., Lemeshow, S. (1989). Best subjects logistic regression. *Biometrics* 45, 1265–1270.

[12] Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition*. Wiley Series in Probability and Mathematical Statistics, New York.

[13] Huettmann, F., Linke, J. (2003). Assessment of different link functions for modeling binary data to derive sound inferences and predictions. In V. Kumar et al. (Eds.) *ICCSA 2003*, LNCS 2669, 43–48.

[14] Koenker, R. (2006). Parametric links for binary response. *Rnews* 6, 32–34.

[15] Landwehr, J.M., Pregibon, D., Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* 79, 61–72.

[16] Lele, S.R., Keim, J.L., Solymos, P. (2013). `ResourceSelection` package manual. R-project.

[17] McCullagh, P., Nelder, J. (1989). *Generalized Linear Models, 2nd edition*. Chapman and Hall/CRC, Boca Raton.

[18] Ni, Y. (2010). `LDdiag` package manual. R-project.

[19] Midi, H., Ariffin, S.B. (2013). Modified standardized Pearson residual for the identification of outliers in logistic regression model. *Journal of Applied Sciences* 13, 828–836.

[20] Nunes, A.R. (2011). *Modelação Espacial de Acidentes Rodoviários na Cidade de Lisboa*. Tese de Mestrado. FCT-UNL.

[21] Pardoe, I., Cook, R.D. (2002). A graphical method for assessing the fit of a logistic regression model. *American Statistician* 56, 263–272.

[22] Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* 29, 15–24.

[23] Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. Available at texttthttp://data.princeton.edu/wws509/notes/.

[24] Sarkar, S.K., Midi, H., Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences* 11, 26–35.

# Distâncias de Mahalanobis, variáveis originais e componentes principais

Jorge Cadima
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) e Matemática/DCEB, Instituto Superior de Agronomia, Universidade de Lisboa, *jcadima@isa.ulisboa.pt*

**Palavras–chave**: Distâncias de Mahalanobis, *outliers*, selecção de variáveis, componentes principais

**Resumo**: Em comunicação anterior [2] mostrou-se a correspondência entre as habituais distâncias amostrais de Mahalanobis e conceitos geométricos simples no espaço das variáveis ($\mathbb{R}^n$). Mostrou-se ainda que essas propriedades geométricas estão subjacentes a algumas recém-descobertas e importantes propriedades das distâncias de Mahalanobis ([4], [1]). Nesta comunicação, utilizam-se esses resultados para definir procedimentos automatizáveis de identificação das variáveis, ou das componentes principais, que mais contribuem para distâncias de Mahalanobis grandes, quer em relação ao centro de gravidade, quer entre indivíduos.

## 1 Introdução

As distâncias de Mahalanobis (DM) foram introduzidas em 1936 por P.C. Mahalanobis [5], como "distâncias estatísticas generalizadas" que levam em conta o padrão de covariâncias num conjunto multivariado de dados. De entre múltiplas utilizações das DM em estatística multivariada, saliente-se o seu papel na identificação de observações atípicas (*outliers*), uma vez que as DM salientam observações que se afastam de tendências de fundo que relacionem a maioria das observações num conjunto de dados multivariado.
Seja $\mathbf{X}$ uma matriz $n \times p$ de dados, cujas $p$ colunas correspondem a variáveis e cujas $n$ linhas correspondem a unidades experimen-

tais sobre as quais se observaram as variáveis. Seja $\mathbf{S}$ a matriz de (co)variâncias amostral correspondente à matriz de dados $\mathbf{X}$. Seja ainda $\mathbf{x}_{[i]} \in \mathbb{R}^p$ o vector correspondente à $i$-esima linha da matriz $\mathbf{X}$ e $\overline{\mathbf{x}} \in \mathbb{R}^p$ o vector das médias amostrais das $p$ variáveis (colunas de $\mathbf{X}$). As mais importantes distâncias (ao quadrado) amostrais de Mahalanobis são a *distância de Mahalanobis do indivíduo i ao centro de gravidade*:

$$d_i^2 \;=\; (\mathbf{x}_{[i]} - \overline{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x}_{[i]} - \overline{\mathbf{x}}) \tag{1}$$

e a *distância de Mahalanobis entre os indivíduos i e j*:

$$d_{ij}^2 \;=\; (\mathbf{x}_{[i]} - \mathbf{x}_{[j]})^t \mathbf{S}^{-1} (\mathbf{x}_{[i]} - \mathbf{x}_{[j]}). \tag{2}$$

A interpretação das DM no espaço das variáveis ($\mathbb{R}^n$), onde cada eixo corresponde a uma unidade experimental observada e cada uma das $p$ variáveis é representada pelos vectores colunas da matriz de dados $\mathbf{X}$, é sugerida, de forma natural, pelo facto das definições acima se poderem re-escrever em termos da matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X}_c) \subset \mathbb{R}^n$, gerado pelas colunas da matriz *centrada* dos dados, $\mathbf{X}_c$, cuja $i$-ésima linha é $(\mathbf{x}_{[i]} - \overline{\mathbf{x}})^t$. De facto, tem-se (ver [2] para mais pormenores):

$$\begin{aligned} d_i^2 &= {}_{(n-1)}\, \mathbf{e}_i^t \mathbf{P}_c \mathbf{e}_i \;=\; {}_{(n-1)}\, \|\mathbf{P}_c \mathbf{e}_i\|^2 \;=\; {}_{(n-1)}\, p_{ii} && (3)\\ d_{ij}^2 &= {}_{(n-1)}\, (\mathbf{e}_i - \mathbf{e}_j)^t \mathbf{P}_c (\mathbf{e}_i - \mathbf{e}_j) \;=\; {}_{(n-1)}\, \|\mathbf{P}_c (\mathbf{e}_i - \mathbf{e}_j)\|^2 && (4)\\ &= {}_{(n-1)}\, (p_{ii} + p_{jj} - 2\, p_{ij}) \;, && (5) \end{aligned}$$

onde $\mathbf{P}_c = \mathbf{X}_c (\mathbf{X}_c^t \mathbf{X}_c)^{-1} \mathbf{X}_c^t$, é a matriz de projecção ortogonal sobre o espaço das colunas de $\mathbf{X}_c$, $\mathcal{C}(\mathbf{X}_c) \subset \mathbb{R}^n$; $p_{ij}$ é o elemento genérico de $\mathbf{P}_c$; e $\mathbf{e}_i$ é o $i$-ésimo vector da base canónica de $\mathbb{R}^n$.
Em Cadima [2] introduziram-se várias expressões alternativas para as DM ao centro de gravidade, $d_i^2$:

- Sendo $\theta_i$ o ângulo entre $\mathbf{e}_i$ e $\mathcal{C}(\mathbf{X}_c)$, tem-se

$$d_i^2 = {}_{(n-1)}\, \cos^2 \theta_i \tag{6}$$

- Sendo $\theta_i^*$ o ângulo entre o vector centrado $\mathbf{e}_i^*$ correspondente a $\mathbf{e}_i$ (cujo elemento $i$ é $1-\frac{1}{n}$, sendo os restantes $-\frac{1}{n}$), e o subespaço $\mathcal{C}(\mathbf{X}_c)$, tem-se:

$$d_i^2 \;=\; \tfrac{(n-1)^2}{n}\,\cos^2\theta_i^*\;; \tag{7}$$

- Sendo $R_i^2$ o coeficiente de determinação da regressão linear múltipla de $\mathbf{e}_i$ sobre as variáveis que definem as colunas da matriz de dados $\mathbf{X}$, tem-se:

$$d_i^2 = \tfrac{(n-1)^2}{n}\,R_i^2\;. \tag{8}$$

Expressões análogas foram dadas para as DM entre indivíduos:

- Sendo $\theta_{ij}$ o ângulo entre $\mathbf{e}_i-\mathbf{e}_j$ e a sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X}_c)$, tem-se

$$d_{ij}^2 \;=\; 2(n\text{–}1)\cos^2\theta_{ij}\;. \tag{9}$$

- Sendo $R_{ij}^2$ o coeficiente de determinação da regressão de $\mathbf{e}_i-\mathbf{e}_j$ sobre as colunas de $\mathbf{X}$, tem-se

$$d_{ij}^2 \;=\; 2(n\text{–}1)R_{ij}^2\;. \tag{10}$$

Nesta comunicação utilizam-se estas expressões e as suas propriedades para identificar componentes principais (CPs) e/ou variáveis originais que mais contribuem para valores elevados das DM.

## 2 Mahalanobis e componentes principais

É sabido que uma matriz de projecção ortogonal sobre um dado subespaço é única, embora possa ser construída a partir de qualquer base do referido subespaço. Assim, dado um subespaço $k$-dimensional $M \subset \mathbb{R}^n$, e qualquer base desse subespaço, colocando

os vectores da base nas colunas duma matriz $\mathbf{B}_{n\times k}$, a matriz de projecção ortogonal sobre $M$ é dada por $\mathbf{P} = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$. No caso da base ser ortonormada, a matriz de projecção ortogonal toma a forma mais simplificada $\mathbf{P} = \mathbf{B}\mathbf{B}^t$. Daqui decorre directamente o seguinte resultado.

**Proposição 2.1** *Seja* $\mathbf{X}$ *uma matriz* $n \times p$ *de dados, e* $\mathbf{X}_c$ *a correspondente matriz de colunas centradas. Seja* $\mathbf{U}$ *uma matriz de elemento genérico* $u_{ij}$*, cujas colunas constituem uma qualquer base ortonormada de* $\mathcal{C}(\mathbf{X}_c)$*, e* $\mathbf{u}_{[i]}^t$ *a sua linha* $i$*. Verificam-se as seguintes expressões para as distâncias de Mahalanobis amostrais do indivíduo* $i$ *ao centro de gravidade, e entre os indivíduos* $i_1$ *e* $i_2$*:*

$$d_i^2 \;=\; {}_{(n-1)}\|\mathbf{u}_{[i]}\|^2 \;=\; {}_{(n-1)}\sum_{j=1}^{p} u_{ij}^2 \tag{11}$$

$$d_{i_1 i_2}^2 \;=\; {}_{(n-1)}\|\mathbf{u}_{[i_1]}-\mathbf{u}_{[i_2]}\|^2 \;=\; {}_{(n-1)}\sum_{j=1}^{p} (u_{i_1 j}-u_{i_2 j})^2 \tag{12}$$

*Dem.*: Da equação (3), e já que $\mathbf{U}^t\mathbf{e}_i = \mathbf{u}_{[i]}$, tem-se $d_i^2 = {}_{(n-1)}\,\mathbf{e}_i^t\mathbf{P}_c\mathbf{e}_i = {}_{(n-1)}\,\mathbf{e}_i^t\mathbf{U}\mathbf{U}^t\mathbf{e}_i = {}_{(n-1)}\,\mathbf{u}_{[i]}^t\mathbf{u}_{[i]} = {}_{(n-1)}\,\|\mathbf{u}_{[i]}\|^2$. A expressão para a distância entre indivíduos sai de forma análoga, uma vez que $\mathbf{U}^t(\mathbf{e}_{i_1} - \mathbf{e}_{i_2}) = \mathbf{u}_{[i_1]}-\mathbf{u}_{[i_2]}$. ■

Assim, as distâncias de Mahalanobis são somas de parcelas associadas a cada um dos vectores da base ortonormada utilizada, sendo fácil identificar os vectores da referida base que mais contribuem para o valor observado duma qualquer distância.
Duas bases ortonormadas naturais para o subespaço $\mathcal{C}(\mathbf{X}_c)$, numa análise de dados multivariados, são dadas pelas componentes principais (normalizadas), quer sobre a matriz de (co)variâncias, quer sobre a matriz de correlações. Estas bases ortonormadas de $\mathcal{C}(\mathbf{X}_c)$ são os vectores singulares esquerdos da matriz $\mathbf{X}_c$, ou os da matriz normalizada de dados, $\mathbf{Z}$, obtida dividindo cada coluna de $\mathbf{X}_c$ pelo respectivo desvio padrão amostral. Assim, os *scores* de cada um

dos $n$ indivíduos, normalizados para que o vector de *scores* tenha norma unitária, permitem seleccionar um subconjunto de componentes principais onde seja evidenciada uma DM elevada.

## 2.1 Um exemplo

Considere-se um conjunto de dados[2] com resultados de análises químicas em $n=178$ vinhos de três castas de Itália. Excluíndo a indicação das castas, sobram $p=13$ variáveis: teor alcoólico (V2); teor de ácido málico (V3); cinzas (V4); alcalinidade das cinzas (V5); teor de magnésio (V6); índice de fenóis totais (V7); teor de flavonóides (V8); teor de outros fenóis (V9); teor de proantocianidinas (V10); intensidade de cor (V11); matiz (V12); razão das densidades ópticas OD280/OD315 (V13); e teor de prolina (V14). Neste exemplo, a DM de qualquer indivíduo ao centro de gravidade não pode exceder o valor máximo $\frac{(n-1)^2}{n} = 176.0056$ (ver equação 7). As maiores distâncias ao centro de gravidade correspondem aos indivíduos 122 ($d^2_{122} = 58.6533$) e 70 ($d^2_{70} = 38.6908$) e não atingem um terço do valor máximo possível. Como já se viu (11), estas DM ao centro são $n-1=177$ vezes as somas de quadrados dos *scores* normalizados de cada indivíduo, quer nas componentes principais sobre os dados normalizados, quer sobre os dados não normalizados. Estes valores são dados na Tabela 1. No que respeita às CPs sobre a matriz de correlações, o subespaço gerado pelas componentes 3, 8 e 13 capta $\frac{19.648+9.028+22.618}{58.6533} \approx 87{,}5\%$ da DM do indivíduo 122 ao centro de gravidade. Para o indivíduo 70, mais de metade da distância ao centro é explicada pela componente 5, e no subespaço gerado pelas CPs 5, 7 e 8 capta-se 70,7% da DM ao centro para todas as variáveis. Estes resultados podem ser visualizados através das nuvens de pontos nas 5 componentes referidas (Figura 1). No que respeita às CPs sobre a matriz de (co)variâncias, as três melhores componentes

[2]Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation . Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy. Dados disponíveis em `http://archive.ics.uci.edu/ml`, com a designação *Wine recognition data.*

| CP | Matriz **S** | | Matriz **R** | |
|---|---|---|---|---|
| | Ind. 122 | Ind. 70 | Ind. 122 | Ind. 70 |
| 1 | 0.799 | 0.008 | 0.377 | 0.825 |
| 2 | 3.481 | **15.450** | 0.026 | 0.784 |
| 3 | 4.456 | 2.648 | **19.648** | 1.054 |
| 4 | 1.175 | 0.459 | 0.062 | 1.437 |
| 5 | **11.822** | 0.337 | 0.065 | **20.428** |
| 6 | 2.972 | 1.530 | 0.994 | 0.377 |
| 7 | 2.571 | 0.840 | 0.651 | **3.828** |
| 8 | **13.013** | **4.948** | **9.028** | **3.101** |
| 9 | 0.645 | **6.587** | 0.596 | 0.482 |
| 10 | **10.845** | 0.211 | 0.251 | 0.686 |
| 11 | 0.114 | 2.629 | 0.092 | 0.236 |
| 12 | 0.363 | 2.453 | 4.245 | 3.078 |
| 13 | 6.395 | 0.591 | **22.618** | 2.374 |
| soma | 58.653 | 38.691 | 58.653 | 38.691 |

Tabela 1: O quadrado dos *scores* normalizados (vezes $n-1=177$) para os indivíduos 122 e 70, nas componentes principais (CPs) sobre a matriz de (co)variâncias **S** e sobre a matriz de correlações **R**.

para cada indivíduo explicam uma percentagem menor da respectiva DM ao centro de gravidade: as CPs 5, 8 e 10 explicam apenas 59,1% da DM do indivíduo 122, enquanto que nas componentes 2, 8 e 9 capta-se 69,7% da distância original do indivíduo 70 ($d^2_{70}=38.691$). No que respeita às DM entre pares de indivíduos, verifica-se que a maior destas distâncias corresponde ao par de indivíduos 60 e 122, para o qual $d^2_{(60,122)}=133.485$, sendo a distância máxima possível $2(n-1)=354$ (equação 9). Com base na equação (12) torna-se possível identificar as componentes principais nas quais são maiores as distâncias entre os *scores* normalizados destes indivíduos. Esses valores são dados na Tabela 2. As CPs 3 e 13 sobre a matriz de correlações explicam, por si só, mais de 80% da DM entre os indivíduos

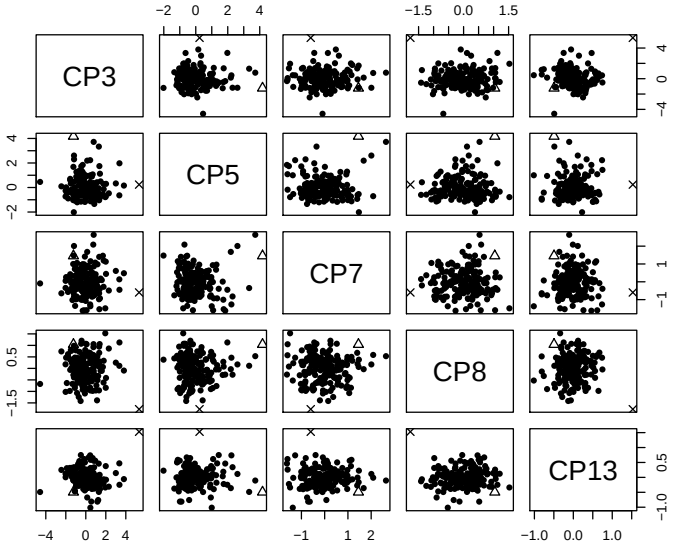



Figura 1: Componentes principais sobre os dados normalizados onde se evidencia a distância de Mahalanobis ao centro dos indivíduos 122 (cruz) e 70 (triângulo).

60 e 122, percentagem que sobe a quase 90% juntando a CP 12.

# 3 Mahalanobis e variáveis originais

De maior interesse interpretativo será a selecção de um subconjunto de variáveis originais onde se evidencie a maior parte do valor da distância de Mahalanobis de um dado indivíduo ao centro de gravidade. Essa selecção não pode ser feita da mesma forma que na secção anterior, uma vez que as variáveis originais não são, em geral, conjuntos ortonormados de variáveis. A partir das expressões

| CP | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **S** | 0.03 | 6.12 | **28.61** | 2.55 | **23.35** | **14.67** | 4.11 | 9.01 | 0.01 | **30.82** | 4.06 | 0.35 | 9.82 |
| **R** | 1.08 | 4.41 | **67.81** | 1.80 | 0.06 | 2.20 | 0.47 | 3.46 | 0.01 | 0.11 | 0.22 | **12.41** | **39.47** |

Tabela 2: Quadrados das diferenças (vezes $n-1 = 177$) entre *scores* normalizados dos indivíduos 60 e 122, nas componentes principais sobre a matriz de (co)variâncias (**S**) e de correlações (**R**). A soma de cada linha é $d^2_{(60,122)} = 133.485$.

(6) a (10) é imediato o seguinte resultado.

**Proposição 3.1** *Seja* $\mathbf{X}_c$ *uma matriz centrada de dados e* $\mathcal{C}(\mathbf{X}_c)$ *o respectivo espaço das colunas. Sejam* $\mathbf{X}_c^1, \ldots, \mathbf{X}_c^k$ *uma sequência de matrizes (com linhas associadas aos mesmos n individuos), cujos espaços-coluna formam uma sucessão encaixada de subespaços:* $\mathcal{C}(\mathbf{X}_c) \supseteq \mathcal{C}(\mathbf{X}_c^1) \supseteq \cdots \supseteq \mathcal{C}(\mathbf{X}_c^k)$. *Então,*

1. *Para cada indivíduo i, as DM ao centro de gravidade* ${d^2}_{i(j)}$ *correspondentes às matrizes* $\mathrm{X}_c^j$, *(*$j = 1, \ldots, k$*), formam uma sucessão decrescente:* ${d^2}_i \geq {d^2}_{i(1)} \geq \cdots \geq {d^2}_{i(k)}$.

2. *Para cada par de indivíduos* $i_1$,$i_2$, *as DM entre indivíduos* ${d^2}_{i_1 i_2(j)}$ *correspondentes às matrizes* $\mathrm{X}_c^j$, *(*$j = 1, \ldots, k$*), formam uma sucessão decrescente:* ${d^2}_{i_1 i_2} \geq {d^2}_{i_1 i_2(1)} \geq \cdots \geq {d^2}_{i_1 i_2(k)}$.

A forma mais natural de gerar as sequências de matrizes referidas na Proposição 3.1 será através da exclusão sucessiva de variáveis do conjunto de dados. Assim, *as distâncias de Mahalanobis formam sequências decrescentes com a exclusão de variáveis do conjunto de dados.* Em particular, para cada indivíduo, uma DM univariada ao centro de gravidade é um limite inferior para a distância de Mahalanobis multivariada de qualquer conjunto que contenha essa variável. Uma vez que a distância de Mahalanobis univariada do indivíduo *i*

ao centro de gravidade é dada (ver [4]) por $\frac{(x_i-\overline{x})^2}{s_x^2}$, temos

$$d_i^2 \ \geq \ \max_j \ z_{ij}^2 \ , \tag{13}$$

sendo $z_{ij}$ o elemento genérico da matriz normalizada de dados, $\mathbf{Z}$.
Um indivíduo cuja DM ao centro de gravidade seja considerada elevada pode ser considerado um indivíduo atípico (*outlier*) *univariado* caso o valor de $d_i^2$ seja próximo do limite inferior (13). Para indivíduos atípicos cujo limite inferior (13) seja muito inferior ao valor de $d_i^2$, haverá que identificar mais do que uma variável que contribua para essa atipicidade. A resposta pode ser dada a partir da expressão (8): sendo $R_i^2$ o coeficiente de determinação do vector canónico $\mathbf{e}_i$ sobre as $p$ variáveis do conjunto de dados, é possível usar toda a gama usual de métodos de selecção de subconjuntos de preditores numa regressão linear múltipla (veja-se [3]) para seleccionar um subconjunto de preditores que preserve o fundamental do valor de $R_i^2$ (ou seja, o fundamental da DM do indivíduo $i$ ao centro de gravidade).
De forma análoga, e a partir da equação (10), métodos de selecção de subconjuntos de preditores, aplicados à regressão da diferença $\mathbf{e}_i - \mathbf{e}_j$ permitem detectar subconjuntos de variáveis que expliquem o fundamental da DM original entre os indivíduos $i$ e $j$.

## 3.1 Um exemplo

Considerando ainda os dados do exemplo da subsecção 2.1, já se viu que a DM do indivíduo 122 ao centro de gravidade é $d_{122}^2 = \frac{(n-1)^2}{n} R_{122}^2 = 58.6533$. Normalizando a matriz de dados, verifica-se que o maior valor ao quadrado na linha 122, a que corresponde este indivíduo, é 9.906, o valor associado à variável $V4$. Assim, esta maior DM ao centro de gravidade não é reflexo dum valor extremo numa única variável. O coeficiente de determinação na regressão do vector $\mathbf{e}_{122}$ da base canónica de $\mathbb{R}^{178}$ é $R_{122}^2 = 0.3332$. Aplicando o tradicional algoritmo de exclusão sequencial para seleccionar um

subconjunto de preditores, obtem-se a sequência indicada na Tabela 3. Assim, as sete variáveis V2, V5, V7, V8, V9, V10 e V11 são responsáveis pelo fundamental do valor da DM ao centro de gravidade revelada pelo indivíduo 122 no conjunto original de $p=13$ variáveis. A DM deste indivíduo ao centro de gravidade, no conjunto de dados de dimensão $178 \times 7$, definido apenas por estas sete variáveis, é $d^{2[*]}_{122} = 50.8668$. A Tabela 3 mostra como as DM univariadas são pouco informativas para identificar estes subconjuntos de variáveis, estando a maior DM univariada associada à primeira exclusão.

| Exclusão | $R^2_{122}$ | $d^2_{122}$ | $d^{2[u]}_{122}$ univariadas |
|---|---|---|---|
| nenhuma | 0.3332 | 58.6533 | — |
| -V4 | 0.3332 | 58.6533 | 9.9064 |
| -V3 | 0.3331 | 58.6249 | 0.0657 |
| -V13 | 0.3308 | 58.2178 | 2.3067 |
| -V14 | 0.3216 | 56.6073 | 0.8013 |
| -V12 | 0.3073 | 54.0781 | 0.0144 |
| -V6 | 0.2890 | 50.8668 | 1.8182 |
| -V9 | 0.2671 | 47.0104 | 0.7551 |
| -V7 | 0.2430 | 42.7700 | 1.9991 |
| -V10 | 0.2024 | 35.6164 | 0.2378 |
| -V11 | 0.1600 | 28.1553 | 0.1651 |
| -V2 | 0.1451 | 25.5399 | 3.1490 |
| -V5 | 0.0530 | 9.3282 | 7.2710 |

Tabela 3: Distâncias de Mahalanobis ao centro do indivíduo 122 ($d^2_{122}$) após cada passo dum algoritmo de exclusão sequencial na regressão múltipla do vector canónico $\mathbf{e}_{122}$ sobre as variáveis, e respectivos coeficientes de determinação ($R^2_{122}$). Na coluna da direita indicam-se as DM univariadas, para as variáveis excluídas (sendo a DM univariada para a última variável retida, $V8$, $d^{2[u]}_{122}=9.3282$).

Um procedimento análogo revela que o fundamental da DM do indivíduo 70 ao centro de gravidade, no conjunto original de variáveis

($d^2_{70} = 38.6908$), é preservado num subconjunto de apenas cinco variáveis (V4, V6, V8, V10 e V13), com $d^{2[*]}_{70} = 34.2365$.
As variáveis que mais contribuem para a DM entre os indivíduos 60 e 122 (a maior DM entre qualquer par de indivíduos) identificam-se através dum algoritmo de selecção de subconjuntos de preditores, partindo da regressão linear multipla do vector $\mathbf{e}_{60} - \mathbf{e}_{122}$ sobre os $p = 13$ preditores. Os resultados obtidos com o algoritmo de exclusão sequencial são dados na Tabela 4. Como se pode constatar, nas

| Exclusão | $R^2_{60,122}$ | $d^2_{60,122}$ | $d^{2[u]}_{60,122}$ univariadas |
|---|---|---|---|
| nenhuma | 0.3771 | 133.4851 | — |
| -V3 | 0.3746 | 132.5979 | 0.9872 |
| -V10 | 0.3715 | 131.4936 | 6.4180 |
| -V14 | 0.3664 | 129.7141 | 0.0305 |
| -V4 | 0.3599 | 127.3929 | 46.4614 |
| -V13 | 0.3512 | 124.3318 | 6.9371 |
| -V12 | 0.3416 | 120.9089 | 0.2756 |
| -V6 | 0.3313 | 117.2795 | 4.7110 |
| -V2 | 0.3164 | 111.9928 | 0.9955 |
| -V9 | 0.2893 | 102.4034 | 2.3307 |
| -V11 | 0.2514 | 88.9916 | 3.0519 |
| -V7 | 0.2131 | 75.4366 | 3.6764 |
| -V8 | 0.0812 | 28.7294 | 20.3866 |

Tabela 4: DM entre os indivíduos $i_1 = 60$ e $i_2 = 122$ ($d^2_{60,122}$) em cada passo dum algoritmo de exclusão sequencial na regressão múltipla do vector $\mathbf{e}_{60} - \mathbf{e}_{122}$, e respectivos coeficientes de determinação ($R^2_{60,122}$). Na coluna final ($d^{2[u]}_{60,122}$) estão as DM univariadas para as variáveis excluídas. Na última variável retida ($V5$) tem-se $d^{2[u]}_{60,122} = 28.7294$.

variáveis originais V2, V5, V7, V8, V9 e V11, a DM entre este par de indivíduos é $d^{2[*]}_{60,122} = 117.2795$, ou seja, quase 88% do valor $d^2_{60,122} = 133.485$ da distância correspondente na totalidade das $p =$

13 variáveis originais.

# 4 Uma pequena discussão final

Com base nas propriedades discutidas em Cadima [2], foram apresentadas técnicas computacionalmente baratas de identificação de subconjuntos de variáveis originais ou de componentes principais que mais contribuem para os valores das distâncias de Mahalanobis de um indivíduo ao centro de gravidade, ou entre pares de indivíduos, num conjunto $p$-variado de dados. Tais técnicas permitem identificar subconjuntos de variáveis onde determinados indivíduos sejam atípicos, ou que distingam pares de indivíduos em conjuntos de dados multivariados.
Foi ainda mostrado que a exclusão sucessiva de variáveis dum conjunto de dados multivariados gera uma sequência decrescente de distâncias de Mahalanobis correspondentes aos mesmos indivíduos.

# Agradecimentos

Este trabalho foi financiado por Fundos Nacionais, através da FCT - Fundação para a Ciência e a Tecnologia, no âmbito dos projectos PEst-OE/MAT/UI0006/2011 e PEst-OE/MAT/UI0006/2014.

# Referências

[1] Branco, J.A., Pires, A.M. (2011). Travelling through multivariate data spaces with Mahalanobis distance. *JOCLAD 2011*, Vila Real.

[2] Cadima, J. (2013). A geometria das distâncias de Mahalanobis no espaço das variáveis. *Atas do XX Congresso da Sociedade Portuguesa de Estatística*, Porto 2012, Edições SPE, pgs. 81-94.

[3] Draper, N.R. & Smith, H. (1998), *Applied Regression Analysis* (3a. edição), Wiley.

[4] Gath, E.G., Hayes, K. (2006). Bounds for the largest Mahalanobis distance. *Linear Algebra and its Applications* 419, 93-106.

[5] Mahalanobis, P.C. (1936). On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences of India (Calcutta)* 2, 49-55.

# Generalized linear models, generalized additive models and generalized estimating equations to capture-recapture closed population models

Md. Abdus Salam Akanda
Department of Statistics, Biostatistics & Informatics, University of Dhaka, Bangladesh, *akanda_du@yahoo.com*

Russell Alpizar-Jara
Research Center in Mathematics and Applications, CIMA-U.E. and Department of Mathematics, University of Évora, 7000-671 Évora, Portugal, *alpizar@uevora.pt*

**Palavras–chave**: Capture-recapture experiment, generalized linear models, generalized additive models, generalized estimating equations, population size estimation

**Abstract**: Estimation of animal population parameters is an important issue in ecological statistics. In this paper generalized linear models (GLM), generalized additive models (GAM) and generalized estimating equations (GEE) are used to account for individual heterogeneity, modelling capture probabilities as a function of individual observed covariates. The GEE also accounts for a correlation structure among capture occasions. We are interested in estimating closed population size, where only heterogeneity is considered, there is no time effect or behavioral response to capture, and the capture probabilities depend on covariates. A real example is used for illustrative purposes. Conditional arguments are used to obtain a Horvitz-Thompson-like estimator for estimating population size. A simulation study is also conducted to show the performance of the estimation procedure and for comparison between methodologies. The GEE approach performs better than GLM or GAM approaches for estimating population size. The simulation study highlight the

importance of considering correlation among capture occasions.

# 1 Introduction

A capture-recapture experiment consists of a sequence of capture occasions upon which individuals are captured. If they have been previously marked their mark is recorded, otherwise they are given a unique mark. In addition various individual characteristics such as age, weight, gender, condition, and size may be measured. Many estimation methods have been developed for the capture-recapture closed population models [11], [22]. The most general closed population model is denoted $M_{tbh}$, where time (t), behavioral response (b), or heterogeneity between individuals (h) may affect the capture probabilities [11], [14]. Our interest is in the submodel $M_h$, where only heterogeneity is considered and there is no time effect or behavioral response to capture. There are two main approaches to model $M_h$. The first is through mixture models that may be parametric or nonparametric. These models date at least to Sanathanan [18], who considered both unconditional and conditional inference on the population size. More recently, Mao [10] considered finite mixture models; see Pledger and Phillpot [13] for a comprehensive review of these models. Dorazio and Royle [6] used the beta-binomial model and Coull and Agresti [4] used the logit normal model. Link [9] also showed that without strong assumptions on the underlying distribution, the population size is not identifiable under this model. Alternatively, one can model capture probabilities as functions of covariates [7]. Pollock [15] pointed out that the consideration of covariates in capture-recapture studies enabled more parsimonious parameterizations for models and there may be inherent ecological importance in understanding the nature of the relationships between the parameters and covariates.

For closed population studies, a conditional likelihood [7] approach is a generalized linear models (GLM) for the positive binomial distribution [12]. A popular extension of GLM are generalized additive

models (GAM) [21], where the linear predictor is a sum of smooth functions of the covariates. It is desirable to be able to employ similar models in capture-recapture analyses. For example, recently Akanda and Alpizar-Jara [1] developed generalized estimating equations approach which accounts for heterogeneity due to observed individual covariates and also dependency among capture occasions, modelling capture probabilities as a function of individual observed covariates. They also showed that the performance of estimating population size of the GEE approach is better than the mixed effects approach in the capture-recapture closed population study [2]. The main purpose of this paper is to estimate population size using GLM, GAM and GEE under closed population model, $M_h$. We also compare performances of the estimation procedures by means of a Monte Carlo simulation study.
In the next section, we describe the notation and models considered in this work. Section 3, illustrate the methodology for a real data set with discussion. A simulation study is given in Section 4 to evaluate the performance of GLM, GAM and GEE in capture-recapture methodology. Finally we provide some concluding remarks in Section 5.

## 2 Notation and Models

Consider a population consisting of $N$ animals in a capture-recapture experiment over $m$ capture occasions, $j = 1,2,\ldots,m$. Let $Y_{ij}$ be a binary outcome, equalling 1 if the $i^{th}$ animal is being caught on the $j^{th}$ capture occasion and 0 otherwise. Let $Y_i = (Y_{i1},Y_{i2},\ldots,Y_{im})'$ be a random vector with the capture history of individual $i$. Let $T_i = \sum_{j=1}^{m} Y_{ij}$ be the number of times the $i^{th}$ animal has been caught in the course of the trapping closed population study. Let $t_i$ be the time the $i^{th}$ individual is first captured. Heterogeneity in captured probabilities is often explained by observed individual covariate $x_i$, such as age, sex, weight, etc. Let the probability that

the $i^{th}$ animal is captured on any trapping occasion $j$, be

$$p_{ij}(\beta) = \mu_{ij} = \Pr(Y_{ij} = 1|X_i) = h(X_i\beta); i = 1,2,\ldots,N, j = 1,2,\ldots,m, \tag{1}$$

where

$$X_i = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_i & x_i & \ldots & x_i \end{bmatrix}'; (i = 1,2,\ldots,N),$$

is the design matrix, $\beta = (\beta_0, \beta_1)'$ is the vector of parameters associated with the covariates, and $h(u) = (1 + \exp(-u))^{-1}$ is the logistic function. We can consider that $p_{ij}(\beta) = h(\beta_0 + \beta_1 x_i) = p_i(\beta)$, because variation in capture probabilities among individuals is explained by the observed individual covariate $x_i$ only. There is no equivalent model of [11], although this model is a restricted version of their model $\mathrm{M_h}$. The probability of not capturing the $i^{th}$ individual on the $j^{th}$ occasion is $(1-p_i(\beta))$ and the variance of $Y_{ij}$ is $p_i(\beta)(1-p_i(\beta))$. Then $T_i \sim Bin(m, p_i(\beta))$ and $\pi_i(\beta) = 1-(1-p_i(\beta))^m$ is the probability of individual $i$ being captured at least once, given the covariate $x_i$. Let the set of distinct individuals captured at least in one occasion be indexed by $i = 1,2,\ldots,n$ and uncaptured individuals would be indexed by $i = n+1,\ldots,N$ without loss of generality. A partial likelihood ($l_p$) is the first product term of the following equation (2), which is the likelihood of the number of recaptures after the first capture [20].

$$\prod_{i=1}^{n} p_i(\beta)^{T_i-1}\{1-p_i(\beta)\}^{m-t_i-(T_i-1)} \prod_{i=1}^{n}\left[\frac{\{1-p_i(\beta)\}^{t_i-1}p_i(\beta)}{\pi_i(\beta)}\right]. \tag{2}$$

For a given $t_i$, $(T_i - 1)|t_i \sim Bin(m - t_i, p_i(\beta))$, which is used to estimate the parameters $\beta$. Once an estimate of $\beta$ is obtained ($\hat{\beta}$), the Horvitz-Thompson estimator $\hat{N} = \sum_{i=1}^{n} 1/\pi_i(\hat{\beta})$ may be used as in Huggins [7].

## 2.1 Generalized Additive Models

In generalized additive models (GAM), continuous covariates can be represented as a sum of smooth functions [21]. Suppose that $p_i(\beta) = h(X_i\beta)$ for a vector of parameters $\beta$ and $X_i = [s_1(x_i),s_2(x_i),\ldots,s_k(x_i)]$, where the $s(.)$'s are spline basis functions and $k$ is the basis dimension. The penalized partial log likelihood is $\log P_l = \log l_p(\beta) - \frac{\lambda}{2}\beta' Z\beta$ for a smoothing parameter $\lambda$ and a matrix $Z$ of known coefficients. Penalized iterative re-weighted least squares (P-IRLS) is used to estimate $\beta$ and a generalized cross-validation procedure is applied to select an optimal value for $\lambda$ [20],[21]. Following the method of Stoklosa and Huggins [20], the variance of $\hat{N}$ for a given $\lambda$ may be estimated by

$$\widehat{\mathrm{Var}}(\hat{N},\lambda) = \sum_{i=1}^{n} \frac{1-\pi_i(\beta)}{\pi_i(\beta)^2} + \{X'\eta(\beta)\}'\mathrm{Var}(\hat{\beta})\{X'\eta(\beta)\}, \quad (3)$$

where the matrix $X$ has $i^{th}$ row $X_i$, $\eta(\beta)$ is a vector with $\eta_i(\beta) = \pi_i(\beta)^{-2} m p_i\{1-\pi_i(\beta)\}$, and all quantities are evaluated at $\hat{\beta}$.

## 2.2 Generalized Estimating Equations

Let $V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$ be the covariance matrix of $Y_i$, where, $A_i = \mathrm{diag}[\mathrm{Var}(Y_{i1}),\mathrm{Var}(Y_{i2}),\ldots,\mathrm{Var}(Y_{im})]$ is a $m \times m$ diagonal matrix and $R_i(\alpha)$ is known as the working correlation structure among $Y_{i1},Y_{i2},\ldots,Y_{im}$ to describe the average dependency of individuals being captured from occasion to occasion. A GEE approach permits several types of working correlation structure $R_i(\alpha)$ (for details, see [5], [3]), but GEE yields consistent estimates even with misspecification of the working correlation matrix [8]. Hence we consider autoregressive working correlation structure among capture occasions in our analysis. We also assume that $Y_{ij}$ is conditional on the captured individuals ($n$) (i.e., $T_i \geq 1$) with the corresponding observed individual covariates similar to Huggins [7]. Let $D_i$ be the

matrix of derivatives $\partial\mu_i/\partial\beta'$, where, $\mu_i = (\mu_{i1},\mu_{i2},\ldots,\mu_{im})'$, and $D_i = A_i X_i$. Therefore, the vector of parameters $\beta$ can be estimated by solving the following generalized estimating equations:

$$U(\beta) = \sum_{i=1}^{n} D_i' V_i^{-1}(Y_i - \mu_i) = 0. \tag{4}$$

The equation (4) can be solved by an iterative procedure. At each iteration, the correlation parameter $\alpha$ can be estimated from the current Pearson's residuals defined by

$$\hat{r}_{ij} = \frac{Y_{ij} - \hat{\mu}_{ij}}{[\hat{\mu}_{ij}(1-\hat{\mu}_{ij})]^{1/2}}, \tag{5}$$

where, $\hat{\mu}_{ij}$ depends upon the current value for $\beta$. Liang and Zeger [8] estimated $\alpha$ as following,

$$\hat{\alpha} = \sum_{i=1}^{n}\sum_{j>l} \hat{r}_{ij}\hat{r}_{il} \Big/ \left\{\frac{1}{2}nm(m-1) - p\right\}, \tag{6}$$

where $p$ is the number of parameters in the model. An estimate of the variance of $\hat{N}$ is given by $\widehat{\mathrm{Var}}(\hat{N}) = \sum_{i=1}^{n} \pi_i(\beta)^{-2}\big(1-\pi_i(\beta)\big) + \Delta(\beta)'\Gamma(\beta)^{-1}\Delta(\beta)$ where $\Gamma(\hat{\beta})$ represents the conditional information matrix of $\beta$ and the vector $\Delta(\hat{\beta}) = \sum_{i=1}^{n} \pi_i(\beta)^{-2}\partial\pi_i(\beta)/\partial\beta$ with all quantities are evaluated at $\hat{\beta}$.

# 3 Illustrative Example

Our example concerns the captures of deer mice (*Peromyscus maniculatus*). The data set collected by V. Reid at East Stuart Gulch Colorado associated with covariates age, sex, and weight (g). A rectangular grid of $9 \times 11$ traps was used, with 50-foot (15.2-m) trap spacing. The data are well known and have been analysed in numerous capture-recapture literature, for example see [11] and [19].

The data set consists of $n = 38$ distinct deer mice. We consider weight (min. = 5.00, mean = 14.53, max. = 24.00, SD= 4.84) as a continuous covariate in our analysis. The numbers of deer mice caught for $m = 6$ occasions ($n_1$ to $n_6$) were 15, 20, 16, 19, 25, 25 and $\sum n_k = 120$. The recorded capture frequencies ($f_1$ to $f_6$) were 9, 6, 7, 6, 6, 4. Our estimation results are summarized in Table 1. We consider fitting a GLM and GAM using the partial likelihood, and quasi-likelihood is used in GEE approach [1], [7], and [20]. A *B*-spline basis function [21] was chosen with 10 knots in GAM approach and autoregressive working correlation structure in GEE approach. The GLM yields a much smaller population size standard error estimate (0.37) compared to the GAM (2.24) and GEE (1.68), but the estimated population size approximately same of the captured individuals. Moreover, the GLM approach does not account for correlation among capture occasions. The estimate of population size using GEE approach is 40.47, this is quite close to the estimates given by [11], using their model $\mathrm{M_b}$ ($\hat{N} = 41$, $\mathrm{SE}(\hat{N}) = 3.01$) producing lower standard error 1.68. The GEE approach also produces lower standard error compare to GAM and it jointly takes into account heterogeneity in capture probabilities and correlation among capture occasions, suggesting the GEE approach may be more suitable.

Tabela 1: Comparison of population size estimates and standard error for deer mice data after fitting models with a covariate (weight)

| Model no. | Model | $\hat{N}$ | $\mathrm{SE}(\hat{N})$ |
|---|---|---|---|
| 1. | PL GLM | 38.13 | 0.37 |
| 2. | PL GAM | 39.93 | 2.24 |
| 3. | QL GEE | 40.47 | 1.68 |

PL: partial likelihood; QL: quasi-likelihood; GLM: generalized linear models; GAM: generalized additive models; GEE: generalized estimating equations.

# 4 Simulation Study

A simulation study was conducted in order to evaluate the performance of the estimators. Factors used in the simulation were population size, $N = 100$ and 300; mean capture probability, $\bar{p} = 0.3$, and 0.5; number of capture occasions, $m = 10$; and correlation coefficient, $\alpha = -0.3, 0, 0.3$ for the autoregressive correlation structure. The effect of heterogeneity among observed individuals was modelled using a covariate, weight. For each individual, we assigned weight from a normal distribution with mean 15 and variance 4. Correlated capture history ($Y_{ij}$) are generated following the method of Qaqish [16]. For each simulation scenario, GLM, GAM and GEE approaches were used for data analyses, and to assess estimators performances. The simulation study was carried out with 1,000 Monte Carlo replicates using program R [17].

We present the averages of the numbers of captured individuals, ($\bar{n}$); the average estimates of population size, (AVE($\hat{N}$)); standard errors of the estimated population size, (SE($\hat{N}$)); percentage relative bias, (PRB $= 100 \times (\mathrm{E}(\hat{N}) - N) \div N$, where $\mathrm{E}(\hat{N})$ is estimated by AVE($\hat{N}$)); root mean square error, (RMSE $= \sqrt{\widehat{\mathrm{Var}}(\hat{N}) + \mathrm{Bias}^2}$); percentage coefficient of variation, (CV $= 100 \times \mathrm{SE}(\hat{N}) \div \mathrm{E}(\hat{N})$) and confidence interval coverage (%), (CIC) for the estimates of population size. The simulation results are given in Table 2 and Table 3. Estimation of average animal abundance (AVE($\hat{N}$)) is almost the same as true population size ($N$) when there is no correlation ($\alpha = 0$) for a average fixed capture probability. For the average capture probability, the estimated population size and its standard error vary depending on the strength of linear correlation among capture occasions and related factors. For a fixed average capture probability, the estimated population size and its standard error is higher for negative correlation comparatively to the same strength of positive correlation. The performance of GEE estimators is better than GLM and GAM estimators producing in general, lower coefficient of variation, absolute value of percentage relative bias, root

mean square error and slightly higher confidence interval coverage.

Tabela 2: Simulation results (1000 repetitions) when capture occasions, $m = 10$ and population size, $N = 100$

| $\bar{p}$ | $\alpha$ | $\bar{n}$ | AVE($\hat{N}$) | SE($\hat{N}$) | PRB | CV | RMSE | CIC |
|---|---|---|---|---|---|---|---|---|
| Partial likelihood based on generalized linear models | | | | | | | | |
| 0.30 | -0.3 | 97 | 103.89 | 1.33 | 3.85 | 2.38 | 4.11 | 85.8 |
| 0.30 | 0.0 | 93 | 100.83 | 1.91 | 0.83 | 2.09 | 2.08 | 97.5 |
| 0.30 | 0.3 | 92 | 94.71 | 2.68 | -5.29 | 1.63 | 5.93 | 22.5 |
| 0.50 | -0.3 | 98 | 100.16 | 0.08 | 0.16 | 0.41 | 0.18 | 99.5 |
| 0.50 | 0.0 | 97 | 100.07 | 0.26 | 0.07 | 0.37 | 0.27 | 94.4 |
| 0.50 | 0.3 | 95 | 99.25 | 0.93 | -0.75 | 0.31 | 1.19 | 45.4 |
| Partial likelihood based on generalized additive models | | | | | | | | |
| 0.30 | -0.3 | 97 | 103.90 | 1.25 | 3.90 | 2.39 | 4.09 | 85.5 |
| 0.30 | 0.0 | 93 | 100.98 | 1.95 | 0.98 | 2.12 | 2.18 | 97.9 |
| 0.30 | 0.3 | 92 | 95.03 | 2.93 | -4.97 | 1.91 | 5.77 | 29.5 |
| 0.50 | -0.3 | 98 | 100.17 | 0.05 | 0.17 | 0.41 | 0.18 | 99.9 |
| 0.50 | 0.0 | 97 | 100.06 | 0.29 | 0.06 | 0.38 | 0.30 | 93.8 |
| 0.50 | 0.3 | 95 | 99.27 | 0.94 | -0.73 | 0.36 | 1.19 | 50.3 |
| Quasi-likelihood based on generalized estimating equations | | | | | | | | |
| 0.30 | -0.3 | 97 | 103.52 | 1.19 | 3.52 | 1.15 | 3.72 | 86.7 |
| 0.30 | 0.0 | 93 | 100.74 | 1.77 | 0.74 | 1.76 | 1.92 | 98.3 |
| 0.30 | 0.3 | 92 | 95.87 | 2.59 | -4.13 | 2.70 | 4.87 | 36.8 |
| 0.50 | -0.3 | 98 | 100.12 | 0.04 | 0.12 | 0.04 | 0.13 | 99.6 |
| 0.50 | 0.0 | 97 | 100.04 | 0.23 | 0.04 | 0.23 | 0.23 | 95.2 |
| 0.50 | 0.3 | 95 | 99.42 | 0.85 | -0.58 | 0.85 | 1.03 | 57.4 |

# 5 Conclusion

Individual heterogeneity and time dependence are fundamentally important in real-life applications of capture-recapture studies. The

Tabela 3: Simulation results (1000 repetitions) when capture occasions, $m = 10$ and population size, $N = 300$

| $\bar{p}$ | $\alpha$ | $\bar{n}$ | AVE($\hat{N}$) | SE($\hat{N}$) | PRB | CV | RMSE | CIC |
|---|---|---|---|---|---|---|---|---|
| Partial likelihood based on generalized linear models | | | | | | | | |
| 0.30 | -0.3 | 291 | 311.44 | 2.11 | 3.81 | 1.36 | 11.6 | 5.00 |
| 0.30 | 0.0 | 279 | 302.46 | 3.24 | 0.82 | 1.19 | 4.07 | 94.0 |
| 0.30 | 0.3 | 276 | 283.21 | 4.68 | -5.60 | 0.91 | 17.4 | 50.0 |
| 0.50 | -0.3 | 293 | 300.47 | 0.12 | 0.16 | 0.23 | 0.49 | 98.6 |
| 0.50 | 0.0 | 290 | 300.14 | 0.51 | 0.05 | 0.21 | 0.53 | 97.2 |
| 0.50 | 0.3 | 285 | 297.66 | 1.63 | -0.78 | 0.17 | 2.85 | 24.4 |
| Partial likelihood based on generalized additive models | | | | | | | | |
| 0.30 | -0.3 | 291 | 311.47 | 2.16 | 3.82 | 1.36 | 11.7 | 4.80 |
| 0.30 | 0.0 | 279 | 302.43 | 3.26 | 0.81 | 1.19 | 4.07 | 94.7 |
| 0.30 | 0.3 | 276 | 283.62 | 4.73 | -5.46 | 0.94 | 17.1 | 10.0 |
| 0.50 | -0.3 | 293 | 300.47 | 0.12 | 0.16 | 0.23 | 0.49 | 98.3 |
| 0.50 | 0.0 | 290 | 300.15 | 0.52 | 0.05 | 0.21 | 0.54 | 97.0 |
| 0.50 | 0.3 | 285 | 297.75 | 1.57 | -0.75 | 0.18 | 2.75 | 26.1 |
| Quasi-likelihood based on generalized estimating equations | | | | | | | | |
| 0.30 | -0.3 | 291 | 310.32 | 2.07 | 3.44 | 0.67 | 10.5 | 8.00 |
| 0.30 | 0.0 | 279 | 302.17 | 3.15 | 0.72 | 1.04 | 3.83 | 95.3 |
| 0.30 | 0.3 | 276 | 286.57 | 4.33 | -4.50 | 1.51 | 14.1 | 35.0 |
| 0.50 | -0.3 | 293 | 300.15 | 0.09 | 0.05 | 0.03 | 0.17 | 99.1 |
| 0.50 | 0.0 | 290 | 300.03 | 0.47 | 0.01 | 0.16 | 0.47 | 97.6 |
| 0.50 | 0.3 | 285 | 298.55 | 1.43 | -0.50 | 0.48 | 2.04 | 35.8 |

main purpose of this study was to compare estimates of population size and their standard error using statistical techniques such as, quasi-likelihood for GEE and partial likelihood for GLM and GAM. The quasi-likelihood GEE approach seems to perform better than GLM and GAM approaches because the standard error of the estimated population size are consistently lower. The estimators perform well when average capture probabilities are high ($\bar{p} = 0.50$),

but it is difficult to obtain reliable estimates of GLM and GAM approach for low capture probabilities ($\bar{p} = 0.30$). For cases where only a small proportion of individuals are captured, the GEE approach provides better RMSE and is robust to violation of the assumption of independence among capture occasions. This approach also provides means of exploring factors thought to be responsible for differences in capture probability among individuals. The simulation study also shows that the estimated population size varies on the nature of existing correlation among capture occasions. Hence, it is important to consider the type of correlation structure among capture occasions when estimating animal population parameters in capture-recapture studies. Researchers may apply these approaches for open population model to estimate unknown animal abundance in capture-recapture studies.

# Acknowledgement

This research was funded by EMMA in the framework of the EU Erasmus Mundus Action 2 and Fundação Nacional para a Ciência e Tecnologia (FCT), Portugal under the project PEst-OE/MAT/UI0117/2011. The authors are very grateful to the anonymous referees for their careful reading of the manuscript and several suggestions that considerably improved the presentation.

# Referências

[1] Akanda, M.A.S., Alpizar-Jara, R. (2014a). A generalized estimating equations approach for capture-recapture closed population models. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-014-0274-7.

[2] Akanda, M.A.S., Alpizar-Jara, R. (2014b). Estimation of capture probabilities using generalized estimating equations and mixed effects approaches. *Ecology and Evolution* 4, 1158–1165.

[3] Akanda, M.A.S., Khanam, M. (2011). *Goodness-of-fit tests for GEE model: methods and applications*. VDM Verlag Dr. Müller, ISBN: 978-3-639-35122-4, Saarbrücken, Germany.

[4] Coull, B.A., Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* 55, 294–301.

[5] Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.T. (2013). *Analysis of Longitudinal Data*. 2nd Edition. Oxford University Press, New York.

[6] Dorazio, R.M., Royle, J.A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* 59, 351–364.

[7] Huggins, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika* 76, 133–40.

[8] Liang, K.Y., Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

[9] Link, W.A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* 59, 1123–1130.

[10] Mao, C.X. (2007). Estimating population sizes for capture-recapture sampling with binomial mixtures. *Computational Statistics & Data Analysis* 51, 5211–5219.

[11] Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monograph* 62, 1–135.

[12] Patil, G.P. (1962). Maximum likelihood estimation for generalized power series distributions and its application to a truncated binomial distribution. *Biometrika* 49, 227–237.

[13] Pledger, S., Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal* 50, 1022–1034.

[14] Pollock, K.H. (1974). *The Assumption of Equal Catchabi1ity of Animals in Tag-recapture Experiments.* Ph.D. Thesis, Cornell University, Ithaca, New York 14850.

[15] Pollock, KH. (2002). The use of auxiliary variables in capture-recapture modelling: An overview. *Journal of Applied Statistics* 29, 85–102.

[16] Qaqish, B.F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 90, 455–463.

[17] R Core Team, (2014).*R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org/.

[18] Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* 43, 142–152.

[19] Stanley, T.R., Richards, J.D.(2005). Software review: a program for testing capture-recapture data for closure. *Wildlife Society Bulletin* 33, 782–785.

[20] Stoklosa, J., Huggins, R.M. (2012). A robust P-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics & Data Analysis* 56, 408–417.

[21] Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman and Hall/CRC.

[22] Williams, B.K., Nichols, J.D., Conroy, M.J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, California.

# Intervalos de amostragem adaptativos inicialmente predefinidos para um risco cumulativo constante

Manuel do Carmo
Universidade Europeia, Laureate International Universities,
CIMA-UE, *manuel.carmo@europeia.pt*

Paulo Infante
CIMA-UE e DMAT, ECT da Universidade de Évora,
*pinfante@uevora.pt*

Jorge M. Mendes
ISEGI-NOVA,Universidade Nova de Lisboa,
CEAUL-FCUL, *jmm@isegi.unl.pt*

**Palavras–chave**: Método de amostragem adaptativa e predefinida, risco cumulativo constante, controlo da qualidade

**Resumo**: Neste trabalho, analisamos a eficiência estatística de um novo método de amostragem, num contexto de controlo da qualidade, que combina dois outros métodos: um método que define os instantes de amostragem em função da estatística amostral (adaptativo), e outro cujos parâmetros são definidos no início do processo, mas que permanecem constantes ao longo do mesmo (predefinido). Desta forma, os instantes de amostragem, inicialmente calendarizados de acordo com as expectativas de ocorrência de uma alteração, tomando como base a distribuição do tempo de vida do sistema, são adaptados em função do valor da estatística amostral calculada no instante anterior.
Utilizando cartas de controlo para a média do tipo *Shewhart*, efectuamos uma análise do desempenho estatístico do método apresentado, considerando diferentes alterações da média, diferentes valores para a dimensão amostral e dois pesos para os instantes de amostragem. Os resultados obtidos, nesta fase por simulação, permitem concluir

que este método apresenta um bom desempenho estatístico para diferentes alterações da média (inclusivé grandes alterações) quando comparado com outros esquemas amostrais, em termos do tempo médio de mau funcionamento do sistema.

# 1 Introdução

Qualquer organização, de produtos ou serviços, deve fazer uma avaliação rigorosa da qualidade dos bens ou serviços que disponibiliza ao consumidor, recorrendo a técnicas estatísticas adequadas. Para monitorizar a qualidade associada a esses bens ou serviços, a carta de controlo é uma das ferramentas mais utilizadas, possibilitando a distinção entre variação intrínseca ao processo e variação com origem em causas externas.
Neste contexto, são vários os estudos que mostram que uma carta de médias com parâmetros adaptativos tem melhor desempenho, em particular, quando as alterações na média são reduzidas e moderadas. Assim, ao longo dos tempos têm sido apresentados, e analisados, diversos métodos de amostragem cujos parâmetros variam em função da estatística amostral (adaptativos) (p.e., [2], [4] e [5]) e outros cujos parâmetros são definidos no início do controlo do processo, mas que permanecem constantes ao longo do mesmo (predefinidos) (p.e., [1] e [6]).
Considerando uma carta de controlo para a média, em [6] foi estudado um método de amostragem cujos instantes são definidos, no início do controlo do processo, de modo a que a taxa cumulativa de risco, definida em 2.1, seja constante entre dois quaisquer instantes consecutivos, denominado PSI (*"Predetermined Sampling Intervals"*). Os autores mostram que o método PSI é sempre mais eficaz que o método periódico clássico, sendo tanto mais eficaz quanto menos amostras são recolhidas no período de controlo, quanto menor for a magnitude da alteração e quanto mais acentuadamente crescente for a taxa de risco do sistema. Quando comparado com outros métodos, revelou-se igualmente eficaz a detectar reduzidas e grandes

alterações da média, em particular para sistemas com taxas de risco crescente.

Em [7], é apresentado um método de amostragem que combina um método adaptativo, no qual os instantes de amostragem são obtidos segundo a função densidade da distribuição normal standard, com o método PSI. Neste método, os instantes de amostragem são definidos pela média ponderada entre os instantes obtidos com o método adaptativo e o método PSI. Considerando o uso simultâneo de carta para médias e carta para amplitudes, os resultados obtidos, em termos de AATS (*"Adjusted Average Time to Signal"*), mostraram um elevado potencial do método, quando comparado com os métodos FSI (*"Fixed Sampling Intervals"*) e VSI (*"Variable Sampling Intervals"*, [5]).

Em [2] foi apresentado e analisado um método de amostragem adaptativo. Neste método, denominado LSI (*"Laplace Sampling Intervals"*), os instantes de amostragem são actualizados ao longo do processo, dependendo da informação recolhida na amostra anterior, segundo a função densidade da distribuição de *Laplace* standard. O método apresenta um bom desempenho, em particular para alterações moderadas da média, quando comparado com os métodos FSI e VSI.

Neste trabalho, utilizando uma carta de controlo para a média, analisamos a eficiência estatística de um novo método que define os instantes de amostragem com base numa média ponderada dos métodos PSI e LSI, dando maior peso ao método LSI para alterações moderadas (onde PSI é menos eficaz) e maior peso ao método PSI nos restantes casos (onde LSI é menos eficaz). Desta forma, os instantes de amostragem, inicialmente calendarizados de acordo com as expectativas de ocorrência de uma alteração tomando como base a distribuição do tempo de vida do sistema, são adaptados em função do valor da estatística amostral calculada no instante anterior.

Quando os métodos envolvem apenas instantes de amostragem adaptativos, o desempenho estatístico é, usualmente, medido pelo tempo médio de mau funcionamento do sistema, AATS, o que se faz neste trabalho. Outras medidas de desempenho estatístico podiam ser

usadas, por exemplo o ANSIC ("*Average Number of Sample In Control*"), apresentada, e utilizada, numa perspectiva económica estatística em [3].
Na secção seguinte, fazemos uma breve descrição dos métodos LSI e PSI e apresentamos algumas propriedades do novo método. Posteriormente, recordamos em que consistem os métodos FSI e VSI e comparamos, em termos de AATS, o desempenho do novo método com o desempenho dos métodos FSI, VSI, PSI e LSI.
Por fim, são efectuadas algumas considerações e apresentadas ideias de trabalho em curso ou a desenvolver no futuro.

# 2 Novo método de amostragem: CAPSI

Nesta secção descrevem-se algumas das principais características do novo método e dos métodos de amostragem que lhe servem como suporte.
Sejam $\mu_0$ e $\sigma_0$, respectivamente, a média e o desvio padrão da característica da qualidade $X$, que se admite ter distribuição aproximadamente normal.

Seja, no método LSI, $t_i$, $i = 1,2,\ldots$, o instante de amostragem de ordem $i$ e $\bar{x}_i$ a média da amostra analisada nesse instante. De acordo com o método LSI, o próximo instante de amostragem (ordem *i+1*) é definido por

$$t_{i+1} = t_i + \frac{k \cdot e^{-|u_i|}}{2}, \tag{1}$$

onde $u_i = \frac{\bar{x}_i - \mu_0}{\sigma_0}\sqrt{n}$, $t_0 = 0$, $\bar{x}_0 = \mu_0$ e $-L < u_i < L$, representando $n$ a dimensão da amostra, $k$ uma constante conveniente de escala, dependente, em particular, de custos associados ao processo produtivo e $L$ o múltiplo nos limites de controlo.
Sendo $u_i$ a média amostral reduzida, quando $|u_i| > L$ estamos numa situação de fora de controlo ou de falso alarme. Assim, os intervalos de amostragem, $d_i = t_i - t_{i-1} = k \cdot l(u_{i-1})$, $i = 1,2,3,\ldots$, onde $l(\cdot)$ é a f.d.p. da distribuição de *Laplace* reduzida, são i.i.d. com a mesma

distribuição da uma variável genérica $D$, definida por

$$D = t_{i+1} - t_i = \frac{k \cdot e^{-|u_i|}}{2}. \tag{2}$$

A ideia implícita ao método, adaptativo, é diminuir a frequência de amostragem quando as médias estão próximas da linha central e aumentá-la quando é maior a probabilidade de alteração da qualidade. Na prática, ao contrário de outros métodos adaptativos, necessitamos apenas de determinar a constante de escala $k$ (considerando os limites de controlo fixos). Considerando os pressupostos para (1) e (2), uma carta para médias, e que, após alteração do processo, $\mu_0$ e $\sigma_0$ podem assumir valores $\mu_1 = \mu_0 \pm \lambda\sigma_0$ e $\sigma_1 = \rho\sigma_0$, onde $\lambda$ e $\rho$ são, respectivamente, os coeficientes da alteração da média e do desvio padrão, obtemos para intervalo médio de amostragem, *E(D)*, a expressão

$$E(D|\lambda,\rho,n) = \frac{k}{2\beta}\left[e^{\lambda\sqrt{n}+\frac{\rho^2}{2}} \cdot A(L,\lambda,\rho,n) + e^{-\lambda\sqrt{n}+\frac{\rho^2}{2}} \cdot B(L,\lambda,\rho,n)\right],$$

onde $\beta$ é a probabilidade de cometer um erro de tipo II,

$$A(L,\lambda,\rho,n) = \Phi\left(\frac{-\rho^2 - \lambda\sqrt{n}}{\rho}\right) - \Phi\left(\frac{-L - \rho^2 - \lambda\sqrt{n}}{\rho}\right),$$

$$B(L,\lambda,\rho,n) = \Phi\left(\frac{L + \rho^2 - \lambda\sqrt{n}}{\rho}\right) - \Phi\left(\frac{\rho^2 - \lambda\sqrt{n}}{\rho}\right)$$

e $\Phi(u)$ é a função distribuição da normal reduzida. A expressão (3) é função de $n$, do coeficiente dos limites de controlo $L$, de $\lambda$ e de $\rho$, mas não depende, directamente, dos valores da média nem do desvio padrão da qualidade. Considerando o processo sob controlo, $\lambda = 0$ e $\rho = 1$, e igualando (3) ao intervalo fixo (sem perda de generalidade,

$d = 1$ em *FSI*), obtemos $k$ dado por

$$k = \frac{\beta}{e^{1/2}\ [\Phi(L+1) - \Phi(1)]}, \tag{3}$$

sendo o seu valor igual a 3,8134 quando *L=3* (usuais limites "*3-sigma*") e $\beta = 0{,}9973$. Desta forma o método fica completamente definido, podendo-se consultar a expressão que nos permite obter o $AATS_{LSI}$ em [2].

Considere-se um sistema cujo tempo de vida é uma variável aleatória $T$ com função densidade de probabilidade f(t) contínua e função distribuição F(t).
Define-se taxa cumulativa de risco do sistema H(t) através da relação $H(t) = -lnR(t)$, onde R(t) é a função de fiabilidade do sistema.
De acordo com o método PSI, os instantes de amostragem $t_i$, $i =$ 0,1,2,..., com $t_0 = 0$, são determinados pela relação

$$H(t_i) = i\Delta H, \tag{4}$$

obtendo-se $H(t_{i+1}) - H(t_i) = \Delta H$.
Assim, os instantes de amostragem $t_i$ são determinados de modo a que a taxa cumulativa de risco entre quaisquer intervalos de amostragem consecutivos seja constante, ou seja, que a probabilidade de ocorrência de uma falha do processo num intervalo de amostragem, condicionada ao facto de nenhuma falha ter ocorrido até ao início desse intervalo, é constante para todos os intervalos. Considerando a definição de taxa cumulativa de risco e (3), os instantes de amostragem, em PSI, são dados por

$$t_i = R^{-1}\left(e^{-i\Delta H}\right), \tag{5}$$

com $t_0 = 0$. Quando o tempo de vida do sistema segue uma distribuição de *Weibull*, os instantes de inspecção são definidos por

$$t_i = \alpha\ (i\Delta H)^{\frac{1}{\delta}}, i = 1,2,\ldots, \tag{6}$$

onde $\alpha$ é parâmetro de escala e $\delta$ o parâmetro de forma da distribuição de *Weibull*, sendo os intervalos de amostragem definidos pela expressão

$$\Delta t_i = \left[ i^{\frac{1}{\delta}} - (i-1)^{\frac{1}{\delta}} \right] t_1, i = 1,2,\ldots, \tag{7}$$

obtida em [1].
Considerando o mesmo número médio de amostras recolhidas, sob controlo, em FSI e em PSI, em [6] foi considerada uma aproximação para o parâmetro $\Delta H$, de modo a que PSI fique completamente definido, dada por

$$\Delta H = \frac{d}{E(T)}, \tag{8}$$

onde $d$ é o intervalo de amostragem em FSI e E(T) o tempo médio de vida do sistema. A expressão para calcular o tempo médio de mau funcionamento do sistema, em PSI, é dada em [6].

O novo método de amostragem, CAPSI, combina os intervalos de amostragem definidos pelos métodos LSI e PSI.
Designem-se por $t_i^{LSI}$ os instantes de amostragem obtidos com o método LSI, dados por (1), e por $t_i^{PSI}$ os instantes de amostragem obtidos com o método PSI, dados por (4).
De acordo com o método combinado proposto, denominado CAPSI (*"Combined Adaptive and Predetermined Sampling Intervals"*), o instante de amostragem de ordem *i+1* é dado por

$$\begin{aligned} t_{i+1} &= \theta\, t_{i+1}^{LSI} + (1-\theta)\, t_{i+1}^{PSI} \\ &= \theta \left[ t_i^{LSI} + k\, l(u_i) \right] + (1-\theta) R^{-1} \left[ e^{-\Delta H(i+1)} \right], \end{aligned} \tag{9}$$

com

$$t_0 = 0,\ t_1 = \theta\, \frac{k}{2} + (1-\theta)\, R^{-1} \left( e^{-\Delta H} \right),\ 0 \le \theta \le 1, \tag{10}$$

onde $l(\cdot)$ é a f.d.p. da distribuição de *Laplace* reduzida e $\theta$ o peso atribuído ao instante de amostragem do método LSI.

Então, os intervalos de amostragem são definidos pela expressão

$$\Delta t_i = \theta \, \frac{k \, e^{-|u_i|}}{2} + (1-\theta) \left[ i^{\frac{1}{\delta}} - (i-1)^{\frac{1}{\delta}} \right] \alpha \big(\Delta H\big)^{\frac{1}{\delta}}, 0 \leq \theta \leq 1, \quad (11)$$

com

$$\Delta t_1 = \theta \, \frac{k}{2} + (1-\theta) \, \alpha \left(\Delta H\right)^{\frac{1}{\delta}}, 0 \leq \theta \leq 1. \quad (12)$$

Assim, o método que propomos permite melhorar as características menos boas de LSI (eficácia em reduzidas e grandes alterações) e de PSI (eficácia em moderadas alterações), tornando-se numa alternativa aos vários métodos existentes na literatura. Pela sua simplicidade, porque só depende dos parâmetros $k$ e $\Delta H$ e, em particular, pelos valores obtidos para os menores intervalos de amostragem (valores próximos de 0.1 que é muito utilizado nos métodos de intervalos adaptativos), estamos convictos da mais valia do seu contributo.
Nesta fase, os resultados de $AATS_{CAPSI}$ foram obtidos por simulação, mas estamos a trabalhar na obtenção de expressões algébricas que permitam obter o $AATS_{CAPSI}$, bem como de outras propriedades estatísticas do método.

# 3 Avaliação do desempenho de CAPSI

## 3.1 Os métodos FSI e VSI

No método FSI retiram-se amostras em instantes fixos, sendo a dimensão amostral e os múltiplos do desvio padrão, nos limites de controlo, também fixos. Em [6] é apresentada uma aproximação para obter o tempo médio de mau funcionamento deste esquema de amostragem periódico.
Em [5] é proposto o método VSI. Considerando dois intervalos de amostragem $d_1$ e $d_2$ ($d_1 < d < d_2$), onde $d$ representa o intervalo de amostragem do método periódico clássico, a região de continuação é dividida em duas sub-regiões, $]-w, w[$ e $]-L, -w] \cup [w, L[$, e o método permite antecipar ou retardar a recolha da amostra seguinte. Quando o intervalo médio de amostragem, em VSI, é igual

a $d$, sob controlo, [8] apresentam uma expressão para obter $w$ e [5] uma expressão para obter $AATS_{VSI}$. O método VSI é particularmente eficaz em detectar alterações reduzidas da média.

## 3.2 Comparação do método CAPSI com os métodos FSI, PSI e LSI

Para comparar a eficácia dos métodos, em termos de AATS, consideramos as expressões dadas em [2], [5] e [6] e os valores obtidos, por simulação, para o método CAPSI, tomando os métodos nas mesmas condições sob controlo, com $E(D) = 1$ e $L = 3$. Consideram-se, também, que o tempo de vida do sistema segue uma distribuição de *Weibull* com $E(T) = 1000$, taxas de risco crescente e o rácio $Q_{CAPSI/MB}$ que representa a variação relativa, em %, do $AATS_{CAPSI}$ relativamente ao AATS de um dos outros métodos (MB na expressão pode representar FSI, VSI, PSI ou LSI), dado por

$$Q_{CAPSI/MB} = \frac{AATS_{MB} - AATS_{CAPSI}}{AATS_{MB}} \times 100\%. \qquad (13)$$

Dos resultados obtidos optámos, devido a limitações de espaço, por apenas apresentar duas dimensões amostrais e dois pesos ($\theta$) dos intervalos de LSI.

Assim, da Tabela 1 podemos concluir que: **a)** CAPSI é sempre mais eficaz do que FSI, melhorando a eficácia quando aumenta a taxa de risco do sistema; **b)** quando aumentamos a dimensão amostral (para $n = 9$), o método melhora o desempenho, relativamente a FSI, para $\lambda = 0.5$ e piora para os restantes valores de $\lambda$; **c)** quando $n = 5$, o método CAPSI é sempre mais eficaz do que o método PSI para $\lambda \geq 1$; quando $n = 9$, CAPSI é sempre melhor do que PSI para $\lambda \leq 1.5$; a eficácia de CAPSI aumenta quando aumentamos $\delta$ até 4, mas decresce para valores superiores; **d)** o método CAPSI é sempre melhor do que o método LSI quando $\delta \geq 3$; quando aumentamos $n$, os valores do rácio aumentam à medida que a taxa de risco é mais acentuadamente crescente e para $\lambda \geq 1$, mas diminuem

quando $\lambda = 0.5$.

| $E(T) = 1000$ | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0.6$ | | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| $(n, \rho)$ | $\lambda$ | $Q_{CAPSI/FSI}$ | | | | $Q_{CAPSI/PSI}$ | | | | $Q_{CAPSI/LSI}$ | | | |
| (5, 1) | 0,5 | 27,0 | 35,4 | 42,1 | 45,4 | 13,8 | 8,1 | 2,4 | -6,7 | 3,2 | 14,3 | 23,2 | 27,6 |
| | 1,0 | 40,7 | 50,4 | 56,1 | 60,0 | 37,0 | 41,3 | 41,6 | 40,6 | -19,6 | 0,0 | 11,4 | 19,3 |
| | 1,5 | 30,3 | 40,2 | 45,6 | 48,9 | 28,2 | 33,9 | 34,2 | 32,8 | -0,5 | 13,8 | 21,5 | 26,4 |
| | 2,0 | 9,0 | 18,1 | 23,2 | 26,0 | 7,4 | 12,1 | 11,2 | 7,3 | 16,1 | 24,6 | 29,3 | 31,8 |
| (9, 1) | 0,5 | 33,2 | 43,6 | 48,7 | 51,6 | 28,4 | 25,6 | 24,3 | 18,4 | -5,9 | 10,5 | 18,5 | 23,2 |
| | 1,0 | 36,8 | 46,1 | 52,0 | 55,1 | 34,6 | 38,4 | 43,4 | 39,1 | -8,5 | 7,3 | 17,6 | 22,8 |
| | 1,5 | 9,5 | 17,5 | 21,6 | 25,6 | 7,8 | 10,7 | 13,7 | 10,5 | 17,1 | 24,5 | 28,2 | 31,9 |
| | 2,0 | 1,4 | 9,3 | 15,3 | 18,6 | -1,2 | 3,7 | 7,6 | -5,4 | 19,3 | 25,9 | 30,7 | 33,4 |

Tabela 1: Valores de $Q_{CAPSI/MB}$, em função de $\lambda$, com $\theta = 0.6$.

Quando reduzimos o peso dos intervalos de LSI, Tabela 2, podemos concluir que: **a)** CAPSI continua a ser, sempre, mais eficaz do que FSI e os valores do rácio aumentam quando aumenta a taxa de risco do sistema; a eficácia de CAPSI, relativamente a FSI, diminui com o aumento da dimensão amostral para $\lambda \geq 1$ e aumenta para $\lambda = 0.5$;

| $E(T) = 1000$ | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0.4$ | | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| $(n, \rho)$ | $\lambda$ | $Q_{CAPSI/FSI}$ | | | | $Q_{CAPSI/PSI}$ | | | | $Q_{CAPSI/LSI}$ | | | |
| (5, 1) | 0,5 | 24,6 | 36,3 | 47,3 | 52,2 | 10,9 | 9,4 | 11,2 | 6,6 | 0,0 | 15,6 | 30,1 | 36,6 |
| | 1,0 | 33,0 | 48,7 | 55,4 | 61,6 | 28,8 | 39,3 | 40,7 | 43,0 | -35,1 | -3,5 | 10,0 | 22,6 |
| | 1,5 | 24,9 | 38,9 | 49,8 | 55,1 | 22,7 | 32,5 | 39,4 | 40,8 | -8,2 | 11,9 | 27,6 | 35,2 |
| | 2,0 | 11,1 | 23,6 | 33,1 | 38,6 | 9,5 | 18,0 | 22,6 | 23,1 | 18,1 | 29,7 | 38,3 | 43,4 |
| (9, 1) | 0,5 | 29,1 | 42,2 | 51,3 | 56,5 | 24,0 | 23,8 | 28,2 | 26,7 | -12,5 | 8,2 | 22,8 | 31,0 |
| | 1,0 | 30,3 | 43,2 | 50,6 | 58,3 | 27,8 | 35,0 | 41,6 | 43,5 | -19,8 | 2,3 | 15,1 | 28,3 |
| | 1,5 | 12,0 | 22,8 | 34,1 | 38,4 | 10,4 | 16,4 | 27,5 | 25,9 | 19,4 | 29,3 | 39,6 | 43,6 |
| | 2,0 | 6,6 | 16,7 | 28,3 | 33,8 | 4,1 | 11,5 | 21,7 | 14,3 | 23,6 | 31,9 | 41,3 | 45,9 |

Tabela 2: Valores de $Q_{CAPSI/MB}$, em função de $\lambda$, com $\theta = 0.4$.

**b)** o método CAPSI é sempre mais eficaz do que o método PSI; o aumento da dimensão amostral diminui os valores do rácio quando $\lambda \geq 1$, excepto quando $\lambda = 1$ e $\delta = 4$ ou $\delta = 5$, mas aumenta os valores do rácio quando $\lambda = 0.5$; **c)** CAPSI é sempre mais eficaz do que LSI quando $\delta = 4$ ou $\delta = 5$; o aumento da dimensão amostral melhora a eficácia do método CAPSI quando $\lambda \geq 1.5$, mas reduz ligeiramente a sua eficácia quando $\lambda \leq 1$ e $\delta \leq 3$; em geral, a eficácia de CAPSI, relativamente a LSI, aumenta com a dimensão da amostra para $\lambda \geq 1$; **d)** as conclusões retiradas, a partir das duas tabelas,

estão de acordo com as expectativas, pois refletem os pressupostos em que assentam os métodos base e o peso atribuído aos intervalos de LSI.

## 3.3 Comparação do método CAPSI com o método VSI

Na comparação do desempenho de CAPSI com VSI, consideramos quatro pares de intervalos de amostragem em VSI e as mesmas condições das comparações anteriores. Assim, no rácio (12) substituimos MB (método base de comparação) por VSI. Os resultados obtidos são apresentados na Tabela 3, a partir da qual podemos concluir que: **a)** a eficácia do método CAPSI, relativamente a VSI, aumenta com a taxa de risco do sistema; **b)** quando $d_1 = 0.5$ em VSI, o método CAPSI é sempre mais eficaz; **c)** em geral, o desempenho de CAPSI melhora, relativamente a VSI, para $\lambda \geq 1.5$ com o aumento da dimensão da amostra.

| $E(T) = 1000$ | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0.6$ | | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| $(n, \rho)$ | $\lambda$ | $(d_1, d_2) = (0.1, 2)$ | | | | $(d_1, d_2) = (0.1, 1.5)$ | | | | $(d_1, d_2) = (0.5, 2)$ | | | | $(d_1, d_2) = (0.5, 1.5)$ | | | |
| (5, 1) | 0,5 | -11,5 | 1,3 | 11,5 | 16,6 | -6,0 | 6,2 | 15,8 | 20,7 | 8,4 | 18,9 | 27,3 | 31,5 | 9,6 | 20,0 | 28,2 | 32,4 |
| | 1,0 | -35,7 | -13,5 | -0,6 | 8,4 | -42,8 | -19,5 | -5,8 | 3,6 | 12,6 | 26,9 | 35,2 | 41,0 | 9,9 | 24,6 | 33,2 | 39,2 |
| | 1,5 | 26,8 | 37,2 | 42,9 | 46,4 | 6,8 | 20,1 | 27,2 | 31,8 | 28,3 | 38,5 | 44,0 | 47,5 | 18,6 | 30,2 | 36,5 | 40,4 |
| | 2,0 | 45,3 | 50,8 | 53,8 | 55,5 | 28,5 | 35,7 | 39,7 | 41,8 | 33,5 | 40,2 | 43,9 | 45,9 | 20,9 | 28,9 | 33,3 | 35,7 |
| (9, 1) | 0,5 | -31,3 | -11,0 | -1,0 | 4,8 | -23,4 | -4,3 | 5,1 | 10,5 | 6,5 | 20,9 | 28,0 | 32,2 | 7,5 | 21,8 | 28,8 | 32,9 |
| | 1,0 | 12,5 | 25,3 | 33,6 | 37,8 | -8,0 | 7,8 | 18,0 | 23,2 | 25,1 | 36,0 | 43,1 | 46,7 | 17,2 | 29,3 | 37,2 | 41,1 |
| | 1,5 | 45,9 | 50,8 | 53,2 | 55,6 | 29,4 | 35,6 | 38,8 | 42,0 | 34,2 | 40,0 | 42,9 | 45,9 | 21,7 | 28,7 | 32,2 | 35,7 |
| | 2,0 | 48,0 | 52,2 | 55,3 | 57,0 | 31,8 | 37,3 | 41,5 | 43,7 | 34,1 | 39,5 | 43,4 | 45,6 | 21,0 | 27,4 | 32,2 | 34,8 |

Tabela 3: Valores de $Q_{CAPSI/VSI}$, em função de $\lambda$, com $\theta = 0.6$.

Quando reduzimos $\theta$, aumentamos o peso dos intervalos de PSI. Este facto, tal como anteriormente, pode ter alguma influência na eficácia do método CAPSI. Com $\theta = 0.4$, os resultados são apresentados na Tabela 4, a partir dos quais podemos concluir que: **a)** em geral, os valores do rácio diminuem quando diminui $d_2$ em VSI (tal já acontecia anteriormente); **b)** em todas as situações, os valores do rácio aumentam quando aumenta a taxa de risco do sistema; **c)**

em geral, quando aumenta a dimensão amostral a eficácia de CAPSI aumenta; **d)** no geral, a redução do peso dos intervalos de LSI afecta, de forma positiva mas ligeira, o desempenho do método CAPSI, reflectindo-se, em particular, em sistemas com uma taxa de risco acentuadamente crescente ($\delta \geq 4$).

| $E(T) = 1000$ | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | | $\delta$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0.4$ | | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| $(n, \rho)$ | $\lambda$ | $(d_1, d_2) = (0.1, 2)$ | | | | $(d_1, d_2) = (0.1, 1.5)$ | | | | $(d_1, d_2) = (0.5, 2)$ | | | | $(d_1, d_2) = (0.5, 1.5)$ | | | |
| (5, 1) | 0,5 | -15,2 | 2,7 | 19,4 | 26,9 | -9,5 | 7,5 | 23,4 | 30,5 | 5,3 | 20,0 | 33,8 | 39,9 | 6,6 | 21,1 | 34,7 | 40,7 |
| | 1,0 | -53,3 | -17,4 | -2,1 | 12,2 | -61,4 | -23,6 | -7,4 | 7,6 | 1,2 | 24,4 | 34,3 | 43,5 | -1,8 | 22,0 | 32,2 | 41,7 |
| | 1,5 | 21,2 | 35,9 | 47,3 | 52,8 | -0,3 | 18,3 | 32,9 | 39,9 | 22,8 | 37,2 | 48,4 | 53,8 | 12,4 | 28,7 | 41,4 | 47,5 |
| | 2,0 | 46,5 | 54,1 | 59,7 | 63,1 | 30,1 | 40,0 | 47,4 | 51,8 | 35,0 | 44,2 | 51,1 | 55,1 | 22,8 | 33,7 | 41,9 | 46,7 |
| (9, 1) | 0,5 | -39,4 | -13,8 | 4,3 | 14,5 | -31,0 | -6,9 | 10,1 | 19,6 | 0,7 | 19,0 | 31,8 | 39,1 | 1,8 | 19,8 | 32,5 | 39,7 |
| | 1,0 | 3,5 | 21,3 | 31,6 | 42,2 | -19,2 | 2,8 | 15,5 | 28,6 | 17,3 | 32,6 | 41,4 | 50,5 | 8,6 | 25,5 | 35,2 | 45,3 |
| | 1,5 | 47,4 | 53,9 | 60,6 | 63,2 | 31,3 | 39,7 | 48,5 | 51,9 | 36,0 | 43,8 | 52,0 | 55,2 | 23,9 | 33,2 | 43,0 | 46,7 |
| | 2,0 | 50,7 | 56,1 | 62,1 | 65,1 | 35,4 | 42,4 | 50,4 | 54,2 | 37,6 | 44,4 | 52,1 | 55,8 | 25,2 | 33,3 | 42,5 | 47,0 |

Tabela 4: Valores de $Q_{CAPSI/VSI}$, em função de $\lambda$, com $\theta = 0.4$.

# 4 Considerações finais

Este trabalho apresenta um novo método de amostragem em controlo da qualidade o qual adapta, em função do valor da estatística amostral e baseando-se na função densidade da distribuição de *Laplace*, os instantes previamente calendarizados de forma a que a taxa cumulativa de risco do sistema permaneça constante entre dois quaisquer instantes de amostragem consecutivos. Apenas para alterações moderadas da média o novo método tem menor eficácia do que o método VSI. Para melhorar este aspecto, pensamos considerar o peso ($\theta$), atribuído aos instantes de amostragem do método LSI, como uma função da probabilidade de ocorrência de uma alteração da média ($\lambda$) e determinar qual o peso óptimo que minimiza um custo total médio por unidade de tempo.

Pensamos também realizar estudos comparativos com outros esquemas adaptativos, estender a sua aplicação à utilização simultânea de uma carta para a média e de uma carta para o desvio padrão e analisar a robustez deste método quando a característica da qualidade a ser monitorizada se afasta da distribuição normal.

## Agradecimentos

Os dois primeiros autores são membros do CIMA-U.E., centro de investigação financiado pelo Programa FEDER e por financiamentos plurianuais da FCT.

## Referências

[1] Banerjee, P.K., Rahim, M.A. (1988). Economic design of $\overline{X}$ control charts under Weibull shock models. *Technometrics* 30, 407–414.

[2] Carmo, M., Infante, P., Mendes, J.M. (2013). Alguns resultados da robustez de um método de amostragem adaptativo em controlo de qualidade. In Maia, M., Campos, P. e Silva, P. D. (Eds.), *Estatística: Novos Desenvolvimentos e Inspirações*, SPE, 95–108.

[3] Carmo, M., Infante, P., Mendes, J.M. (2014). A different and simple approach for comparing sampling methods in quality control. *International Journal Quality & Reliability Management* 31, 478–499.

[4] Reynolds, M.R. (1996). Variable sampling interval control charts with sampling at fixed times. *IIE Transactions* 28, 497–510.

[5] Reynolds, M.R., Amin, R.W., Arnold, J.C., Nachlas, J.A. (1988). $\overline{X}$ charts with variables sampling intervals. *Technometrics* 30, 181–192.

[6] Rodrigues Dias, J., Infante P. (2008). Control charts with predetermined sampling intervals. *International Journal of Quality and Reliability Management* 25, 423–435.

[7] Rosmaninho, E., Infante P. (2007). Métodos de Amostragem com Parâmetros Predefinidos Adaptáveis: Uma análise estatística e económica, In Ferrão, M. E., Nunes, C. e Braumann, C. A. (Eds.), *Estatística: Ciência Interdisciplinar*, SPE, 659–708.

[8] Runger, G.C., Pignatiello, J.J. (1991). Adaptive sampling for process control. *Journal of Quality Technology* 23, 133–155.

# Modelo Bayesiano de equações simultâneas para a estimação dos parâmetros da área basal e da mortalidade

Marco Marto
Universidade de Aveiro, *marcovmarto@gmail.com*
Isabel Pereira
Universidade de Aveiro, CIDMA, *isabel.pereira@ua.pt*
Margarida Tomé
Instituto Superior de Agronomia, CEF, *magatome@isa.utl.pt*

**Palavras–chave**: Estimação NSUR, métodos MCMC, modelo globulus

**Resumo**: Este estudo apresenta uma alternativa bayesiana para a estimação clássica dos modelos de equações simultâneas, considerando uma metodologia NSUR com recurso a métodos de Monte Carlo baseados em cadeias de Markov, mais concretamente o algoritmo de Gibbs com o passo de Metropolis. Desta forma generaliza-se o trabalho anteriormente feito por Marto [3], comparando os resultados de estimação obtidos pelas abordagens clássica e bayesiana e analisando qual o melhor modelo preditivo. A validação do modelo é feita usando como medida de desvio a raíz quadrada do erro quadrático médio.

## 1 Introdução

Os modelos de crescimento florestais para a espécie *Eucalyptus globulus* têm sido objeto de estudo e aperfeiçoamento pelo Centro de Estudos Florestais do Instituto Superior de Agronomia desde a década de 70 (Tomé, Ribeiro e Soares [4], [5]). Foram utilizados métodos de sistemas de equações simultâneas, nomeadamente o *Nonlinear Seemingly Unrelated Regression*- NSUR para possibilitar a estima-

ção de parâmetros nas equações não lineares usadas para simular a evolução das variáveis biológicas e selecionadas para variáveis de estado do modelo, por exemplo a área basal ($G$) e o número $N$ de árvores por hectare (ha). No contexto da metodologia bayesiana Zellner [6] popularizou a inferência bayesiana e descreveu o SUR. De entre vários trabalhos posteriores pode destacar-se Griffiths [2].

# 2 Modelo *globulus 2.1*

Comece-se por definir o modelo *globulus 2.1*, considerando apenas dados de primeira rotação, não diferenciando o ajustamento por regiões climáticas em Portugal e considerando apenas espaços de tempo unitários. Cada modelo considera dois módulos, o de inicialização e o de projeção. O módulo de projeção é constituído por um conjunto de equações que atualizam o valor das variáveis de estado no instante $t_i$ para o instante $t_{i+1}$. Uma das funções que tem sido usada para modelar o crescimento de plantações de eucaliptos é a função de Lundqvist, definida por

$$y = A\mathrm{e}^{-kt^{-n}},$$

onde o parâmetro $A$ representa a assíntota superior do crescimento e $k$ o parâmetro com a taxa de crescimento (declive da curva).

## 2.1 Modelo de mortalidade

Este modelo baseia-se no modelo SOP de Amaro [1]. A variável de mortalidade $N_t$ representa o número de árvores por ha na parcela florestal. Nalguns casos temos a primeira medição da variável (inicialização) que ocorre a uma dada idade ($t$) e nos restantes casos o modelo usa o valor estimado para as variáveis no ano anterior (projeção com base em equações às diferenças), daí a necessidade das equações de inicialização e de projeção.

**Equação de inicialização**

$$N_t = N_{pl}\mathrm{e}^{-amnp\frac{N_{pl}}{1000}t}, t \in \mathbb{N},$$

onde $N_t$, o número de árvores no povoamento, é a variável a explicar em função de $N_{pl}$, número de árvores por ha à plantação e de $t$.

**Equação de projeção**

$$N_t = N_{t-1}\mathrm{e}^{-amnp\frac{N_{pl}}{1000}}, t \in \mathbb{N}, t > 1.$$

O parâmetro a estimar no modelo de mortalidade é *amnp*.

## 2.2 Modelo de área basal

A área basal é medida pela soma das áreas das árvores da parcela a 1,30 m de altura em $m^2$/ha. Para a equação de inicialização foi escolhida a função de Lundqvist na sua forma integral com os parâmetros $A_g$ e $n_g$ iguais aos da respetiva função de crescimento e com o parâmetro $k_g$ expresso em função de um conjunto razoável de variáveis de controlo e suas interações; enquanto que para a equação de projeção foi usada a função de Lundqvist-k com o parâmetro de forma $n_g$.

**Equação de inicialização**

$$G_t = A_g\mathrm{e}^{-k_g t^{-n_g}}, t \in \mathbb{N},$$

com

$$A_g = A_{gq}(Iqe)^2, k_g = k_{gq}(Iqe)^{-1}, n_g = n_{g0} + n_{gq}\ln(Iqe) + n_{gn}\frac{N_t}{1000},$$

onde a variável endógena $G_t$ representa a área basal do povoamento e $Iqe$ é o índice de qualidade da estação para aquela parcela florestal.

**Equação de projeção**

$$G_t = A_g \left( \frac{G_{t-1}}{A_g} \right)^{\frac{(t-1)^{n g_{t-1}}}{t^{n g_t}}}, t \in \mathbb{N}, t > 1,$$

com
$A_g = A_{gq}(Iqe)^2, n_g = n_{g0} + n_{gq}\ln(Iqe) + n_{gn}\frac{N_t}{1000}$.

Os parâmetros a estimar no modelo da área basal são $A_{gq}$, $k_{gq}$, $n_{g0}$, $n_{gq}$ e $n_{gn}$.

# 3 Inferência bayesiana

Considera-se o modelo *NSUR* caracterizado por

$$y_j = h_j(\mathbf{X}; \beta) + \varepsilon_j, j = 1, \ldots, l,$$

onde $h_j$ são as funções não lineares, $\mathbf{X}$ representa o conjunto de variáveis explicativas e $\beta$ o vetor dos parâmetros desconhecidos. Aplicando ao caso em estudo $l = 2$ então as dimensões de $y_j$,$h_j$ e $\varepsilon_j$ são $n \times 1$. Denote-se por

$$\mathbf{Y} = [h_1(\mathbf{X},\beta) \quad h_2(\mathbf{X},\beta)]^T = [G \quad N]$$

onde

$$G{=}[G_i]{=}\left[ (A_{g_i} \mathrm{e}^{-k_{g_i}(\frac{1}{t_i})^{n g_i}}) I(t_i{=}i_1) + A_{g_i}(\frac{G_{i-1}}{A_{g_i}})^{\frac{(t_i-1)^{n g_{i-1}}}{t_i^{n g_i}}} \{1{-}I(t_i{=}i_1)\} \right],$$

$$N{=}[N_i]{=}\left[ (Npl_i \mathrm{e}^{-amnp\frac{Npl_i}{1000}t_i}) I(t_i{=}i_1) + N_{i-1}\mathrm{e}^{-amnp\frac{Npl_i}{1000}} \{1{-}I(t_i{=}i_1)\} \right],$$

$I(\cdot)$ é a função indicatriz e admitindo que a inicialização se verifica no instante $t_i = i_1$.

Pressupõe-se ainda que o nº total de registos na amostra é $n$, que ela é constituída por $m$ parcelas e que cada uma delas tem $n_k$ registos, tal que $\sum_{i=1}^{m} n_k = n$. O modelo pode escrever-se matricialmente por $\mathbf{Y} = h(\mathbf{X},\beta) + \varepsilon$, onde $\mathbf{Y}, h(\mathbf{X},\beta)$ e $\varepsilon$ têm dimensão $(2n \times 1)$ e $\beta = (amnp, A_{gq}, n_{g0}, n_{gq}, n_{gn}, k_{gq})$ representa o vetor dos parâmetros; assume-se ainda que $\varepsilon \sim N(\mathbf{0}, \Sigma \otimes I_n)$ e supõe-se que os erros são não correlacionados dentro de cada equação e homocedásticos. Segundo Zellner [6](cap. 8), considera-se como distribuição *a priori*

$$\pi(\beta, \Sigma) \propto |\Sigma|^{-3/2}.$$

A distribuição *a posteriori* pode ser escrita na forma

$$\begin{aligned} f(y|\beta,\Sigma) &= (2\pi)^{-n}|\Sigma|^{-\frac{n+3}{2}} \mathrm{e}^{-\frac{1}{2}(y-h(\mathbf{X},\beta))^T(\Sigma^{-1}\otimes I_n)(y-h(\mathbf{X},\beta))} \\ &= (2\pi)^{-n}|\Sigma|^{-\frac{n+3}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(A\Sigma^{-1})}, \end{aligned}$$

com

$$A = [A_{ij}]_{2\times 2} = [y_i - h_i(\mathbf{X},\beta)]^T[y_j - h_j(\mathbf{X},\beta)].$$

Atendendo à complexidade da distribuição *a posteriori*, recorre-se a métodos de Monte Carlo via Cadeias de Markov (MCMC) para se obter uma amostra aleatória dos valores dos parâmetros dos modelos da área basal e de mortalidade, usando o algoritmo de Gibbs com o passo de Metropolis. A distribuição condicional completa de $\beta$ é

$$\pi(\beta|y;\Sigma) \propto |\Sigma|^{-\frac{n+3}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(A\Sigma^{-1})}.$$

Para facilitar o tratamento consideraram-se os valores da matriz $\Sigma$ obtidos através da estimação clássica. Como distribuições proponentes (ou envelopes) utilizaram-se distribuições da família exponencial com a particularidade de terem valor esperado nas vizinhanças das estimativas clássicas e desvios-padrão baixos na ordem dos 15,33%.

# 4 Aplicação

Os dados para a estimação foram obtidos de 114 parcelas de eucalipto em primeira rotação, constituindo uma amostra de dimensão 1477 registos. A abordagem da estimação pela metodologia clássica é descrita detalhadamente em Marto [3], pelo que apenas se apresentam os principais resultados e conclusões. Segundo esta metodologia rejeitaram-se as hipóteses de nulidade de cada um dos parâmetros estimados de cada equação, com probabilidades críticas inferiores a 0.0001, pelo que se rejeita a hipótese de nulidade conjunta dos parâmetros de cada uma das equações (pois bastava que pelo menos um dos parâmetros fosse estatísticamente diferente de zero). Adicionalmente o modelo apresenta eficiências de modelação $R^2 = 0.9356$ e $R^2 = 0.9862$, respetivamente, para as equações da mortalidade e da área basal, podendo traduzir uma alta percentagem de explicação da variabilidade das variáveis endógenas pelas variáveis exógenas.
Para a estimação segundo a metodologia bayesiana, geraram-se 7000 réplicas, considerando um período de aquecimento de 3000; a convergência foi verificada através do teste de Geweke, usando o *BOA* do programa *R*, conduzindo aos resultados expressos na Tabela 1:

| Parâmetros | médias amostrais | Desvios padrão | medianas |
|---|---|---|---|
| $amnp$ | 0.000152 | <0.0001 | 0.000118 |
| $A_{gq}$ | 0.120025 | <0.0001 | 0.119906 |
| $n_{g0}$ | 1.900603 | 0.0009 | 1.900511 |
| $n_{gq}$ | -0.399612 | 0.0004 | -0.399289 |
| $n_{gn}$ | 0.053481 | <0.0001 | 0.053476 |
| $k_{gq}$ | 125.0256 | 0.0184 | 125.0411 |

Tabela 1: Estimativas bayesianas dos parâmetros do modelo da mortalidade e da área basal pelo método NSUR.

Também neste caso não existem covariáveis suscetíveis de poderem ser consideradas nulas e portanto serem retiradas do modelo.

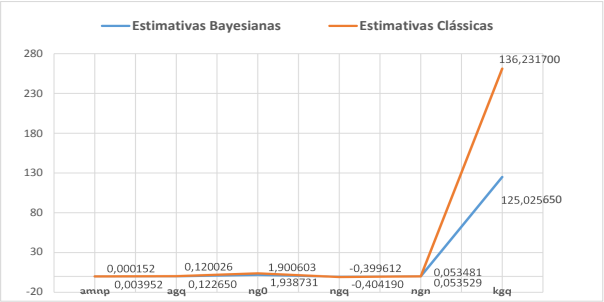



Figura 1: Perfil comparativo das estimativas dos parâmetros.

Analisando o gráfico da Figura 1 que apresenta uma comparação entre as estimativas obtidas pelas duas abordagens constata-se que elas são bastante próximas para todos os parâmetros, à exceção do parâmetro $kgq$. As estimativas bayesianas são, em valor absoluto, inferiores às clássicas correspondentes.

Seguidamente as duas metodologias foram comparadas em termos da capacidade preditiva do modelo, utilizando-se para esse efeito, um conjunto de dados constituído por parcelas independentes com 316 registos. Para medir os desvios recorre-se à raíz quadrada do erro quadrático médio das estimativas dos modelos correspondentes, dada por

$$\sqrt{EQM_M} = \sqrt{\sum_{i=1}^{316} (\hat{M}_i - M_i)^2 / 316},$$

onde $M_i$ representa os valores do modelo $M$, com $M = N$ ou $M = G$. Fazendo a aplicação a este conjunto de dados, cujos resultados estão na Tabela 2, conclui-se que a abordagem bayesiana teve claramente melhores capacidades preditivas para a Mortalidade enquanto que relativamente à Área Basal não se pode dizer que uma das metodologias é substancialmente melhor que a outra.

| modelo | met. clássica | met. bayesiana |
|---|---|---|
| mortalidade ($N$) | 60.4903 | 43.8037 |
| área basal ($G$) | 2.8322 | 2.8942 |

Tabela 2: Raíz quadrada do *EQM* dos dados usados na validação.

# Agradecimentos

Agradece-se a ALTRI Florestal a autorização da utilização de dados de parcela permanentes e inventário florestal. Este trabalho foi parcialmente subsidiado por fundos portugueses através do CIDMA e da FCT, inseridos no projeto PEst-OE/MAT/UI4106-/2014.

# Referências

[1] Amaro, A.P.N. (1997). *Modelação do Crescimento de Povoamentos de Eucalyptus Globulus Labill de 1ª Rotação em Portugal.* Tese de Doutoramento em Engenharia de Sistemas. IST/UTL, Lisboa.

[2] Griffiths, W.E. (2003). Bayesian inference in the Seemingly Unrelated Regression Model. In Giles, D. (ed.): *Computer-Aided Econometrics*, Marcel Dekker.

[3] Marto, M. (2011). *Estimação Clássica e Bayesiana de Parâmetros do Modelo Globulus 2.1.* Tese de Mestrado, Departamento de Matemática, Universidade de Aveiro.

[4] Tomé, M., Ribeiro, F., Soares, P. (1998). *O modelo Globulus 2.1.* Relatórios Tecnico-científicos do GIMREF nº 1/2001, Instituto superior de Agronomia, Lisboa.

[5] Tomé, M., Ribeiro, F., Soares, P. (2001). Silvipastoral systems in Portugal. Em Pukkala,T. e Eerikainen, K. (eds): *Modelling the grouth of tree plantations and agroforestry systems in south and east Africa.* Tiedonantoja Research Notes 80, 23–33.

[6] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley.

# Análise da fiabilidade de centros de maquinação - um caso de estudo

Maria João Dias
Departamento de Matemática, Universidade de Aveiro, *moreira-dias@ua.pt*
Adelaide Freitas
Departamento de Matemática & CIDMA, Universidade de Aveiro, *adelaide@ua.pt*
Constantino Pinto
Renault C.A.C.I.A., Aveiro, *constantino.pinto@renault.com*

**Palavras–chave**: Curva *loess*, fiabilidade, manutenção, trajetórias

**Resumo**: Dois instrumentos gráficos para análise da fiabilidade em função do plano da manutenção implementado, em máquinas existentes numa unidade elementar de trabalho da fábrica Renault C.A.C.I.A., são desenhados. Primeiro, sugere-se a regressão localmente ponderada para estimar uma curva de ajustamento do tempo médio sem intervenção entre duas manutenções preventivas consecutivas dado o tempo entre as duas preventivas. Em seguida, sugere-se a regressão linear para ajustar trajetórias do processo estocástico definido pela razão entre o número de intervenções preventivas e o número de intervenções (preventivas e corretivas) ocorridas num mesmo intervalo de tempo. O objetivo é providenciar instrumentos de fácil leitura gráfica da influência do plano de manutenções preventivas na fiabilidade das máquinas, e assim sugerir melhores práticas de manutenções preventivas.

## 1 Introdução

A inovação e a enorme competitividade que as empresas conhecem nos dias de hoje levam-nas a uma constante procura de técnicas

que permitam conhecer o correto funcionamento dos equipamentos e, consequentemente, definir as melhores práticas e processos de optimização. Em empresas como a Renault C.A.C.I.A. (em Cacia, Aveiro), que detém um avançado processo de maquinação e montagem de componentes mecânicos com elevada precisão, é de extrema importância garantir o correto funcionamento dos seus equipamentos, sem paragens inesperadas ou avarias. Neste sentido, a caraterização da fiabilidade dos equipamentos, considerando o plano de ações de manutenção preventiva aplicado, é um instrumento de avaliação indispensável.

De janeiro a julho de 2013, a primeira autora realizou um estágio curricular na Renault C.A.C.I.A. no âmbito do plano de estudos do seu curso de Mestrado em Matemática e Aplicações da Universidade de Aveiro, tendo como principal objetivo de investigação: (i) analisar as intervenções preventivas e corretivas realizadas no período de 2009 a 2012 de 19 máquinas pertencentes aos centros de maquinação de marca GROB posicionadas no departamento de componentes mecânicos daquela unidade fabril; e (ii) fornecer uma ferramenta que providencie uma avaliação da fiabilidade de cada equipamento dos centros de maquinação, em função do plano de manutenção aplicado, e então sugerir recomendações para o plano de manutenções preventivas das máquinas GROB.

Neste trabalho são descritos dois instrumentos gráficos para visualizar influências das práticas preventivas de uma máquina na sua fiabilidade. Assim, as seguintes secções deste artigo estão organizadas do seguinte modo: na Secção 2 abordam-se os conceitos de fiabilidade e de manutenção. Na Secção 3, descreve-se o algoritmo associado à regressão local ponderada e justifica-se a sua escolha como metodologia a usar na obtenção de um dos instrumentos gráficos. Por fim, na Secção 4 e com os dados de uma das máquinas, são ilustrados os dois instrumentos propostos e como interpretá-los.

## 2 Fiabilidade e Manutenção

A fiabilidade de um equipamento exprime o grau de confiança que se deposita no seu correto funcionamento (sem falha). Denotando por $T$ a variável aleatória que representa o tempo em funcionamento sem falha de um equipamento, em termos formais define-se fiabilidade no instante $t$ por $P(T > t)$. Na prática, esta noção probabilística de fiabilidade é preterida por estimativas de parâmetros associados à distribuição de $T$ com vista a facilitar a avaliação, em campo, da fiabilidade do equipamento. Na literatura especializada uma diversidade de indicadores para a fiabilidade são apontados como seja o tempo médio para a primeira falha, a taxa de falha, o tempo médio entre manutenções, a disponibilidade, entre outros.

A manutenção de um equipamento corresponde a qualquer atividade desenvolvida no seu ciclo de vida com o objetivo de manter ou repor o seu correcto funcionamento nas melhores condições de custo, disponibilidade e segurança. Uma manutenção pode ser corretiva ou preventiva. A manutenção corretiva pretende corrigir avarias quando estas ocorrem aleatoriamente no tempo, com o objetivo de recolocar a máquina em funcionamento o mais rápido possível. Trata-se de uma ação não periódica devida a danos atuais ou não iminentes. É o tipo de manutenção mais comum podendo implicar perda de produção e reduzir o tempo de vida útil do equipamento. A manutenção preventiva corresponde a qualquer ação de manutenção planeada em cronograma com vista a evitar a falha do equipamento.

Para avaliar o efeito do plano de preventivas programado sobre a fiabilidade de cada uma das 19 máquinas GROB, é proposto relacionar o tempo que cada máquina permanece sem receber qualquer tipo de manutenção com o tempo decorrido entre ações preventivas consecutivas. Assim, considera-se o tempo médio de funcionamento sem intervenções como indicador da fiabilidade e o tempo decorrido entre preventivas consecutivas como variável explanatória associada ao plano das preventivas. A Figura 1 esquematiza os dados originais, fornecidos pela empresa, correspondendo às datas (dia, mês, ano), em que ocorreu cada intervenção (devida a falha da máquina ou a uma

ação de prevenção programada). Com estes dados calculou-se o número de dias entre cada duas preventivas consecutivas e, para cada período, quantos dias há sem qualquer intervenção.

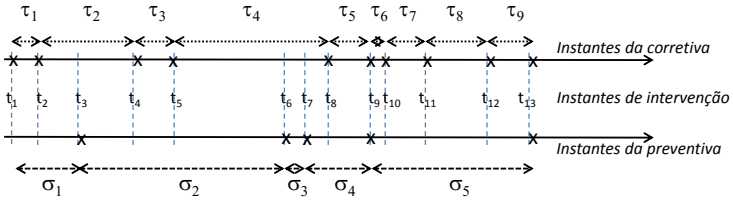



Figura 1: Pares de dados em análise: (tempo entre duas preventivas consecutivas, tempo sem intervenções entre essas duas preventivas). Por exemplo, os pares $(\sigma_1, t_2 - t_1)$, $(\sigma_1, t_3 - t_2)$, $(\sigma_2, t_4 - t_3)$.

# 3 Metodologia *loess*

A regressão local ponderada, conhecida por *loess* (*LOcal regrESSion*) é um método não-paramétrico útil para obter representações de curvas de ajustamento de possíveis relações entre duas variáveis. O aspeto central é a visualização gráfica da relação, o que vai ao encontro dos objetivos propostos. O algoritmo do método *loess* encontra-se implementado em vários pacotes estatísticos (R, S-PLUS, SAS e outros). Resumidamente, os seguintes passos descrevem o algoritmo do método *loess* aplicado sobre cada ponto $(x_i,y_i)$ de um conjunto de $n$ observações $(x_i,y_i) \in \mathbb{R}^2$, $i = 1,2,\ldots,n$:

**Passo 1.** Calcular a ponderação para cada $(x_k,y_k)$ usando uma função cujo domínio $D_h$ depende de um parâmetro de suavização ou amplitude $h$. Por exemplo, a função tricúbica dada por:

$$w_k \equiv w_k(x_i) = \left(1 - \left(\frac{|x_i - x_k|}{\max_{x_k \in D_h} |x_i - x_k|}\right)^3\right)^3$$

**Passo 2.** Aplicar o método dos mínimos quadrados ponderados para estimar os parâmetros do modelo polinomial de grau $d$ a ajustar. Por exemplo, para $d = 1$ será $\widehat{y} = a + bx$, com:

$$b = \frac{\sum w_k x_k y_k - \frac{\sum w_k x_k \sum w_k y_k}{\sum w_k}}{\sum w_k x_k^2 - \frac{(\sum w_k x_k)^2}{\sum w_k}} \text{ e } a = \frac{\sum w_k y_k}{\sum w_k} - b\frac{\sum w_k x_k}{\sum w_k}$$

onde $\sum$ representa a soma para todo o $k : \ x_k \in D_h$.

**Passo 3.** Obter o valor suavizado dado pelo valor predito na regressão ponderada: $\widehat{y_i}$.

Para o método *loess* é necessário escolher: o parâmetro $h$, a função peso $w$, e o grau $d$ do polinómio a ajustar. O parâmetro de suavização $h$ determina o tamanho da vizinhança de cada ponto $x$ na qual a função peso será aplicada. Este parâmetro afeta a variabilidade e o viés da estimativa da resposta para cada ponto. Para $h$ elevado (pequeno), a estimativa terá um viés elevado (pequeno, resp.)) e uma variabilidade pequena (elevada, resp.). Poderá ser escolhido um parâmetro local, $h(x)$, de forma a conter um número específico de pontos. A escolha do grau do polinómio local a ajustar é, em geral, feita por inspeção visual do gráfico com os dados originais e com a estimativa de regressão local.

Neste trabalho, o parâmetro de suavização foi escolhido dentro de um conjunto de valores de $h$ e corresponde ao valor que minimiza a soma dos quadrados dos erros do modelo ajustado e, simultaneamente, a soma dos quadrados dos erros do modelo estimado por validação cruzada para cada observação (para detalhes dos comandos do *R* ver [3], Anexo I). Optou-se por ajustar polinómios de grau $d = 1$ porque não se visualizaram melhorias quando a complexidade do modelo era aumentada ($d > 1$). Por fim, a função de ponderação, responsável por atribuir pesos às observações na vizinhança de cada ponto foi a função tricúbica a qual devolve uma boa suavização na maior parte dos casos (veja-se [1]).

Para avaliar a qualidade de ajuste da curva *loess* aos dados, dado o carácter não paramétrico do método, em [2] sugere-se (apenas) gráfi-

cos dos resíduos visualizando, no espaço das variáveis independentes, a existência de viés e a variabilidade das estimativas. Aqui, como se verá, a qualidade de ajuste não é boa; a ideia é sumariar, em termos médios, informação da nuvem. Outros métodos não-paramétricos, mais adaptativos de suavização de nuvens de pontos, poderiam ter sido usados. Por exemplo, a regressão por *splines* seria útil se os dados exibissem regiões locais de variação abrupta, o que não se verifica. A escolha do método *loess* residiu no facto de este basear-se em ajustamentos polinomiais locais com atribuição de pesos maiores na vizinhança de cada instante ignorando o que se passa para tempos mais afastados, o que pareceu ser mais adequado à realidade.

# 4 Resultados

Para visualizar o efeito do plano preventivo de cada das máquinas GROB sobre a sua fiabilidade sugere-se construir, para cada uma, dois instrumentos gráficos. A seguir, usando os dados relativos a uma das máquinas (Máquina nº 2102 que maquina suporte de injetores) são ilustrados esses instrumentos (Figura 2) e como são interpretados. Para a construção dos gráficos recorreu-se ao programa estatístico *R* (versão 3.0.1). O primeiro instrumento (gráfico à esquerda na Figura 2) corresponde ao diagrama de dispersão dos pares observados $(x,y)$, com $x$=tempo entre duas preventivas consecutivas e $y$=tempo sem intervenções entre essas duas preventivas, e a curva *loess* de ajustamento do tempo médio sem intervenções entre preventivas consecutivas. O segundo instrumento (gráfico à direita na Figura 2) é dado pela trajetória do processo estocástico $\{R(t),\, t > 0\}$, onde $R(t)$ representa a razão entre o número de intervenções preventivas ocorridas num intervalo de tempo de amplitude $t$ e o número de intervenções (preventivas e corretivas) ocorridas nesse mesmo intervalo de tempo. Esta trajetória permite visualizar a evolução, ao longo do tempo, da percentagem de intervenções preventivas e, por complementaridade, da percentagem de intervenções corretivas.
Da Figura 2, no primeiro gráfico, observam-se pontos sobre a reta

$y = x$. Tal significa que entre duas preventivas consecutivas não houveram intervenções corretivas ou falhas. Tal correu para preventivas espaçadas de 0, 7, 14, 21 e 28 dias. Também, quanto mais perto se encontra a curva *loess* da reta $y = x$ significa que se estima que mais perto da próxima intervenção preventiva programada ocorre uma falha. Para pares de preventivas espaçadas de mais de 14 dias, a curva *loess* sofre uma quebra afastando-se da reta $y = x$. Este comportamento da curva estima que, quando as intervenções preventivas são espaçadas por mais de 14 dias, o número de avarias na máquina aumenta e o tempo que decorre entre avarias diminui.

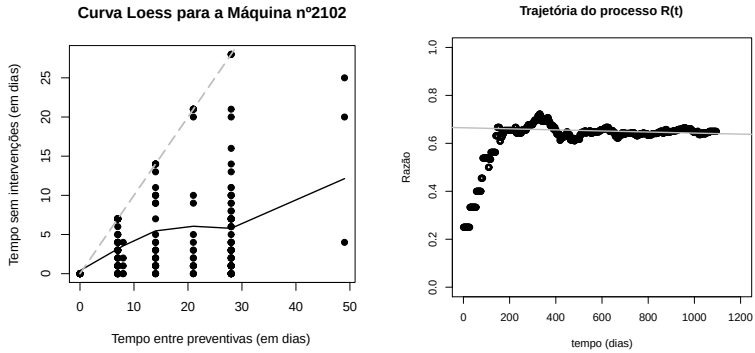



Figura 2: Instrumentos para avaliar graficamente o efeito do plano de manutenção sobre a fiabilidade da Máquina nº 2102. À esquerda, a cheio a curva *loess* e a tracejado a reta $y = x$. À direita, a cheio a reta de regressão estimada por $r(t) = -0.00002213t + 0.6649$, para $t > 200$ (os coefientes da reta são signficativamente diferentes de zero, com valor-p $< 0.2 \times 10^{-16}$, para ambos).

Da Figura 2, no segundo gráfico, observa-se que no início do registo das ações (preventivas e corretivas), até cerca de 200 dias ($t < 200$) há um crescimento inicial da razão $r(t)$, sendo que ultrapassa o valor

0.5 para $t$ próximo de 90 dias. Tal significa que, nesta fase inicial, há um aumento de ações preventivas face ao total de intervenções e, ao fim de cerca de 90 dias, o número de intervenções programadas é superior ao número de intervenções corretivas. Quando $t > 200$, a razão $r(t)$ tende a estabilizar em torno da reta $r(t) = -0.00002213t + 0.6649$. Tal significa que ao fim de 200 dias estima-se que cerca de 65% das intervenções são de carácter preventivo estabilizando praticamente o plano de ações da Máquina 2102 nesta rotina. Apesar do declive daquela reta estimada ser estatisticamente signficativo, o seu valor reduzido não parece ser relevante em termos práticos.
Em resumo, pode dizer-se que, à data final do registo dos dados, cerca de 35% das intervenções da Máquina 2102 são por avaria desta e que esta apresenta um plano de manutenção preventivo estabilizado em termo de respostas da fiabilidade do equipamento. Contudo, realça-se que a curva *loess* sugere repensar as intervenções preventivas espaçadas de 21 ou mais dias por ocorrerem avarias em um menor espaço de tempo médio.

# Agradecimentos

Este trabalho foi parcialmente financiado por fundo portugueses através do CIDMA (*Center for Research and Development in Mathematics and Applications*) e a FCT (*Fundação para a Ciência e a Tecnologia*) dentro do projeto PEst-OE/MAT/UI4106/-2014.

# Referências

[1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74, 829–836.

[2] Cleveland, W., Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 403, 596–610.

[3] Dias, M.J. (2013). *Fiabilidade de Centros de Maquinação - um Caso de Estudo*. Dissertação de Mestrado. Universidade de Aveiro.

# Sobrevivência a longo prazo de doentes com cancro do cólon e do reto

Mariana Rodrigues
Unidade de Investigação, SESARAM, *marianacfr@gmail.com*
Carina Alves
NGDEstatística, SESARAM, *anacarina.alves@gmail.com*
Ana Maria Abreu
CCCEE e CCM, Universidade da Madeira, *abreu@uma.pt*

**Palavras–chave**: Análise de sobrevivência, estimador de Kaplan-Meier, modelo de Cox, modelo de cura

**Resumo**: Uma das especialidades da Medicina onde é muito comum a aplicação da Análise de Sobrevivência é a Oncologia. O conhecimento dos fatores que influenciam o diagnóstico, o tratamento e a sobrevivência tem sido o objetivo de vários estudos. O objetivo do nosso estudo consiste numa pequena contribuição para um melhor conhecimento de fatores que possam influenciar a sobrevivência e a taxa de cura de doentes com cancro do cólon e do reto. Foi considerada uma base de dados de 800 indivíduos, aos quais foi diagnosticado cancro do cólon e reto, entre 2000 e 2009, na Região Autónoma da Madeira, com um *follow-up* mínimo de três anos e meio. Os resultados obtidos não foram animadores pois, dos indivíduos com estadio conhecido, 22% estavam no estadio IV, ou seja, no estadio mais grave da doença. O modelo de cura foi aplicado aos indivíduos nos estadios IV e Desconhecido, pois só nestes casos o tempo de *follow-up* foi suficiente. As covariáveis quimioterapia e cirurgia foram as mais importantes para a sobrevivência dos indivíduos, quer através do modelo de Cox quer através do modelo de cura.

# 1 Introdução

Na Análise de Sobrevivência, analisam-se tempos de vida, também designados por tempos de sobrevivência. Este tempo tem um sentido muito vasto pois, por exemplo, pode representar o tempo até à morte ou o tempo até à cura da doença. Os tempos de vida são registados para um grupo de indivíduos e são calculados considerando o tempo decorrido desde um instante inicial até à ocorrência de um acontecimento de interesse. Para alguns indivíduos esse acontecimento pode não ser observado durante o período em que estão em observação, caso em que ocorre censura à direita, que é a mais usual e aquela que vamos considerar. Para outros esse acontecimento pode não vir a ocorrer, sendo os mesmos designados por indivíduos imunes ou curados. Para esta situação foram desenvolvidos os modelos de cura, sendo os mais comuns os de mistura ([1], [4]).
No caso do cancro do cólon e do reto, alguns indivíduos conseguem ficar curados [8], pelo que se torna adequado aplicar um modelo de cura para complementar a informação proveniente da Análise de Sobrevivência clássica. Segundo o GLOBOCAN, [5], o cancro do cólon e do reto é o mais incidente e o que apresenta uma maior taxa de mortalidade de entre o cancros do aparelho digestivo, tendo ocorrido 6952 novos casos e 3691 óbitos em Portugal em 2008, ou seja, em média, cerca de 20 novos casos e 10 óbitos por dia. A mesma fonte prevê que, em 2020, os valores ascendam a 8178 novos casos e a 4376 óbitos. Estes números, sendo elevados, incentivam ao aparecimento de mais estudos e a novas abordagens, como a que aqui se apresenta.

# 2 Metodologia

Os dados deste estudo, fornecidos pelo ROR-Sul, referem-se a 800 doentes da Região Autónoma da Madeira (RAM), aos quais foi diagnosticado tumor maligno do cólon e reto, entre 01 de janeiro de 2000 e 31 de dezembro de 2009. O *follow-up* foi feito até junho de 2013. As covariáveis consideradas foram: Género, Grupo etário,

Quimioterapia, Cirurgia e Estadio.
Foram calculadas as estimativas de Kaplan-Meier das funções de sobrevivência (f.s.) e, através do modelo de regressão de Cox [3], foi determinada a influência das covariáveis no tempo de vida. O conhecido modelo de regressão de Cox pode ser escrito à custa da função de risco, no instante $t$, na forma que se segue,

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) = h_0(t)\exp(\beta_1 z_1 + \cdots + \beta_p z_p),$$

onde $\mathbf{z} = (z_1,\ldots,z_p)'$ representa o vetor de covariáveis associado a cada indivíduo, $\beta_1,\ldots,\beta_p$ os correspondentes coeficientes de regressão e $h_0(t)$ a função de risco subjacente.
Foi ainda implementado o modelo de cura de mistura [4], considerando os estadios IV e Desconhecido, pois só nestes casos o tempo de *follow-up* foi suficiente [7], ou seja, foi maior do que o correspondente valor da mediana do tempo de vida dos indivíduos suscetíveis. Foram consideradas covariáveis quer na f.s. quer na taxa de cura [1], ou seja, foi usado o modelo

$$S(t|\mathbf{x},\mathbf{z}) = p(\mathbf{z}) + (1 - p(\mathbf{z}))S_d(t|\mathbf{x}),$$

onde $S(t|\mathbf{x},\mathbf{z})$ designa a f.s. populacional da variável aleatória $T$, $S_d(t|\mathbf{x})$ a f.s. correspondente aos indivíduos doentes, $p(\mathbf{z})$ a taxa de cura e $\mathbf{x}$ e $\mathbf{z}$ vetores de covariáveis. Note-se que uma covariável que seja importante para a f.s. pode não o ser para a taxa de cura e vice-versa.
Para os indivíduos doentes foi usado o modelo de riscos proporcionais e para a taxa de cura o modelo de regressão logístico.
Os dados foram analisados usando o programa de análise estatística *PASW Statistics for Windows*, Versão 18.0, bem como o *software* R [6], em particular, o *package smcure* [2].

# 3 Resultados

## 3.1 Análise preliminar

A amostra estudada evidencia que o número de casos ao longo dos anos tem tido uma tendência crescente, tendo havido um grande aumento em 2009, como se pode ver na Figura 1.

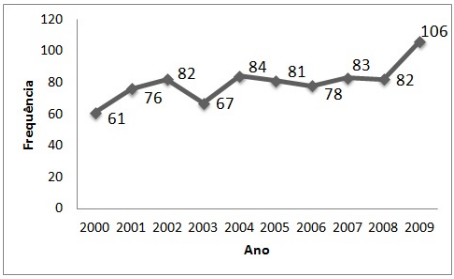



Figura 1: Número de casos entre 2000 e 2009.

Na Tabela 3.1 encontra-se uma breve descrição da amostra. Verifica-se que a maior incidência ocorre no grupo etário dos 70 aos 79 anos, logo seguido do grupo dos 60 aos 69 anos: em conjunto constituem quase 60% da amostra. Nota-se ainda que a quase totalidade dos doentes foi sujeita a cirurgia (87%).
Na Figura 2 observa-se que a sobrevivência é tanto melhor quanto menor for o estadio. Por exemplo, o estadio I é o único em que mais de metade dos indivíduos se encontram vivos no fim do estudo. Além disso, como as curvas dos estadios I e II estão próximas até aos 5 anos, significa que, neste período, a probabilidade de sobrevivência dos doentes no estadio II também é elevada.

## 3.2 Modelo de regressão de Cox

Como se pode observar na Tabela 3.2, a cirurgia tem um efeito protetor pois os doentes que são operados têm 61% do risco de morte

Tabela 1: Características dos doentes no que diz respeito ao género, grupo etário e tratamentos, por estadio.

| **Var.** | **Estadio** | | | | | |
|---|---|---|---|---|---|---|
| | I n=123 | II n=183 | III n=187 | IV n=179 | Desc. n=128 | $p^*$ |
| **Género** | | | | | | |
| Masc. | 60 (48,8) | 106 (57,9) | 92 (49,2) | 94 (52,5) | 67 (52,3) | 0,455 |
| **Grupo etário** (anos) | | | | | | |
| $< 50$ | 13 (10,6) | 11 (6,0) | 26 (13,9) | 21 (11,7) | 10 (7,8) | 0,021 |
| 50 a 59 | 23 (18,7) | 34 (18,6) | 29 (15,5) | 23 (12,8) | 16 (12,5) | |
| 60 a 69 | 31 (25,2) | 62 (33,9) | 49 (26,2) | 46 (25,7) | 28 (21,9) | |
| 70 a 79 | 36 (29,3) | 51 (27,9) | 63 (33,7) | 54 (30,2) | 40 (31,3) | |
| $\geq 80$ | 20 (16,3) | 25 (13,7) | 20 (10,7) | 35 (19,6) | 34 (26,6) | |
| **Quimioterapia** | | | | | | |
| Fez | 50 (40,7) | 120 (65,6) | 149 (79,7) | 111 (62,0) | 33 (25,8) | <0,001 |
| **Cirurgia** | | | | | | |
| Fez | 120 (97,6) | 181 (98,9) | 183 (97,9) | 131 (73,2) | 82 (64,1) | <0,001 |

**Legenda:** Var. - Variáveis; Desc. - Desconhecido; Masc. - Masculino; os valores entre parêntesis indicam percentagens.

***** Teste de Qui-quadrado.

dos que não são ($e^{-0,5} = 0,606$). O mesmo acontece em relação à quimioterapia, onde os doentes que fazem este tratamento têm apenas 81% do risco de morte dos que não fazem. Em relação ao estadio e à idade, como existe interação, a interpretação não é tão direta. Por exemplo, para comparar dois indivíduos em estadios diferentes, digamos III e IV, temos que fixar a idade; se considerarmos dois indivíduos com 50 anos, então o risco de morte de um indivíduo no estadio IV é cerca de 3 vezes ($e^{4,791-0,043\text{x}50}/e^{3,26-0,036\text{x}50} = 3,258$) superior ao de um indivíduo no estadio III. Analogamente, dois indivíduos no estadio II, com idades de 50 e 65 anos, o mais velho tem cerca do dobro ($e^{0,052\text{x}65-0,008\text{x}65}/e^{0,052\text{x}50-0,008\text{x}50} = 1,935$) do risco de morte do mais novo.

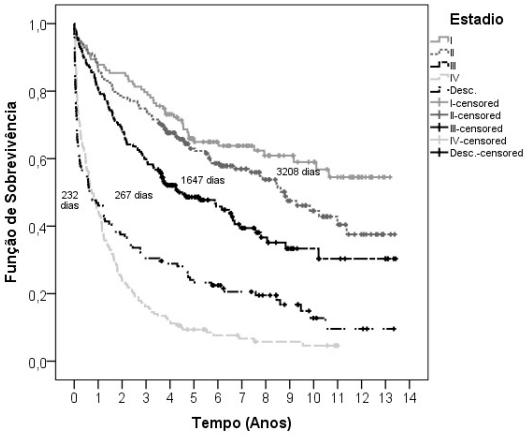



Figura 2: Estimativas de K-M da f.s. por estadio.

Tabela 2: Resultados obtidos pelo modelo de regressão de Cox.

| **Covariáveis** | **Estimativa de $\beta$** | **Erro padrão** | $p$ |
|---|---|---|---|
| Estadio II | 0,898 | 1,222 | 0,462 |
| Estadio III | 3,260 | 1,100 | 0,003 |
| Estadio IV | 4,791 | 1,055 | <0,001 |
| Estadio Desconhecido | 2,273 | 1,129 | 0,044 |
| Cirurgia | -0,500 | 0,132 | <0,001 |
| Quimioterapia | -0,217 | 0,110 | 0,048 |
| Idade | 0,052 | 0,013 | <0,001 |
| Estadio II*Idade | -0,008 | 0,017 | 0,633 |
| Estadio III*Idade | -0,036 | 0,015 | 0,019 |
| Estadio IV*Idade | -0,043 | 0,015 | 0,003 |
| Estadio Desconhecido*Idade | -0,015 | 0,016 | 0,338 |

## 3.3 Modelo de cura

As medianas do tempo de vida dos indivíduos nos estadios IV e Desconhecido, obtidas através das estimativas das f.s. de Kaplan-

Meier, foram de 267 e 232 dias, respetivamente, e o tempo mínimo de *follow-up* de cerca de 3 anos e meio (1261 dias) em ambos os estadios, cumprindo assim o requisito referido na metodologia para se poder aplicar o modelo de cura. Foi usado o modelo de riscos proporcionais para os indivíduos doentes com as covariáveis Quimioterapia e Cirurgia; para a taxa de cura foi usado o modelo de regressão logístico apenas com a covariável Quimioterapia uma vez que a estimativa de Kaplan-Meier da f.s. da categoria "Não fez" da covariável Cirurgia atinge o valor zero.
Na Figura 3, verifica-se que as curvas diferem na parte inicial (onde existem muitos doentes) mas não na parte final, quando as curvas estabilizam. Assim, no estadio IV as taxas de cura foram de 3,6% para os indivíduos que não fizeram quimioterapia e de 4,4% para os que fizeram, valores estes obtidos apenas à custa da influência das covariáveis no tempo de vida dos indivíduos doentes (Tabela 3.3), ou seja, a quimioterapia não influencia a taxa de cura. No estadio Desconhecido, as taxas de cura foram de 6,9% e de 7,3%, respetivamente, sendo que neste caso nenhuma das covariáveis se revelou significativa em qualquer dos dois modelos. No entanto, a razão da aplicação deste modelo a este último estadio foi essencialmente ilustrativo e não de interesse clínico.

Tabela 3: Resultados obtidos pelo modelo de cura para o estadio IV.

| **Covariáveis** | Estimativas | $Z$ | $p$ |
|---|---|---|---|
| Quimioterapia | -1,277 | -4,276 | <0,001 |
| Cirurgia | -0,578 | -1,775 | 0,076 |

# 4 Conclusão

O número de casos de cancro do cólon e reto, na RAM, tem vindo a aumentar, seguindo a tendência mundial). A maior incidência

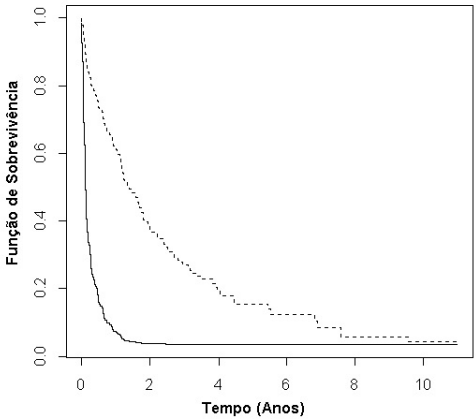



Figura 3: Estimativas da f.s. obtidas pelo modelo de cura para o estadio IV (⋯ Fez quimioterapia, — Não fez).

verifica-se entre os 60 e os 79 anos. A cirurgia e a quimioterapia têm um efeito protetor: os doentes que as realizam têm apenas 61% e 81%, respetivamente, do risco de morte dos que não fazem, de acordo com os resultados do modelo de Cox. Quanto maior a idade e o estadio, maior é o risco de morte. A taxa de cura no estadio IV é reduzida, situando-se em cerca de 4%. A influência da covariável Quimioterapia no tempo de vida era de esperar tendo em conta, por exemplo, que a mediana do tempo de vida para quem não fez este tratamento é de apenas 46 dias e, para quem fez, de 491 dias. As covariáveis quimioterapia e cirurgia foram as mais importantes para a sobrevivência dos indivíduos, quer através do modelo de Cox quer através do modelo de cura.

# Agradecimentos

Investigação parcialmente financiada pela FCT – Fundação para a Ciência e a Tecnologia, projeto PEst-OE/MAT/UI0219/2011 – Projeto Estratégico do CCM (Centro de Ciências Matemáticas).

# Referências

[1] Abreu, A.M e Rocha, C.S. (2013). A Parametric Cure Model with Covariates. Em: Lita da Silva, J., Caeiro, J., Natário, I., Braumann, C.A. (eds.):*Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications* 37–45, Springer–Verlag, Berlin Heidelberg.

[2] Cai, C., Zou, Y., Peng, Y. e Zhang, J. (2013). smcure: Semiparametric mixture cure model. R package version 2.0.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society.* B, 34, 187–220.

[4] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 39, 1–38.

[5] GLOBOCAN 2008. Cancer Incidence, Mortality and Prevalence Worldwide in 2008, URL http://globocan.iarc.fr/. Acesso em: julho 2013.

[6] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-00051-07-0, URL http://www.r-project.org/.

[7] Yu, B., Tiwari, R.C., Cronin, K.A., Feuer, E.J. (2004). Cure fraction from the mixture cure models for grouped survival data. *Statistics in Medicine* 23, 1733–1747.

[8] Yu, X.Q., De Angelis, R., Andersson, T.M.L., Lambert, P.C., O'Connell, D.L., Dickman, P.W. (2013). Estimating the proportion cured of cancer: Some practical advice for users. *Cancer Epidemiology* 37, 836–842.

# On the protection of $\alpha$-thalassaemia from malaria infection in northeast Tanzania

Nuno Sepúlveda
London School of Hygiene and Tropical Medicine, United Kingdom, and CEAUL, Portugal, *nuno.sepulveda@lshtm.ac.uk*

Alphaxard Manjurano
National Institute for Medical Research and Joint Malaria Programme, Tanzania, *amanjurano@yahoo.co.uk*

Chris J Drakeley
London School of Hygiene and Tropical Medicine, United Kingdom, *chris.drakeley@lshtm.ac.uk*

Taane G Clark
London School of Hygiene and Tropical Medicine, United Kingdom, *taane.clark@lshtm.ac.uk*

**Palavras–chave**: Multiple imputation, genetic association, logistic regression

**Abstract**: In northeast Tanzania, the $\alpha$-thalassaemia gene is of great research interest due to its association with clinical malaria and altitude, two key evolutionary drivers of the human genome. To understand this association better, we analysed data from about 7000 asymptomatic individuals living in different altitude-related transects. We first used multiple imputation to tackle the high percentage of missing genotypes in the sample. We then assessed the association of this gene with malaria infection using Logistic regression and adjusting the results for putative confounding effects of altitude, ethnicity, age and gender. We found that the $\alpha$-thalassaemia gene seems indeed to protect from malaria infection and, thus, likely to be under genetic selection in the region.

# 1 Introduction

Selection reflects the evolutionary process by which the genetic variants increasing the survival or the reproduction capacity of the individuals tend to increase in frequency in the respective population. Human genome seems to be under selection due to, for example, altitude or malaria. In most extreme settings, altitude induces strong stress responses to oxygen deprivation (hypoxia). Although high altitude acclimatization might occur after some time, the long-term exposure to hypoxic conditions leads to several medical disorders, affecting the overall fitness of the population. Therefore, the human populations living permanently in highlands are undergoing selection on genes related to oxygen sensing and usage [1]. Conversely, genes that protect from malaria have been selected throughout human evolution and in different parts of the globe [5]. Sickle-cell and $\alpha$-thalassaemia are two classical examples of genes under selection in subsaharan Africa that, although affecting the hemoglobin concentration in the blood, protect individuals from most severe forms of malaria [10]. Interestingly, malaria-driven selection works in combination with other factors, including altitude. This is particularly evident in high altitude regions where climate conditions are not ideal for malaria transmission. In those situations, lowland and highland populations are under different malaria selection pressures, as it happens in northeast Tanzania in East Africa.

Northeast Tanzania extends from the coastal plains of Tanga to the high altitude mountains of Kilimanjaro, Usambara and Pare. Because of this natural variation in altitude range and, thus, in malaria exposure, a large cross-sectional survey was conducted on about 12,000 individuals from 24 villages in 6 altitude-related transects [2]. Altitude was found to be negatively correlated not only with malaria prevalence [2], but also with the underlying transmission intensity [3]. Similar negative correlation was also found between altitude and the prevalence of sickle-cell and $\alpha$-thalassaemia traits, but using a restricted subset of individuals from 13 villages [4]. Here we focused our analysis on the $\alpha$-thalassaemia gene as it seems protective from

clinical malaria in the area [6], but it is unclear whether that protection is also exerted on malaria infection. The corresponding data set is interesting in terms of statistical analysis, because of the high frequency of missing genotypes due to different genotyping efforts attempted across the villages. Discarding missing data is known to introduce estimation bias or decrease statistical power. The goals of the analysis are two fold: (1) to perform multiple imputation on the corresponding missing data, and (2) to study the association between that gene and malaria parasite positivity, adjusting for putative altitude effects. A more detailed description of this work can be found elsewhere [8].

## 2 Data and statistical methodology

Sampling was performed on 24 villages divided into 6 altitude-related transects and matched for age and gender [2]. Data at hand refers to genetic, phenotypic, and clinical information on about 7000 individuals with age between 6 months and 45 years old. Genetic data encompasses the genotypes of 110 single nucleotide polymorphisms (SNPs) — genetic markers with known DNA sequences but differing in a single position across the population — predominantly from malaria candidate genes and the information on the number of deletions in the $\alpha$-thalassaemia gene. In particular, individuals with 0 or at least one deletion were classified as negative and positive for the $\alpha$-thalassaemia trait, respectively. Conversely phenotypic and clinical data refer to information on ethnicity, gender, age, hemoglobin (Hb) concentration, malaria parasite positivity, and the respective parasite density.

As mentioned above, there is a high frequency of missing genotypes of the $\alpha$-thalassaemia gene (Table 1). The occurrence of such missing data is explained by a combination of factors, such as budget restrictions, time constraints and blood sample availability for genotyping. However, because the sampling was matched for age, gender, and performed in different altitude-related transects, missing genotypes

from some villages can in theory be informed by fully observed data from other villages located in the same altitude and, thus, with similar malaria selective pressure. We used multiple imputation based on chained equations (MICE) to replace missing genotypes by plausible guesses. In theory, MICE requires the construction of an appropriate imputation model that includes information on different auxiliary variables thought to be important for the imputation process [9, 11]. In our data, we used the following Multinomial Logistic model using a set of $m$ imputation covariates

$$\log \frac{p_{1,i}}{p_{0,i}} = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \text{ and } \log \frac{p_{2,i}}{p_{0,i}} = \beta_0^* + \sum_{j=1}^{m} \beta_j^* x_{ij},$$

where $p_{k,i}$ is the probability of the $i$-th individual having $k = 0,1,2$ deletions in the $\alpha$-thalassaemia gene, $x_{ij}$ is the corresponding value of $j$-th imputation covariate, $\beta_j$ and $\beta_j^*$ are the main effects in the respective log-odds. Previously we performed a simulation study where we artificially generated missing data from complete cases and assessed the quality of imputation [8]. We concluded that the imputation model with least association bias was the one including the following covariates: four SNPs (rs1800629, rs3211938, rs334, and rs542998), Hb concentration, mild anaemia and malaria positivity, and transect. It is worth mentioning that ethnicity, albeit associated with the $\alpha$-thalassaemia gene, is highly correlated to transect. To avoid putative collinearity problems, this variable was not included in competing imputation models.

To generate each imputed data set, we used the following algorithm: (i) generate initial values for the missing data, (ii) estimate the Multinomial Logistic regression described above, (iii) generate new random guesses using the respective estimates for the missing data, (iv) repeat previous two steps until convergence of an adequate summary statistics for the missing data. A good convergence of the algorithm was obtained after 100 iterations while tracing the underlying proportion of $\alpha$-thalassaemia individuals. We used a total number of 100 imputed data sets in the analysis. We tested a higher number of

imputed data sets but the corresponding post-imputation standard errors did not change significantly.

The next step of the analysis was to fit two nested Logistic regression models to each imputed data set: one including only confounding effects and another including both confounding and the effect of $\alpha$-thalassaemia. We considered the following confounders: ethnicity (Wapare, Wasambaa, Wachaga and Others), age (in years), altitude (in meters), and transect (Kilimanjaro, North Pare, South Pare, West Usambara 1, 2, and 3). To assess the association between $\alpha$-thalassaemia and malaria positivity, we compared these models using the average of all likelihood ratio test statistics obtained from the analysis of each imputed data set. This average was used to calculate the corresponding $-\log_{10}$(p-value), as frequently reported in genetic association studies. Large values of this quantity are indicative of a strong association between $\alpha$-thalassaemia and malaria positivity.

The final step of the analysis is to combine the parameter estimates generated from each imputed data set [7]. Let $\lambda_1$ and $\lambda_2$ be the main genetic effects of one and two deletions in the $\alpha$-thalassaemia gene, respectively. The post-imputation estimates and respective standard errors ($se$) were calculated as follows

$$\bar{\lambda}_i = \frac{\sum_{j=1}^{k} \hat{\lambda}_{ij}}{k}$$

and

$$se\left(\bar{\lambda}_i\right) = \sqrt{\frac{\sum_{j=1}^{k} \mathrm{var}(\hat{\lambda}_{ij})}{k} + \frac{k+1}{k} \times \frac{\sum_{j=1}^{k} (\hat{\lambda}_{ij} - \bar{\lambda}_i)^2}{k-1}},$$

where $k$ is the total number of imputed data sets (in our case, $k = 100$) and $\hat{\lambda}_{ij}$ is the estimate of $\lambda_i$ using the $j$-th imputed data set. For a sufficiently large number of imputed data sets, the combined estimates $\bar{\lambda}_i$ follow approximately a Gaussian distribution.

# 3 Results

Table 1 presents the basic data description of the 24 villages in the study. In brief, altitude ranges from 196m (Mgome, West Usambara 2) to 1845 (Emmao, West Usambara 3). The odds of malaria positivity and $\alpha$-thalassaemia correlates inversely with altitude, as reported elsewhere [2, 4]. There are 11 villages where genotyping of the $\alpha$-thalassaemia gene was effectively not attempted. Where genotyping was indeed attempted, the proportion of missing $\alpha$-thalassaemia status varies from 3.0% in Mgome (West Usambara 2) to 51.9% in Mokala (Kilimanjaro). Note that missing data is at a low proportion for the remainder variables in the data set and, thus, removed from the analysis. In terms of putative population structure, a specific ethnic group predominates in each transect: Wachaga in Kilimanjaro, Wapare in North and South Pare, Wasambaa in West Usambara transects.

Before performing imputation on real data, we first conducted a simulation study where we assessed the quality of the post-imputation inferences. Specifically, we focused on complete case data and generated different missing genotype patterns. A more detailed description of this study can be found elsewhere [8]. In brief, MICE was able to produce unbiased estimates for the genetic effects and association signals using those from the complete case analysis (CCA) as the reference of bias. However, it failed to predict the correct genotypes accurately. For that we would need a set of genetic markers strongly correlated with each other (*e.g.*, in linkage disequilibrium), as demonstrated by Souverein et al [9]. Strong correlation is typically found among genetic markers closely located in the same chromosome. In our data, rs2230739 is the closest genetic marker to the $\alpha$-thalassaemia gene ($\approx$ 3.8 Mb distance), but not close enough to be in linkage disequilibrium.

We then performed genotype imputation on real missing data and used the corresponding results to study genetic association between $\alpha$-thalassaemia and malaria positivity. To this end we considered three data settings: imputation of missing data from the 13 villa-

| Village | altitude (m) | malaria (%) | $\alpha$-thalassaemia | | |
|---|---|---|---|---|---|
| | | | − | + | missing |
| Kilimanjaro | | | | | |
| Mokala | 1702 | 4.5 | 154 | 28 | 196 |
| Machame Aleni | 1421 | 1.7 | 168 | 25 | 49 |
| Ikuini | 1160 | 10.7 | 157 | 33 | 128 |
| Kileo | 723 | 6.6 | 175 | 58 | 9 |
| North Pare | | | | | |
| Kilomeni | 1556 | 2.8 | — | — | 322 |
| Lambo | 1187 | 2.5 | — | — | 275 |
| Ngulu | 831 | 6.0 | — | — | 386 |
| Kambi ya Simba | 745 | 10.3 | — | — | 234 |
| South Pare | | | | | |
| Bwambo | 1598 | 3.2 | 178 | 40 | 157 |
| Mpinji | 1445 | 2.8 | 175 | 58 | 128 |
| Goha | 1162 | 10.9 | 172 | 66 | 151 |
| Kadando | 528 | 23.9 | 136 | 99 | 146 |
| West Usambara 1 | | | | | |
| Kwadoe | 1523 | 7.7 | 166 | 48 | 190 |
| Funta | 1279 | 24.1 | 129 | 82 | 92 |
| Tamota | 1176 | 24.8 | 130 | 92 | 181 |
| Mgila | 432 | 38.9 | 125 | 103 | 154 |
| West Usambara 2 | | | | | |
| Magamba | 1685 | 3.2 | — | — | 218 |
| Ubiri | 1216 | 16.6 | — | — | 165 |
| Kwemasimba | 662 | 24.3 | — | — | 242 |
| Mgome | 196 | 48.9 | 100 | 129 | 7 |
| West Usambara 3 | | | | | |
| Emmao | 1845 | 3.7 | — | — | 190 |
| Handei | 1425 | 25.8 | — | — | 383 |
| Tewe | 1049 | 33.4 | — | — | 347 |
| Mngalo | 416 | 47.8 | — | — | 373 |

Tabela 1: Summary data on mean altitude, malaria prevalence, $\alpha$-thalassaemia status across 24 villages located in 6 altitude-related transects.

ges studied where genotyping was actually attempted (scenario I), imputation of missing data from previous 13 villages and an additional village where $\alpha$-thalassaemia genotyping was not attempted

(scenario II), and imputation of all missing data from the 24 villages included in the study (scenario III). In general, the post-imputation genetic effect estimates did not vary dramatically with the type of analysis conducted ($\hat{\lambda}_1 \approx 0.23$ and $\hat{\lambda}_2 \approx 0.18$; Table 2).

| data scenario | $-\log_{10}$(p-value) | $\hat{\lambda}_1$ (SE) | $\hat{\lambda}_2$ (SE) |
|---|---|---|---|
| I. 13 villages | 1.47 | -0.23 (0.12) | 0.18 (0.27) |
| II. 14 villages | | | |
| North Pare | | | |
| Kilomeni | 1.47 | -0.22 (0.13) | 0.19 (0.27) |
| Lambo | 1.59 | -0.24 (0.13) | 0.18 (0.27) |
| Ngulu | 1.62 | -0.23 (0.13) | 0.17 (0.28) |
| Kambi ya Simba | 1.65 | -0.23 (0.13) | 0.20 (0.26) |
| West Usambara 2 | | | |
| Magamba | 1.56 | -0.23 (0.12) | 0.20 (0.27) |
| Ubiri | 1.77 | -0.25 (0.13) | 0.17 (0.27) |
| Kwemasimba | 1.83 | -0.24 (0.13) | 0.18 (0.27) |
| West Usambara 3 | | | |
| Emmao | 1.51 | -0.23 (0.13) | 0.17 (0.27) |
| Handei | 1.67 | -0.23 (0.12) | 0.21 (0.27) |
| Tewe | 1.88 | -0.23 (0.13) | 0.20 (0.27) |
| Mngalo | 1.75 | -0.22 (0.13) | 0.21 (0.27) |
| III. All villages | 2.70 | -0.23 (0.12) | 0.18 (0.25) |

Tabela 2: Post-imputation inferences under different data scenarios where $-\log_{10}$(p-value) is the association signal between $\alpha$-thalassaemia and malaria infection, $\lambda_1$ and $\lambda_2$ are the corresponding genetic effects of 1 and 2 deletions in the $\alpha$-thalassaemia gene, respectively.

In particular, the log-odds of malaria infection appears to decrease 0.23 in individuals with a single $\alpha$-thalassaemia deletion in relation to those without any deletion. Conversely there is a 0.18 increase in the same log-odds but for individuals with two deletions. However, that increase lacks statistical support due to large standard errors resulting from a low number of individuals with that genotype. With respect to the genetic association, the corresponding signals — defined by $-\log_{10}$(p-value) — increased from 0.80 in the CCA to 2.70 in the analysis including all data from the 24 villages. Therefore, the $\alpha$-

thalassaemia gene seems indeed protective against malaria positivity but yet its corresponding effect is moderate at best.

# 4 Concluding remarks

Up to date two studies revealed a protective effect of $\alpha$-thalassaemia from clinical malaria in East Africa [6, 12]. Here we extended evidence of protection to non-clinical malaria. This protection, if true, might also be present in neighbouring areas or countries, such as Kenya or Rwanda, whose malaria epidemiology might be governed by similar environmental drivers. This avenue of research should be followed but definitely requires the use of a large sample size as a moderate association between $\alpha$-thalassaemia and malaria infection was found in this study.

The last word goes to the question on how reliable post-imputation inferences are for the entire data set. In a previous paper, we demonstrated that imputation is useless to predict the missing genotypes accurately [8]. However, post-imputation association signals and corresponding genetic effect estimates are reasonably unbiased and thus trustworthy. We believe that this good performance of imputation on genetic association was achieved because of our well-designed study. In particular, the use of transects and their close relationship with ethnicity was invaluable to minimise putative biases due to latent population structure in the study area. The use of altitude-based sampling has also ensured the control of malaria endemicity levels across villages. The additional matching for gender and age has minimised further confounding effects. Data imputation seems facilitated in this scenario because genetic association could be well informed by borrowing information from individuals with similar sociological, environmental and malaria epidemiology characteristics. We then conclude that future studies should be well thought not only to control putative confounding effects, but also to ensure data imputation produces reliable association results.

# Acknowledgements

Nuno Sepúlveda was funded by the Wellcome Trust grant number 091924 and Fundação para a Ciência e Tecnologia through the project Pest-OE/MAT/UI0006/2011. Taane Clark is funded by the UK Medical Research Council and Wellcome Trust. Alphaxard Manjurano is funded by the Royal Society-The Leverhulme Africa Award. Medical Research Council has funded the main study UK Medical Research Council (grant G9901439), MalariaGEN supported genotyping of $\alpha$-thalassaemia locus.

# Referências

[1] Beall, C.M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integrative and Comparative Biology* 46, 18–24.

[2] Drakeley, C.J., Carneiro, I., Reyburn, H., Malima, R., Lusingu, J.P.A., Cox, J., Theander, T.G., Nkya, W.M.M.M., Lemnge, M.M., Riley, E.M. (2005). Altitude-Dependent and -Independent Variations in Plasmodium falciparum Prevalence in Northeastern Tanzania. *Journal of Infectious Diseases* 191, 1589–1598.

[3] Drakeley, C.J., Corran, P.H., Coleman, P.G., Tongren, J.E., McDonald, S.L., Carneiro, I., Malima, R., Lusingu, J., Manjurano, A., Nkya, W.M., Lemnge, M.M., Cox, J., Reyburn, H., Riley, E.M. (2005). Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure. *Proceedings of the National Academy of Sciences USA* 102, 5108–5113.

[4] Enevold, A., Alifrangis, M., Sanchez, J.J., Carneiro, I., Roper, C., Borsting, C., Lusingu, J., Vestergaard, L.S., Lemnge, M.M., Morling, N., Riley, E., Drakeley, C.J. (2007). Associations between $\alpha^+$-thalassemia and Plasmodium falciparum malarial infection in northeastern Tanzania. *Journal of Infectious Diseases* 196, 451–459.

[5] Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics* 77, 171–192.

[6] Manjurano, A., Clark, T.G., Nadjm, B., Mtove, G., Wangai, H., Sepúlveda, N., Campino, S.G., Maxwell, C., Olomi, R., Rockett, K.R., Jeffreys, A., MalariaGen Consortium, Riley, E.M., Reyburn,

H., Drakeley, C. (2012). Candidate Human Genetic Polymorphisms and Severe Malaria in a Tanzanian Population. *PLoS One* 7, e47463.

[7] Rubin, D.B. (1996). Multiple Imputation after 18+ years. *Journal of the American Statistical Association* 91, 473–489.

[8] Sepúlveda, N., Manjurano, A., Drakeley, C., Clark, T.G. (2014). On the performance of multiple imputation based on chained equations in tackling missing data of the African $\alpha^{3.7}$-globin deletion in a malaria association study. *Annals of Human Genetics* 78, 277–289.

[9] Souverein, O.W., Zwinderman, A.H., Tanck, M.W.T. (2006). Multiple Imputation of Missing Genotype Data for Unrelated Individuals. *Annals of Human Genetics* 70, 372–381.

[10] Taylor, S.M., Cerami, C., Fairhurst, R.M. (2013). Hemoglobinopathies: Slicing the Gordian Knot of Plasmodium falciparum Malaria Pathogenesis. *PLoS Pathog.* 9:e1003327.

[11] van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16, 219–242.

[12] Williams, T.N., Wambua, S., Uyoga, S., Macharia, A., Mwacharo, J.K., Newton, C.R., Maitland, K. (2005). Both heterozygous and homozygous $\alpha^{+}$-thalassaemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood* 106, 368–371.

# Porque duram tanto tempo algumas dissertações de Mestrado?

Rita Freitas
MMEAD e ECT, Universidade de Évora, *ritabf8@gmail.com*
Paulo Infante
CIMA-UE e ECT, Universidade de Évora, *pinfante@uevora.pt*
Gonçalo Jacinto
CIMA-UE e ECT, Universidade de Évora, *gjcj@uevora.pt*
Fernanda Figueiredo
CEAUL e FEP, Universidade do Porto, *otilia@fep.up.pt*
João Dias
Serviços Académicos da Universidade de Évora, *jpsd@uevora.pt*

**Palavras–chave**: Análise de sobrevivência, controlo de qualidade, tempo até conclusão de uma dissertação de mestrado

**Resumo**: O tempo até conclusão dos cursos superiores assume uma particular importância, pois tem impacto na projeção da Universidade para o exterior. Neste trabalho, com base nos dados recolhidos pelos Serviços Académicos da Universidade de Évora, pretende-se estudar o tempo até conclusão de uma dissertação de Mestrado. Recorrendo à análise de sobrevivência conclui-se que a idade, o ano de ingresso, a fase de candidatura, a duração da parte curricular e o estatuto de trabalhador estudante são fatores significativos na duração de uma dissertação de Mestrado. Como os trâmites administrativos também podem influenciar significativamente o tempo de duração do Mestrado, mostramos como a implementação de cartas de controlo pode ajudar a identificar causas que fazem aumentar ou diminuir esse tempo.

# 1 Introdução

O tempo para concluir um curso superior pode ter impacto na decisão de reformulação dos cursos. Neste trabalho pretende-se identificar fatores com impacto significativo no tempo até conclusão de uma dissertação de Mestrado e analisar o tempo desde a sua entrega nos serviços académicos (SAC) até à sua discussão pública.
Para modelar o tempo até conclusão da dissertação recorre-se à análise de sobrevivência, considerando preditores sócio-demográficos e académicos. Esta temática foi abordada em [3], onde os autores recorreram à análise de sobrevivência para estudar o tempo de permanência dos estudantes num curso de Física e em [4] onde se modela o tempo até conclusão de um curso de 1.º Ciclo.
Após a entrega da dissertação, outros procedimentos administrativos podem atrasar ainda mais o tempo até conclusão do Mestrado. Neste sentido, recorremos ao controlo estatístico de qualidade, implementando cartas de controlo restrospetivas, com o objetivo de dar *feedback* aos serviços e de motivar a sua utilização para monitorização em tempo real. Com esta abordagem pretende-se identificar causas assinaláveis que são responsáveis por tempos mais longos. Como alguns dos tempos desde a entrega da dissertação até à discussão pública eram muito assimétricos, aplicámos cartas de controlo baseadas na distribuição normal assimétrica e desenvolvidas em [1].

# 2 Caracterização da amostra

A amostra é constituída pelos alunos que ingressaram na Universidade de Évora entre os anos letivos de 2007/2008 e 2012/2013, tendo sido seguidos até Julho de 2013. O tempo até conclusão de uma dissertação foi definido como o intervalo de tempo entre a entrega do projeto de dissertação e a entrega da dissertação. A amostra é constituída por 1428 alunos que terminaram ou podiam ter terminado a dissertação no final do período de *follow-up*. Para esta amostra 60% dos alunos ingressaram em cursos da área das Ciências Sociais (CS),

39% em cursos da área das Ciências e Tecnologia (CT) e os restantes em cursos da área de Artes (A). Têm idades compreendidas entre 20 e 60 anos, sendo a maioria destes alunos do sexo feminino (60%), ingressaram na Universidade na 1ª ou 2ª fase (95%), não beneficiaram do estatuto de trabalhador estudante (TE) (84%) e terminaram a parte curricular em não mais de 3 semestres (54%).

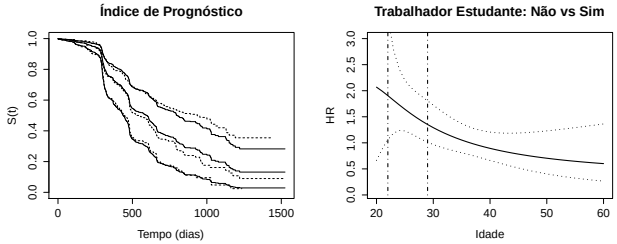
# 3 Modelação do tempo até conclusão de uma dissertação de Mestrado

Atendendo a que existem dados censurados (alunos ativos e inativos que estavam a realizar a dissertação no final do período de *follow-up*), recorreu-se a um modelo de Cox, tendo sido ajustado de acordo com os passos sugeridos em [2] e recorrendo ao *software R Project 3.0.1.* Através deste modelo, pretende-se estudar o tempo até a conclusão de uma dissertação de Mestrado, estimando a razão de riscos (*hazard ratio*, *HR*) de um determinado grupo de indivíduos em relação a outro. O modelo semiparamétrico de riscos proporcionais de Cox ajusta a função de risco $h(t)$, considerando um risco basal $h_0(t)$ e incluindo o vetor de covariáveis $\mathbf{x}$, de forma que

$$h(t|\mathbf{x}) = h_0(t)exp(x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p) = h_0(t)exp(\mathbf{xB}).$$

Esta formulação assume que as covariáveis têm um efeito multiplicativo na função de risco e deste modo a razão entre o risco de ocorrência do evento para dois indivíduos é constante no tempo. O pressuposto dos riscos proporcionais foi validado via teste de Harrell, via análise de resíduos de Schoenfeld e via representação das curvas log-log. O índice de prognóstico (IP) associado a este modelo está representado na Figura 1(a), traduzindo-se num modelo aceitável. As linhas a cheio referem-se à função de sobrevivência empírica e as linhas a tracejado às estimativas de Kaplan-Meier, obtidas para grupos de alto, médio e baixo IP. Recorde-se que o IP é o preditor linear do modelo de Cox calculado para cada indiví-

duo utilizando as covariáveis observadas e as estimativas dos coeficientes de regressão do modelo ajustado. Como a variável idade não verificava o pressuposto da linearidade, o método dos polinómios fracionários sugeriu a transformação: $Idade1 = (idade/10)^{-2}$ e $Idade2 = (idade/10)^{-2} log(idade/10)$. No modelo final, cujos coe-




(a) (b)

Figura 1: (a) Índice de prognóstico, $S(t)$, do modelo final; (b) $HR$ para alunos que não obtiveram o estatuto de TE.

ficientes estão representados na Tabela 1, agrupámos a variável ano de ingresso em três categorias. Optámos por agrupar os alunos com ingresso em 2007 e 2008, devido ao número reduzido de dados para o ano 2007, e na fase de modelação foram agrupadas as categorias correspondentes aos anos 2009 e 2010, pois os coeficientes correspondentes no modelo de Cox não eram significativamente diferentes entre si (verificado pelo teste de razão de verosimilhanças com valor $p = 0{,}91$). Os tempos até conclusão da parte curricular apresentavam alguns dados censurados (alunos em condições de terminar a dissertação, mas que não terminaram a parte curricular), optando-se por categorizar a variável usando como ponto de corte 3 semestres.
Para valores fixos das restantes variáveis podemos concluir que um aluno que ingressou na Universidade na primeira ou segunda fase tem um risco quase duas vezes superior ($1/HR = exp(0{,}58) = 1{,}78$; $IC_{95\%} = (1{,}27; 2{,}49)$) de concluir a dissertação mais cedo. Na Figura

Tabela 1: Coeficientes estimados ($\hat{\beta}$) do modelo de Cox, respetivos desvios padrão estimados ($\hat{\sigma}_{\hat{\beta}}$) e valores $p$ (teste de Wald) associados, para o modelo final (coef. de concordância = 0,62; $R^2$ = 0,11).

| Covariável | $\hat{\beta}$ | $\hat{\sigma}_{\hat{\beta}}$ | Valor p |
|---|---|---|---|
| Idade1 (I1) | 13,39 | 2,95 | <0,001 |
| Idade1 (I2) | -14,49 | 5,52 | 0,009 |
| Ano de Ingresso (AI) | | | |
| 2007/2008 (referência) | | | |
| 2009/2010 | 0,1 | 0,09 | 0,301 |
| 2011 | 0,78 | 0,22 | <0,001 |
| Fase de Candidatura (FC) | | | |
| 1ª ou 2ª (referência) | | | |
| 3ª ou 4ª | -0,58 | 0,17 | <0,001 |
| Trabalhador Estudante (TE) | | | |
| Sim (referência) | | | |
| Não | 1,06 | 1,06 | 0,315 |
| Duração da parte Curricular (DC) | | | |
| $\leq$ 3 semestres (referência) | | | |
| $>$ 3 semestres | -0,77 | 0,13 | <0,001 |
| I1 x TE (sim) | 0,88 | 11,73 | 0,940 |
| I2 x TE (sim) | -11,59 | 19,73 | 0,557 |
| AI (2009/2010) x DC ($>$ 1,5 anos) | 0,42 | 0,16 | 0,007 |
| AI (2011) x DC ($>$ 1,5 anos) | 0,67 | 0,41 | 0,102 |

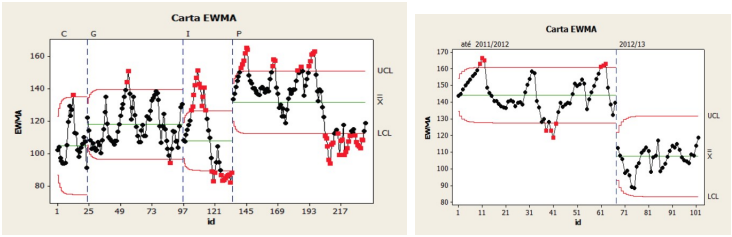
1(b), representamos a *HR* para os alunos que não obtiveram o estatuto de TE relativamente aos que beneficiaram deste estatuto, em função da idade (as linhas verticais marcam as idades entre as quais a diferença é significativa). Podemos observar que para os alunos com idades compreendidas entre os 22 e 29 anos, os que não obtiveram o estatuto de TE têm um risco maior em terminar a dissertação mais cedo. Também para interpretar a variável ano de ingresso é necessário fixar a variável tempo da parte curricular e vice-versa. No caso dos alunos que demoraram mais de 3 semestres a termi-

nar a parte curricular, os que ingressaram em 2009 ou 2010 têm um risco quase duas vezes superior ($HR = exp(0{,}1 + 0{,}42) = 1{,}68$; $IC_{95\%} = (1{,}31;\ 2{,}15)$) de concluir a tese mais cedo relativamente aos que ingressaram até 2008, aumentando esse risco para mais de 4 vezes ($HR = exp(0{,}78 + 0{,}67) = 4{,}28$; $IC_{95\%} = (2{,}14;\ 8{,}56)$) para os que ingressaram em 2011. Para os alunos que terminaram a parte curricular em não mais de 3 semestres apenas se registaram diferenças significativas entre os que entraram até 2008 e os que entraram em 2011 ($HR = exp(0{,}78) = 2{,}19$; $IC_{95\%} = (1{,}43;\ 3{,}35)$). Por outro lado, podemos concluir que para os alunos que ingressaram até 2010, quem concluiu a parte curricular mais cedo tem também maior risco de concluir a dissertação mais cedo, sendo o risco mais acentuado para os que ingressaram até 2008 ($1/HR = exp(0{,}77) = 2{,}16$; $IC_{95\%} = (1{,}68;\ 2{,}77)$) do que para os que ingressaram em 2009 ou 2010 ($1/HR = exp(0{,}77 - 0{,}42) = 1{,}42$; $IC_{95\%} = (1{,}18;\ 1{,}70)$).

# 4 Controlo de qualidade do tempo entre a entrega e discussão da dissertação

Nesta secção vamos analisar o intervalo de tempo que decorre entre a entrega da dissertação e a sua discussão pública, o qual definimos por tempo até discussão (TD). Uma análise exploratória do TD permitiu identificar tempos excessivamente elevados (superiores a 180 dias), pelo que nesta fase optámos por estudar os TD inferiores a esse valor (que já inclui tempos 60 dias superior ao máximo estabelecido pelo regulamento dos mestrados), resultando numa amostra de 816 alunos diplomados que concluíram o Mestrado até ao ano letivo de 2012/13. A variável em estudo apresenta algum afastamento em relação à distribuição normal, o que nos levou a aplicar cartas de controlo EWMA, por serem mais robustas relativamente a pequenos afastamentos da normalidade e também mais eficazes na detecção de pequenas alterações da média. Foram elaboradas várias cartas em que se estratificou pelo ano de conclusão do Mestrado, Escola e curso de Mestrado. Algumas cartas obtidas para os TD de alguns

cursos estão representadas na Figura 2(a).




(a) (b)

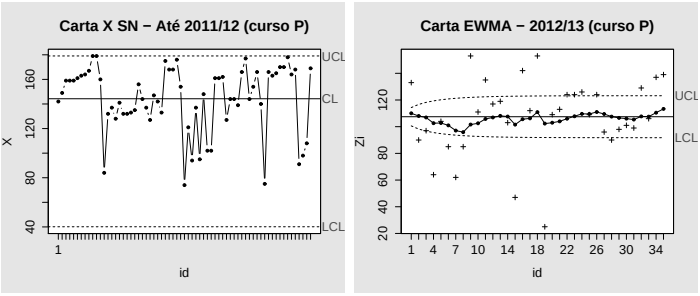
Figura 2: Cartas de Controlo EWMA: (a) valores TD estratificando pelos cursos C, G, I e P; (b) valores TD do curso P.

Uma análise a estas cartas sugeriu que o processo não estava sob controlo. Em particular, para o curso aqui denominado por P, devido a confidencialidade dos dados, podemos observar que a carta de controlo da Figura 2(b) reflete comportamentos distintos em dois horizontes temporais. Investigada a causa responsável por tal comportamento, verificou-se que no final do ano letivo de 2011/2012, houve alterações relacionadas com processos de formação do júri, conduzindo a uma redução do TD. Em seguida, implementámos cartas de controlo para cada um dos períodos do curso P. Dada a assimetria muito acentuada dos TD do primeiro período (Mestrado concluído até 2011/2012), ajustámos uma distribuição Skew-Normal (SN) [1]. Para os parâmetros de interesse (Tabela 2), o teste de ajustamento do qui-quadrado (valor $p = 0{,}09$) permitiu assumir a distribuição dos tempos como SN, pelo que foi construída uma carta X-SN (Figura 3(a)), por métodos de Bootstrap (limites de controlo obtidos: (LCL=40,1;UCL=179,1)), usando uma adaptação do algoritmo definido em [1]. Ao analisar as Figuras 2(b) e 3(a), podemos observar comportamentos distintos entre as cartas EWMA e X-SN do primeiro período do curso P, revelando a importância da distribuição assumida para a característica da qualidade em estudo. No caso

particular do curso P, apesar do processo estar sob controlo (algo que não concluímos com a carta EWMA anterior), a sua capacidade para cumprir as especificações de 120 dias é muito baixa. Neste caso, a probabilidade do TD durar mais de 120 dias é muito elevada (0,82), que corresponde a um índice de capacidade do processo cpku=0,17. Já para o segundo período, representado por uma carta EWMA (uma vez que neste caso a distribuição dos tempos é normal) na Figura 3(b), essa probabilidade é igual a 0,32 e o cpku=0,16. É pois necessário implementar medidas para tornar o processo mais capaz e para tal estamos a estudar os tempos parciais que compõem o TD.

Tabela 2: Parâmetros da distribuição SN para os TD do 1.º período.

| Loc. ($\lambda$) | Escala ($\delta$) | Forma ($\alpha$) | Média | sd | Assim. |
|---|---|---|---|---|---|
| 179,14 | 43,82 | -861,21 | 144,18 | 26,41 | -0,99 |




(a) (b)

Figura 3: (a) Carta de Controlo X-SN para os valores TD do curso P do primeiro período; (b) Carta de Controlo EWMA para os valores TD do curso P do segundo período.

# 5 Considerações finais

Este trabalho permitiu identificar alguns fatores que influenciam o tempo até conclusão de uma dissertação e que têm uma influência direta na conclusão de um curso de Mestrado. Podemos concluir que os alunos que ingressaram na Universidade depois da segunda fase, nos anos de 2007 ou 2008, que não terminaram a parte curricular no tempo previsto e que beneficiaram do estatuto de TE quando ainda jovens têm um risco muito elevado de demorarem mais tempo a terminar a sua dissertação. Apresentámos também um exemplo de monitorização do TD para um Mestrado, onde se exemplifica não só a mais valia de implementação de cartas de controlo nestes serviços, como a importância da seleção da carta adequada em função da distribuição da característica a ser monitorizada. A monitorização do processo por cartas de controlo aos tempos parciais poderá identificar as causas que tornem o processo mais capaz, reduzindo o tempo de espera por prestação de provas e, consequentemente, melhorando a qualidade do serviço desta Universidade.

# Referências

[1] Figueiredo, F., Gomes M. (2013). The Skew-Normal Distribution in SPC. *REVSTAT - Statistical Journal* 11, 83–104.

[2] Hosmer, D., Lemeshow, S. (2008). *Applied Survival Analysis* (2nd Edition). John Wiley & Sons, New York.

[3] Junior, P., Silveira, F., Ostermann, F. (2012). Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em Física: um exemplo de uma universidade brasileira. *Revista Brasileira de Ensino de Física* 34, 1403.1-1403.10.

[4] Teixeira, V., Infante, P., Dias, J. (2012). Modelação do tempo até conclusão de um curso superior. *Estatística Novos Desenvolvimentos e Inspirações* (Eds. Maia, M., Campos, P. e Duarte Silva, P.), Edições SPE, pp. 271-284.

# Sobrevivência relativa do cancro colo-rectal e do estômago no sul de Portugal

Ricardo São João
CEAUL & Escola Superior de Gestão e Tecnologia de Santarém-Instituto Politécnico de Santarém, *ricardo.sjoao@esg.ipsantarem.pt*

Ana Luisa Papoila
CEAUL & Faculdade de Ciências Médicas-Universidade Nova de Lisboa, *ana.papoila@fcm.unl.pt*

Ana Miranda
Registo Oncológico Regional Sul-ROR Sul, *amiranda@ipolisboa.min-saude.pt*

**Palavras–chave**: Sobrevivência relativa, cancro colo-rectal, cancro do estômago, ROR-Sul

**Resumo**: A análise de sobrevivência é de importância ímpar na área das Ciências Biomédicas, em particular na oncologia onde constitui um instrumento de monitorização das atividades de controlo do cancro. Dado que a morte de um doente poderá ficar a dever-se a outras causas que não o cancro e que a causa específica de morte pode ser desconhecida, foi proposta uma medida objetiva da sobrevivência designada por sobrevivência relativa, no sentido de identificar qual a contribuição da neoplasia em si para a sobrevivência. Não dependendo da causa específica de morte, esta é uma medida útil na monitorização da sobrevivência, permitindo comparações entre grupos étnicos, regiões e registos de cancro de base populacional. Com base nos casos diagnosticados entre 1998 e 2006, registados pelo Registo Oncológico Regional Sul (ROR-Sul), o presente estudo tem como objetivo estimar a sobrevivência relativa considerando 11859 e 21575 doentes diagnosticados com cancro do estômago e colo-rectal, respetivamente. Foi considerado um período de *follow-up* de cinco anos. Procurar-se-á identificar a existência de diferenças significativas na sobrevivência relativa quanto ao tipo de neoplasia, idade,

género e região geográfica.

# 1 Sobrevivência Relativa

O interesse de alguns dos estudos baseados em registos de cancro de base populacional (RCBP) recai sobre a mortalidade associada ao diagnóstico de determinado tipo de neoplasia. Nesse sentido, a mortalidade por causa específica é habitualmente utilizada para estimar a mortalidade atribuível apenas ao cancro em estudo (*net survival*). No entanto, existem muitas situações em que a causa de morte é desconhecida e, mesmo que esta informação esteja disponível através dos certificados de óbito, não é fácil distinguir os casos em que a principal causa de morte é devida ao cancro. Assim sendo, a sobrevivência relativa surge como uma medida objetiva que não necessita do conhecimento da causa específica da morte para o seu cálculo, permitindo assim controlar as diferenças na mortalidade por causas de morte que não o cancro (Pohar e Stare, 2006)[6]. Considerando a variável aleatória $T$ que representa o tempo até à morte por cancro, a função de sobrevivência relativa acumulada no instante $t$ define-se como

$$r(t) = \frac{S_O(t)}{S_E(t)}, \tag{1}$$

onde $S_O(t)$ representa a sobrevivência observada no grupo de doentes com cancro e $S_E(t)$ representa a sobrevivência esperada num grupo comparável da população geral, livre da doença (grupo de correspondência). Dada a dificuldade em obter uma coorte de indivíduos sem cancro, a $S_E(t)$ é estimada a partir das tábuas de mortalidade que representam a sobrevivência da população em geral. Estas tábuas consideram as mortes pelo cancro em estudo mas, ainda assim, Ederer et al., (1961) [4] mostram que este facto é irrelevante na medida em que estas mortes constituem uma ínfima parte da mortalidade global. A $S_E(t)$ é estimada por métodos que diferem entre si quanto à mensuração do tempo em risco e quanto à definição do grupo de correspondência (Dickman et al., 2013)[2]. Neste estudo foi adotado

o método Ederer II [3] baseado nas tábuas de mortalidade disponibilizadas pelo Instituto Nacional de Estatística. Estas tábuas, estratificadas por sexo, região [3] e grupos etários anuais dos 0 aos 99 anos, referem-se ao período 1998-2010. Da equação (1) resulta que $r(t)$ só admite valores positivos sendo habitualmente expressa em percentagem (SR). Usualmente, a SR toma valores inferiores a 100%, refletindo o excesso de mortalidade nos doentes diagnosticados com cancro face à restante população considerada livre da doença. Menos usual é o caso de valores de SR superiores a 100% que podem, esporadicamente, resultar da potencial cura dos doentes ou da sua maior monitorização comparativamente à restante população.

## 2 Análise por período

A esperança de vida pode ser estimada com base na utilização de tábuas de mortalidade por coorte (*cohort life tables*) ou por período (*period life tables*). As primeiras são de aplicação limitada pois contêm a probabilidade de morte de coortes de indivíduos. Tal facto implica que as estimativas da sobrevivência esperada sejam baseadas em tábuas de mortalidade relativas a anos longínquos, não traduzindo o progresso recente nas taxas de sobrevivência como reflexo de avanços na deteção e tratamento do cancro. Em oposição, as tábuas de mortalidade por período irão contemplar doentes de todas as idades num período fixo, não se limitando a uma geração, tornando mais atuais as estimativas de sobrevivência. Face ao exposto, efetuou-se uma análise por período. A determinação do período em análise deve ser feita tendo em conta a informação mais recente disponibilizada pelo registo oncológico e que permita maximizar o número de anos de diagnóstico [1, 5]. Neste estudo, este período corresponde aos anos de $2003 - 2006$ onde estão contemplados todos os anos de diagnóstico (1998 a 2006) para as referidas neoplasias.

[3] segundo o local de residência NUTS-2002 para as regiões que constituem o registo ROR-Sul: Lisboa, Alentejo, Algarve e Região Autónoma da Madeira (RAM).

# 3 Modelos de regressão de sobrevivência relativa

O risco individual de cada doente $\lambda$ é expresso em função do risco esperado $\lambda^*$ (devido à sua idade, sexo, ano de diagnóstico ou outra combinação de variáveis incluídas nas tábuas de mortalidade) e do risco específico de cancro $\nu$. A forma como $\lambda$ é decomposto irá ditar a natureza do modelo. Se expresso numa soma de riscos, $\lambda = \lambda^* + \nu$, então tratar-se-à de um modelo aditivo; se expresso num produto de riscos, então teremos um modelo multiplicativo. Os modelos aditivos são os mais utilizados em oncologia.

Os modelos de regressão podem ser implementados através de uma abordagem convencional [1] ou, alternativamente, através de uma abordagem recentemente proposta por Holleczek & Brenner (2013)[5]. Esta "herda" a estrutura dos modelos lineares generalizados permitindo ter em conta o efeito de covariáveis adicionais para além do tempo de *follow-up*. Este facto ditou a sua implementação no presente estudo. Por uma questão de simplicidade definiremos o modelo apenas com os preditores ano de *follow-up* ($y$) e grupo etário ($agr$), sendo extensível a outras covariáveis. Foi considerado um período de *follow-up* de 5 anos e o grupo etário apenas com duas categorias ($agr$= 0:<65 anos, 1:$\geq$ 65 anos). O modelo proposto tem como variável resposta o número de óbitos para cada combinação de $y$ e $agr$.

Considere-se $d_{y,agr}$ o número de óbitos, $l_{y,agr}$ o número de pessoas em risco, $e_{y,agr}$ o número esperado de óbitos e $\lambda_{y,agr}$ a taxa de mortalidade, onde

- $d_{y,agr}$ segue uma distribuição Poisson com parâmetro $\mu_{y,agr} = \lambda_{y,agr} \times l_{y,agr}$ ;
- $\mathbf{x} = (x_y, x_{agr})$ é o vetor de covariáveis;
- $\alpha_y$ e $\beta$ são os coeficientes associados a $y$ e a $agr$, respetivamente;

- a função de ligação é $ln\left(\mu_{y,agr} - d^*_{y,agr}\right)$ onde

$$d^*_{y,agr} = -\left(l_{y,agr} - \frac{d_{y,agr}}{2}\right) \times ln\left(\frac{l_{y,agr} - e_{y,agr}}{l_{y,agr}}\right);$$

- $ln\left(l_{y,agr} - \frac{d_{y,agr}}{2}\right)$ é o *offset* do modelo.

O modelo é dado por

$$ln\left(\mu_{y,agr} - d^*_{y,agr}\right) = ln\left(l_{y,agr} - \frac{d_{y,agr}}{2}\right) + \sum_{y=1}^{5} \alpha_y x_y + \beta x_{agr}, \quad (2)$$

A interpretação dos coeficientes do modelo passa pela dupla exponenciação dos mesmos. A sobrevivência relativa no grupo de doentes mais jovens para o ano $y$ é dada por

$$r_{y,agr=0} = exp(-exp\left(\alpha_y\right)),$$

e por

$$r_{y,agr=1} = exp(-exp\left(\alpha_y + \beta\right))$$

para o grupo de doentes mais idosos. Para um *follow-up* de 5 anos, a sobrevivência relativa acumulada em cada um dos grupos etários é dada por

$$R_{y,agr} = \prod_{y=1}^{5} r_{y,agr}.$$

Para além do grupo etário e do tempo de *follow-up*, foram ainda incluídas a região e o sexo, pela mesma razão. No que diz respeito às categorias de referência, consideraram-se o grupo etário mais jovem, a região do Alentejo e o sexo masculino. Para a implementação do modelo foi utilizado o pacote `periodR` [5, 7] e rotinas em linguagem `S` desenvolvidas pelos autores.

# 4 Resultados

Na análise da sobrevivência relativa é necessário que a percentagem de registos com data de diagnóstico coincidente com a data do certificado de óbito (DCO-*Death Certificate Only*) não ultrapasse 15% do número total de casos. Neste estudo, o número total de casos diagnosticados no período 1998-2006 para os cancros do cólon, recto e estômago é respetivamente de 14149, 7426 e 11859, com apenas 0.70%, 0.33% e 0.86% de DCOs. Estes casos foram excluídos da análise uma vez que o tempo de *follow-up* é nulo, não trazendo qualquer contributo ao estudo (Holleczek et al., 2013)[5]. Foram considerados para a análise doentes com idades à data do diagnóstico entre os 15 e os 99 anos.

## 4.1 SR no cancro do cólon

A SR estimada no primeiro ano de *follow-up* de doentes com idade inferior a 65 anos e residentes no Alentejo é de 77.1%: exp[-exp(-1.3465)]=0.771 e a acumulada ao fim de 5 anos é de 55.21%. Considerando ainda o primeiro ano de *follow-up*, idades mais avançadas

| Preditores | Estimativa | Erro Padrão | z | Pr(>\|z\|) |
|---|---|---|---|---|
| *follow-up* 1 | -1.3465 | 0.0498 | -27.04 | <0.001 |
| *follow-up* 2 | -2.0844 | 0.0574 | -36.29 | <0.001 |
| *follow-up* 3 | -2.4213 | 0.0681 | -35.56 | <0.001 |
| *follow-up* 4 | -2.6299 | 0.0821 | -32.04 | <0.001 |
| *follow-up* 5 | -3.0273 | 0.1150 | -26.34 | <0.001 |
| $\geq$ 65 anos | 0.2926 | 0.0324 | 9.04 | <0.001 |
| Algarve | -0.0374 | 0.0666 | -0.56 | 0.575 |
| Lisboa | -0.1483 | 0.0460 | -3.23 | <0.001 |
| RAM | 0.1259 | 0.0960 | 1.31 | 0.190 |

Tabela 1: Resultados do ajustamento do modelo aos dados do cancro do cólon.

influenciam de forma negativa a sobrevivência, com uma redução de cerca de 6.5% na estimativa da SR: exp[-exp(-1.3465+.2926)]=0.706. Os doentes residentes na região de Lisboa e com idade inferior a 65

anos, possuem um pequeno acréscimo de 2.8% na estimativa da SR: exp[-exp(-1.3465-0.1483)]=0.799, comparativamente aos doentes residentes no Alentejo com a mesma idade. Constatou-se que o sexo não tem influência na SR. No cálculo da SR, poderiam ainda ter sido contempladas outras combinações entre as diversas categorias de cada covariável.

### 4.2 SR no cancro do recto

A SR estimada no primeiro ano de *follow-up* de doentes com idade inferior a 65 anos e residentes no Alentejo é de 83.2% e a acumulada é de 59.28% ao fim de 5 anos. Verificou-se que o grupo de doentes com pelo menos 65 anos de idade apresenta uma redução de cerca de 7.3% na estimativa da SR. Constatou-se que o sexo não tem influência na sobrevivência.

### 4.3 SR no cancro do estômago

A estimativa da SR é reduzida no primeiro ano após o diagnóstico de cancro do estômago (56.76%), bem como a acumulada (34.5%) ao fim de 5 anos. Uma maior idade influencia de forma negativa a sobrevivência, com uma redução de cerca de 11.0% na estimativa da SR (45.9%). Na região de Lisboa detetou-se um acréscimo de 6% na SR (62.8%) comparativamente ao Alentejo. Verificou-se que no sexo feminino a estimativa da SR é superior à do sexo masculino (5.2%).

## 5 Conclusões

Os resultados evidenciaram diferenças significativas na SR quanto ao tipo de neoplasia, idade, género e região. O cancro do recto é aquele que possui melhor prognóstico de sobrevivência. Em relação à idade, o grupo etário dos doentes mais idosos apresentou uma estimativa da SR inferior. No que diz respeito ao género, encontraram-se diferenças na neoplasia do estômago, com valores superiores para o sexo feminino. A região não influencia a SR no cancro do recto.

## Agradecimentos

Este trabalho é financiado pela FCT, Fundação para a Ciência e Tecnologia, Projecto PEst-OE/MAT/UI0006/2014.

# Referências

[1] Brenner, H., Hakulinen, T., Gefeller, O. (2004). Period analysis for up-to-date cancer survival data: theory, empirical evaluation, computacional realization and applications. *European Journal of Cancer* 40, 326–335.

[2] Dickman, P.W., Coviello, E., Hills, M. (2013). Estimating and modeling relative survival. *The Stata Journal* VV, Number ii, 1–25.

[3] Ederer, F., Heise, H. (1959). *Instructions to IBM 650 programmers in processing survival computations*. Methodological note no. 10. End Results Evaluation Section. National Cancer Institute, Bethesda(MD).

[4] Ederer, F., Axtell, L.M., Cutler, S.J. (1961). The relative survival rate: A statistical methodology. *National Cancer Institute Monograph* 6, 101–121.

[5] Holleczek, B., Brenner, H. (2013). Model based period analysis of absolute and relative survival with R: Data preparation, model fitting and derivation of survival estimates. *Computer Methods and Programs in Biomedicine* 110, 192–202.

[6] Pohar, M., Stare, J. (2006). Relative survival analysis in R. *Computer Methods and Programs in Biomedicine* 81, 272–278.

[7] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/

# Um estudo de simulação para avaliar a performance de estimadores para a taxa de prevalência usando testes compostos

Ricardo Sousa
Escola Superior de Tecnologia da Saúde de Lisboa, Instituto Politécnico de Lisboa, CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, *ricardo.sousa@estesl.ipl.pt*

Rui Santos
Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, *rui.santos@ipleiria.pt*

João Paulo Martins
Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, *jpmartins@ipleiria.pt*

**Palavras–chave**: Sensibilidade, especificidade, taxa de prevalência, testes compostos

**Resumo**: Este trabalho analisa a performance de estimadores pontuais (analiticamente) e intervalares (via simulação) para a taxa de prevalência, quer na ausência quer na presença de erros de classificação, com o objetivo de comparar os resultados obtidos utilizando testes compostos com os obtidos recorrendo aos testes individuais.

## 1 Introdução

Os testes compostos, testes aplicados a uma mistura homogénea de amostras recolhidas a partir de $n$ indivíduos, surgem na literatura estatística com o trabalho de Dorfman [3]. Este tinha por objetivo determinar um valor ótimo para $n$ em função da taxa de prevalência, no sentido de minimizar o número esperado de testes necessários

para a identificação de todos os indivíduos infetados. Todavia, o recurso a análises conjuntas não visa unicamente a classificação de indivíduos, podendo ser igualmente útil na estimação da respetiva taxa de prevalência. Para este fim, perante um resultado positivo do teste composto, não é necessária a posterior realização de testes individuais (uma vez que o objetivo já não passa pela classificação dos indivíduos mas apenas pela estimação da taxa de prevalência, tendo ainda a vantagem de manter o anonimato dos indivíduos infetados). Além disso, os estimadores da prevalência obtidos pela aplicação de testes compostos têm, sob determinadas condições, melhor comportamento que os estimadores tradicionais baseados em testes individuais, cf. [4, 9, 13]. O enviesamento, a eficiência e a robustez destes estimadores foram examinados em diversos artigos, tais como os de [2, 5, 6]. Assim, os estimadores baseados em testes compostos permitem, para taxas de prevalência baixas, não apenas a obtenção de ganhos monetários (diminuindo o número de testes efetuados) mas também a obtenção de estimativas mais precisas, em comparação com as obtidas com base em testes individuais. Na maioria das aplicações de testes compostos é considerado que os testes classificam sempre corretamente os indivíduos, uma simplificação que é pouco realista em muitas aplicações. Por conseguinte, nesta análise irá ser tida em consideração a possibilidade de ocorrência de erros de classificação. Todavia, iremos supor que as probabilidades associadas aos erros de classificação são iguais nos testes individuais e nos testes compostos (não dependem da dimensão do grupo) cf. [7], i.e., que não ocorre (ou que é negligenciável) o efeito de rarefação/diluição (o qual pode implicar o aumento da probabilidade de ocorrência de falsos negativos com o aumento da dimensão do grupo, cf. [11]). Deste modo, o principal objetivo deste artigo consiste em comparar a performance de diferentes metodologias para estimar a taxa de prevalência, utilizando testes individuais e testes compostos, em situações onde se admite a ocorrência de erros de classificação. Para tal, na secção 2 serão analisados estimadores pontuais, sendo as suas características determinadas analiticamente, e na Secção 3 são comparados diversos estimadores intervalares via simulação.

# 2 Estimação pontual

Consideremos uma população composta por $N$ indivíduos distribuídos por $m$ grupos de dimensão $n$, na qual a probabilidade de infeção de um indivíduo é igual a $p$. Deste modo, podemos caraterizar os membros da população através das variáveis aleatórias (v.a.) $X_i$, $i = 1,2,\ldots,N$, com distribuição de Bernoulli de parâmetro $p$, onde cada v.a. $X_i$ assume os valores 1 ou 0 consoante o $i$-ésimo indivíduo esteja ou não infetado. Assumindo que a constituição de cada grupo é determinada aleatoriamente (indivíduos independentes dentro do grupo) então o número de indivíduos infetados em cada grupo é modelado por $I^{[n]} \frown \text{Binomial}\,(n,p)$, sendo um grupo considerado infetado se contiver pelo menos um indivíduo infetado ($I^{[n]} \geq 1$).

## 2.1 Ausência de erros de classificação

Denotando por $X_n^+$ o número de testes positivos quando $m$ grupos de dimensão $n$ são analisados, tem-se que $X_n^+ \frown \text{Binomial}\,(m,\pi_n)$ onde $\pi_n = \mathbb{P}\left(I^{[n]} \geq 1\right) = 1-(1-p)^n$. Aplicando a propriedade da invariância dos estimadores de máxima verosimilhança obtém-se o estimador para a taxa de prevalência

$$\widehat{p} = 1 - \left(1 - \frac{X_n^+}{m}\right)^{\frac{1}{n}}. \tag{1}$$

O número de testes positivos em $m$ testes individuais ($n = 1$) é caraterizado pela v.a. $X_1^+ \frown \text{Binomial}\,(m,p)$, logo $\widehat{p}$ é um estimador centrado para $p$ com variância igual ao limite inferior de Cramér-Rao tendo, por conseguinte, variância uniformemente mínima cf. [12].

Para comparar a performance de $\widehat{p}$ em diferentes taxas de prevalência (para valor de $p$ baixos, que correspondem aos casos para os quais a aplicação dos testes compostos é aconselhada, cf. [3]) e dimensões do grupo ($n$) fixamos o número de testes realizados ($m$) e analisamos o comportamento para diversos valores de $m$ entre 30 e 10000. Tendo em consideração o comportamento análogo observado nos diferentes valores de $m$, bem como a limitação de espaço, iremos unicamente apresentar os resultados para $m = 100$. Refira-se que, como seria expectável, quanto maior for o valor de $m$ menor serão os valores do viés bem como do erro quadrático médio (EQM) do

Tabela 1: Valores do viés ($m = 100$)

| $p$ | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|
| $n = 1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n = 2$ | .0004 | .0003 | .0001 | $6e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $3e^{-6}$ |
| $n = 3$ | .0006 | .0004 | .0002 | $8e^{-5}$ | $3e^{-5}$ | $2e^{-5}$ | $3e^{-6}$ |
| $n = 4$ | .0007 | .0004 | .0002 | .0001 | $4e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 5$ | .0009 | .0005 | .0002 | .0001 | $4e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 6$ | .0010 | .0006 | .0002 | .0001 | $4e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 7$ | .0011 | .0006 | .0003 | .0001 | $4e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 8$ | .0013 | .0007 | .0003 | .0001 | $5e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 9$ | .0014 | .0007 | .0003 | .0001 | $5e^{-5}$ | $2e^{-5}$ | $4e^{-6}$ |
| $n = 10$ | .0016 | .0008 | .0003 | .0001 | $5e^{-5}$ | $2e^{-5}$ | $5e^{-6}$ |
| $n = 15$ | .0031 | .0011 | .0003 | .0001 | $5e^{-5}$ | $2e^{-5}$ | $5e^{-6}$ |
| $n = 20$ | .0206 | .0017 | .0004 | .0002 | $5e^{-5}$ | $3e^{-5}$ | $5e^{-6}$ |
| $n = 25$ | .1483 | .0030 | .0005 | .0002 | $5e^{-5}$ | $3e^{-5}$ | $5e^{-6}$ |
| $n = 50$ | .8234 | .5307 | .0016 | .0003 | $6e^{-5}$ | $3e^{-5}$ | $5e^{-6}$ |
| $n = 100$ | .8500 | .8975 | .5215 | .0009 | $9e^{-5}$ | $3e^{-5}$ | $5e^{-6}$ |
| $n = 1000$ | .8500 | .9000 | .9500 | .9750 | .9857 | .5101 | $9e^{-6}$ |

estimador.

Na Tabela 1 são apresentados os valores[4] do viés de $\widehat{p}$ referentes à realização de 100 testes. No caso de amostras individuais ($n = 1$) o estimador tem obviamente viés nulo, mas para $n > 1$ o estimador é positivamente enviesado pelo que o valor médio do estimador será superior ao verdadeiro valor do parâmetro. O enviesamento do estimador é consequência do desconhecimento do número de indivíduos infetados num grupo infetado. Para cada taxa de prevalência, o aumento da dimensão do grupo origina o aumento da probabilidade do grupo estar infetado (bem como do viés), pelo que a partir de determinada dimensão $n$ o viés aumenta consideravelmente (pois a maioria dos grupos estarão infetados e $\widehat{p} \to 1$ quando $n \to +\infty$). Refira-se ainda que, regra geral, o viés também aumenta com a taxa de prevalência uma vez que $\pi_n$ é uma função crescente com $p$.

A utilização de testes compostos permite, para prevalências reduzidas, obter uma redução significativa no EQM, cf. Tabela 2. Note-se

[4] Nos valores apresentados nas tabelas $e^{-k}$ representa $\times 10^{-k}$.

Tabela 2: Valores do EQM ($m = 100$)

| $p$ | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|
| $n = 1$ | .0013 | .0009 | .0005 | .0002 | .0001 | $5e^{-5}$ | $e^{-5}$ |
| $n = 2$ | .0007 | .0005 | .0002 | .0001 | $5e^{-5}$ | $3e^{-5}$ | $5e^{-6}$ |
| $n = 3$ | .0005 | .0003 | .0002 | $8e^{-5}$ | $3e^{-5}$ | $2e^{-5}$ | $3e^{-6}$ |
| $n = 4$ | .0004 | .0003 | .0002 | $6e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $3e^{-6}$ |
| $n = 5$ | .0004 | .0002 | .0001 | $5e^{-5}$ | $2e^{-5}$ | $e^{-5}$ | $2e^{-6}$ |
| $n = 6$ | **.0003** | .0002 | $9e^{-5}$ | $4e^{-5}$ | $2e^{-5}$ | $8e^{-6}$ | $2e^{-6}$ |
| $n = 7$ | .0003 | .0002 | $8e^{-5}$ | $4e^{-5}$ | $e^{-5}$ | $7e^{-6}$ | $e^{-6}$ |
| $n = 8$ | .0003 | .0002 | $7e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $6e^{-6}$ | $e^{-6}$ |
| $n = 9$ | .0003 | .0002 | $7e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $6e^{-6}$ | $e^{-6}$ |
| $n = 10$ | .0003 | .0002 | $6e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $5e^{-6}$ | $e^{-6}$ |
| $n = 15$ | .0005 | **.0001** | $5e^{-5}$ | $2e^{-5}$ | $7e^{-6}$ | $3e^{-6}$ | $7e^{-7}$ |
| $n = 20$ | .0144 | .0002 | $\mathbf{4e}^{-5}$ | $2e^{-5}$ | $5e^{-6}$ | $3e^{-6}$ | $5e^{-7}$ |
| $n = 25$ | .1277 | .0007 | $4e^{-5}$ | $e^{-5}$ | $5e^{-6}$ | $2e^{-6}$ | $4e^{-7}$ |
| $n = 50$ | .7016 | .4833 | .0004 | $\mathbf{e}^{-5}$ | $3e^{-6}$ | $e^{-6}$ | $2e^{-7}$ |
| $n = 100$ | .7225 | .8079 | .4984 | .0003 | $\mathbf{2e}^{-6}$ | $\mathbf{7e}^{-7}$ | $e^{-7}$ |
| $n = 1000$ | .7225 | .8100 | .9025 | .9506 | .9759 | .5078 | $\mathbf{2e}^{-8}$ |

que comparativamente aos testes individuais, os testes compostos permitem a obtenção de um EQM inferior em todas as taxas de prevalência analisadas ([14] refere que tal ocorre para $p \in (0, 0.58)$, mas a obtenção de um viés reduzido só é garantido para $p \approx 0$, cf. Tabelas 1 e 3, pelo que restringimos a análise a $p \in (0{,}0.15]$). Os valores a negrito identificam, para cada taxa de prevalência, os grupos de dimensão ótima (entre as dimensões que figuram na tabela), no sentido que minimizam o EQM do estimador (apesar de, em termos de viés, o ótimo é utilizar testes individuais, i.e., $n = 1$).

## 2.2 Presença de erros de classificação

Ao relaxar a hipótese de ausência de erros de classificação, surge a necessidade de estender os conceitos de sensibilidade, $\varphi_s$, e especificidade, $\varphi_e$, de um teste individual para a realização de testes compostos. Denote-se por $X^{[+,n]}$ [respetivamente $X^{[-,n]}$] a ocorrência de um resultado positivo [respetivamente negativo] num teste

composto realizado a um grupo de dimensão $n$. A sensibilidade de um teste composto é definida como a probabilidade de obter um teste positivo numa amostra infetada, i.e. $\varphi_s^{[n]} = \mathbb{P}\left(X^{[+,n]}|I^{[n]} \geq 1\right)$. Por outro lado, a especificidade de um teste composto é definida como a probabilidade de obter um teste negativo numa amostra limpa, i.e. $\varphi_e^{[n]} = \mathbb{P}\left(X^{[-,n]}|I^{[n]} = 0\right)$. Nas simulações realizadas considerou-se que estas probabilidades não dependem de $n$ (ausência de efeito de diluição, cf. [7]) pelo que $\varphi_s^{[n]} = \varphi_s$ e $\varphi_e^{[n]} = \varphi_e$, cf. [11, 12]. Uma vez mais se tem $X_n^+ \frown \text{Binomial}\,(m, \pi_n)$, na qual o parâmetro $\pi_n$, com $\pi_n = \varphi_s + (1-p)^n (1 - \varphi_e - \varphi_s)$, representa a probabilidade do teste composto ser positivo. No caso de $\varphi_e = \varphi_s = 1$ tem-se $\pi_n = 1 - (1-p)^n$, que representa a probabilidade da amostra combinada conter pelo menos um elemento contaminado na situação clássica de ausência de erros de classificação (analisada na subsecção anterior). No caso especial de $n = 1$ obtém-se $\pi_1 = p\varphi_s + (1-p)(1-\varphi_e)$ que representa a probabilidade de um teste individual ter resultado positivo na presença de erros de classificação. Facilmente se verifica que $\pi_n$ é uma função monótona crescente com $p$ caso $(1 - \varphi_e - \varphi_s) < 0$ (i.e., se $\varphi_e + \varphi_s > 1$, único caso que iremos considerar doravante uma vez que o contrário corresponderia a testes com uma probabilidade de má classificação extremamente elevada). Assim, quanto maior for a taxa de prevalência maior será a probabilidade do teste à amostra combinada dar positivo. Atendendo ao facto de $p$, $\varphi_s$ e $\varphi_e$ tomarem valores no intervalo [0,1], tem-se que $1 - \varphi_e \leq \pi_n \leq \varphi_s$, em que a igualdade no limite inferior do intervalo ocorre quando $p = 0$ e no limite superior quando $p = 1$. O estimador de máxima verosimilhança de $\pi_n$ é dado por $\widehat{\pi_n} = \varphi_s\left[1 - (1-\widehat{p})^n\right] + (1-\widehat{p})^n (1 - \varphi_e)$. Dado que $\widehat{\pi_n} = \frac{1}{m} X_n^+$ tem-se

$$\widehat{p} = 1 - \left(\frac{\varphi_s - X_n^+/m}{\varphi_s + \varphi_e - 1}\right)^{1/n}. \tag{2}$$

Para que $\widehat{p} \in [0,1]$ é contudo necessário que $1 - \varphi_e \leq \frac{1}{m} X_n^+ \leq \varphi_s$. Com o intuito de contornar as dificuldades computacionais impostas por esta condição recorreu-se ao estimador (cf. [7])

Tabela 3: Valores do viés ($m = 100$ *e* $\varphi_s = \varphi_e = 0.95$)

| $p$ | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|
| $n = 1$ | -.0003 | .0002 | .0011 | .0023 | .0052 | .0076 | .0092 |
| $n = 2$ | .0004 | .0002 | .0005 | .0004 | .0017 | .0030 | .0046 |
| $n = 3$ | .0006 | .0005 | .0003 | .0005 | .0006 | .0015 | .0029 |
| $n = 4$ | .0010 | .0007 | .0004 | $9e^{-5}$ | .0003 | .0009 | .0020 |
| $n = 5$ | .0011 | .0009 | .0003 | .0003 | .0001 | .0007 | .0016 |
| $n = 6$ | .0009 | .0007 | .0003 | .0001 | $9e^{-5}$ | .0004 | .0012 |
| $n = 7$ | .0015 | .0010 | .0003 | .0001 | .0002 | .0002 | .0010 |
| $n = 8$ | .0017 | .0008 | .0004 | .0002 | $4e^{-5}$ | .0002 | .0009 |
| $n = 9$ | .0018 | .0010 | .0003 | .0002 | .0002 | .0002 | .0007 |
| $n = 10$ | .0023 | .0010 | .0003 | .0001 | .0002 | .0002 | .0006 |
| $n = 15$ | .0127 | .0013 | .0004 | .0002 | $6e^{-5}$ | $2e^{-5}$ | .0003 |
| $n = 20$ | .1173 | .0033 | .0005 | .0002 | $7e^{-5}$ | $2e^{-5}$ | .0002 |
| $n = 25$ | .2943 | .0229 | .0006 | .0002 | $4e^{-5}$ | $3e^{-5}$ | .0001 |
| $n = 50$ | .4970 | .4632 | .0160 | .0003 | $10e^{-5}$ | $5e^{-5}$ | $3e^{-5}$ |
| $n = 100$ | .4811 | .5345 | .4924 | .0141 | $8e^{-5}$ | $6e^{-6}$ | $9e^{-6}$ |
| $n = 1000$ | .4664 | .5206 | .5675 | .5999 | .6036 | .5047 | $e^{-7}$ |

$$\widehat{p} = 1 - \left[ \frac{\varphi_s - \min\left\{\varphi_s, \max\left(1-\varphi_e, \frac{X_n}{m}\right)\right\}}{\varphi_s + \varphi_e - 1} \right]^{1/n} \tag{3}$$

com variância dada por

$$\mathrm{Var}\,(\widehat{p}) = \frac{\{\varphi_s - r(1-p)^n\}\{r(1-p)^n + 1 - \varphi_s\}}{m\, r^2\, n^2\, (1-p)^{2(n-1)}}, \quad r = \varphi_s + \varphi_e - 1.$$

Fixando a dimensão dos grupos e as taxas de erro de classificação $\varphi_s$ e $\varphi_e$ a variância do estimador decresce com o número de grupos ($m$). Assim, a precisão de $\widehat{p}$ poderá ser melhorada aumentando o número de grupos, o que poderá não ser viável devido às restrições de custos. A presença de erros de classificação provoca alterações no viés e no EQM de $\widehat{p}$. Para amostras de dimensão um (Tabela 3), ao contrário do que sucede no modelo binomial, $\widehat{p}$ não é um estimador centrado de $p$. Relativamente ao EQM (Tabela 4), é visível que mesmo na presença de erros de classificação é possível obter uma redução significativa do EQM recorrendo à utilização de testes compostos sendo, para tal, necessário construir grupos de dimensão inferior à dimensão ótima utilizada no modelo binomial. Os valores a negrito (Tabela 4) identificam a dimensão ótima em termos de EQM para cada valor de $p$. Apesar de não apresentarmos resultados

Tabela 4: Valores do EQM ($m = 100$ e $\varphi_s = \varphi_e = 0.95$)

| $p$ | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|
| $n = 1$ | .0019 | .0015 | .0010 | .0006 | .0004 | .0004 | .0003 |
| $n = 2$ | .0009 | .0007 | .0004 | .0003 | .0001 | .0001 | $9e^{-5}$ |
| $n = 3$ | .0006 | .0004 | .0002 | .0002 | $8e^{-5}$ | $6e^{-5}$ | $4e^{-5}$ |
| $n = 4$ | .0005 | .0003 | .0002 | .0001 | $6e^{-5}$ | $4e^{-5}$ | $2e^{-5}$ |
| $n = 5$ | .0005 | .0003 | 0001 | $8e^{-5}$ | $4e^{-5}$ | $3e^{-5}$ | $2e^{-5}$ |
| $n = 6$ | **.0004** | **.0002** | .0001 | $6e^{-5}$ | $3e^{-5}$ | $2e^{-5}$ | $e^{-5}$ |
| $n = 7$ | .0004 | .0002 | .0001 | $5e^{-5}$ | $3e^{-5}$ | $2e^{-5}$ | $8e^{-6}$ |
| $n = 8$ | .0004 | .0002 | $9e^{-5}$ | $5e^{-5}$ | $2e^{-5}$ | $e^{-5}$ | $7e^{-6}$ |
| $n = 9$ | .0004 | .0002 | $9e^{-5}$ | $4e^{-5}$ | $2e^{-5}$ | $e^{-5}$ | $5e^{-6}$ |
| $n = 10$ | .0005 | .0002 | $8e^{-5}$ | $4e^{-5}$ | $2e^{-5}$ | $e^{-5}$ | $5e^{-6}$ |
| $n = 15$ | .0073 | .0002 | $6e^{-5}$ | $3e^{-5}$ | $e^{-5}$ | $6e^{-6}$ | $2e^{-6}$ |
| $n = 20$ | .1006 | .0010 | $\mathbf{5e}^{-5}$ | $2e^{-5}$ | $8e^{-6}$ | $4e^{-6}$ | $e^{-6}$ |
| $n = 25$ | .2590 | .0177 | $5e^{-5}$ | $2e^{-5}$ | $6e^{-6}$ | $3e^{-6}$ | $e^{-6}$ |
| $n = 50$ | .4492 | .4288 | .0133 | $\mathbf{e}^{-5}$ | $3e^{-6}$ | $e^{-6}$ | $4e^{-7}$ |
| $n = 100$ | .4503 | .5038 | .4737 | .0128 | $\mathbf{2e}^{-6}$ | $\mathbf{e}^{-8}$ | $2e^{-7}$ |
| $n = 1000$ | .4525 | .5046 | .5568 | .5928 | .5999 | .5027 | $\mathbf{2e}^{-8}$ |

para outros valores de $m$, o aumento do número de testes implica uma redução significativa do viés. Note-se que a probabilidade de todos os testes serem positivos, dada por $[1-(1-p)^n]^m$, é uma função decrescente com $m$, o que provoca uma redução do viés de $\widehat{p}$ e consequentemente uma redução do EQM do estimador, cf. [12].

# 3 Estimação intervalar

Para avaliar as vantagens decorrentes da utilização de amostras conjuntas na construção de intervalos de confiança para a taxa de prevalência $p$, recorreu-se a simulação de Monte Carlo e ao *package binGroup* para o R, cf.[1], que engloba um conjunto de estimadores intervalares. Foram efetuadas $10^5$ réplicas e o nível de confiança utilizado foi de 0.95. Para cada valor de $p$ foram registados, para grupos de dimensão ótima (cf. Tabelas 2 e 4) e para amostras individuais, os valores dos seguintes parâmetros: ML (comprimento médio do intervalo em percentagem), $\widehat{\alpha_L}$ (taxa de erro interior, i.e., a percentagem de observações inferiores ao limite inferior do intervalo) e $\widehat{\alpha_U}$ (taxa de erro superior que denota a percentagem de observações superiores ao limite superior do intervalo). Os estimadores intervalares utiliza-

Tabela 5: Estimação intervalar ($m = 100$)

| | $p$ | .10 | | .05 | | .01 | | $10^{-3}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | $n$ | 1 | 10 | 1 | 20 | 1 | 100 | 1 | 100 |
| CP | ML | 12.57 | 5.11 | 9.43 | 2.64 | 5.23 | 0.54 | 3.80 | 0.14 |
| | $\widehat{\alpha_L}$ | 2.05 | 2.21 | 1.12 | 2.31 | 1.91 | 1.67 | 0.50 | 1.36 |
| | $\widehat{\alpha_U}$ | 2.41 | 2.29 | 0.56 | 2.37 | 0 | 2.11 | 0 | 1.12 |
| B | ML | 12.05 | 4.98 | 8.89 | 2.58 | 4.95 | 0.53 | 3.71 | 0.13 |
| | $\widehat{\alpha_L}$ | 2.05 | 2.21 | 2.76 | 2.31 | 1.91 | 2.85 | 0.50 | 2.78 |
| | $\widehat{\alpha_U}$ | 2.41 | 2.29 | 0.56 | 2.37 | 0 | 2.12 | 0 | 1.12 |
| AC | ML | 12.14 | 4.81 | 9.45 | 2.49 | 5.86 | 0.51 | 4.60 | 0.13 |
| | $\widehat{\alpha_L}$ | 2.05 | 2.21 | 2.76 | 2.31 | 1.91 | 2.85 | 0.50 | 2.78 |
| | $\widehat{\alpha_U}$ | 0.78 | 2.29 | 0.56 | 2.37 | 0 | 2.12 | 0 | 1.12 |
| S | ML | 11.77 | 4.80 | 8.85 | 2.48 | 5.10 | 0.51 | 3.86 | 0.13 |
| | $\widehat{\alpha_L}$ | 3.94 | 2.21 | 2.76 | 2.31 | 7.99 | 2.85 | 9.54 | 2.78 |
| | $\widehat{\alpha_U}$ | 2.41 | 2.29 | 0.56 | 2.37 | 0 | 2.12 | 0 | 1.12 |
| SOC | ML | 11.65 | 4.87 | 8.54 | 2.52 | 4.20 | 0.52 | 2.52 | 0.13 |
| | $\widehat{\alpha_L}$ | 2.05 | 2.21 | 1.12 | 2.31 | 0.33 | 2.85 | 0.02 | 2.78 |
| | $\widehat{\alpha_U}$ | 2.41 | 2.29 | 3.66 | 2.37 | 0 | 2.11 | 0 | 3.33 |
| W | ML | 11.59 | 4.85 | 8.29 | 2.51 | 3.00 | 0.52 | 0.38 | 0.13 |
| | $\widehat{\alpha_L}$ | 0.98 | 1.28 | 0.42 | 1.33 | 0.04 | 1.67 | 0.001 | 0.65 |
| | $\widehat{\alpha_U}$ | 5.74 | 3.68 | 11.68 | 3.75 | 36.79 | 3.40 | 90.4 | 7.72 |

dos foram: CP (Clopper-Pearson), B (Blaker), AC (Agresti-Coull), S (*score* de Wilson), SOC (*Second order corrected*) e W (Wald), cf. [10]. Os resultados obtidos, em percentagem, são apresentados nas Tabelas 5 (ausência de erros de classificação) e 6 (na presença de erros de classificação).

## 3.1 Ausência de erros de classificação

Ao analisar os resultados que figuram na Tabela 5 verificamos que, para todo o valor de $p$, o comprimento médio do intervalo de confiança (ML) para grupos de dimensão ótima é sempre inferior ao valor obtido para amostras individuais. Esta conclusão já era expectável uma vez que, para o mesmo número de testes, o número de indivíduos analisados na construção do intervalo recorrendo a amostras conjuntas é superior ao número de indivíduos analisados recorrendo amostras individuais e, por conseguinte, quanto mais informação existir menor será a amplitude do intervalo e consequentemente maior será a sua precisão. Refira-se ainda que a referida redução do comprimento médio do intervalo é mais acentuada para taxas de prevalência mais pequenas, uma vez que a dimensão ótima do grupo aumenta com a redução da taxa de prevalência.

Tabela 6: Estimação intervalar ($m = 100$ e $\varphi_s = \varphi_e = 0.95$)

| | $p$ | .10 | | .05 | | .01 | | $10^{-3}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | $n$ | 1 | 10 | 1 | 20 | 1 | 100 | 1 | 100 |
| CP | ML | 15.81 | 5.81 | 12.50 | 2.99 | 8.19 | 0.61 | 6.93 | 0.18 |
| | $\widehat{\alpha_L}$ | 1.78 | 1.94 | 1.24 | 2.1 | 1.4 | 2.16 | 1.22 | 1.42 |
| | $\widehat{\alpha_U}$ | 2.11 | 1.86 | 1.16 | 1.9 | 1.67 | 1.63 | 0.5 | 1.37 |
| B | ML | 15.30 | 5.66 | 12.10 | 2.92 | 7.87 | 0.60 | 6.61 | 0.17 |
| | $\widehat{\alpha_L}$ | 1.78 | 1.94 | 2.69 | 2.1 | 1.4 | 2.16 | 2.92 | 2.53 |
| | $\widehat{\alpha_U}$ | 2.11 | 2.84 | 1.16 | 2.91 | 1.67 | 2.7 | 0.5 | 1.37 |
| AC | ML | 15.17 | 5.46 | 12.28 | 2.81 | 8.33 | 0.58 | 7.13 | 0.17 |
| | $\widehat{\alpha_L}$ | 3.23 | 1.94 | 2.69 | 2.1 | 3.46 | 2.16 | 2.92 | 2.53 |
| | $\widehat{\alpha_U}$ | 2.11 | 2.84 | 1.16 | 2.91 | 0.22 | 2.7 | 0.5 | 1.37 |
| S | ML | 14.90 | 5.45 | 11.98 | 2.81 | 8.01 | 0.58 | 6.79 | 0.16 |
| | $\widehat{\alpha_L}$ | 3.23 | 1.94 | 2.69 | 2.1 | 3.46 | 2.16 | 2.92 | 2.53 |
| | $\widehat{\alpha_U}$ | 2.11 | 2.84 | 1.16 | 2.91 | 1.67 | 2.7 | 0.5 | 1.37 |
| SOC | ML | 14.79 | 5.53 | 11.48 | 2.85 | 7.14 | 0.58 | 5.87 | 0.16 |
| | $\widehat{\alpha_L}$ | 1.78 | 3.15 | 2.69 | 3.2 | 1.4 | 2.16 | 1.22 | 2.53 |
| | $\widehat{\alpha_U}$ | 2.11 | 2.84 | 3.69 | 2.91 | 1.67 | 2.7 | 3.5 | 3.07 |
| W | ML | 14.65 | 5.50 | 10.79 | 2.83 | 6.07 | 0.58 | 4.78 | 0.16 |
| | $\widehat{\alpha_L}$ | 0.91 | 1.94 | 1.24 | 2.1 | 0.51 | 1.24 | 0.45 | 0.74 |
| | $\widehat{\alpha_U}$ | 4.45 | 2.84 | 8.01 | 2.91 | 5.87 | 4.27 | 11.3 | 6.47 |

## 3.2 Presença de erros de classificação

Na presença de erros de classificação é necessário recalcular os limites de confiança apresentados anteriormente. Denotemos por $[L_1,U_1]$ os limites de confiança dum intervalo na ausência de erros de classificação. Recorrendo à equação (1) determinamos o valor de $\frac{1}{m}X_n^+$ que daria origem à estimativa $L_1$ (isto é, considerando que $\widehat{p} = L_1$) e consequentemente $\frac{1}{m}X_n^+ = 1-(1-L_1)^n$. Aplicando raciocínio análogo em $U_1$ podemos determinar os limites $[L_2,U_2]$ do intervalo de confiança na presença de erros de classificação recorrendo à equação (3), obtendo-se

$$L_2 = 1 - \left[\frac{\varphi_s - \min\{\varphi_s, \max(1-\varphi_e, 1-(1-L_1)^n)\}}{\varphi_s+\varphi_e-1}\right]^{1/n},$$

$$U_2 = 1 - \left[\frac{\varphi_s - \min\{\varphi_s, \max(1-\varphi_e, 1-(1-U_1)^n)\}}{\varphi_s+\varphi_e-1}\right]^{1/n}.$$

Para um número fixo de testes ($m = 100$), a existência de erros de classificação provoca alterações significativas nos valores dos parâmetros analisados (cf. Tabela 6). É notório um aumento do comprimento médio dos intervalos se compararmos com os resultados apresentados na Tabela 5. Apesar de não serem apresentados resultados para valores de $m$ distintos, o aumento do número de testes traduz-se numa redução das amplitudes dos intervalos e na estabili-

zação da taxa de erro, $\widehat{\alpha_L} + \widehat{\alpha_U}$, em torno do nível de significância utilizado. Note-se que na ausência de erros de classificação, o nível de confiança dos estimadores intervalares utilizados é de 0.95. Apesar de não podermos garantir que o nível de confiança dos novos intervalos se mantenha, podemos no entanto referir que os resultados das simulações apontam para uma estabilização do nível de confiança em torno de 0.95, isto é, que as transformações aplicadas permitem obter intervalos de confiança na presença de erros de classificação.

# 4 Conclusões

Este estudo analisou a performance de estimadores pontuais e intervalares para a taxa de prevalência, quer os baseados em testes individuais quer os apoiados em testes compostos, na presença e na ausência de erros de classificação. Nesta análise comparativa apenas foi considerado o custo associado à realização de cada teste e, por conseguinte, o custo relativo à junção das amostras não foi contemplado (este custo é, em muitas aplicações, considerado negligenciável em comparação com o custo de realização dos testes, cf. [8]). Sob estas condições, os resultados apresentados mostram que, fixando o número de testes, os estimadores baseados em testes compostos têm, para prevalências reduzidas, menor erro quadrático médio que os estimadores baseados em testes simples. Há, contudo, ainda alguns obstáculos na sua aplicação, uma vez que para a determinação da dimensão ótima $n$ de cada grupo é necessário dispor previamente de um valor aproximado da taxa de prevalência e, caso existam erros de classificação, é igualmente necessário conhecer o valor da sensibilidade e da especificidade (ou, pelo menos, boas estimativas desses valores). A análise da robustez dos estimadores para a taxa de prevalência relativamente à utilização de estimativas pouco precisas da sensibilidade e da especificidades serão realizados brevemente.

# Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst-OE/MAT/UI0006/2014.

# Referências

[1] Bilder, C.R., Zhang, B., Schaarsehmidt, F., Tebbs, J. (2010). Bingroup: a package for group testing, *The R Journal* 2, 56-60.

[2] Chen, C.L., Swallow, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* 46, 1035–1046.

[3] Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* 14, 436–440.

[4] Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R.(1989). Probability estimation with sample compositing techniques. *Journal of Official Statistics* 5, 365–374.

[5] Hung, M., Swallow, W.H. (1999). Robustness of group testing in the estimation of proportions. *Biometrics* 55, 231–237.

[6] Lancaster, V.A., Keller-McNulty, S. (1998). A review of composite sampling methods. *Journal of the American Statistical Association* 93, 1216–1230.

[7] Liu, A., Liu, C., Zhang, A., Albert, P.S. (2012). Optimality of group testing in the presence of misclassification, *Biometrika* 99, 245–251.

[8] Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., and Yang, T.C. (2011). Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors, *Annals of Applied Biology* 159, 491–502.

[9] Loyer, M.W. (1983). Bad probability, good statistics, and group testing for Binomial estimation. *American Statistician* 37, 57–59.

[10] Pires, A.M., Amado, C. (2008). Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT* 6, 165–197.

[11] Santos, R., Pestana, D., Martins, J.P. (2013). Extensions of Dorfman's Theory. In P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics, Studies in Theoretical and Applied Statistics*, 179–189.

[12] Sousa, R. (2012). *Testes Conjuntos. Extensões da Teoria de Dorfman.* Tese de doutoramento, Universidade de Lisboa.

[13] Sobel, K.M., Elashoff, R.M. (1975). Group testing with a new goal, estimation. *Biometrika* 62, 181–193.

[14] Swallow, W.H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *American Phytopathological Society* 75, 882–889.

# Medidas para avaliar a utilização de testes compostos

Rui Santos
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, *rui.santos@ipleiria.pt*

João Paulo Martins
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, *jpmartins@ipleiria.pt*

Miguel Felgueiras
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, *mfelg@ipleiria.pt*

**Palavras–chave**: Testes compostos, classificação, média, extremos, sensibilidade, especificidade.

**Resumo**: A aplicação de testes compostos (ou conjuntos) em análises clínicas, ou em amostragem de aceitação, permite poupar muitos recursos. Todavia, o uso deste tipo de testes deve ser efetuado com precaução, de forma a evitar a existência de uma elevada probabilidade de má classificação. Neste trabalho, através de um estudo comparativo via simulação, concluímos que o peso das caudas da distribuição, que caracteriza a quantidade da substância em análise, é relevante para a avaliação da adequação da aplicação de algoritmos de classificação hierárquicos e não hierárquicos, sendo garantida uma sensibilidade elevada para distribuições com cauda pesada.

## 1 Introdução e metodologias

Seja $p$ a taxa de prevalência da infeção em análise numa população com $N$ indivíduos. As variáveis aleatórias (v.a.) $X_i$, com

$i = 1, \ldots, N$, denotam a presença ($X_i = 1$) ou ausência ($X_i = 0$) da infeção no $i$-ésimo indivíduo da população, i.e. $X_i \sim \text{Ber}(p), i = 1, \ldots, N$. Para a classificação do $i$-ésimo indivíduo é analisado, por exemplo, um mililitro (ml) de sangue, no qual a quantidade $Y_i$ da substância que permite a identificação da infeção é caracterizada por uma distribuição conhecida $\mathbf{D}_\theta$, i.e. $Y_i \sim \mathbf{D}(\theta)$, $i = 1, \ldots, N$. Cada indivíduo está infetado se a quantidade $Y_i$ for superior a um determinado limiar predefinido $t$, i.e. $Y_i > t \Rightarrow X_i = 1$ (a desigualdade oposta pode ser aplicada de forma análoga). Caso contrário, o indivíduo é considerado saudável ou não infetado ($Y_i \leq t \Rightarrow X_i = 0$). Neste trabalho é considerado que não há erros de classificação nos testes individuais.

Para a realização de testes compostos é necessário subdividir aleatoriamente a população em grupos de $n$ indivíduos. Por conseguinte, o número de indivíduos infetados em cada grupo $I^{[n]} = \sum_{i=1}^{n} X_i$ tem distribuição binomial, $I^{[n]} \sim \mathbf{B}(n,p)$, e as v.a. $Y_1, \ldots, Y_n$ são independentes e identicamente distribuídas. Após a junção de um ml de sangue de cada elemento do grupo, os $n$ ml são bem misturados de forma a tornar o fluido homogéneo, no qual a quantidade de substância é descrita pela v.a. $B_n = \sum_{i=1}^{n} Y_i$. Posteriormente é retirado aleatoriamente um ml desta mistura para a realização do teste composto, no qual a quantidade de substância é descrita pela v.a. $B_1$. Se a distribuição $\mathbf{D}_\theta$ for discreta (e.g. número de bactérias) então a distribuição da v.a. $B_1$ pode ser determinada recorrendo a modelos hierárquicos (utilizando um filtro binomial, com probabilidade $n^{-1}$, em cada uma das $B_n$ bactérias presentes nos $n$ mililitros de sangue), portanto $B_1 \sim \mathbf{B}\left(B_n, \frac{1}{n}\right)$, cf. [11]. No caso de a distribuição $\mathbf{D}_\theta$ ser absolutamente contínua não é possível utilizar esta metodologia. Contudo, considerando a mistura homogénea (o que atualmente é garantido pelos meios informáticos e mecânicos disponíveis), então a média do grupo será uma boa aproximação de $B_1$, i.e. $B_1 \approx \overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

No teste (ao sangue) composto, o resultado negativo ($X^-$) significa que não há qualquer indivíduo infetado no grupo ($\sum_{i=1}^{n} X_i = 0$), como tal $Y_i \leq t$ para $i = 1, \ldots, n$ e o máximo $M_n = \max(Y_1, \ldots, Y_n)$

verifica $M_n \leq t$. Caso contrário, se o teste composto for positivo ($X^+$), então pelo menos um elemento do grupo está infetado ($\sum_{i=1}^{n} X_i \geq 1$) e, deste modo, o máximo do grupo irá ultrapassar o limiar $t$ ($M_n > t$). O objetivo destes testes é identificar se o máximo do grupo é superior ao limiar $t$ utilizando a média do grupo como única informação, o que permitirá eventualmente classificar $n$ indivíduos com um único teste e, por conseguinte, diminuir o número de testes necessários para a classificação de uma população, conforme detalharemos na Secção 2.

A realização destes testes pode originar erros de classificação, usualmente avaliados através da sensibilidade e da especificidade. Deste modo, seja $\varphi_s \in (0,1]$ a sensibilidade do teste composto – probabilidade de se obter um teste positivo num grupo infetado, i.e. $\mathrm{P}\left(X^+ | \sum_{i=1}^{n} X_i \geq 1\right)$; e seja $\varphi_e \in (0,1]$ a especificidade do teste composto – probabilidade de se obter um teste negativo num grupo saudável, i.e. $\mathrm{P}\left(X^- | \sum_{i=1}^{n} X_i = 0\right)$. Deste modo, $1-\varphi_s$ é a probabilidade de um falso negativo e $1-\varphi_e$ a de um falso positivo (nos testes compostos).

Na metodologia tradicional de realização de um teste composto são consideradas as hipóteses (metodologia $\mathbf{T_1}$):

$$\mathbf{H}_0 : \sum_{i=1}^{n} X_i = 0 \ \ (M_n \leq t) \ \ \textit{versus} \ \mathbf{H}_1 : \sum_{i=1}^{n} X_i > 0 \ \ (M_n > t)\,, \quad (\mathbf{T_1})$$

para as quais o nível de significância $\alpha$ corresponde à probabilidade de um falso positivo e, como tal, nesta aplicação estamos a controlar a especificidade do teste pelo facto de $\varphi_e = 1-\alpha$ (sendo a sensibilidade relegada para segundo plano). Todavia, em muitas aplicações (nomeadamente se houver problemas de contágio) o controlo da sensibilidade é fundamental de forma a minimizar a probabilidade de falsos negativos. Por esta razão introduzimos uma nova metodologia, já proposta em [9], mas ainda sem ter sido avaliado o seu desempenho, que corresponde à permuta das hipóteses em $\mathbf{T_1}$, originando a metodologia $\mathbf{T_2}$ cujas hipóteses correspondem a

$$\mathbf{H}_0 : \sum_{i=1}^{n} X_i > 0 \ \ (M_n > t) \ \ \textit{versus} \ \mathbf{H}_1 : \sum_{i=1}^{n} X_i = 0 \ \ (M_n \leq t)\,, \quad (\mathbf{T_2})$$

para as quais $\varphi_{_s} = 1 - \alpha$ e, por conseguinte, a sensibilidade do teste é por nós fixada *a priori*. Além disso, uma vez que para taxas de prevalência baixas a existência de pelo menos um indivíduo infetado no grupo implica, quase certamente, um único infetado (cf. [11]), então para a determinação do limiar da região crítica podemos utilizar (na prática) as hipóteses:

$$\mathbf{H}_0 : \sum_{i=1}^{n} X_i = 1 \ \textit{versus} \ \mathbf{H}_1 : \sum_{i=1}^{n} X_i = 0.$$

Notemos que, por outro lado, estamos a considerar o pior cenário possível (pois a probabilidade de um falso negativo será menor se existir mais do que um indivíduo infetado no grupo, como consequência da rarefação/diluição da substância) e, por conseguinte, ao fixarmos a probabilidade deste erro em $\alpha$, será expectável que o nível de significância observado seja ligeiramente inferior a $\alpha$ (isto é, que a sensibilidade seja ligeiramente superior a $1 - \alpha$). Refira-se que as hipóteses não estão formalizadas relativamente aos parâmetros ou aos pressupostos de distribuição como as usuais hipóteses estatísticas, contudo elas referem-se a uma decisão sobre uma medida amostral desconhecida ($M_n$) e, portanto, a mesma linguagem técnica pode ser empregue. Para a definição da região crítica destes testes será fundamental conhecer a distribuição da v.a. $\overline{Y}_n$ que pode ser determinada analiticamente (para algumas distribuições), ou utilizando processos aproximados (nomeadamente via simulação Monte Carlo).

# 2 Metodologias de classificação

Neste trabalho foram considerados os três tipos de metodologias de classificação mais usuais: testes individuais, algoritmos hierárquicos e não hierárquicos. Para cada metodologia $\mathcal{M}$ foi determinada a sua eficiência através do quociente

$$\text{eff}\,(\mathcal{M}) = \text{E}\,(T_{_N})\,/N,$$

onde $\text{E}(\cdot)$ representa o valor esperado e $T_{_N}$ o número de testes necessários para classificar uma população de dimensão $N$ usando a me-

todologia $\mathcal{M}$. Deste modo, eff $(\mathcal{M})$ indica o número médio de testes necessários para classificar cada indivíduo quando é aplicada a metodologia $\mathcal{M}$ (na usual aplicação de testes individuais será necessário realizar um teste por indivíduo pelo que eff $(\mathcal{M}) = 1$). Pretende-se, com esta medida, analisar o custo associado à aplicação de cada metodologia ([4] apresenta uma visão geral da evolução da eficiência em diferentes metodologias), sendo considerado unicamente o custo associado à realização de cada teste, uma vez que o custo de misturar as amostras é geralmente negligível (cf. [8]).

Para avaliar a precisão dos resultados foram utilizados os conceitos de sensibilidade e especificidade de uma metodologia $\mathcal{M}$ (cf. [11]). Seja $X_i^+$ [respetivamente $X_i^-$] a representação do $i$-ésimo indivíduo ser classificado como infetado [respetivamente não infetado]. Por conseguinte, a sensibilidade da metodologia $\mathcal{M}$, $\varphi_s^{\mathcal{M}}$, é a probabilidade de um indivíduo infetado ser corretamente classificado pela metodologia $\mathcal{M}$; e a especificidade da metodologia $\mathcal{M}$, $\varphi_e^{\mathcal{M}}$, é a probabilidade de um indivíduo saudável ser corretamente classificado pela metodologia $\mathcal{M}$, isto é,

$$\varphi_s^{\mathcal{M}} = \mathrm{P}_{\mathcal{M}}\left(X_i^+ | X_i = 1\right) \qquad \text{e} \qquad \varphi_e^{\mathcal{M}} = \mathrm{P}_{\mathcal{M}}\left(X_i^- | X_i = 0\right).$$

Na aplicação de **testes individuais** realiza-se um único teste por indivíduo, obtendo-se eff $(\mathcal{M}) = \varphi_s^{\mathcal{M}} = \varphi_e^{\mathcal{M}} = 1$ uma vez que é observado o valor $Y_i$ de cada indivíduo. Dorfman [1] foi o primeiro a sugerir a utilização de testes compostos (a grupos de $n$ indivíduos) com o objetivo de poupar recursos através da redução do número esperado de testes. Se o resultado for negativo todos os indivíduos são classificados como saudáveis. Caso contrário, haverá pelo menos um indivíduo infetado e, consequentemente, será necessário realizar testes individuais a todos os elementos do grupo de forma a identificar os infetados. Posteriormente, surgiram algoritmos mais complexos, envolvendo a divisão sucessiva das amostras compostas, cujo resultado laboratorial fosse positivo, em subamostras de menor dimensão até que, em última análise, sejam realizados testes individuais (**algoritmos hierárquicos**). A investigação de novas metodologias desde o trabalho seminal de Dorfman tem sido particularmente ativa

tendo dado origem a algoritmos mais complexos, mas que partilham o princípio fundamental da metodologia de Dorfman que é iniciar a deteção de infetados com testes compostos e apenas realizar testes individuais nos indivíduos suspeitos. Assim, o processo de Dorfman que tem 2 etapas pode ser alargado a 3 ou mais etapas em que em cada uma delas é efetuado um teste composto, sendo que as amostras onde o resultado é positivo são novamente testadas em grupos de menor dimensão até que, chegando à última etapa, são realizados testes individuais. Informações mais detalhadas sobre algoritmos hierárquicos podem ser encontradas em [2, 6, 7] entre outros.

Os **algoritmos não hierárquicos**, baseados em tabelas (*arrays*), são a alternativa mais comum, cf. [5, 10, 13]. A sua versão mais simples corresponde a uma tabela quadrada, denotada por $\mathrm{A2}\,(n:1)$, na qual $n^2$ indivíduos são dispostos numa matriz de dimensão $n \times n$. Em seguida, são realizados $2n$ testes conjuntos: a todos os indivíduos situados na mesma linha e a todos os situados na mesma coluna. Sejam $\mathrm{P}_r$ e $\mathrm{P}_c$ o número de linhas e colunas positivas, respetivamente. Se $\max(\mathrm{P}_r,\mathrm{P}_c) = 0$ todos os $n^2$ indivíduos são classificados como saudáveis. Se $\min(\mathrm{P}_r,\mathrm{P}_c) \geq 1$ realizam-se testes individuais a todos os indivíduos situados nas interseções de linhas e colunas infetadas. Por fim, se $\min(\mathrm{P}_r,\mathrm{P}_c) = 0$ e $\max(\mathrm{P}_r,\mathrm{P}_c) \geq 1$ será necessário testar individualmente todos os indivíduos na(s) linha(s) (ou coluna(s)) positiva(s). Uma variante desta metodologia, denotada por $\mathrm{MA2}\left(n^2:n:1\right)$, inclui a realização prévia de um teste composto ao conjunto dos $n^2$ indivíduos, designado por teste global (*masterpool*). Se o resultado for negativo todos os $n^2$ indivíduos são classificados como não infetados. Caso contrário, é aplicado o procedimento $\mathrm{A2}\,(n:1)$.

# 3 Simulações: resultados e comentários

A performance de cada metodologia de classificação depende de diversos fatores, tais como a taxa de prevalência $p$, a dimensão do grupo $n$ e as propriedades da distribuição $\mathbf{D}_\theta$. Santos, Felgueiras e Martins [12] salientam o peso da cauda direita da distribuição $\mathbf{D}_\theta$

como fator capital para a determinação da performance de testes compostos. Para analisarmos a influência do peso da cauda direita da distribuição $\mathbf{D}_\theta$, foram determinados (Tabelas 1 e 2) os coeficientes $\tau$ (crescentes com o peso da cauda) definidos em [3],

$$\tau = \left( \frac{F^{-1}_{\mathbf{D}_\theta}(0.99) - F^{-1}_{\mathbf{D}_\theta}(0.5)}{F^{-1}_{\mathbf{D}_\theta}(0.75) - F^{-1}_{\mathbf{D}_\theta}(0.5)} \right) \left( \frac{\Phi^{-1}(0.99) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)} \right)^{-1},$$

onde $F^{-1}_{\mathbf{D}_\theta}$ representa a inversa generalizada da função de distribuição de $\mathbf{D}_\theta$ e $\Phi$ a função de distribuição da normal padrão.

Nas Tabelas 1 e 2 estão apresentados, em percentagem, alguns dos resultados para a eficiência e a sensibilidade (valores apresentados entre parêntesis) obtidos via simulação de Monte Carlo recorrendo ao *software* **R**. Sublinhemos que a especificidade é sempre igual a um (100%) pois um indivíduo só é classificado como infetado depois da realização de um teste individual, o qual, conforme referido previamente, consideramos ao longo deste estudo que não está sujeito a erros e, como tal, a probabilidade de falsos positivos é nula. Foram simuladas populações com $10^6$ indivíduos, aplicadas as três metodologias de classificação apresentadas na Secção 2 (Dorfman, $\mathrm{A2}\,(n:1)$ e $\mathrm{MA2}\left(n^2:n:1\right)$) e as duas metodologias de teste ($\mathbf{T_1}$ e $\mathbf{T_2}$) propostas na Secção 1, utilizando um nível de significância $\alpha = 0.05$. Nesta análise foram utilizadas diferentes dimensões dos grupos $n \in \{3{,}10{,}20\}$, taxas de prevalência $p \in \{0.1{,}0.01{,}0.001\}$, bem como distribuições contínuas. As distribuições foram ordenadas por ordem crescente do valor de $\tau$, sendo expostas na Tabela 1 a normal padrão (representada por $\mathcal{N}$), a exponencial padrão (**Exp**), a quiquadrado com um grau de liberdade ($\chi^2_{(1)}$), a log-normal ($\ln\mathcal{N}$), a Weibull com parâmetro de forma igual a 0.5 ($\mathbf{W}_{(0.5)}$) e a normal dividida (**SN**, obtida pelo quociente entre duas v.a. independentes, uma com distribuição normal padrão e a outra com distribuição uniforme no intervalo [0,1]); na Tabela 2 são apresentados os resultados para a $t$ de Student com um grau de liberdade ($t_{(1)}$), o valor absoluto de uma $t$ de Student com um grau de liberdade ($|t_{(1)}|$), o valor absoluto de uma normal dividida ($|\mathbf{SN}|$), a Pareto com parâmetro de forma igual a 1 ($\mathbf{P}_{(1)}$), a Weibull com parâmetro 0.25 ($\mathbf{W}_{(0.25)}$) e a Lévy

Tabela 1: Simulação da eficiência e sensibilidade (entre parêntesis)

| | | $\mathcal{N}$ | | **Exp** | | $\chi^2_{(1)}$ | | $\ln \mathcal{N}$ | | $\mathbf{W}_{(0.50)}$ | | **SN** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | | 1 | | 1.64 | | 2.06 | | 2.78 | | 4.17 | | 7.87 | |
| $p$ | $n$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ | $\mathbf{T_1}$ | $\mathbf{T_2}$ |
| **Metodologia de Dorfman** | | | | | | | | | | | | | |
| .1 | 3 | 49.4 | 101. | 56.7 | 74.5 | 58.9 | 68.4 | 59.3 | 68.9 | 62.2 | 63.7 | 57.7 | 121. |
| | | (50.9) | (95.8) | (74.8) | (95.9) | (82.9) | (96.0) | (84.2) | (96.0) | (93.7) | (96.0) | (77.0) | (95.7) |
| .1 | 10 | 33.0 | 102. | 48.6 | 94.3 | 53.8 | 90.1 | 55.7 | 90.2 | 62.7 | 84.7 | 47.6 | 105. |
| | | (42.4) | (97.8) | (67.8) | (98.1) | (75.1) | (98.1) | (77.5) | (98.1) | (85.6) | (98.1) | (62.7) | (95.9) |
| .1 | 20 | 36.4 | 101. | 57.9 | 101. | 65.7 | 99.6 | 67.8 | 99.5 | 76.9 | 97.9 | 51.6 | 101. |
| | | (46.3) | (98.7) | (73.1) | (99.3) | (80.9) | (99.3) | (82.7) | (99.3) | (90.0) | (99.2) | (61.7) | (96.2) |
| .01 | 3 | 39.8 | 76.8 | 40.9 | 42.7 | 41.1 | 39.7 | 41.0 | 37.8 | 41.1 | 37.1 | 41.1 | 36.4 |
| | | (50.8) | (95.0) | (89.9) | (95.2) | (99.5) | (95.0) | (100.) | (95.3) | (100.) | (95.4) | (98.7) | (95.1) |
| .01 | 10 | 16.9 | 87.7 | 18.9 | 56.8 | 19.7 | 45.5 | 21.4 | 35.3 | 22.2 | 29.1 | 23.5 | 56.1 |
| | | (25.1) | (95.5) | (47.4) | (95.3) | (56.2) | (95.6) | (74.4) | (95.7) | (82.3) | (95.7) | (91.3) | (95.4) |
| .01 | 20 | 12.2 | 91.1 | 14.8 | 73.9 | 16.1 | 65.5 | 19.5 | 55.0 | 20.3 | 45.7 | 23.7 | 93.6 |
| | | (19.1) | (95.8) | (35.8) | (95.9) | (42.9) | (95.8) | (59.7) | (96.1) | (64.9) | (96.4) | (79.8) | (95.6) |
| .001 | 3 | 38.5 | 61.4 | 38.6 | 35.0 | 38.5 | 34.3 | 38.7 | 33.7 | 38.6 | 33.7 | 38.4 | 33.6 |
| | | (66.3) | (94.8) | (100.) | (94.6) | (100.) | (95.1) | (100.) | (94.3) | (100.) | (94.8) | (99.8) | (95.9) |
| .001 | 10 | 15.2 | 79.4 | 15.6 | 31.6 | 15.6 | 21.9 | 15.9 | 13.0 | 16.0 | 12.5 | 16.0 | 10.9 |
| | | (27.2) | (94.9) | (65.7) | (95.9) | (52.3) | (94.3) | (99.9) | (94.7) | (100.) | (95.4) | (99.7) | (95.5) |
| .001 | 20 | 10.3 | 85.4 | 10.7 | 51.1 | 11.1 | 37.9 | 11.5 | 15.3 | 11.9 | 13.4 | 11.6 | 6.84 |
| | | (19.3) | (95.3) | (40.6) | (94.5) | (52.3) | (95.8) | (86.8) | (95.2) | (92.0) | (95.5) | (98.7) | (95.1) |
| **Metodologia baseada em tabelas quadradas com teste global** | | | | | | | | | | | | | |
| .1 | 3 | 31.0 | 121. | 44.4 | 90.3 | 49.6 | 92.9 | 51.1 | 83.4 | 57.6 | 75.6 | 44.5 | 152. |
| | | (24.5) | (90.7) | (51.3) | (92.6) | (62.6) | (93.4) | (65.7) | (93.2) | (79.9) | (93.8) | (52.5) | (90.2) |
| .1 | 10 | 21.8 | 105. | 36.3 | 92.7 | 41.4 | 85.7 | 43.2 | 86.2 | 49.9 | 77.5 | 26.3 | 108. |
| | | (17.6) | (95.6) | (46.7) | (96.3) | (58.0) | (96.1) | (61.5) | (96.2) | (74.3) | (96.1) | (32.8) | (89.7) |
| .1 | 20 | 20.7 | 104. | 38.9 | 101. | 47.9 | 99.9 | 50.4 | 99.7 | 61.7 | 96.4 | 26.2 | 99.2 |
| | | (21.8) | (97.6) | (54.2) | (98.5) | (66.2) | (98.6) | (69.2) | (98.5) | (81.0) | (98.6) | (32.3) | (90.3) |
| .01 | 3 | 17.0 | 87.6 | 18.7 | 46.7 | 19.3 | 37.3 | 20.5 | 29.1 | 21.0 | 24.6 | 21.6 | 32.6 |
| | | (18.7) | (88.6) | (48.2) | (93.4) | (59.2) | (93.8) | (77.2) | (94.2) | (85.0) | (93.8) | (91.8) | (94.1) |
| .01 | 10 | 3.33 | 79.8 | 4.73 | 42.4 | 5.76 | 33.3 | 7.41 | 27.1 | 8.37 | 24.1 | 10.0 | 42.0 |
| | | (3.35) | (89.0) | (11.5) | (89.4) | (18.0) | (89.0) | (33.8) | (90.5) | (43.0) | (90.6) | (55.8) | (88.2) |
| .01 | 20 | 2.08 | 82.6 | 4.04 | 58.4 | 4.68 | 47.2 | 7.14 | 35.4 | 8.25 | 27.4 | 8.45 | 85.5 |
| | | (1.78) | (90.0) | (7.86) | (91.5) | (12.2) | (91.7) | (27.8) | (91.4) | (36.2) | (91.9) | (44.8) | (87.0) |
| .001 | 3 | 15.8 | 73.2 | 15.8 | 25.5 | 15.9 | 18.8 | 15.9 | 13.1 | 15.9 | 12.7 | 15.7 | 11.7 |
| | | (29.0) | (90.1) | (68.1) | (95.0) | (79.9) | (92.8) | (100.) | (40.5) | (100.) | (94.9) | (99.7) | (94.6) |
| .001 | 10 | 2.48 | 66.3 | 2.41 | 23.5 | 2.64 | 20.8 | 2.93 | 15.1 | 2.99 | 13.7 | 3.85 | 9.05 |
| | | (3.49) | (88.1) | (10.4) | (85.6) | (19.4) | (90.5) | (40.5) | (92.6) | (41.4) | (92.3) | (92.5) | (93.8) |
| .001 | 20 | 1.05 | 72.1 | 1.15 | 30.5 | 1.29 | 20.6 | 1.40 | 11.1 | 1.64 | 10.8 | 2.95 | 9.85 |
| | | (0.96) | (88.8) | (4.39) | (88.4) | (7.73) | (88.9) | (21.3) | (88.2) | (24.3) | (89.1) | (67.9) | (91.9) |
| **Metodologia baseada em tabelas quadradas sem teste global** | | | | | | | | | | | | | |
| .1 | 3 | 25.7 | 129. | 44.5 | 104. | 51.1 | 99.7 | 51.7 | 100. | 59.3 | 95.6 | 46.2 | 156. |
| | | (54.0) | (92.6) | (74.3) | (94.8) | (82.0) | (95.2) | (83.4) | (95.1) | (93.0) | (95.5) | (76.7) | (91.7) |
| .1 | 10 | 24.4 | 104. | 36.8 | 92.3 | 41.6. | 85.8 | 43.5 | 86.3 | 50.3 | 77.9 | 36.1 | 111. |
| | | (21.3) | (95.6) | (47.2) | (96.3) | (58.3) | (96.3) | (61.7) | (96.2) | (74.5) | (96.3) | (42.6) | (92.4) |
| .1 | 20 | 20.7 | 104. | 38.8 | 1.02 | 47.9 | 99.9 | 50.4 | 99.6 | 61.9 | 96.8 | 32.6 | 102. |
| | | (21.4) | (97.6) | (53.9) | (98.5) | (66.2) | (98.6) | (69.1) | (99.6) | (81.3) | (98.6) | (39.1) | (92.3) |
| .01 | 3 | 7.56 | 118. | 12.2 | 86.2 | 13.8 | 83.3 | 15.2 | 81.3 | 16.5 | 80.4 | 16.8 | 79.4 |
| | | (65.5) | (93.1) | (94.9) | (97.4) | (99.8) | (97.9) | (100.) | (98.1) | (100.) | (97.7) | (99.5) | (98.7) |
| .01 | 10 | 7.24 | 82.3 | 10.3 | 44.6 | 11.7 | 35.3 | 14.7 | 29.6 | 15.5 | 27.1 | 16.9 | 43.7 |
| | | (22.5) | (90.8) | (40.6) | (91.3) | (49.4) | (91.0) | (67.6) | (92.3) | (75.8) | (92.5) | (87.1) | (31.1) |
| .01 | 20 | 7.29 | 84.6 | 9.26 | 58.3 | 10.1 | 47.8 | 11.9 | 35.8 | 12.4 | 27.6 | 14.1 | 89.1 |
| | | (9.02) | (91.2) | (17.8) | (91.7) | (24.0) | (92.0) | (40.5) | (92.0) | (47.5) | (91.9) | (66.9) | (90.7) |
| .001 | 3 | 5.31 | 99.0 | 7.99 | 80.0 | 8.98 | 78.9 | 10.2 | 78.1 | 10.8 | 78.0 | 11.6 | 77.9 |
| | | (82.4) | (94.9) | (1.00) | (99.1) | (100.) | (98.5) | (100.) | (98.5) | (100.) | (97.9) | (100.) | (97.3) |
| .001 | 10 | 5.10 | 70.4 | 6.15 | 28.3 | 6.39 | 26.2 | 7.41 | 23.3 | 7.83 | 22.9 | 9.25 | 21.2 |
| | | (30.2) | (90.4) | (65.6) | (91.7) | (80.3) | (91.5) | (99.8) | (96.6) | (100.) | (97.9) | (99.8) | (97.7) |
| .001 | 20 | 5.00 | 75.3 | 5.75 | 32.2 | 6.13 | 21.6 | 6.91 | 12.7 | 7.15 | 12.5 | 7.77 | 10.6 |
| | | (13.1) | (92.0) | (33.2) | (90.2) | (43.4) | (92.2) | (81.1) | (91.1) | (88.8) | (92.3) | (98.4) | (96.5) |

(**Lévy**). O valor absoluto foi utilizado em algumas distribuições de forma a evitar o uso de distribuições com caudas pesadas em ambos os lados, nas quais um valor muito elevado pode ser disfarçado por um valor muito baixo (pois unicamente conhecemos a média do grupo). Da análise aos valores obtidos via simulação salientamos as seguintes conclusões:

- quanto maior for o valor do coeficiente $\tau$, associado a $\mathbf{D}_\theta$, melhor tende a ser a sensibilidade de todas as metodologias;
- regra geral, dada uma distribuição $\mathbf{D}_\theta$, melhorar o valor de eff $(\mathcal{M})$ implica piorar $\varphi_s^{\mathcal{M}}$ (e vice-versa);
- $\mathbf{T_2}$ garante uma sensibilidade com valores moderadamente elevados, situados (em quase todos os casos) perto de 95% (objetivo pretendido ao fixarmos $\alpha = 0.05$). Contudo, nas distribuições com cauda direita pesada, $\mathbf{T_1}$ tende a obter melhor performance (obtendo sensibilidades superiores a 95%);
- $\mathbf{T_1}$ obtém por vezes eficiências bastantes mais favoráveis, contudo em alguns casos com $\varphi_s^{\mathcal{M}}$ muito baixa (para $\tau$ reduzido);
- nas distribuições com caudas pesadas em ambos os lados ($t_{_{(1)}}$ e **SN**) a sensibilidade tende a ser prejudicada (teste conjunto com maior probabilidade de ocorrência de falsos negativos), nomeadamente quando utilizada a metodologia de teste $\mathbf{T_1}$. Todavia, quando utilizamos o valor absoluto o problema não se coloca (passamos a ter uma distribuição unicamente com uma cauda direita pesada);
- com teste global tende-se a obter melhor eficiência, mas em muitos casos com menor sensibilidade (se $n^2$ for um valor elevado a probabilidade de um falso negativo no teste teste global é mais elevada devido à diluição);
- todas as metodologias apresentam situações nas quais evidenciam alcançar melhor performance que as restantes metodologias, o que está de acordo com resultados obtidos por [5] sobre a eficiência dos algoritmos hierárquicos e não hierárquicos.

Tabela 2: Simulação da eficiência e sensibilidade (entre parêntesis)

| | | $t_{(1)}$ | | $\lvert t_{(1)} \rvert$ | | $\lvert SN \rvert$ | | $P_{(1)}$ | | $W_{(0.25)}$ | | Lévy | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | | 9.23 | | 12.9 | | 13.2 | | 14.2 | | 37.6 | | 241 | |
| $p$ | $n$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ |
| | | **Metodologia de Dorfman** | | | | | | | | | | | |
| .1 | 3 | 58.0 (79.2) | 120. (95.7) | 62.8 (96.1) | 62.9 (96.1) | 62.7 (95.7) | 63.0 (96.1) | 63.2 (97.2) | 62.2 (95.9) | 64.3 (100.) | 60.2 (95.8) | 64.4 (100.) | 60.2 (96.0) |
| .1 | 10 | 48.2 (64.0) | 105. (95.9) | 66.2 (89.2) | 83.1 (98.0) | 66.0 (89.0) | 83.4 (98.1) | 67.3 (90.2) | 82.0 (98.0) | 72.2 (95.3) | 77.1 (98.1) | 73.6 (96.5) | 76.7 (98.1) |
| .1 | 20 | 52.2 (62.5) | 101. (96.4) | 80.5 (92.7) | 97.4 (99.3) | 80.3 (92.5) | 97.2 (99.3) | 82.4 (93.8) | 97.1 (99.3) | 87.2 (96.9) | 94.4 (99.3) | 89.2 (97.7) | 94.4 (99.4) |
| .01 | 3 | 41.0 (98.4) | 36.4 (94.9) | 41.2 (100.) | 36.2 (95.4) | 41.2 (100.) | 36.2 (95.2) | 41.2 (100.) | 36.3 (95.3) | 41.2 (100.) | 36.2 (95.1) | 41.0 (100.) | 36.1 (94.9) |
| .01 | 10 | 23.3 (91.5) | 52.0 (95.4) | 24.1 (100.) | 20.4 (95.2) | 23.9 (100.) | 20.2 (95.4) | 24.1 (100.) | 20.4 (95.3) | 24.1 (100.) | 20.5 (95.6) | 23.7 (100.) | 18.8 (95.5) |
| .01 | 20 | 23.3 (79.9) | 93.3 (95.3) | 26.5 (95.7) | 26.6 (95.9) | 26.3 (95.2) | 26.5 (95.6) | 26.4 (95.3) | 26.4 (95.7) | 26.1 (94.7) | 26.9 (96.0) | 27.2 (100.) | 22.7 (95.7) |
| .001 | 3 | 38.7 (99.8) | 33.6 (93.8) | 38.7 (100.) | 33.6 (94.3) | 38.6 (100.) | 33.7 (95.2) | 38.5 (100.) | 33.6 (94.6) | 38.7 (100.) | 33.6 (93.8) | 38.8 (100.) | 33.6 (94.6) |
| .001 | 10 | 15.8 (99.6) | 10.9 (94.1) | 16.0 (100.) | 11.0 (95.4) | 15.9 (100.) | 11.1 (95.5) | 15.8 (100.) | 10.9 (95.6) | 15.9 (100.) | 11.1 (94.6) | 16.2 (100.) | 11.1 (95.5) |
| .001 | 20 | 11.9 (99.1) | 7.01 (95.0) | 11.8 (100.) | 6.98 (96.3) | 12.3 (100.) | 7.27 (94.3) | 12.0 (100.) | 6.95 (94.5) | 11.8 (100.) | 7.28 (94.7) | 12.1 (100.) | 7.08 (96.0) |
| | | **Metodologia baseada em tabelas quadradas com teste global** | | | | | | | | | | | |
| .1 | 3 | 44.6 (54.4) | 151. (89.9) | 60.4 (85.3) | 73.7 (93.8) | 60.0 (85.1) | 73.9 (94.0) | 61.1 (87.7) | 71.9 (94.1) | 65.3 (95.4) | 66.7 (94.1) | 66.3 (96.6) | 66.1 (94.3) |
| .1 | 10 | 25.9 (33.1) | 107. (89.1) | 54.3 (80.9) | 75.6 (96.2) | 53.7 (80.5) | 75.7 (96.1) | 54.7 (82.7) | 73.4 (96.3) | 60.9 (91.6) | 67.2 (96.4) | 62.5 (93.4) | 66.3 (96.3) |
| .1 | 20 | 25.5 (31.5) | 98.4 (89.4) | 68.0 (86.4) | 96.0 (98.7) | 68.2 (86.2) | 96.0 (98.7) | 70.0 (87.9) | 94.8 (98.7) | 78.3 (94.0) | 90.4 (98.7) | 81.1 (95.4) | 90.1 (98.7) |
| .01 | 3 | 21.5 (92.2) | 31.1 (94.2) | 21.9 (100.) | 18.4 (93.6) | 21.9 (100.) | 18.5 (94.6) | 22.0 (100.) | 18.5 (94.0) | 21.9 (100.) | 18.7 (94.5) | 21.7 (100.) | 17.6 (94.7) |
| .01 | 10 | 9.80 (55.9) | 38.4 (87.5) | 13.9 (84.9) | 18.7 (92.9) | 13.9 (85.7) | 18.6 (94.7) | 14.3 (86.5) | 18.7 (92.9) | 13.3 (82.3) | 18.9 (92.9) | 15.6 (96.1) | 15.3 (93.6) |
| .01 | 20 | 8.15 (43.9) | 85.6 (87.8) | 14.6 (89.9) | 15.7 (93.1) | 14.3 (89.5) | 15.6 (93.0) | 14.1 (88.8) | 15.3 (92.3) | 13.7 (85.4) | 15.6 (92.2) | 15.6 (99.4) | 14.0 (93.8) |
| .001 | 3 | 16.0 (99.4) | 11.8 (93.5) | 16.0 (100.) | 11.8 (94.3) | 16.1 (100.) | 11.9 (94.1) | 15.8 (100.) | 11.7 (95.3) | 15.9 (100.) | 11.9 (94.7) | 15.9 (100.) | 11.8 (93.8) |
| .001 | 10 | 4.03 (90.5) | 11.7 (92.9) | 4.30 (100.) | 3.33 (92.9) | 4.50 (99.8) | 3.64 (93.6) | 4.12 (100.) | 3.15 (93.2) | 3.94 (88.4) | 4.42 (93.0) | 4.33 (100.) | 3.11 (94.6) |
| .001 | 20 | 3.01 (65.5) | 9.93 (91.2) | 3.82 (86.5) | 4.77 (92.2) | 4.21 (85.1) | 5.39 (94.2) | 3.95 (85.9) | 4.88 (91.9) | 2.73 (62.8) | 6.74 (91.3) | 4.65 (99.0) | 3.98 (93.6) |
| | | **Metodologia baseada em tabelas quadradas sem teste global** | | | | | | | | | | | |
| .1 | 3 | 47.6 (79.2) | 157. (91.8) | 61.0 (95.5) | 94.7 (95.5) | 61.0 (95.3) | 95.1 (95.7) | 62.4 (97.0) | 94.2 (95.6) | 66.2 (100.) | 91.8 (95.5) | 66.3 (100.) | 91.7 (95.5) |
| .1 | 10 | 36.7 (44.0) | 111. (92.0) | 54.2 (80.9) | 75.4 (96.1) | 53.7 (80.4) | 75.7 (96.0) | 55.0 (82.4) | 73.7 (96.2) | 61.3 (91.6) | 67.6 (96.5) | 63.1 (93.7) | 66.9 (96.5) |
| .1 | 20 | 32.9 (39.9) | 102. (92.8) | 68.0 (86.4) | 95.6 (98.7) | 67.8 (86.1) | 95.8 (98.6) | 70.2 (87.9) | 94.9 (98.7) | 78.5 (94.0) | 90.6 (98.7) | 80.7 (95.4) | 89.6 (98.7) |
| .01 | 3 | 16.9 (99.6) | 79.3 (98.7) | 18.5 (100.) | 79.0 (96.4) | 18.5 (100.) | 79.0 (96.7) | 18.4 (100.) | 79.0 (96.4) | 19.0 (100.) | 79.0 (96.5) | 18.9 (100.) | 78.8 (95.3) |
| .01 | 10 | 16.7 (88.4) | 39.8 (91.2) | 18.5 (100.) | 23.8 (94.4) | 18.3 (100.) | 23.7 (94.7) | 18.7 (100.) | 23.7 (94.3) | 18.4 (100.) | 23.8 (95.0) | 19.1 (100.) | 22.8 (95.1) |
| .01 | 20 | 13.8 (67.9) | 88.7 (90.9) | 15.3 (92.5) | 15.6 (92.9) | 15.2 (92.5) | 15.5 (93.1) | 15.4 (92.8) | 15.5 (92.8) | 15.2 (91.4) | 15.8 (93.4) | 15.6 (100.) | 13.9 (94.2) |
| .001 | 3 | 12.1 (100.) | 77.9 (96.8) | 13.1 (100.) | 77.9 (95.0) | 12.9 (100.) | 77.9 (96.4) | 12.5 (100.) | 77.9 (94.9) | 13.1 (100.) | 77.9 (97.3) | 13.8 (100.) | 77.9 (93.5) |
| .001 | 10 | 8.86 (99.5) | 21.1 (97.9) | 9.65 (100.) | 21.1 (94.5) | 9.91 (100.) | 21.1 (97.0) | 9.70 (100.) | 21.1 (96.3) | 9.52 (100.) | 21.3 (96.6) | 10.3 (100.) | 21.1 (95.6) |
| .001 | 20 | 7.82 (98.4) | 10.6 (96.4) | 7.91 (100.) | 10.5 (94.4) | 8.33 (100.) | 10.5 (96.5) | 8.31 (100.) | 10.5 (95.5) | 8.05 (100.) | 10.9 (95.5) | 8.46 (100.) | 10.4 (94.6) |

# 4 Conclusão

Os resultados obtidos neste estudo permitem concluir que o valor de $\tau$ da distribuição $\mathbf{D}_\theta$ é uma medida reveladora da qualidade do desempenho da aplicação de metodologias de classificação baseadas em testes compostos. Por outro lado, a utilização de testes compostos corresponde sempre a uma poupança de recursos (para taxas de prevalência baixas, as únicas analisadas neste trabalho) garantido igualmente, no caso de utilização da metodologia $\mathbf{T_2}$, uma sensibilidade com valores moderadamente elevados. Contudo, a escolha da metodologia mais adequada a aplicar em cada caso não é consensual, dependendo muito das características da distribuição $\mathbf{D}_\theta$ e da taxa de prevalência $p$, bem como do custo que o utilizador associar à realização de mais testes (eficiência) e à existência de uma maior probabilidade de ocorrência de má classificação (nomeadamente da sensibilidade). Assim, permanece em aberto a definição de um processo que permita a seleção da metodologia mais adequada para uma dada distribuição, bem como do tipo de contexto no qual a aplicação da metodologia $\mathbf{T_1}$ se revele mais favorável, embora tais definições dependam inequivocamente dos pesos atribuídos pelo investigador à eficiência *versus* fiabilidade da metodologia.

# Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst-OE/MAT/UI0006/2014.

# Referências

[1] Dorfman, R. (1943). The detection of defective members in large populations. *Annals of Mathematical Statistics* 14, 436–440.

[2] Gastwirth, J.L. (2000). The efficiency of pooling in the detection of rare mutations. *American Journal of Human Genetics* 67, 1036–1039.

[3] Hoaglin D.M., Mosteller F., Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley.

[4] Hughes-Oliver, J.M. (2006). Pooling experiments for blood screening and drug discovery, Screening - Methods for Experimentation in Industry, Drug Discovery, and Genetics, Springer, 48–68.

[5] Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C. (2007). Comparison of group testing algorithms for case identification in the presence of testing errors. *Biometrics* 63, 1152–1163.

[6] Johnson, N.L., Kotz, S., Wu, X. (1991).*Inspection Errors for Attributes in Quality Control*, Chapman and Hall.

[7] Litvak, E., Tu, X.M., Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* 89, 424–434.

[8] Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, T.C. (2011). Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors. *Annals of Applied Biology* 159, 491–502.

[9] Martins, J.P., Santos, R., Sousa, R. (2014). Testing the maximum by the mean in quantitative group tests. In Pacheco, A. *et al.* (eds.): *New Advances in Statistical Modeling and Applications, Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies*, Springer-Verlag, pp. 55–63.

[10] Phatarfod, R.M., Sudbury, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine* **13**, 2337–2343.

[11] Santos, R., Pestana, D., Martins, J.P. (2013). Extensions of Dorfman's theory. In Oliveira, P.E. et al. (eds.): *Studies in Theoretical and Applied Statistics, Recent Developments in Modeling and Applications in Statistics, Selected Papers of the Statistical Societies*, 179–189, Springer.

[12] Santos, R., Felgueiras, M., Martins, J.P. (2014). Known mean, unknown maxima? Testing the maximum knowing only the mean. *Communications in Statistics – Simulation and Computation* (Published online: 23 Jan 2014)

[13] Woodbury, C.P., Fitzloff, J.F., Vincent, S.S. (1995). Sample multiplexing for greater throughput in HPLC and related methods. *Analytical Chemistry* **67**, 885–890.

# Modelação de grandes incêndios em Portugal

Alexandra Ramos
FEP e CMUP, Universidade do Porto, *aramos@fep.up.pt*

**Palavras–chave**: teoria de valores extremos bivariados, modelos para a dependência assintótica extremal, modelos de markov, grandes incêndios

**Abstract**: O objetivo deste estudo é a análise das caraterísticas da distribuição dos grandes incêndios em Portugal. A modelação da cauda conjunta da distribuição de um par de valores consecutivos de áreas ardidas $(X_i,X_{i+1})$ será aqui realizada através de uma cadeia de Markov estacionária de primeira ordem onde a dependência temporal é descrita pelo modelo bivariado apresentado em Ramos e Ledford (2009).

## 1 Introdução

Os grandes incêndios podem ter um efeito devastador no meio ambiente, economia e na vida humana. Recentemente, os incêndios têm causado danos significativos no clima, no eco-sistema e até resultado em perdas de vidas humanas. Em trabalhos como de Zea Bermudez *et al.* [1] e Mendes *et al.* [2], os grandes incêndios em Portugal são modelados usando a distribuição generalizada de Pareto e modelos bayesianos hierárquicos. A caracterização da distribuição da dimensão dos incêndios em Portugal é aqui feita tomando como os dados as áreas ardidas por incêndio e usando uma abordagem que admite dependência entre observações extremas consecutivas. Nesta abordagem adoptamos o modelo paramétrico para a cauda conjunta de uma distribuição bivariada construído em Ramos e Ledford [3]. Usando este flexível modelo paramétrico, analisamos a dependência temporal de curta duração numa série temporal estacionária,

caracterizando o comportamento extremal da série. Este tipo de caracterização tem sido feita, na sua grande maioria, com base no pressuposto da série temporal ter a propriedade de Markov, visto a classe das cadeias de Markov de ordem $d$ ser suficientemente geral e tratável. A modelação da cauda conjunta da distribuição de um par de valores consecutivos $(X_i, X_{i+1})$ de uma cadeia de Markov estacionária de primeira ordem será então aqui realizada através do modelo descrito em Ramos e Ledford [3]. Esta análise requer a avaliação conjunta dos extremos da estrutura de dependência temporal, bem como da cauda da distribuição marginal. Em particular, estudamos o grau de dependência extremal na cauda conjunta superior de duas observações consecutivas e o peso da cauda da série.

## 2 Dados

Tendo os incêndios florestais se tornado um problema grave de segurança interna, em 2010, foi promovida uma reformulação ao Sistema de Gestão de Incêndios Florestais (SGIF) de forma a se agilizar os procedimentos entre os diversos intervenientes. Este processo exigiu uma análise exaustiva da base de dados dos incêndios florestais desde a sua implementação em 2001. A base de dados aqui considerada contém 571582 ocorrências de incêndio de qualquer dimensão em Portugal de 1980 a 2011 e tem como fonte a Autoridade Florestal Nacional de acordo com o SGIF.

A caraterística em estudo será a área ardida (em ha) num incêndios de qualquer dimensão, sendo que estes se classificam como fogachos, se a área ardida for inferior a 1 ha (correspondendo a 407303 ocorrências num total de 71.3% dos casos); incêndio, se a área ardida for superior ou igual a 1 ha (correspondendo a 164279 ocorrências num total de 28.7% dos casos) e classificam-se como grandes incêndios se a área ardida for superior ou igual a 100 ha (correspondendo a 4905 ocorrências num total de 0.9% dos casos). Dada a importância dos incêndios de grandes dimensões, o estudo e a caracterização dos valores extremos deste tipo de dados mostra-se fundamental. De

facto, um pequeno número de incêndios de grandes dimensões é responsável pela maior parte da área ardida e pelos maiores estragos sociais e ambientais causados. Por exemplo, os 10% maiores incêndios (acima de 3 ha) são responsáveis por 93.29% da área ardida total. Ou ainda, os 5% / 2% maiores incêndios (acima de 7.2 ha / 30 ha) são responsáveis por 89.81% / 82.54% da área ardida total. A existência e importância de valores extremos é ainda evidente na assimetria à direita presente na distribuição das áreas ardidas que apresenta uma pesada cauda superior e onde valores superiores a 11 ha são considerados outliers (ver a caixa de bigodes da Fig. 1).

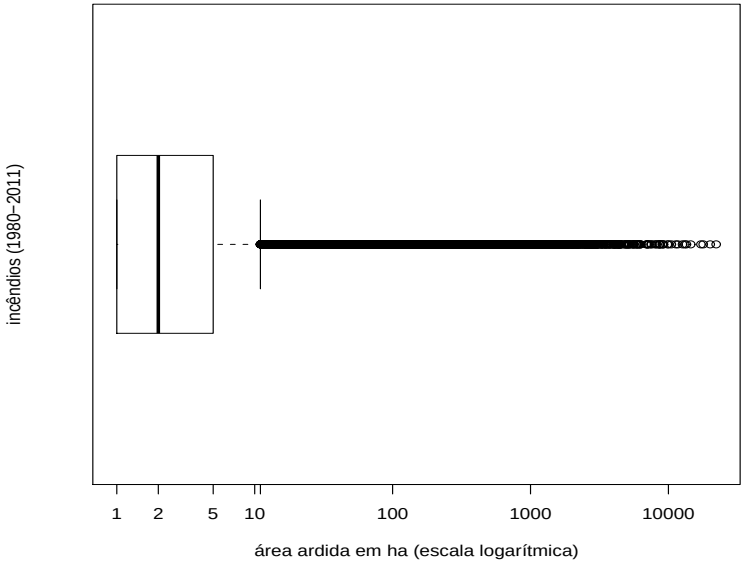



Figura 1: Caixa de bigodes da distribuição da área ardida.

A série temporal de áreas ardidas em incêndios ao longo do período de 1980 a 2011 é apresentada na Fig. 2. A presença de dependência local em níveis extremos é evidente através da observação de

*clusters* de valores extremos elevados. Este comportamento de *clustering* pode ter importantes implicações na prática e corresponde à ocorrência de elevadas observações consecutivas. Note que, apesar de valores extremos consecutivos da série poderem ocorrer em diferentes regiões do país, estes mostram uma forte relação entre eles, facto que poderá ser justificado por dias de muito calor afetarem usualmente todo o país. A estacionaridade da série é suportada pelo seu comportamento aparente na Fig. 2 ou pela observação da similaridade entre as caixas de bigodes para os valores das áreas ardidas dos diferentes anos (não reportado), visto não ser evidente nenhuma tendência da série ao longo do tempo.

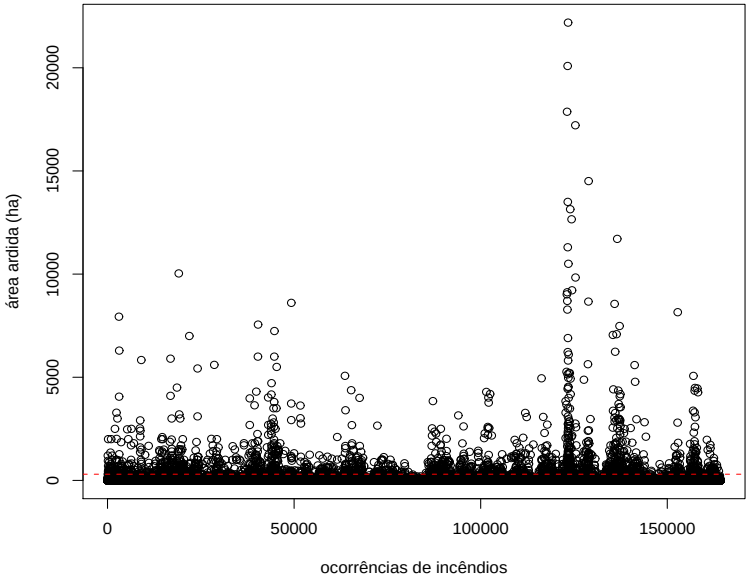



Figura 2: Valores das áreas ardidas em incêndios ocorridos entre 1980 e 2011 apresentados por ordem cronológica de ocorrência.

# 3 Modelos para as caudas de cadeias de Markov

Os avanços na teoria dos valores extremos multivariados conduziram a um aperfeiçoamento das técnicas de caracterização do comportamento extremal de séries temporais estacionárias. Em particular, tem sido dada atenção ao comportamento dentro de *clusters* de extremos de uma série temporal, que é determinado pela dependência temporal de curta duração. Grande parte desta caracterização tem sido feita baseada na suposição de que a série temporal é uma cadeia de Markov, devido ao facto das cadeias de Markov serem suficientemente gerais e tratáveis. Uma abordagem usual para modelar estatisticamente a cauda de uma cadeia de Markov consiste na utilização de uma determinada distribuição conjunta de valores extremos $d$-dimensional para modelar a estrutura de dependência entre variáveis consecutivas $X_i$, ..., $X_{i+d}$ que excedam um nível elevado fixo $u$, abordagem que se baseia na suposição de que o comportamento limite da cadeia é igualmente válido acima de um nível elevado $u$.
Seja $\{X_n\}_{n\geq 1}$ uma cadeia de Markov estacionária de $1^a$ ordem com espaço de estados contínuos e denote-se a sua função de distribuição conjunta por $F(x_i,x_{i+1})$, e a sua distribuição marginal por $F(x) = \Pr(X_i \leq x)$ para todo o $i \geq 1$. Em seguida, considera-se a distribuição generalizada de Pareto (GPD) para descrever o comportamento da cauda univariada acima de um nível elevado $u_1$, nível na escala original da variável $X_i$, ou seja a distribuição

$$F(x) = \begin{cases} 1 - \lambda_1 \left\{1 + \xi(x-u_1)/\sigma\right\}_+^{-1/\xi}, & x \geq u_1 \\ 1 - \lambda_1, & x < u_1 \end{cases} \qquad (1)$$

onde $s_+ = \max(s,0)$, $\xi$ e $\sigma > 0$ são, respetivamente, os parâmetros de forma e escala e $\lambda_1$ denota a probabilidade de excedência ao nível. A distribuição conjunta de $(X_i,X_{i+1})$ numa região de cauda $R_{11} = (u_1,\infty) \times (u_1,\infty)$ é obtida de forma análoga. Adotamos aqui o modelo logístico $\eta$-assimétrico dado em [3] como modelo da cauda conjunta para descrever a dependência entre observações elevadas

consecutivas da cadeia, modelo dado por

$$F(x_i,x_{i+1}) = F(x_i) + F(x_{i+1}) - 1 + \frac{\lambda}{N_\varrho}\left[\left(\frac{\varrho y_i}{u_f}\right)^{-\frac{1}{\eta}} + \left(\frac{y_{i+1}}{\varrho u_f}\right)^{-\frac{1}{\eta}} - \left\{\left(\frac{\varrho y_i}{u_f}\right)^{-\frac{1}{\alpha}} + \left(\frac{y_{i+1}}{\varrho u_f}\right)^{-\frac{1}{\alpha}}\right\}^{\frac{\alpha}{\eta}}\right] \tag{2}$$

para $x_i$,$x_{i+1} > u_1$ e onde $\eta, \alpha \in (0,1]$ e $\varrho > 0$, $N_\varrho = \varrho^{-1/\eta} + \varrho^{1/\eta} - (\varrho^{-1/\alpha} + \varrho^{1/\alpha})^{\alpha/\eta}$, $y_j = -1/\log F(x_j)$ $(j = i,i+1)$, $F(x)$ como definido em (1) e $u_f = -1/\log(1-\lambda_1)$ é o nível elevado na escala Fréchet unitária.

O ajustamento do modelo é feito através do método da máxima verosimilhança com múltipla censura, sendo observações abaixo de um nível elevado censuradas para cada margem, ver *e.g.* Smith *et al.* [4].

# 4 Aplicação aos dados

Nesta secção, aplica-se a metodologia dos modelos das cadeias de Markov descritos na Secção 3 à série $\{X_n\}$, das áreas ardidas apresentada na Fig. 2. Para a cauda conjunta superior dos pares $(X_i,X_{i+1})$ será examinada a dependência extremal (via $\eta$) e o peso da cauda (via $\xi$). Esta análise requer a estimação da estrutura da dependência temporal extremal, bem como da cauda da distribuição marginal. O modelo de Markov para caudas conjuntas definido em (2) foi ajustado à série para o nível elevado $u_1 = 300$, representado na Fig. 2, que corresponde ao quantil empírico de ordem 0.99 da distribuição marginal de $\{X_n\}$. As estimativas obtidas através do método da máxima verosimilhança referido acima para os parâmetros de dependência e respectivos desvios padrão dados entre parêntesis são os seguintes: $\widehat{\alpha} = 1.056$ (0.09), $\widehat{\eta} = 0.809$ (0.05) e $\widehat{\varrho} = 1.65$ (0.56) e as estimativas para os parâmetros marginais são: $\widehat{\xi} = 0.758$ (0.05) e $\widehat{\sigma} = 272.543$ (12.38). Estes valores foram obtidos usando programa-

ção no *software* R. Uma estimativa para o coeficiente de dependência assintótica na cauda, $\eta$, significativamente menor do que 1 indica que a série é assintoticamente independente, embora haja uma forte dependência temporal extremal (positiva) entre observações consecutivas ($X_i$,$X_{i+1}$) acima de níveis elevados sub-assintóticos, visto o valor para $\eta$ ser superior a 0.5 e próximo de 1. Por outro lado, uma estimativa para o índice de cauda, $\xi$, superior a 0 indica que a distribuição marginal tem uma cauda pesada. A qualidade do ajuste do modelo (2) aos dados pode ser verificada na Fig. 3, onde estão representadas as curvas de nível da densidade do modelo ajustado à cauda conjunta das observações ($X_i$,$X_{i+1}$), sobrepostas na figura. Note-se que observações consecutivas extremas correspondem a grandes incêndios ocorridos numa determinada época de um mesmo ano (usualmente no verão).

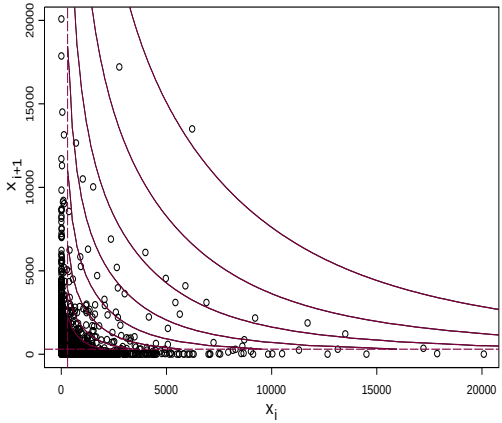



Figura 3: Curvas de nível da densidade do modelo ajustado à cauda conjunta das observações ($X_i$,$X_{i+1}$). O nível elevado escolhido $u_1$ está também identificado.

Para se verificar a estabilidade das estimativas obtidas, ajustou-se o modelo anterior acima de vários níveis $u_1$. Analisando a Fig. 4, verifica-se uma estabilidade das estimativas de $\eta$ em torno do valor 0.8 e das de $\xi$ em torno do valor positivo 0.75. Intervalos de confiança a 95% estão também incluídos na figura. Estes valores continuam a sugerir uma forte dependência temporal extremal e caudas marginais pesadas.

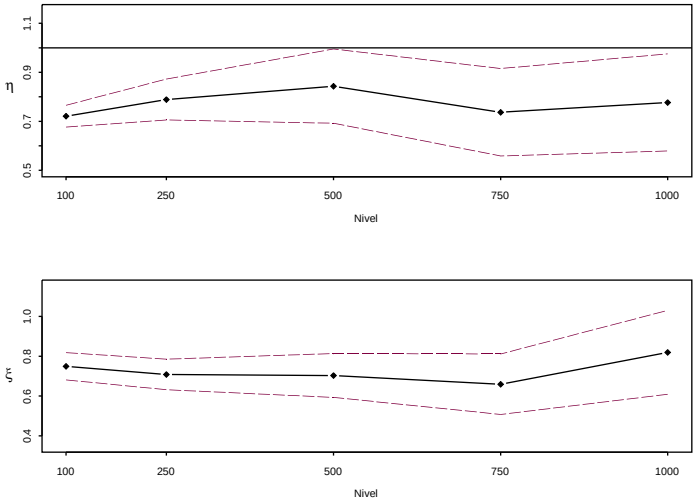



Figura 4: Estimativas de $\eta$ (acima) e de $\xi$ (abaixo) obtidas ajustando o modelo (2) para vários valores do nível $u_1$ e respectivos intervalos de confiança a 95% obtidos pelo método delta.

# 5 Conclusões e comentários

Os valores das estimativas dos parâmetros sugerem que a série de valores de área ardida em grandes incêndios em Portugal para o período de 1980 a 2011 apresenta caudas marginais pesadas, o que

traduz uma agressividade na observação de incêndios cada vez de maiores dimensões, e uma forte dependência temporal extremal, revelando que grandes incêndios tendem a ocorrer seguidos.
Muito importante seria também a análise da componente espacial na modelação da área ardida, ou seja a sua análise por região em Portugal. Adotamos no entanto por não a incluir neste trabalho pois, como referido em de Zea Bermudez *et al.* [1] e outros trabalhos por eles referenciados, o padrão geográfico dos incêndios de dimensões extremas imita o da área total ardida em Portugal para a maioria das regiões.

# Agradecimentos

Este trabalho é financiado por Fundos FEDER através do Programa Operacional Factores de Competitividade - COMPETE e por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projecto PEst-C/MAT/UI0144/2011.

# Referências

[1] de Zea Bermudez, P., Mendes, J., Pereira, J.M.C., Turkman, K.F., Vasconcelos, M.J.P. (2009). Spatial and temporal extremes of wildfire sizes in Portugal (1984–2004). *International Journal of Wildland Fire* 18, 983–991.

[2] Mendes, J., de Zea Bermudez, P., Pereira, J.M.C., Turkman, K.F., Vasconcelos, M.J.P. (2010). Spatial extremes of wildfire sizes: Bayesian hierarchical models for extremes. *Environmental and Ecological Statistics* 17, 1–28.

[3] Ramos, A., Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society B* 71, 219–241.

[4] Smith, R.L., Tawn, J.A., Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika* 84, 249–268.

# Índice de Autores


SOCIEDADE PORTUGUESA
DE ESTATÍSTICA