XXII CONGRESSO SPE 2015
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Estatística: Progressos e Aplicações

Atas do XXII Congresso da Sociedade Portuguesa de Estatística

Editores:
Clara Cordeiro, Conceição Ribeiro, Carlos Sousa, Maria Helena Gonçalves, Nelson Antunes e Maria Eduarda Silva.

UAlg
UNIVERSIDADE DO ALGARVE

SPE
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

ESTATÍSTICA:
Progressos e Aplicações

Atas do XXII Congresso da
Sociedade Portuguesa de Estatística

Olhão, 07 a 10 de outubro de 2015

Editores
Clara Cordeiro
Conceição Ribeiro
Carlos Sousa
Maria Helena Gonçalves
Nelson Antunes
Maria Eduarda Silva

Novembro, 2016
Edições SPE

Editores: Clara Cordeiro, Conceição Ribeiro, Carlos Sousa, Maria Helena Gonçalves, Nelson Antunes e Maria Eduarda Silva

Título: Estatística: Progressos e Aplicações
Atas do XXII Congresso da Sociedade Portuguesa de Estatística

Editora: Sociedade Portuguesa de Estatística

Conceção Gráfica da Capa: Ludovico Silva, Gabinete de Comunicação e Protocolo da Universidade do Algarve

Impressão: Instituto Nacional de Estatística

Tiragem: 200 Exemplares

ISBN: 978-972-8890-39-1

Depósito Legal: 417937/16

Prefácio

Este é o Livro de Atas do XXII Congresso da Sociedade Portuguesa de Estatística (SPE) e o seu conteúdo é o resultado de um excelente trabalho de revisão de artigos apresentados durante o congresso e submetidos a apreciação para publicação nestas Atas.

O congresso realizou-se no Centro de Congressos Ria Formosa, sediado no Real Marina Hotel & SPA em Olhão, de 07 a 10 de outubro de 2015, e decorreu de forma excelente, tendo reunido perto de 200 participantes, na grande maioria portugueses mas também com a presença de outras nacionalidades. Desta forma, foi possível, mais uma vez, fomentar o desenvolvimento da investigação na área da Probabilidade e da Estatística, promover a sua implantação junto da sociedade civil e o intercâmbio científico através do diálogo e colaboração entre os participantes.

O programa científico contemplou 4 sessões plenárias, 7 sessões temáticas com 21 comunicações orais, 82 comunicações orais livres e 30 posters. Para as sessões plenárias foram convidados os oradores Luzia Gonçalves, da Universidade Nova de Lisboa, James W. Taylor, da Universidade de Oxford, Manuel Scotto, da Universidade de Lisboa e Peter Müller, da Universidade do Texas, Austin. As sessões temáticas foram organizadas pelos colegas Fátima Salgueiro, do Instituto Universitário de Lisboa e Business Research Unit, Henrique Cabral, da Universidade de Lisboa & MARE, Inês Sousa, da Universidade do Minho, Irene Oliveira, da Universidade de Trás-os Montes e Alto Douro, Luísa Canto e Castro, da Universidade de Lisboa, Pedro Fernandes do Instituto Gulbenkian de Ciência, Regina Bispo e Tiago Marques, da Startfactor.

O início dos trabalhos foi precedido pelo habitual minicurso do congresso da SPE, este ano, intitulado "Estatística Bayesiana Computacional - uma introdução", e lecionado pela Professora Doutora Maria Antónia Amaral Turkman, CEAUL e Faculdade de Ciências da Universidade de Lisboa e pelo Professor Doutor Carlos Daniel Paulino, CEAUL e Instituto Superior Técnico da Universidade de Lisboa.

No decorrer do congresso realizou-se também a atribuição do Prémio SPE 2015 à aluna do Doutoramento em Ciências - especialidade em Matemática, da Escola de Ciências da Universidade do Minho, Ana Isabel Borges, com o trabalho "Modelação Conjunta de Dados Longitudinais e de Sobrevivência de Cancro da Mama". Este prémio é atribuído anualmente e tem por objetivo estimular o estudo e a investigação científica em Probabilidade e Estatística entre os jovens. Um dos momentos mais marcantes do XXII Congresso da SPE foi a atribuição do Prémio Carreira SPE à Professora Doutora Maria Antónia Amaral Turkman, pela sua obra científica e pela sua dedicação ao desenvolvimento e divulgação da Estatística em Portugal.

A consecução das tarefas inerentes à realização deste congresso deveu-se às respetivas Comissão Organizadora (CO) e Comissão Científica (CC), mas também à passagem de testemunho de elementos de comissões anteriores, a todos queremos expressar o nosso profundo agradecimento. Um agradecimento especial é devido à Margarida Silva do CEAUL, pela prontidão e disponibilidade que sempre manifestou em apoiar-nos. Queremos também agradecer aos autores dos artigos submetidos a apreciação para publicação nestas atas e, em especial, a todos os revisores. O nosso agradecimento é extensivo ao INE por mais uma vez ter aceite encarregar-se da impressão deste documento no âmbito da frutuosa colaboração que mantém com a SPE. Só com o envolvimento de todos os intervenientes foi possível concluir mais um volume das Atas da SPE e divulgar, por esta via, parte da produção científica da comunidade estatística portuguesa.

Por fim, queremos expressar o nosso reconhecimento a todos os congressistas pelos trabalhos apresentados durante o congresso, na certeza de ser a divulgação do que de melhor se faz em **Estatística** que promove, na sociedade, os **Progressos e Aplicações** desta ciência.

Faro, novembro de 2016
Os Editores

Agradecimentos

Aos seguintes colegas, pelo generoso trabalho de revisão:

Adelaide Figueiredo, Faculdade de Economia da Universidade do Porto
Alexandra Ramos, Faculdade de Economia da Universidade do Porto
Ana Isabel Carita, Secção Autónoma de Métodos Matemáticos e CIPER, Faculdade de Motricidade Humana, Universidade de Lisboa
Ana Pires, Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa
Anabela Afonso, CIMA, IIFA, Departamento de Matemática, ECT, Universidade de Évora
A. Manuela Gonçalves, Centro de Matemática (CMAT), Departamento de Matemática e Aplicações (DMA), Universidade do Minho
A. Rita Gaio, Departamento de Matemática, Faculdade de Ciências da Universidade do Porto
Carlos Daniel Paulino, CEAUL e Instituto Superior Técnico, Universidade de Lisboa
Carlos Tenreiro, Departamento de Matemática da FCTUC, Universidade de Coimbra
Conceição Costa, Departamento de Matemática e CIDMA, Universidade de Aveiro
Cristina Rocha, CEAUL e Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa
Dário Ferreira, Departamento de Matemática e Centro de Matemática e Aplicações da Universidade da Beira Interior
Dinis Pestana, CEAUL e Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa
Dora Prata Gomes, Centro de Matemática e Aplicações (CMA), e Departamento de Matemática, FCT, Universidade Nova de Lisboa
Dulce Pereira, Departamento de Matemática, Escola de Ciências e Tecnologia, Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora
Elisabete Carolino, Escola Superior de Tecnologia de Saúde de Lisboa, Instituto Politécnico de Lisboa

Agradecimentos

Agradecemos às seguintes entidades o valioso apoio concedido para a realização do XXII Congresso da SPE

Associação de Turismo do Algarve, ATA

Banco de Portugal

Câmara Municipal de Olhão, CMO

Centro de Estatística e Aplicações da Universidade de Lisboa, CEAUL

Edições Sílabo

Escolar Editora

Fundação para a Ciência e a Tecnologia

Fidelidade

HPZ

Hubel

Instituto Nacional de Estatística, INE

João Mendes & Rita, Lda

Junta de Freguesia de Olhão

Junta de Freguesia de Pechão

Manná

Multicópias

Produtos e Serviços de Estatística, PSE

Real Marina Hotel & SPA

Região Turismo do Algarve, RTA

Salexpor

Untapped Events, Organização e Gestão de Eventos

Universidade do Algarve

- **Centro de Ciências do Mar, CCMAR**
- **Centro de Estudos em Património, Paisagem e Construção, CEPAC**
- **Departamento de Engenharia Civil, Instituto Superior de Engenharia**
- **Departamento de Matemática, Faculdade de Ciências e Tecnologia**
- **Gabinete de Comunicação e Protoclo**

Um agradecimento especial é devido aos colegas da Direção da Sociedade Portuguesa de Estatística que colaboraram diretamente na realização deste congresso e aos colegas das Comissões Científica e Organizadora do Congresso.

Direção SPE

- **Isabel Pereira**, Universidade de Aveiro
- **Maria Eduarda Silva**, Universidade do Porto
- **Patrícia de Zea Bermudez**, Universidade de Lisboa

Comissão Científica

- **Carlos Daniel Paulino**, Universidade de Lisboa
- **Clara Cordeiro**, Universidade do Algarve
- **Conceição Ribeiro**, Universidade do Algarve
- **Maria Antónia Turkman**, Universidade de Lisboa
- **Maria Eduarda Silva**, Universidade do Porto
- **Maria Manuela Neves**, Universidade de Lisboa

Comissão Organizadora

- **Clara Cordeiro**, CEAUL e Universidade do Algarve
- **Conceição Ribeiro**, CEAUL e Universidade do Algarve
- **Carlos Sousa**, CEPAC e Universidade do Algarve
- **Mª Helena Gonçalves**, CEAUL e Universidade do Algarve
- **Nelson Antunes**, CEMAT e Universidade do Algarve

Índice

Análise de Sobrevivência e Valores Extremos em R

Ana Maria Abreu
FCEE, Universidade da Madeira e CIMA, *abreu@staff.uma.pt*

Délia Gouveia-Reis
FCEE, Universidade da Madeira e CEAUL, *delia@uma.pt*

Palavras–chave: Análise de sobrevivência, *software* R, valores extremos

Resumo: O *software* R é uma ferramenta extremamente útil para a investigação estatística. No entanto, a proliferação de bibliotecas (na ordem dos milhares) dificulta o rápido e eficiente acesso a todas as possibilidades em cada uma das áreas desta ciência. Uma forma de limitar esta procura é aceder à *task view* correspondente, se existir. Pelos motivos descritos, neste trabalho procura-se compilar informação relevante nas áreas de análise de sobrevivência e de valores extremos, de modo a minimizar as dificuldades referidas. A abordagem na análise de sobrevivência, que possui *task view*, será sobretudo através de exemplos. Nos valores extremos será dada uma visão geral do que existe, uma vez que nesta área não há *task view*.

1 Introdução

O R é uma linguagem que surge pela criação da *R Foundation for Statistical Computing* [7], com o objetivo de fornecer uma ferramenta gratuita e de utilização livre, para o tratamento e análise de dados e para a construção de gráficos. Em 1993, Robert Gentleman e Ross Ihaka, na Universidade de Auckland, deram origem à linguagem R e tornaram-na *open source* em 1995.
O R é uma ferramenta bastante abrangente, com boas capacidades ao nível da programação e um conjunto bastante vasto (e em cons-

tante crescimento) de bibliotecas (livrarias) que acrescentam inúmeras potencialidades à já poderosa versão base do R. O *download* do R é gratuito e pode ser feito a partir da página principal do *R Project for Statistical Computing* em http://www.r-project.org/ ou do *Comprehensive R Archive Network* (CRAN) em http://cran.r-project.org/. Uma biblioteca muito utilizada é o *R Commander* (abreviadamente Rcmdr, desenvolvido em 2003, por John Fox) pois possui um interface gráfico que torna a interação com o utilizador muito mais amigável do que na consola do R. Além de possuir menus, permite a escrita de código e engloba as restantes funcionalidades existentes no R original. Existem ainda os *plugins* do *R Commander* que adicionam funcionalidades aos menus.

Uma forma de tornar eficiente a utilização do R, consiste em aceder à *task view* correspondente à área em estudo, se existir, pois estas *task views* são excelentes guias para encontrar as bibliotecas e funções adequadas ao propósito do investigador. Atualmente existem 33 *task views*, abrangendo áreas tão diversas como inferência bayesiana, ensaios clínicos, genética, otimização e programação matemática, análise de sobrevivência, séries temporais, entre outras. Contudo, ainda há várias áreas para as quais não existe esta funcionalidade, como sejam, valores extremos, análise em componentes principais, modelos com equações estruturais ou controlo de qualidade. Assim, neste trabalho irá ser feita uma breve revisão das bibliotecas existentes para a análise de sobrevivência e para os valores extremos, procurando contribuir para um eficiente acesso às potencialidades do R nestas áreas. A abordagem na análise de sobrevivência ([5, 9]) será sobretudo através de exemplos. Nos valores extremos ([1, 3]) será dada uma visão geral do que existe, uma vez que nesta área não há *task view*.

2 Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística que possui *task view* no R (https://cran.r-project.org/web/views/Survival.html),

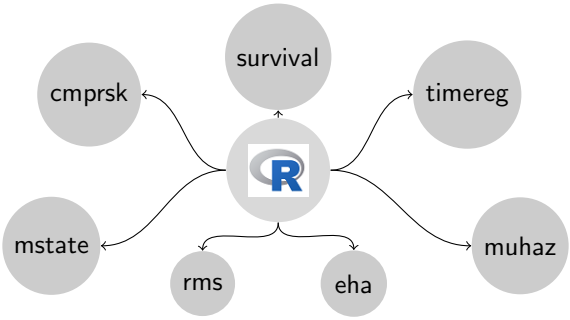


Figura 1: Bibliotecas *core*.

a qual se encontra organizada por temas e por ordem alfabética das bibliotecas. Importa notar desde logo que esta *task view* não esgota todas as funcionalidades que existem no R para esta área, mas cobre a maior parte. Allignol e Latouche (os responsáveis pela manutenção da *task view*) identificam sete bibliotecas *core* (Figura 1), sendo a survival, a rms e a eha as mais abrangentes.

Para além destes, há muitos outros (mais de cem), cujas particularidades seriam impossíveis de enumerar aqui de forma suficientemente abreviada. De qualquer modo, a já referida organização por temas permite uma escolha rápida da biblioteca apropriada para o objetivo pretendido como, por exemplo, a estimação da função de sobrevivência, a realização de testes ou a obtenção de modelos de regressão (bibliotecas survival, rms, eha e timereg) ou ainda a estimação da função de risco (bibliotecas rms, eha e muhaz). As bibliotecas cmprsk e mstate são mais específicas; referem-se, respetivamente, a modelos de riscos competitivos e a modelos multiestado.

O que se segue são pequenos exemplos de representações gráficas usuais (função de sobrevivência e função de risco) com alguns pormenores extra. A amostra aleatória, de dimensão 100, usada nos exemplos que se seguem, foi gerada da seguinte forma: os tempos de vida através da distribuição exponencial de parâmetro $\lambda = 1$, a

censura através da distribuição uniforme no intervalo 0.5 e 2, a idade através da distribuição normal de parâmetros $\mu = 60$ e $\sigma = 8$, e os estratos através de uma amostra aleatória de valores de 1 a 3. O respetivo código é o que se segue:

```
set.seed(123)
stime < − rexp(100) * 10
cens < − runif(100,.5,2) * 10
sevent < − as.numeric(stime <= cens)
stime < − pmin(stime, cens)
strat < − sample(1:3, 100, replace=TRUE)
idade < − rnorm(100,60,8)
dd < − data.frame("surv.time"=stime, "surv.event"=sevent, "strat"=strat,
"idade"=idade)
ddweights < − array(1, dim=nrow(dd))
```

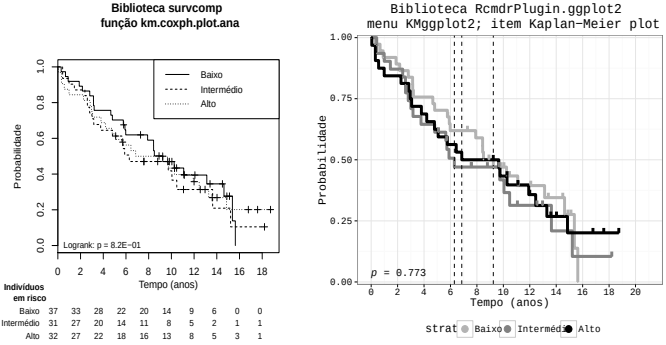


Figura 2: Estimativas de Kaplan-Meier da função de sobrevivência.

Exemplo 2.1 *Tendo presente que qualquer gráfico no R pode ser melhorado através da alteração do seu código base, a Figura 2 exibe dois gráficos da estimativa de Kaplan-Meier com uma alteração mínima do seu código (essencialmente a tradução para português).*

A particularidade da biblioteca survcomp *diz respeito não apenas ao facto de exibir o valor de prova resultante do teste logrank para a igualdade das funções de sobrevivência, mas sobretudo por mostrar o número de indivíduos em risco, para cada categoria, nos valores indicados na escala do tempo. Já no que diz respeito ao plugin* RcmdrPlugin.KMggplot2 *da biblioteca* RCommander *(que também indica o valor p), a principal inovação é a indicação do valor da mediana do tempo de vida através das retas verticais.*

Outro tipo de gráfico muito útil na análise de sobrevivência é o da função de risco pois, além de descrever o risco ao longo do tempo, ajuda na escolha da distribuição para a variável aleatória que representa o tempo de vida. O Exemplo 2.2 refere-se a duas representações desta função.

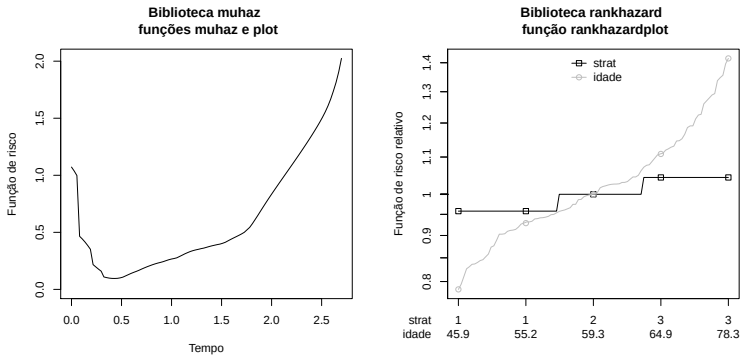


Figura 3: Funções de risco.

Exemplo 2.2 *A biblioteca* muhaz *permite que a representação da função de risco seja feita de um modo bastante simples (Figura 3), recorrendo apenas às funções* muhaz *e* plot*.*
Uma outra perspetiva interessante e inovadora da função de risco, é a que se obtém através da função rankhazardplot*, fornecida pela*

biblioteca rankhazard*, [4]. Esta abordagem indica o risco não ao longo do tempo mas ao longo dos valores das covariáveis presentes no modelo. Assim, num mesmo gráfico é possível observar o risco relativo para cada covariável e representar os valores das covariáveis no eixo horizontal (mínimo,* Q_1*, mediana,* Q_3 *e máximo). Concretamente, em relação às covariáveis apresentadas, verifica-se que a diferença entre os níveis 1 ("Baixo") e 2 ("Intermédio") é sensivelmente a mesma que entre os níveis 2 ("Intermédio") e 3 ("Alto") da covariável "strat" e que, em relação à covariável "idade", o risco diminui com a idade. Note-se que, quando a covariável é quantitativa, o valor de referência (correspondente ao indivíduo padrão) que é considerado por regra é o correspondente à mediana, como acontece com a covariável "idade".*

Por último, mas não menos importante, apresenta-se um outro gráfico de utilização frequente pois permite uma análise visual preliminar da proporcionalidade das funções de risco.

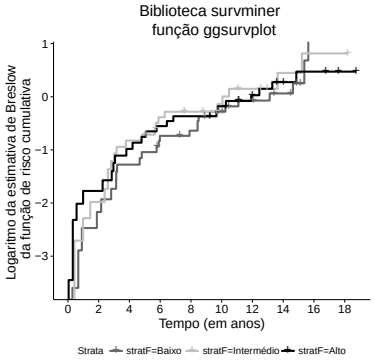


Figura 4: Logaritmo da estimativa de Breslow da função de risco cumulativa para os três estratos.

Exemplo 2.3 *Na Figura 4, o tempo é representado no eixo das abcissas e o logaritmo da estimativa de Breslow da função de risco*

cumulativa no eixo das ordenadas. Este gráfico é obtido duma forma simples através da biblioteca survminer*, utilizando a função* ggsurvplot*. O cruzamento das curvas dá indicação de não haver proporcionalidade das correspondentes funções de risco.*

As particularidades de natureza mais algébrica não foram aqui abordadas. No entanto, só a título de exemplo, refira-se que a função ConvertWeibull da biblioteca SurvRegCensCov faz a conversão da parametrização das estimativas dos parâmetros do modelo Weibull.

3 Valores Extremos

A crescente aplicabilidade da teoria dos valores extremos já era bem visível em 2006, ano no qual surgiu a elaboração de um estado da arte [10] sobre o *software* utilizado no estudo de valores extremos. Os seus autores, Stephenson e Gilleland, além de outros *softwares*, indicam algumas bibliotecas do R tais como ismev, evir, evd e evdBayes. Além disso, apresentam também a biblioteca extRemes como sendo essencialmente um interface gráfico do ismev e referem que muitas das funções da biblioteca fExtremes se baseiam em funções das bibliotecas ismev, evir e evd. Mais recentemente, Gilleland *et al.* [2] direcionam essa escolha para o R, pelo facto de ser o *software* que continha, em 2013, a maior variedade de metodologias na área de valores extremos. De entre estas metodologias, a dos máximos anuais foi a escolhida no estudo efetuado por Penalva *et al.* [6], para exemplificar uma análise de valores extremos no R. Nesse estudo, as autoras mencionam as bibliotecas evir, fExtremes e evdBayes e realizam uma descrição das bibliotecas ismev e evd. Até ao momento não existe qualquer *task view* exclusivamente dedicada à teoria dos valores extremos que facilite o acesso às bibliotecas e funcionalidades apropriadas, mas as *task views* Bayesian, Distributions, Environmetrics, Finance, Spatial incluem pelo menos uma das bibliotecas indicadas na Figura 5.

A existência de algumas funções relativas à teoria dos valores extremos motivou a referência de outras bibliotecas tanto nos esta-

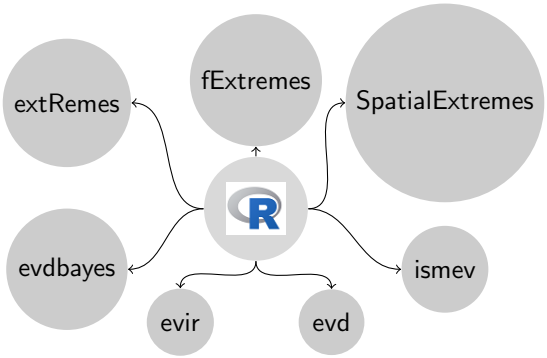


Figura 5: Bibliotecas em *task views*.

dos da arte já referidos como no *website* mantido por Eric Gilleland (http://www.ral.ucar.edu/ ericg/softextreme.php), embora essas bibliotecas não sejam exclusivamente dedicadas à aplicação de metodologias nesta área. Alguns exemplos são as bibliotecas copula, fitdistrplus e RandomFields. A biblioteca fitdistrplus estabelece a ligação mais visível entre as áreas da análise de sobrevivência e dos valores extremos. De facto, ao carregar a biblioteca fitdistrplus, automaticamente também é carregada a biblioteca survival. Por outro lado existem também bibliotecas no R destinadas a conteúdos muito específicos na área de valores extremos tais como as bibliotecas bgeva e spatial.gev.bma. A primeira disponibiliza uma função para modelos de regressão para extremos bivariados enquanto que a segunda permite ajustar um modelo espacial hierárquico a valores extremos. Apesar de mais gerais, as bibliotecas evt0, evmix, MCMC4Extremes e Renext estão focalizadas em certas metodologias específicas da área de valores extremos. A biblioteca evt0 é a única que aborda a metodologia PORT (*Peaks Over Random Threshold*) de entre as bibliotecas do R (pelo menos do conhecimento das autoras). Esta biblioteca, não indicada no *website* mantido por Eric Gilleland, é um produto da escola portuguesa de valores extremos que permite determinar o

índice de valores extremos γ por meio do estimador MOP (média de ordem p). Além disso, esta biblioteca (que requer a biblioteca evd) permite também obter as estimativas para γ pelos estimadores dos momentos, dos momentos mistos e generalizado de Hill. A biblioteca evmix fornece funções para a modelação mista de valores extremos, para a estimação do limiar u e para estimadores de densidade pelo método do núcleo. Apesar de esta biblioteca não requerer qualquer das bibliotecas mencionadas, os seus criadores indicam que existe uma razoável consistência com as funções base da biblioteca evd. O Exemplo 3.1 refere-se a uma dessas funções, cuja aplicação originou os dois gráficos da Figura 6.

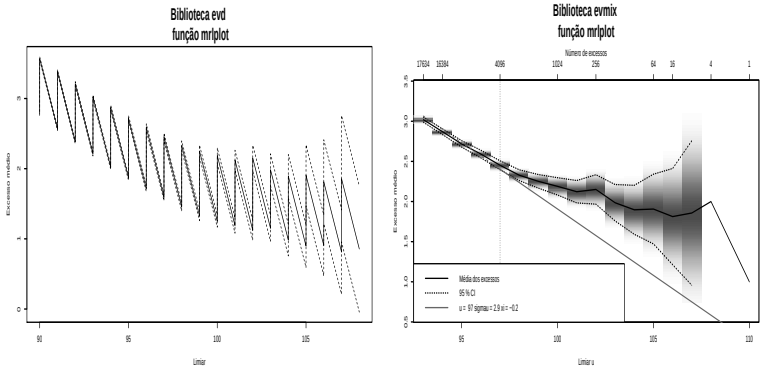


Figura 6: Gráficos de vida residual média.

Exemplo 3.1 *A Figura 6 exibe dois gráficos de vida residual média com uma alteração mínima do seu código (tradução para português e alteração da escala de cores). Os dados utilizados, já analisados por outros autores no contexto da metodologia POT [8], correspondem a idades de mulheres nascidas por volta do ano 1900 que morreram no ano de 1993.*

A particularidade da função mrlplot *da biblioteca* evmix *diz respeito não apenas ao facto de exibir um eixo superior com o número de*

excessos, mas sobretudo por mostrar um valor de referência para o limiar u. Além da indicação gráfica (linha vertical) e numérica de um limiar u, as estimativas de máxima verosimilhança dos parâmetros de escala e forma da correspondente distribuição generalizada de Pareto são também apresentadas. Esta informação pode ainda ser apresentada para três valores a considerar para o limiar u, facilitaando a interpretação do gráfico e a comparação de estimativas. A título exemplificativo, na Figura 7 é também indicado o valor sugerido para o limiar u pelos autores Reiss e Thomas [8].

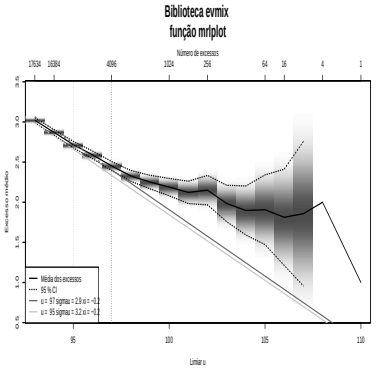


Figura 7: Gráfico de vida residual média com $u = 95$ e $u = 97$.

A biblioteca MCMC4Extremes partilha a metodologia Bayesiana com a biblioteca menos recente evdBayes mas requer a biblioteca evir em vez da biblioteca evd. A biblioteca Renext apareceu em 2010 e inclui a implementação de funções relativas ao denominado *méthode du renouvellement*. Esta abordagem surge como uma generalização da abordagem clássica POT (*Peaks Over Threshold*) ao permitir que os excessos em relação a um dado limiar u sigam uma distribuição de probabilidade diferente da distribuição de Pareto. Para esta biblioteca, que também requer a biblioteca evd, existe um interface gráfico denominado de RenextGUI.

A biblioteca extRemes, que inicialmente foi interface gráfico da biblioteca ismev, tornou-se numa biblioteca de valores extremos por si só, tendo sido criado um interface gráfico para algumas das suas funções, denominado de in2extRemes, cujo tutorial pode ser encontrado no link http://www.ral.ucar.edu/staff/ericg/extRemes. Outra biblioteca é a intitulada *texmex* (*Statistical modelling of extreme values*), que pode ser utilizada na modelação quer de máximos quer de excessos. Além disso, esta biblioteca tem a particularidade de ter funcionalidades específicas da abordagem bayesiana e da análise multivariada de valores extremos. Muito recentemente (fevereiro de 2016), surgiu uma nova biblioteca intitulada *Multivariate Extreme Value Distributions* (mev), a qual é inteiramente direcionada para o estudo dos valores extremos multivariados. Além da implementação de métodos de seleção do limiar, esta biblioteca permite ainda a simulação de processos max-estáveis. A biblioteca eva (*Extreme Value Analysis with Goodness-of-Fit Testing*), que surgiu também recentemente (dezembro de 2015), possui a particularidade de incluir testes de ajustamento para a escolha tanto do limiar u na metodologia de excessos de nível, como do número de observações k na metodologia das maiores observações.
Além das bibliotecas já mencionadas, existem outras mais gerais que englobam várias áreas da estatística e probabilidades, entre as quais a análise dos valores extremos (lmomco e VGAM, por exemplo). Existem ainda outras mais focalizadas numa área de aplicação em que recorrem a esta teoria como sejam, por exemplo, as bibliotecas actuar e QRM.

4 Conclusão

O R é tão dinâmico que qualquer trabalho sobre ele é inevitavelmente incompleto e um pouco desatualizado. Exemplo disso é o facto de à data do início da escrita deste artigo (junho de 2015) haver 6730 bibliotecas e atualmente (setembro de 2016) já haver 9202. Mas precisamente por essa razão, entende-se que uma sistematização

periódica por áreas pode se revelar útil, de modo a que as inúmeras potencialidades deste *software* possam ser plenamente aproveitadas. Embora a biblioteca Rcmdr não conste na *task view* para a análise de sobrevivência, recomenda-se o seu uso pois assim é possível usar os *plugins* relativos a esta área que lhe estão associados: RcmdrPlugin.survival, RcmdrPlugin.KMggplot2, RcmdrPlugin.EZR e RcmdrPlugin.NMBU. Esta abordagem torna a interação mais simples e mantém a mais valia da escrita do código, realidade que é válida para as restantes áreas. A existência da *task view* é uma grande vantagem pois rapidamente se identificam as bibliotecas existentes para áreas específicas, como sejam, por exemplo, modelos multiestado, sobrevivência relativa, modelos de efeitos aleatórios, modelos bayesianos, simulação, entre outros. Além disso, todos as bibliotecas podem ser instaladas simultaneamente, em vez de uma a uma, bastando para tal instalar a biblioteca ctv e proceder de acordo com as instruções existentes em https://cran.r-project.org/web/views/. Neste breve trabalho, tentou-se mostrar algumas das peculiaridades que distinguem este *software* de outros mais comerciais, através das inovações que apresenta nos gráficos mais usuais desta área. No entanto, as particularidades de natureza mais algébrica não foram abordadas.

Neste trabalho, fez-se também uma revisão das bibliotecas do R que podem ser aplicadas na análise de valores extremos. Procurou-se assim organizar uma coletânea de informação sobre estas bibliotecas, tendo como linhas de orientação a sua abrangência relativamente às metodologias da área, as suas interligações e as suas particularidades. Nesta área recomenda-se igualmente a biblioteca Rcmdr em detrimento dos interfaces gráficos in2extRemes e RenextGUI pois permite a utilização simultânea de uma ou mais das bibliotecas mencionadas num mesmo ambiente amigável. Se em 2013 o R já era o *software* que continha a maior variedade de metodologias na área de valores extremos, atualmente esse facto é ainda mais evidente dado o surgimento de novas bibliotecas, bem como o aperfeiçoamento das já existentes. Seria pois bastante útil reunir e organizar as bibliotecas do R sobre análise de valores extremos numa *task view*, segundo

tópicos que permitissem um fácil acesso e manuseamento da grande quantidade de funcionalidades existentes. Essa é a nossa proposta de trabalho futuro.
Em conclusão, trabalhar com o R é estar preparado para uma constante descoberta, acompanhada por muitos momentos de satisfação intercalados por alguns de frustração.

Agradecimentos

Este trabalho é parcialmente financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/MAT/04674/2013 (CIMA) e UID/MAT/00006/2013 (CEAUL).

Referências

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

[2] Gilleland, E., Ribatet, M., Stephenson, A. (2013). A software review extreme value analysis. *Extremes* 16, 103–119.

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

[4] Karvanen, J., Harrell Jr., F.E. (2009). Visualizing covariates in proportional hazards model. *Statistics in Medicine* 28, 1957–1966.

[5] Klein, J.P., Moeschberger, M.L. (1998). *Survival Analysis. Techniques for Censored and Truncated Data*, 2^a impressão. Springer, New York.

[6] Penalva, H., Neves, M., Nunes, S. (2013). Topics in data analysis using R in extreme value theory. *Metodološki zvezki* 10, 17–29.

[7] R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/

[8] Reiss, R.D., Thomas, M. (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser, Basel.

[9] Rocha, C., Papoila, A.L. (2009). *Análise de Sobrevivência*. Edições SPE, Lisboa.

[10] Stephenson, A., Gilleland, E. (2006). Software for the analysis of extreme events: The current state and future directions. *Extremes* 8, 87–109.

Erro de Tipo I no teste de Friedman e nos testes de comparação múltipla

Anabela Afonso
Departamento de Matemática, ECT, Centro de Investigação em Matemática e Aplicações, IIFA, Universidade de Évora, *aafonso@uevora.pt*

Dulce G. Pereira
Departamento de Matemática, ECT, Centro de Investigação em Matemática e Aplicações, IIFA, Universidade de Évora, *dgsp@uevora.pt*

Palavras–chave: ANOVA, medidas repetidas, testes não paramétricos

Resumo: O teste de Friedman é a alternativa não paramétrica à ANOVA de medidas repetidas. Os testes de comparação múltipla são aplicados após a rejeição da hipótese nula do teste Friedman. Neste trabalho, realizamos um estudo de simulação para analisar a probabilidade de erro de Tipo I tanto no teste de Friedman como também nos testes de comparação múltipla mais usuais. Consideraram-se as distribuições discretas vulgarmente utilizadas para modelar dados de contagens, nas áreas da Biologia e da Ecologia. No teste de Friedman a taxa de erro de Tipo I empírica é menor quando se considera a estatística de teste com aproximação ao qui-quadrado. Nos testes de comparação múltipla, a probabilidade de erro de Tipo I aumenta com o aumento do número de blocos, e no teste LSD de Fisher também com o aumento do número de tratamentos. O teste dos Sinais é o mais conservativo e o teste LSD de Fisher o mais liberal.

1 Introdução

O teste de Friedman substitui a ANOVA de um delineamento experimental em blocos casualizados (medidas repetidas para cada bloco em cada nível do fator ou do tratamento). É aplicado quando os

pressupostos de normalidade multivariada e esfericidade são violados. Para o cálculo da estatística de teste são utilizadas as ordens dos dados em vez dos valores observados. Este teste não é tão potente quanto a ANOVA, quando os pressupostos da ANOVA não são violados [1].
No teste de Friedman pretende-se testar se as distribuições das observações da variável dependente nos diversos tratamentos são idênticas contra a hipótese alternativa de que pelo menos uma distribuição difere na localização. A estatística de teste associada a este teste segue aproximadamente uma distribuição Qui-quadrado. No entanto, está demostrado que esta aproximação à distribuição do Qui-quadrado é demasiado conservadora, i.e., com maior probabilidade para cometer o erro do Tipo II. Assim, foi proposta uma estatística de teste alternativa com uma distribuição aproximada F [5].
Perante a rejeição da igualdade das distribuições dos tratamentos, podem aplicar-se testes de comparação múltipla para averiguar quais os tratamentos que diferem entre si. Existe uma grande variedade de testes que podem ser aplicados, não existindo um teste que seja melhor que todos os outros em todos os contextos. Na análise da potência destes testes, Pereira *et al.* [6] concluíram que o teste dos Sinais é muito conservador e os testes LSD de Fisher e HSD de Tukey, calculados com base nas ordens, são os mais liberais.
Este trabalho tem como objetivo estudar a taxa de erro de Tipo I do teste de Friedman e dos testes de comparação múltipla mais usuais quando dispomos de dados de contagens.

2 Métodos

Os dados consistem em n vetores aleatórios $(X_{i1}, X_{i2}, ..., X_{iK})$, $i = 1$, $..., n$, mutuamente independentes designados por blocos (Tabela 1). Para a realização do teste Friedman, e posteriores testes de comparação múltipla, em cada uma das linhas substituem-se os dados originais pelas respetivas ordens quando se considera a ordenação por ordem crescente (Tabela 2). Em caso de empate, usa-se a média

Tabela 1: Dados originais

	Tratamento		
Bloco	1	...	K
1	X_{11}	...	X_{1K}
2	X_{21}	...	X_{2K}
...	...	...	...
n	X_{n1}	...	X_{nK}

Tabela 2: Ordens dos dados

	Tratamento			
Bloco	1	...	K	Total
1	R_{11}	...	R_{1K}	$R_{1.}$
2	R_{21}	...	R_{2K}	$R_{2.}$
...	...	...	...	...
n	R_{n1}	...	R_{nK}	$R_{n.}$
Total	$R_{.1}$	...	$R_{.K}$	$R_{..}$

das ordens. Seja R_{ij} a ordem atribuída a X_{ij} dentro do bloco i, com $1 \leq r_{ij} \leq K$, $i = 1, ..., n$, $j = 1, ..., K$.

2.1 Teste de Friedman

Assumindo que as distribuições das observações da variável dependente nos vários tratamentos são as mesmas, ou que estas distribuições são simétricas com a mesma média embora com variâncias diferentes, as hipóteses em estudo são:

H_0: As distribuições das observações da variável dependente nos diversos tratamentos são idênticas *vs.*

H_1: Pelo menos uma distribuição difere das restantes na localização.

Para este teste foram propostas duas estatísticas de teste:

1. A inicialmente sugerida por Friedman [2]:

$$Fr_1 = \frac{4(K-1)\sum_{j=1}^{K}\left(R_{.j} - \frac{n(K+1)}{2}\right)^2}{4\sum_{i=1}^{n}\sum_{j=1}^{K}R_{ij}^2 - nK(K+1)}.$$

   Para valores de n e K pequenos, os pontos críticos estão tabulados (ver por ex. [9]). Para um dado nível de significância

α, a hipótese nula é rejeitada quando o valor observado para a estatística de teste for superior ou igual ao ponto crítico tabulado. Para valores grandes de K e/ou n a estatística de teste segue aproximadamente uma distribuição qui-quadrado com $(K-1)$ graus de liberdade.

2. Como a estatística de teste anterior é demasiado conservadora, foi proposta uma alternativa que consiste na estatística da ANOVA com dois fatores calculada com base nas ordens R_{ij} [5]:

$$Fr_2 = \frac{(n-1)Fr_1}{K(n-1) - Fr_1},$$

que segue aproximadamente uma distribuição F com $(K-1)$ e $(K-1)(n-1)$ graus de liberdade. A hipótese nula é rejeitada, ao nível de significância α, quando Fr_2 for superior ou igual ao quantil de probabilidade $(1-\alpha)$ da distribuição $F_{K-1;(K-1)(n-1)}$.

2.2 Testes de comparação múltipla

Perante a rejeição de H_0 no teste Friedman, muitas vezes interessa averiguar quais as distribuições que diferem entre si na localização. Para tal, realizam-se os testes de comparação múltipla nos quais se comparam todos os pares de tratamentos. Assim, as hipóteses a testar são:

H_0: As distribuições das observações da variável dependente nos tratamentos j e k são idênticas *vs.*

H_1: As duas distribuições diferem na localização,

com $j, k = 1, ..., K$ e $j \neq k$.
Os testes de comparação múltipla mais usuais são:

- *Teste de Bonferroni-Dunn* [9].
- *Teste LSD de Fisher* [1]: calculado com base nas ordens R_{ij}.

- *Teste de Wilcoxon-Nemenyi-McDonald-Thompson* [3]: equivalente ao teste HSD de Tukey calculado com base nas ordens.
- *Teste de Wilcoxon* [8]: com a correção de Bonferroni, ou Holm, ou Hochberg, ou Hommel [4].
- *Teste dos Sinais* [8]: com a correção de Bonferroni, ou Holm, ou Hochberg, ou Hommel.

3 Simulação

No estudo de simulação levado a cabo, consideraram-se modelos lineares aditivos da forma:

$$X_{ik} = \theta + \beta_i + \tau_k + \epsilon_{ik},$$

onde θ é a mediana global, β_i o efeito do bloco i, τ_k o efeito do tratamento k e ϵ_{ik} o efeito aleatório do bloco i e do tratamento k, com $i = 1, ..., n$ e $k = 1, ..., K$.
Foram consideradas várias combinações para n e K, nomeadamente $n = 3, 4, 5, 6, 10, 15, 20, 30$ blocos e $K = 3, 4, 5, 6, 10$ tratamentos, e diferentes cenários distribucionais, onde se considerou a mesma distribuição de probabilidade para os efeitos dos blocos e dos erros:

- Binomial: $\beta_i \frown B(10; 0{,}5)$, $\epsilon_{ik} \frown B(10; 0{,}5)$ sendo a distribuição dos tratamentos:
  - assimétrica positiva:
    $\tau_k \frown B(N_k; 0{,}2)$ com $N_k = 25, 50, 75, 100$;
  - simétrica:
    $\tau_k \frown B(N_k; 0{,}5)$ com $N_k = 10, 20, 30, 40$;
  - assimétrica negativa:
    $\tau_k \frown B(N_k; 0{,}8)$ com $N_k = 6, 12, 19, 25$.
- Binomial Negativa: $\beta_i \frown BN(10; 0{,}5)$, $\tau_k \frown BN(R_k; 0{,}5)$ e $\epsilon_{ik} \frown BN(10; 0{,}5)$, com $R_k = 5, 10, 15, 20$.

- Poisson: $\beta_i \frown P(5)$, $\tau_k \frown P(\lambda_k)$ e $\epsilon_{ik} \frown P(5)$, com $\lambda_k = 5, 10, 15, 20$.
- Uniforme: $\beta_i \frown U\{0, ..., 10\}$, $\tau_k \frown U\{0, ..., N_k\}$ e $\epsilon_{ik} \frown U\{0, ..., 10\}$, com $N_k = 10, 20, 30, 40$.

Nos cenários considerados os tratamentos têm o mesmo efeito, i.e., a hipótese nula do teste de Friedman é verdadeira bem como as hipóteses nulas dos testes de comparação múltipla.
Para cada uma das combinações de n, K e cenário distribucional foram realizadas $M = 1500$ replicações, tendo-se contabilizado para cada combinação o número total de réplicas com:

- teste de Friedman significativo, r_F;
- teste de comparação múltipla significativo, r_c;
- teste de Friedman e o teste comparação múltipla significativos, r_{Fc}.

A partir destas contagens, para cada uma das combinações, foram obtidas as taxas de erro de Tipo I empíricas para:

- o teste de Friedman: $\hat{\alpha}_F = r_F/M$,
- cada um dos testes de comparação múltipla: $\hat{\alpha}_c = r_c/M$.

Adicionalmente, foi calculada a proporção de vezes que cada um dos testes de comparação múltipla foi significativo, quando se rejeitou a hipótese nula do teste de Friedman: $r_{c|F} = r_{Fc}/r_F$.
Foi considerado um nível de significância de 5% para todos os testes e foi usado o programa R project [7].

4 Resultados

Em todos os cenários gerados obtiveram-se resultados similares com os vários tipos de assimetria e dispersão considerados. Nas Figuras 1, 2 e 3, apresenta-se apenas uma seleção desses resultados.

De um modo geral, a taxa de erro de Tipo I empírica do teste de Friedman obtido com o uso da estatística de teste proposta por Friedman (Fr_1) é inferior à alternativa com distribuição aproximada F (Fr_2), especialmente perante um número reduzido de blocos (Figura 1). Nestes casos a taxa de erro de Tipo I obtida com a estatística de teste Fr_2 é superior ao nível de significância α definido. À medida que aumenta o número de blocos e/ou tratamentos as taxas de erro de Tipo I empíricas das duas estatísticas de teste aproximam-se, e para um número elevado de blocos coincidem.

O teste dos Sinais nunca rejeitou a hipótese nula, sendo por isso o que apresentou menor probabilidade de erro de Tipo I empírica (Figura 2). Usualmente, o teste de Wilcoxon é o que apresenta a segunda menor taxa de erro de Tipo I; para um número reduzido de tratamentos a correção de Hommel é a que dá origem a maiores valores do erro, mas à medida que se aumenta o número de tratamentos o erro é idêntico em todas as correções. No entanto, perante um número reduzido de tratamentos e elevado de blocos, o teste de Bonferroni-Dunn tende a apresentar uma menor probabilidade de erro de Tipo I do que o teste de Wilcoxon com qualquer uma das correções. O teste de Wilcoxon-Nemenyi-McDonald-Thompson apresentou uma taxa de erro de Tipo I semelhante ao α previamente definido, embora por vezes tenha ligeiramente ultrapassado esse nível. O teste LSD de Fisher foi o teste que mostrou o pior desempenho, com probabilidade de erro de Tipo I empírica muito superior ao α definido e aumenta com o número de tratamentos. A salientar que com 10 tratamentos a taxa de erro de Tipo I empírica deste teste ultrapassa os 50%.

A Figura 3 ilustra o desempenho dos vários testes de comparação múltipla, após a rejeição da hipótese nula do teste de Friedman. Mais concretamente, permite comparar a proporção de vezes que se comete o erro de Tipo I com os testes de comparação múltipla, quando também se cometeu esse erro com o teste de Friedman. Perante os resultados obtidos, é possível ordenar os testes por ordem crescente de proporção de concordância na decisão errada dos tes-

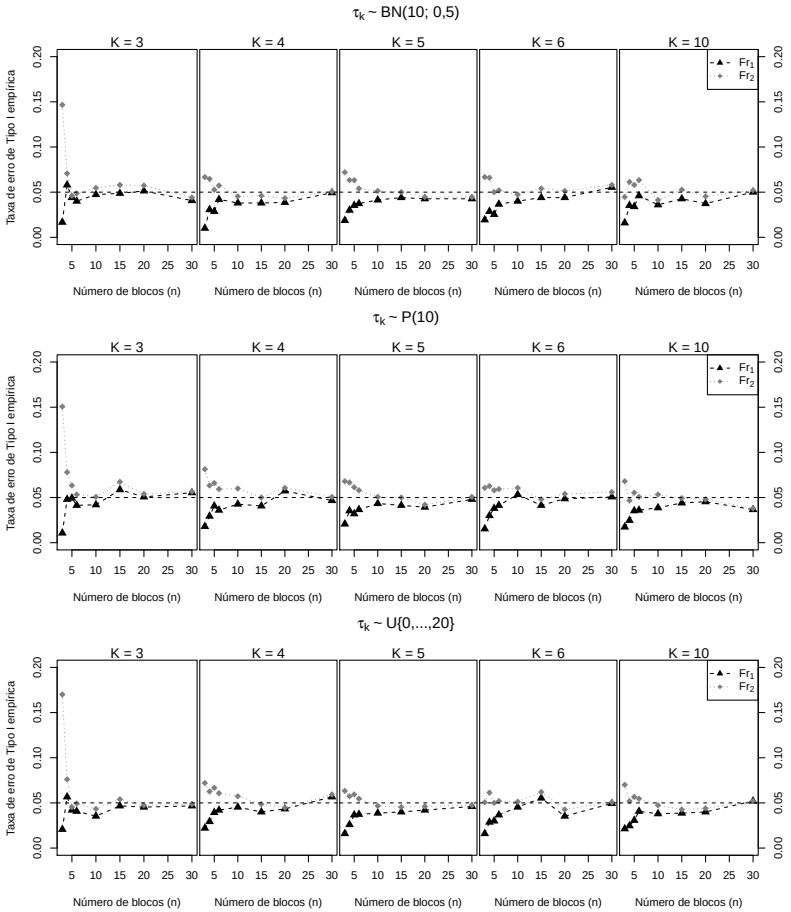


Figura 1: Probabilidade de erro de Tipo I empírica do teste de Friedman com as estatísticas de teste Fr_1 e Fr_2, considerando as distribuições Binomial Negativa, Poisson e Uniforme discreta, com $E(X_{.k}) = 10$, $k = 1, ..., K$. A linha horizontal tracejada representa o nível de significância de 5%.

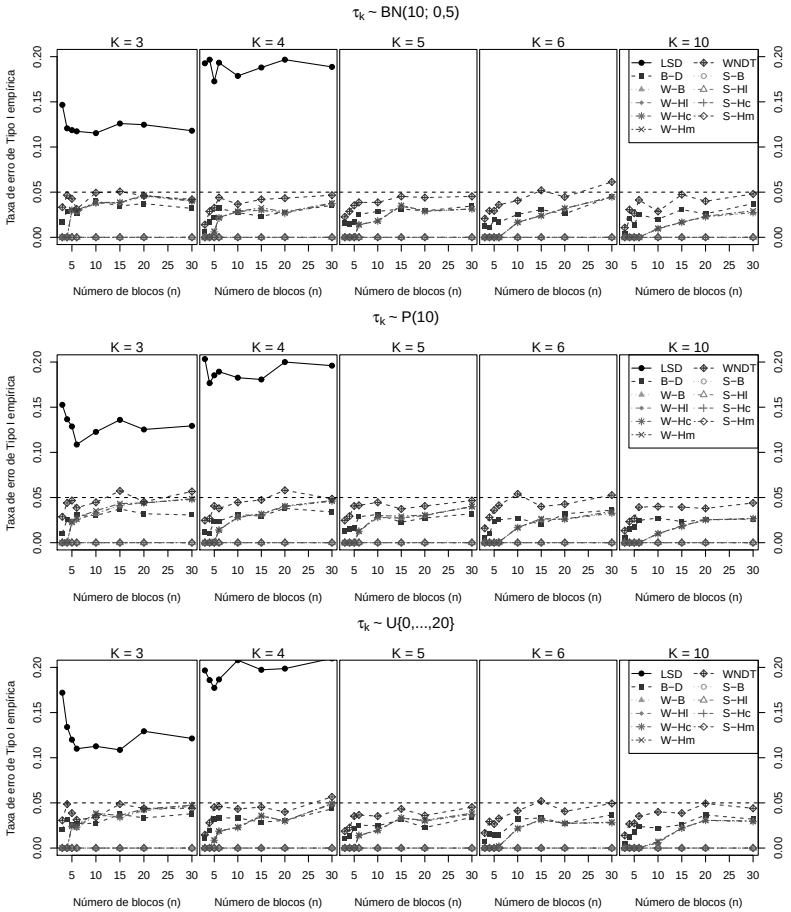


Figura 2: Probabilidade de erro de Tipo I empírica dos testes de comparação múltipla, considerando as distribuições Binomial Negativa, Poisson e Uniforme discreta, com $E(X_{.k}) = 10$, $k = 1, ..., K$. A linha horizontal tracejada representa o nível de significância de 5%. (Testes - LSD: LSD de Fisher; B-D: Bonferroni-Dunn; W-B, W-Hl, W-Hc e W-Hl: Wilcoxon com correção de Bonferroni, Holm, Hochberg e Hommel, respetivamente; WNDT: Wilcoxon-Nemenyi-McDonald-Thompson; S-B, S-Hl, S-Hc e S-Hl: Sinais com correção de Bonferroni, Holm, Hochberg e Hommel, respetivamente)

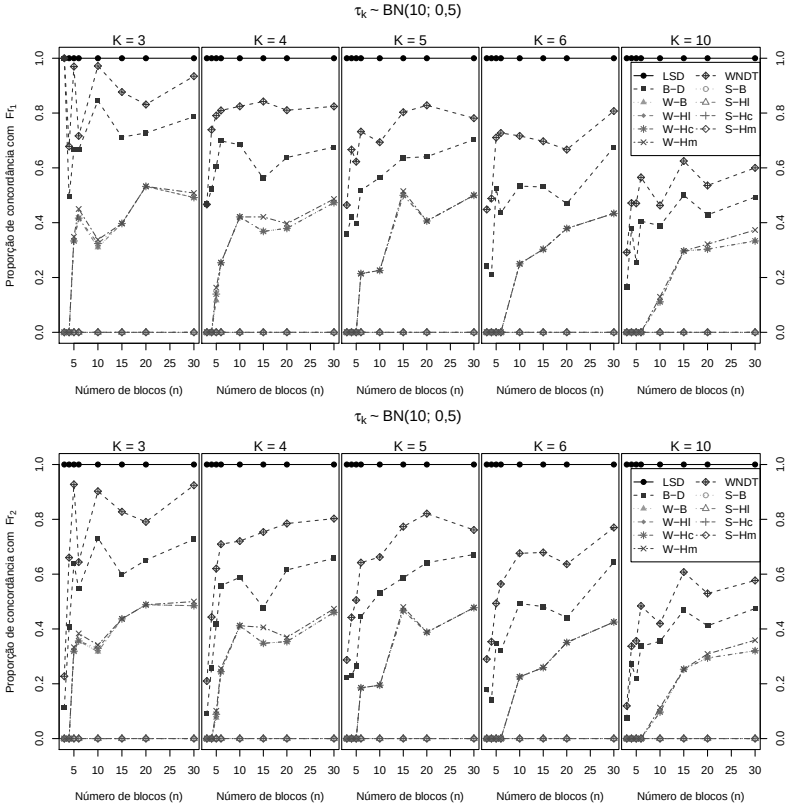


Figura 3: Proporção de vezes que cada um dos testes de comparação múltipla foi significativo, quando se rejeitou a hipótese nula do teste de Friedman, considerando a distribuição Binomial Negativa com $E(X_{.k}) = 10$, $k = 1, ..., K$.

tes de comparação múltipla com o teste de Friedman (para ambas as estatísticas de teste): 1) teste dos Sinais (todas as correções); 2) teste de Wilcoxon (todas as correções); 3) teste de Bonferroni-Dunn; 4) teste de Wilcoxon-Nemenyi-McDonald-Thompson; 5) teste LSD de Fisher. A observar que enquanto que na Figura 2 a comparação entre o desempenho dos testes de Wilcoxon e de Bonferroni-Dunn depende do número de blocos e tratamentos considerados, na Figura 3 o teste de Wilcoxon apresentou sempre um desempenho superior.

5 Conclusão

Na análise da taxa de erro de Tipo I empírica cometido pelo teste de Friedman e pelos testes de comparação múltipla, foram considerados vários cenários distribucionais que incluíram distribuições simétricas e assimétricas, bem como diferentes graus de dispersão, para avaliar se estas características tinham influência no desempenho dos testes. Os resultados obtidos foram similares para todas as distribuições e parâmetros considerados, pelo que a probabilidade de se cometer o erro de Tipo I não é afetada pelas características destas distribuições.
No teste de Friedman a taxa de erro de Tipo I é menor com a estatística de teste proposta por Friedman (Fr_1) do que com a aproximação à distribuição F (Fr_2). No entanto, à medida que aumenta o número de blocos as probabilidades de erro de Tipo I destas estatísticas aproximam-se, uma vez que o erro obtido com a estatística de teste Fr_1 aumenta e o obtido com a estatística de teste Fr_2 diminui.
Nos testes de comparação múltipla, a taxa de erro de Tipo I empírica aumenta com o aumento do número de blocos, e no teste LSD também com o aumento do número de tratamentos. A menor probabilidade de erro de Tipo I é observada no teste dos Sinais e a maior no teste LSD de Fisher.
Perante a decisão incorreta do teste de Friedman, exceto no teste LSD de Fisher, à medida que aumenta o número de tratamentos diminui a concordância dos testes de comparação múltipla com o teste de Friedman na tomada de decisão incorreta.

Em suma, neste estudo da taxa de erro de Tipo I, o teste dos Sinais é o mais conservador e o teste LSD de Fisher o mais liberal.

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto "UID/MAT/04674/2013 (CIMA)".

Referências

[1] Conover, W.J. (1999). *Pratical nonparametric statistics*, Third edition. John Wiley & Sons, New York.

[2] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32, 675-701

[3] Hollander, M. and Wolfe, D.A. (1999). *Nonparametric statistical methods*, Second edition. John Wiley & Sons, New York.

[4] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65-70.

[5] Iman, R.L., Davenport, J.M. (1980). Approximations of the critical region of the Friedman statistics. *Communications in Statistics - Theory and Methods* 9, 571–595.

[6] Pereira, D.G., Afonso, A., Medeiros, F.M. (2015). Overview of Friedman's test and post-hoc analysis. *Communications in Statistics - Simulation and Computation* 44, 2636–2653.

[7] R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*, Fourth edition. Chapman & Hall/CRC, Boca Raton.

[9] Siegel, S. and Castellan Jr., N.J. (1988). *Nonparametric statistics for the Behavioral Sciences*, Second edition. McGraw-Hill, New York.

Modelação Conjunta de Dados Longitudinais e de Sobrevivência de Cancro da Mama

Ana Borges
CIICESI, ESTGF-IPP, DMA-ECUM, Universidade do Minho, *aib@estgf.ipp.pt*

Inês Sousa
DMA-ECUM e CMAT, Universidade do Minho, *isousa@math.uminho.pt*

Palavras–chave: modelos conjuntos, sobrevivência, dados longitudinais, cancro da mama

1 Introdução

Doenças oncológicas são a segunda maior causa de morte em Portugal, e têm um grande impacto social nos pacientes e suas famílias [22]. Na Europa o cancro da mama é o tumor com maior incidência nas mulheres [2]. A publicação de 2003 por Pinheiro, Tyczynski, Bray, Amado, Matos e Parkin [22], refere que desde 1995 a mortalidade devido ao cancro de mama, tem vindo a diminuir em Portugal. Os autores argumentam que essa melhoria é uma consequência de um diagnóstico precoce e melhor qualidade de tratamento. O Plano Estratégico Nacional de Saúde refere, nas recomendações para o cancro de mama, a importância da "medicina baseada em evidências"para estabelecer diretrizes e especificar os protocolos para uma boa medicina na prática.
Atualmente existe um número reduzido de estudos que caracterizam a complexidade desse tipo de cancro na população portuguesa, que são conduzidos principalmente pelos registros de cancro de base populacional, como por exemplo o RORENO (Registo Oncológico

Regional do Norte). Embora de grande importância, uma vez que quantificam a incidência e prevalência da doença na população [10], as análises produzidas apenas incidem em dados agrupados e sem especificidade ao nível do indivíduo. Ainda, a informação que recolhem pode ser de alguma forma escassa em termos de especificidade do tumor ou mesmo, por exemplo, em termos de informação sobre a verdadeira causa da morte. Assim, os dados obtidos diretamente de uma Unidade de Senologia de um hospital permitem investigar estimativas de sobrevivência específica do cancro e o valor prognóstico de certos fatores clínicos que usualmente não podem ser recolhidos pelos registros populacionais [18]. Torna-se, dessa forma, extremamente importante o investimento contínuo em estudos estatísticos e epidemiológicos em doenças oncológicas para a compreensão da evolução da doença em Portugal, centrados se possível nos centros de senologia dos hospitais.

De facto, até agora, a publicação mais recente a que temos acesso é a publicação do RORENO que apresenta resultados de sobrevivência de pacientes diagnosticados com tumores malignos no período de 2007–2008, residentes até à data de diagnóstico na área de influência do RORENO, que inclui Braga, a área de interesse deste trabalho.

O presente trabalho propõe a utilização de métodos estatísticos no âmbito da bioestatística para estudar o cancro da mama, em pacientes da unidade de Senologia no Hospital de Braga. Sendo a bioestatística uma ciência que desenvolve metodologias estatísticas motivadas por questões e problemas científicos nas áreas de medicina, epidemiologia, saúde pública e biologia.

A análise aqui exposta tem como objetivo primordial desenvolver modelos conjuntos para dados longitudinais (medições repetidas ao longo do tempo de marcadores tumorais) e de sobrevivência (tempo até evento de interesse) de pacientes com cancro de mama, sendo a morte por cancro da mama o nosso evento de interesse.

Para tal, num primeiro momento, realizamos uma análise exploratória dos dados recolhidos, seguida por uma análise de sobrevivência independente a fim de compreender quais os possíveis fatores de risco para a morte por cancro de mama, para estes pacientes. Pos-

teriormente procedeu-se a uma análise longitudinal independente de dois marcadores tumorais: o Carcinoma Antigénio 15–3 (CA15–3) e o Antigénio carcino embrionário (CEA), para identificação de fatores de risco relacionados com o aumento dos seus valores. Neste trabalho apenas se apresentará os resultados relativos ao marcador CA15–3. Os resultados da análise longitudinal do marcador CEA foram anteriormente apresentados em Borges, Sousa e Castro [1].

Em muitos estudos longitudinais, as variáveis de interesse registadas em cada indivíduo incluem medidas repetidas em tempos pré-especificados e, também, o tempo até ocorrer um evento de particular interesse: por exemplo, morte, recidiva de um determinado sintoma ou saída do estudo [8]. Em dados médicos, tais como os apresentados neste estudo, onde a informação sobre sobrevivência é simultaneamente recolhida com dados de medições repetidas ao longo do tempo (valores de CA15–3 neste caso particular), é habitual detetar-se que o processo longitudinal está associado ao processo de sobrevivência dos pacientes [19].

Até ao final do século XX, dados desta natureza eram usualmente analisados considerando as componentes de sobrevivência e longitudinal em separado. No entanto, num contexto onde as observações longitudinais poderão estar correlacionadas com as de sobrevivência os modelos conjuntos dos processos longitudinais e de sobrevivência têm sido cada vez propostos, por serem capazes de capturar informações relativamente à associação desses dois processos. Mais, como McCrink, Marshall e Cairns [19] salientam, quando há uma associação entre os dois processos que não é devidamente tida em conta, introduz-se no modelo um enviesamento desnecessário, afetando os resultados e, portanto, as estimativas obtidas.

Como tal, sendo um aumento abrupto dos valores do marcador tumoral CA15-3, acima de um determinado valor de referência, um sinal de alarme para uma possível recidiva do tumor (que poderá estar relacionada com a morte iminente do paciente), o processo de sobrevivência e processo longitudinal de cada marcador tumoral estão de certa forma associados. Dessa forma, torna-se pertinente a construção de um modelo conjunto para estudar a associação entre

a progressão longitudinal dos valores do marcador tumoral CA15–3 e a sobrevivência dos pacientes.
No trabalho apresentado, começamos por descrever sucintamente o conjunto de dados recolhidos dos registos do Hospital de Braga. Posteriormente clarificamos a metodologia estatística implementada na análise de sobrevivência, na análise longitudinal e na análise conjunta. Em seguida, expomos os principais resultados terminando com uma seção de discussão, onde conjuntamente se sugere os trabalhos futuros que poderão ser implementados.
A análise apresentada foi realizada com o suporte do open source software estatístico R [23].

2 Base de dados de cancro da mama do Hospital de Braga

O Hospital de Braga está situado na cidade de Braga, localizada a norte de Portugal. Atualmente serve uma área direta de aproximadamente 275 000 utentes. A sua área de influência inclui os distritos de Braga e Viana do Castelo e funciona como unidade hospitalar de primeira linha para os municípios de Braga, Braga, Póvoa de Varzim, Terras de Bouro, Vieira do Minho e Vila Verde e como unidade hospitalar de segunda linha para a área restante. Sendo um hospital central, que abrange uma população, de acordo com o censo de 2011, de 1 081 641 habitantes. A população feminina com mais de 15 anos dos municípios de primeira linha é de 128 859. Em 2008, foi criada uma unidade de Senologia do Hospital de Braga.
Os dados foram recolhidos diretamente dos registos médicos de cada paciente, listados no sistema informático do hospital de Braga - *Glintt HS*. Teve-se, portanto, acesso a todo o historial clínico de cada paciente: um vasto conjunto de informação, tais como informação pessoal, diagnóstico, pré-operatório, pós-operatório, reuniões de grupo; acompanhamento (*follow-up*) e exames médicos. A autorização para a recolha e uso de dados Senologia foi aprovado pelo Comité de Ética do Hospital de Braga.

Foi reunida informação sobre 596 pacientes, onde 56 pacientes foram excluídos, uma vez que corresponderam a, pelo menos, um dos seguintes critérios de exclusão: (i) sem informação de diagnóstico, tratamento ou de acompanhamento; (ii) género Masculino e (iii) neoplasia benigna da mama.
Portanto, foram selecionados 540 pacientes para a presente análise. Como 19 pacientes apresentaram cancro da mama bilateral que, segundo sugestão do médico consultor do presente estudo, foram tratados como casos independentes, traduziu-se num número total de 559 casos analisados de pacientes do sexo feminino com diagnóstico de tumor maligno no período de 2008 até 2012. No entanto, registaram-se casos diagnosticados antes de 2008, mas todos vivos a 2008 e em acompanhamento. O número total de mortes é de 74, no entanto, o número total de mortes por cancro de mama é de apenas 55.
A partir das informações recolhidas dos inúmeros relatórios médicos foi possível reunir mais de 50 variáveis. As variáveis recolhidas agrupam-se em duas categorias: (i) variáveis explicativas a nível do paciente, que são um grupo de características demográficas, fatores prognósticos e etiológicos [24][28], como por exemplo: idade, menopausa, idade à primeira gravidez a termo, história familiar de cancro de mama, etc.; e (ii) as variáveis explicativas a nível do tumor, que incluem as características do tumor, alguns deles fatores prognósticos importantes já relatado na literatura [11] [3], tais como: classificação TNM, estadiamento, tipo histológico, expressão dos recetores hormonais, invasão vascular ou linfático, valores de marcadores tumorais CEA e CA15–3, entre outros.

3 Metodologia

3.1 Análise de sobrevivência

Quando a variável de resposta de interesse é o tempo desde o diagnóstico do tumor até morte, como na presente análise, deverão ser

utilizados métodos estatísticos de sobrevivência [5], que modelam o risco de morte em cada momento. O modelo de riscos proporcionais de Cox (CPHM) é o modelo comumente utilizado em análise de sobrevivência. No entanto, existem algumas restrições à aplicação deste tipo de modelos ou seja, a suposição de risco proporcional e não à não formulação da função de sobrevivência cumulativa de baseline, que pode ser de interesse médico.
Assim sendo, recorremos a modelos paramétricos flexíveis de sobrevivência, em particular, ao modelo paramétrico flexível de Royston-Parmar (FRPM) [25], para estimar as razões de risco (RR) ao longo do tempo desde o diagnóstico considerando um conjunto de covariáveis estatisticamente significativas. E, também, recorrendo ao conhecido CPHM, apenas com o propósito de comparar as estimativas.
No caso particular desta análise considerámos como evento de interesse a morte por cancro da mama considerando como tempo de referência o tempo, em meses, desde o diagnóstico de tumor maligno até morte por cancro da mama, ou até ao final do estudo onde considerada a data de 30/11/2014.
Para estimar o risco relativo de morte por cancro da mama, procedeu-se inicialmente com o cálculo das estimativas não paramétricas de Kaplan-Meier, para cada variável, estratificadas por categoria, comparando-se estas através da sua representação gráfica.
A estimativa de Kaplan-Meier [16] é uma estimativa não paramétrica de máxima verosimilhança (MLE) da função de sobrevivência, $S(t)$, amplamente utilizada em análises deste tipo. Essa estimativa é uma função em degrau com saltos nos tempos observados do evento (morte por cancro da mama), t_i. A estimativa de Kaplan-Meier da função de sobrevivência é dada por:

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t_1 \leq t \end{cases}, \tag{1}$$

onde $0 < t_1 < t_2 < ... < t_D$, d_i representa o número de indivíduos que tiveram o evento no tempo t_i, e o valor Y_i representa o número de indivíduos em risco no tempo t_i (ou seja, indivíduos que morreram

no momento t_i ou mais tarde).
Para avaliar a significância estatística das diferenças entre as curvas de sobrevivência de cada categoria de variável, fizemos uso da família G^ρ de testes de Harrington e Fleming [13], com pesos em cada morte de $S(t)^\rho$, implementados no software estatístico utilizado. Como se considerou $\rho = 0$, fez-se dessa forma uso do log-rank ou teste de Mantel-Haenszel, onde a hipótese nula a ser testada é: não há diferença entre as (verdadeiras) curvas de sobrevivência.
Considerando as variáveis estatisticamente significativas, obtidas no testes de log-rank, ajustaram-se vários modelos de sobrevivência com múltiplas covariáveis para estimar o efeito conjunto de diversas variáveis independentes na sobrevivência dos pacientes.
O CPHM define a probabilidade de sobrevivência, com base no pressuposto de riscos proporcionais, como uma função do tempo t, para um vetor de covariáveis x_i, da seguinte forma:

$$S(t \mid x_i) = [S_0(t)]^{\exp(x_i^{'}\beta)}, \tag{2}$$

onde $S_0(t)$ é a função de sobrevivência de baseline e β é o vetor de coeficientes a estimar.
A função de risco respetiva é dada por: $h(t \mid x_i) = h_0(t)\exp(x_i^{'}\beta)$, onde $h_0(t)$ é a função de risco de baseline.
A particularidade deste modelo, que pode ser visto como uma das suas vantagens, é que o cálculo dos coeficientes não requer a formulação da função de sobrevivência cumulativa de baseline, uma vez que é absorvida quando os coeficientes são estimados pelo método de log-verosimilhança parcial. No entanto, como Royston e Parmar [25] referem muitas vezes é de interesse, em estudos médicos, a estimativa da função de risco de baseline, uma vez que está diretamente relacionada com o percurso temporal de uma doença. Por isso, é de interesse o uso de um modelo mais flexível, em que a visualização da função de risco de baseline é mais simples.
Optamos, por isso, por trabalhar com a abordagem proposta por Royston e Parmar [25], onde modelam o logaritmo da função de risco cumulativa de baseline como uma função de spline cúbica "natural"de

tempo logarítmico. O FRPM vem da família de funções que têm por base a transformação da função de sobrevivência por uma função de ligação $g(.)$:

$$g\left[S(t \mid x_i)\right] = g\left[S_0(t)\right] + x_i\beta. \tag{3}$$

Como estamos interessados em estimar as RR, e como Royston e Parmar [25] sugerem, usamos splines cúbicos para modelar $g[S_0(t)]$, dentro da família Aranda-Ordaz de funções de ligação:

$$g(x;\theta) = \log\left(\frac{x^{-\theta}-1}{\theta}\right). \tag{4}$$

fazendo $\theta \to 0$, em vez de trabalhar com os valores mais gerais de θ, seguindo a sugestão de Royston e Parmar [25], que explicam que de outra forma a interpretação dos efeitos das covariáveis iria revelar-se obscura.
Assim, a transformação do modelo por ser escrita da seguinte forma:

$$g\left[S(t \mid x_i)\right] = \log\left[H(t \mid x_i)\right] = \eta_i = s(\log(t) \mid \gamma, k) + x_i\beta. \tag{5}$$

onde $H(t \mid x_i)$ é a função de risco cumulativa e s é um spline cúbico natural a atuar na escala logarítmica de tempo t, com um parâmetro γ ajustável e k nós. A escolha do número de nós pode ser feita recorrendo ao valor mínimo do critério de informação de Akaike (AIC).
A respetiva função de sobrevivência pode ser formulada da seguinte forma:

$$S(t \mid x_i) = \exp\left(-\exp(\eta_i)\right). \tag{6}$$

E a função de risco como:

$$h(t \mid x_i) = \left(\frac{ds(\log(t) \mid \gamma, k_0^{'})}{dt}\right)\exp(\eta_i). \tag{7}$$

Ao trabalhar com dados reais, especialmente num estudo em que o tempo de referência, como a data de início do estudo e o fim do

estudo, é de extrema importância, é preciso ter em conta a existência de censura e, até mesmo, de truncatura dos dados. Neste estudo particular detetou-se a necessidade de reconhecer truncatura à esquerda e censura à direita dos dados.

Consideramos como início do estudo a data do diagnóstico de tumor maligno e como final do estudo a data 30/11/2013. Uma vez que os indivíduos entram no estudo em diferentes momentos e o momento final do estudo é predeterminado por nós, estes têm seu próprio específico, e fixo, tempo de censura. Esta forma de censura é designada por censura generalizada de Tipo I [6].

Neste estudo conhece-se tanto o momento em que os indivíduos entram no estudo como o de morte se este ocorrer antes do final do estudo. No entanto, não se pode precisar o momento da morte se esta ocorreu após o final do estudo. Sendo que a única informação que se tem é que o momento do evento é maior ao momento do final do estudo. Estamos, portanto, confrontados com uma situação de censura à direita dos dados [17].

Adotando a notação de Klein e Moeschberger [17], para um específico indivíduo em estudo, assumimos que há um tempo de vida X e um tempo censurado fixo, C_r (C_r para tempo de censura à direita). Os tempos X são considerados independentes e identicamente distribuídos com função densidade de probabilidade $f(x)$ e função de sobrevivência $S(x)$. O tempo de vida X exato de um indivíduo será conhecido se, e apenas se, X é menor ou igual a C_r. Se X for maior do que C_r, o indivíduo é um sobrevivente, e o seu momento do evento é censurado em C_r.

Como mencionado anteriormente, a unidade de senologia do Hospital de Braga só foi criada em 2008, no entanto, temos informações sobre os pacientes que foram diagnosticados com tumor maligno antes desse momento, vivos à criação da Unidade. Como não temos qualquer informação sobre pacientes que foram diagnosticados antes de 2008 e morreram antes desse ano, temos de considerar uma truncatura à esquerda dos dados nesse ano.

Klein e Moeschberger[17] esclarecem que a truncatura de dados de sobrevivência ocorre quando somente aqueles indivíduos cujo mo-

mento do evento se encontra dentro de uma determinada "janela observacional"(YL, YR) são observados. Quando YR é infinito estamos perante truncatura à esquerda. Um indivíduo cujo momento do evento não se encontra neste intervalo não foi observado e, por isso, nenhuma informação sobre este indivíduo está disponível para o investigador. Os autores alertam para o fato de que, por apenas termos conhecimento de indivíduos com tempos de evento dentro da janela observacional, a inferência de dados truncados é restrita a uma estimação condicional.
Especificando para o presente estudo, a função de verosimilhança, tendo em conta censura à direita e truncatura à esquerda dos dados, pode ser construída da seguinte forma:

$$L\left(\theta \mid Y\right) = \prod_{i \in D} \frac{f(x_i)}{S(Y_{Li})} \prod_{i \in R} S(C_r), \tag{8}$$

Onde D é o conjunto de tempos de morte e R é o conjunto de observações censuradas à direita. Uma vez que tanto o CPHM e o FRPM assumem riscos proporcionais realizou-se o teste estatístico proposto por Grambsch e Therneau [27] com base no cálculo do resíduo de Schoenfeld, para avaliar a validade dessa suposição.
O modelo final escolhido foi estimado por "step-wise backwards", ou seja, começando com o modelo saturado considerando todas as variáveis significativas, e, em seguida, eliminando uma-por-uma as variáveis menos significativa.
Finalmente, para diagnóstico do modelo ajustado foi representada graficamente uma sobreposição da curva de sobrevivência não paramétrica (Kaplan-Meier) versus a curva de FRPM versus a curva CPHM, para uma determinada combinação de valores de covariáveis. Uma vez que os resultados apontam que tanto o FRPM e o CPHM devolvem estimativas semelhantes representou-se graficamente os resíduos de Cox-Snell [4], um tipo de resíduos para modelos de sobrevivência, para avaliar a qualidade global do ajuste. Se o modelo ajustado estiver correto, fazendo a transformação integral de probabilidade no verdadeiro momento de morte T, a variável aleatória

resultante tem uma distribuição uniforme sobre o intervalo unitário. Da mesma forma, a variável aleatória $U = H(T_j|x_j)$, onde x_j é um vetor de todas os covariáveis de tempo fixo, tem uma distribuição exponencial com taxa de risco 1. Aqui, $H(T_j|x_j)$ é a função de risco cumulativa verdadeira para um indivíduo. Se as estimativas dos β's do modelo são $b = (b_1,...,b_p)^{'}$, então, os resíduos de Cox-Snell são definidos como [17]:

$$r_{CS_i} = \widehat{H}_0(T_j) \exp\left(\sum_{k=1}^{p}(x_{ijk},b_k)\right), \qquad j = 1,...,n. \tag{9}$$

Onde $\widehat{H}_0(t)$ é a a função de risco de baseline estimada.
Se o modelo estiver correto e os valores dos $b's$ estiverem perto dos verdadeiros valores de β, então os valores dos $r'_{CS_i}s$ deverão ser uma amostra censurada de uma distribuição exponencial.
Uma possível avaliação gráfica da qualidade do ajuste, utilizada na presente análise, é comparar a função de sobrevivência da distribuição exponencial unitária $S_{\exp}(t) = \exp(-t)$, com as estimativas de sobrevivência de Kaplan-Meier do r_{CS_i} [30].
Toda a análise foi realizada com open source software estatístico R [23], em particular fazendo uso dos pacotes Survival [27] e flexsurv [15].

3.2 Análise Longitudinal

Dados longitudinais são geralmente caracterizados como variáveis de resposta que são medidas repetidamente ao longo do tempo para um grupo de indivíduos.
É importante a utilização de métodos longitudinais ao estudar este tipo de dados, pois permitem distinguir, por exemplo, alterações ao longo do tempo dentro de indivíduos e diferenças entre indivíduos nos seus níveis de baseline [9]. A principal característica dos modelos longitudinais é que estes permitem modelar tanto a dependência entre a resposta das variáveis explanatórias como a autocorrelação

entre as respostas. Ignorando correlação em dados longitudinais poderia levar a conclusões incorretas sobre os coeficientes de regressão, a estimativas dos coeficientes ineficientes [9].
Neste estudo em particular, a variável resposta - os valores de CA15–3 - foi analisada fazendo uso de modelos longitudinais definidos por Diggle, Heagerty, Liang e Zeger [9], onde foram testadas diferentes estruturas de correlação.
As mesmas covariáveis utilizadas no modelo final de sobrevivência foram testadas no modelo longitudinal ajustado. O tempo de referência utilizado foi o tempo, em anos, desde diagnóstico de cancro de mama até à data do teste sanguíneo que regista o valor do marcador. De acordo com os procedimentos médicos habituais, os médicos ficam alerta para uma possível recidiva do cancro da mama para pacientes que apresentem valores dos marcadores acima do valor de referência de 37 U/ml.
No geral, denotamos cada paciente pelo índice $i = 1,...,n$. Medidas repetidas dos marcadores para cada paciente i, no momento correspondente t_{ij}, são indicadas por Y_{ij}, onde $j = 1,...,m_i$. Note-se que, neste estudo em particular, as medições não são feitas, para todos os indivíduos, todas nos mesmos momentos, por isso estamos perante um estudo não balanceado. Considere-se $N = \sum_i^n m_i$ como o número total de medições da base de dados.
Iniciamos com uma análise exploratória e estimação pontual ajustando um modelo linear dos mínimos quadrados ordinários saturado (OLS)[9] com todas as variáveis que apresentaram um efeito significativo na sobrevivência dos pacientes, dado por:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \tag{10}$$

onde $E[Y_{ij}] = \mu_{ij}$ e ε_{ij} são N realizações independentes e identicamente distribuídas (i.i.d) de $N(0,\xi^2)$.
Uma vez que o modelo OLS assume independência entre quaisquer duas medições, do mesmo indivíduo ou entre indivíduos diferentes, é importante considerar diferentes modelos no contexto da análise longitudinal, que tenham em conta a correlação que geralmente existe

nas medições do mesmo indivíduo.
Para ter uma perceção da estrutura de correlação a considerar no modelo longitudinal final analisou-se o variograma [9] dos resíduos do modelo OLS saturado. O padrão deste sugeriu a existência de variabilidade entre indivíduos (efeitos aleatórios), e uma possível variabilidade dentro de indivíduos (correlação serial). Assim sendo, mantendo a mesma estrutura média (efeitos fixos) compararam-se dois modelos "aninhados"com as seguintes diferentes estruturas de correlação: (i) efeitos aleatórios, correlação serial exponencial e erro de medição (REE) e (ii) efeitos aleatórios, correlação serial Gaussiana e erro de medição (REG).
No geral, os modelos longitudinais considerados são dados por:

$$Y_{ij} = \mu_{ij} + d_{ij}U_i + W_i(t_{ij}) + Z_{ij}, \tag{11}$$

onde $\mathbf{U}_i$ são realizações i.i.d de $MVN(0,\Sigma)$, que representam os efeitos aleatórios ao nível do indivíduo e $\mathbf{d}_{ij}^{'}$ é um vector de covariáveis para os efeitos aleatórios. $W_i(t_{ij})$ é um processo Gaussiano contínuo com $E[W_i(t_{ij})] = 0$ e $Var[W_i(t_{ij})] = \sigma^2$, que representa a variabilidade dentro dos indivíduos e Z_{ij} são N realizações i.i.d. de $N(0,\tau^2)$, representado o erro de medição (variabilidade não explicada).
Para modelar o termo fixo do modelo longitudinal, μ_{ij}, testou-se a existência de um ponto de mudança no efeito do tempo nos marcadores tumorais. Em termos práticos, o ponto de mudança é o momento em que existe uma alteração no declive da progressão média do marcador. Considerando δ como o ponto de mudança, podemos escrever $E[Y_{ij}] = \mu_{ij}$ como:

$$\mu_{ij} = \begin{cases} X_{ij}\beta + \alpha_1 t_{ij}, & \text{if } t_{ij} < \delta \\ X_{ij}\beta + \alpha_2(t_{ij} - 1) & \text{if } t_{ij} \geq \delta \end{cases}, \tag{12}$$

onde X_{ij} representa o vetor de covariáveis, β o vetor de coeficientes de regressão a estimar, α_1 e α_2 os coeficientes representando o declive antes e depois do ponto de mudança, respetivamente.
Para a estimação dos parâmetros recorreu-se ao método de máxima verosimilhança, com a função de verosimilhança dada por:

$$L(\theta \mid Y) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \frac{1}{2\phi|V_{ij}|} \exp\left\{-\frac{1}{2}(y_{ij} - \mu_{ij})V_{ij}^{-1}(y_{ij} - \mu_{ij})^{'}\right\}, \tag{13}$$

Onde V_{ij} são as posições de variância/covariância da matriz de variância/covariância de todos os dados.
Em seguida, realizou-se, partindo do modelo saturado, uma eliminação das variáveis não significativas, até que a estrutura média ficou definida apenas com covariáveis significativas.
Para o modelo RE a estrutura de correlação apenas engloba os termos U_i e Z_{ij}. Sendo que para os modelos REE e REG a estrutura de correlação fica definida por: $U_i + W_i(t_{ij}) + Z_{ij}$. O que distingue estes dois últimos modelos longitudinais é a forma como diferentes realizações de W_i estão correlacionadas no tempos. Ou seja, se considerarmos a correlação entre os $W_i(t_{ij})$, digamos entre $W(t)$ e $W(t-u)$, determinada pela função de autocorrelação $\rho(u)$, teremos para o modelo REE $\rho(u) = \exp\left(-\frac{1}{\phi}|u|\right)$ e para o modelo REG $\rho(u) = \exp\left(-\frac{1}{\phi}u^2\right)$, onde ϕ é o parâmetro intervalar que especifica a taxa a que a correlação estabiliza.
A validação da estrutura de correlação a considerar no modelo final foi realizada graficamente através da comparação entre o variograma empírico e os variogramas teóricos dos modelos ajustados, e também comparando os valores maximizados da log-verosimilhança de cada modelo.
O variograma [9] de um processo estocástico $Y(t)$ é dado por:

$$V(u) = \frac{1}{2}\mathrm{Var}Y(t) - Y(t-u),\ u \geq 0. \tag{14}$$

Para um processo estacionário, a função de autocorrelação, $\rho(u)$, e a variância de $Y(t)$, σ^2, estão relacionados por: $\gamma(u) = \sigma^2\{1 - \rho(u)\}$.
A estimação do variograma empírico é baseada no cálculo das diferenças de resíduos, $\nu_{ij} = \frac{1}{2}(r_{ij} - r_{ik})^2$, e as correspondentes di-

ferenças temporais, $u_{ijk} = t_{ij} - t_{ik}$, onde $r_{ij} = Y_{ij} - \nu_{ij}$ com $j \leq k = 1,...,m_i$.
A função de autocorrelação em cada intervalo u é estimada do variograma amostral por:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}, \tag{15}$$

onde $\hat{\gamma}(u)$ é a média de todos os $\nu'_{ij}s$ correspondentes aquele valor particular de u, e $\hat{\sigma}^2$ é a variância do processo estimada.
Para a validação gráfica dos pressupostos do modelo longitudinal ajustado são apresentados o gráfico de resíduos padronizados contra os valores ajustados da variável resposta e o gráfico Q-Q dos quantis empíricos dos resíduos observados contra os quantis teóricos duma distribuição Normal Reduzida, para a validação dos pressupostos de variâncias homogéneas e distribuição gaussiana dos erros não explicados, Z_{ij}, respetivamente.
Toda a análise foi realizada com open source software estatístico R[23], em particular fazendo uso dos pacotes *nlme* [21] e *joineR* [20].

3.3 Análise conjunta de dados longitudinais e de sobrevivência

Modelos conjuntos para dados longitudinais e de sobrevivência são modelos estatísticos que nos permitem compreender dois processos de interesse, simultaneamente, longitudinal e sobrevivência, dado que existe uma associação entre eles [26].
Como o principal interesse da presente análise reside no estudo da associação entre o processo longitudinal e o processo de sobrevivência, adotou-se o modelo conjunto de efeitos aleatórios desenvolvido inicialmente por Wulfsohn e Tsiatis [29], dentro da metodologia de máxima verosimilhança para modelação conjunta.
Como Diggle, Sousa e Chetwynd [8] esclarecem, o princípio das metodologias de máxima verosimilhança na modelação conjunta consiste

na especificação da distribuição conjunta de Y, um vetor de medições repetidas, e um único tempo até ao evento F, que denotam por $[Y, F]$. Existem três tipos de metodologias de verosimilhança que diferem na parametrização da função de verosimilhança conjunta dos processos longitudinais e sobrevivência: os modelos de seleção; os modelos de mistura padrão e os modelos de efeitos aleatórios. As suas diferenças podem ser percetíveis pelas seguintes equações, onde U representa o efeito aleatório latente que liga os processos longitudinal e sobrevivência [26]:

$$\text{Modelos de seleção:}[Y,F,U] = [U][Y|U][F|Y]. \tag{16}$$

$$\text{Modelos de mistura padrão:}[Y,F,U] = [U][F|U][Y|F]. \tag{17}$$

$$\text{Modelos de efeitos aleatórios:}[Y,F,U] = [U][Y|U_1][F|U_2], \tag{18}$$

onde $U = (U_1,U_2)$.
Na abordagem de efeito aleatórios para modelar conjuntamente os resultados observados de Y e F são assumidos como condicionalmente independentes dado uma variável latente U, assim a distribuição conjunta de Y e F assume a forma:

$$[Y,F] = \int_U [U][Y,F|U]dU = \int_U [U][Y|U][F|U]dU. \tag{19}$$

Neste tipo de modelos a associação entre as medidas longitudinais e o tempo até ao evento fica completamente determinado pela estrutura de correlação entre os dois efeitos aleatórios U_1 e U_2.
Um método frequentemente utilizado para estimar os parâmetros do modelo conjunto, adotado no presente trabalho, é o método da máxima verosimilhança (ML), que maximiza a log-verosimilhança da distribuição conjunta dada por [19]:

$$\prod_{i=1}^{N} \int f(Y_i|U_i,\theta) f(F_i,\delta_i|U_i,\theta) f(U_i|\theta) dU_i, \tag{20}$$

onde $f(Y_i|U_i,\theta)$, $f(F_i,\delta_i|U_i,\theta)$ e $f(U_i|\theta)$ são as funções de densidade dos processos longitudinal, sobrevivência e efeitos aleatórios, respetivamente, e δ é o indicador do evento, igualando a um se o evento ocorreu e a zero caso contrário.

Usualmente as medições repetidas são modeladas por um modelo linear de efeitos aleatórios, explicado na secção anterior, incorporando efeitos aleatórios ao nível do indivíduo e o modelo para os resultados de sobrevivência trata-se de um modelo de riscos proporcionais de Cox que incorpora uma fragilidade logGaussiana. A dependência estocástica é, então, capturada permitindo que os efeitos aleatórios Gaussianos do modelo linear estejam correlacionados com o termo de fragilidade do modelo de Cox. Um formato típico da variável latente U é considerar o efeito ordenada na origem e declive aleatórios dado por [14]: $U_i(t) = U_{0i} + U_{i1}t_{ij}$, onde U_{0i} e U_{i1} representam a ordenada na origem e do declive aleatórios, respetivamente, para o indivíduo i.

A presente análise conjunta adopta a metodologia de efeitos aleatórios fazendo uso de dois pacotes do software R [23]: *JM* [30], e *joineR* [20]. McCrink, Marshall e Cairns [19] fornecem, na sua revisão dos avanços em modelação conjunta, uma comparação clara da utilização destes dois pacotes.

No entanto, relativamente a esta análise em particular, existem dois grandes constrangimentos no software utilizado: embora o modelo conjunto de efeitos aleatórios permita que se considere simultaneamente censura à direita e truncatura à esquerda dos dados, ambos os pacotes utilizados ainda não são capazes de lidar com truncatura à esquerda de dados. Dessa forma, só consideramos, na modelagem conjunta, um modelo de Cox que incorpora censura à direita no processo de sobrevivência. Assim sendo, apenas é relevante comparar estimativas obtidas para o processo de sobrevivência com o modelo, obtido anteriormente, que considera apenas o mecanismo de censura à direita. O outro constrangimento prende-se com o fato de que esses dois pacotes não são capazes de incorporar, no processo longitudinal, uma correlação de série, entre duas medidas registadas em momentos diferentes para o mesmo indivíduo, a fim de explicar a

existência de uma possível variabilidade dentro de indivíduos. No nosso caso específico, foi comprovada a existência de uma estrutura de correlação exponencial que representa a variabilidade dentro do individuo que não deve ser ignorado. Como tal, decidiu-se, para que a correlação entre duas medidas tomadas em dois momentos diferentes dependa do tempo, considerar o efeito ordenada na origem e declive aleatórios, descrito acima. A escolha entre os dois pacotes depende do foco da pesquisa implementada [19]. O modelo conjunto implementado no pacote *JM* foca-se no processo de sobrevivência e em como este é afetado por uma covariável longitudinal dependente do tempo que é medida com erro. O pacote *joineR* implementa um modelo conjunto onde o foco reside em ambos os processos com o objetivo de inferir sobre a força da ligação entre os dois processos. Sendo que ambos os objetivos são do interesse para a presente análise, fez-se uso de ambos os pacotes retirando as respetivas conclusões com os resultados obtidos.
Mantendo a notação das seções anteriores, ambos os modelos implementados pelos pacotes utilizam o modelo linear de efeitos aleatórios para representar o processo longitudinal, dado por:

$$\text{Processo Longitudinal: } Y_{ij} = \mu_{ij} + U_i + Z_{ij} = X_1\beta + U_{0i} + U_{1i}t_{ij} + Z_{ij}, \tag{21}$$

onde X_1 é a matriz de desenho para os efeito fixos, com os respetivos parâmetros de regressão a estimar β. Na presente análise consideraram-se como covariáveis fixas as que apresentaram efeito significativo na progressão do marcador na análise longitudinal separada, apresentada na secção anterior. A variável latente dada por (U_{0i}, U_{1i}) é a realização de $MVN(0,\Sigma)$ onde $\Sigma = \begin{pmatrix} \nu_1^2 & \nu_{12} \\ \nu_{12} & \nu_2^2 \end{pmatrix}$. Z_{ij} são N realizações i.i.d de $N(0,\tau^2)$, representando o erro de medição (variabilidade não especificada).
A diferença primordial entre os pacotes *joineR* e *JM* reside na forma como incorporam a variável latente aleatória no processo de sobrevivência.

O pacote *JM* incorpora a estimativa precisa do processo longitudinal, $m_i(t) = \hat{\mu}_{ij} + \hat{U}_i$, no processo de sobrevivência da seguinte forma:

$$\text{Processo de Sobrevivência } JM\text{: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2 + \alpha m_i(t)\}, \tag{22}$$

onde X_{2i} representa as covariáveis de baseline, neste estudo em particular as covariáveis com efeito significativo obtidas na analise de sobrevivência separada (secção anterior), e β_2 o vetor dos coeficientes de regressão, a estimar, respetivos. O parâmetro representa o efeito da verdadeira resposta longitudinal no processo de sobrevivência, ou seja, particularizando, representa o efeito do verdadeiro valor do marcador tumoral CA15-3 no processo de sobrevivência.
O pacote *joineR* incorpora os efeitos aleatórios longitudinais no modelo de sobrevivência da seguinte forma:

$$\text{Processo de Sobrevivência } joineR\text{: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2+\gamma_0 U_{0i}+\gamma_1 U_{1i}t\}, \tag{23}$$

onde γ_0 e γ_1 representam o efeito da ordenada na origem e declive aleatórios do processo longitudinal, no processo de sobrevivência.
Sintetizando, utilizando o modelo implementado pelo pacote *JM* vamos ser capazes de determinar os fatores que influenciam as alterações dos valores de CA15-3 nos pacientes, e que efeito tem essa alteração na sua sobrevivência. A utilização do pacote *joineR* é apropriada para a nossa análise uma vez que estamos interessado em determinar o efeito da resposta inicial dos valores de CA15-3, γ_0, e as alterações na resposta ao longo do tempo, γ_1, na sobrevivência dos pacientes.
Para a validação gráfica dos pressupostos do modelo conjunto ajustado são apresentados, para validação do processo longitudinal, o gráfico de resíduos padronizados contra os valores ajustados da variável resposta e o gráfico Q-Q dos quantis empíricos dos resíduos observados contra os quantis teóricos duma distribuição Normal Reduzida, para a validação dos pressupostos de variâncias homogéneas

e distribuição gaussiana dos erros não explicados, Z_{ij}, respetivamente. Para validação do processo de sobrevivência faz-se uso da representação gráfica dos resíduos de Cox-Snell.

4 Principais Resultados

4.1 Análise de sobrevivência

Uma análise exploratória prévia (não apresentada no presente documento por restrições de espaço) reporta que a maioria dos casos de cancro da mama analisados se encontram num estádio inicial. Uma vez que as distribuições de frequência das variáveis estádio, tamanho e grau do tumor se centram em valores mais baixos das categorias destas. Este resultado pode estar relacionado com um aumento do rastreio precoce levando a deteção prematura do tumor. No entanto, é um resultado importante de ser mencionado, uma vez que pode traduzir-se numa elevada probabilidade de sobrevivência destes pacientes.

De facto, numa primeira abordagem exploratória, em termos de sobrevivência, revela que a estimativa de KaplanMeier para uma sobrevivência a 10 anos (120 meses) para estes pacientes (Figura1) está bastante próxima dos 70%. Ainda, a estimativa não paramétrica parece apontar xque para o tempo total de followup a probabilidade de sobrevivência está acima dos 50%. A importância em se considerar truncatura à esquerda destes dados em particular está bem expressa nas diferenças entre as curvas de Kaplan-Meier (Figura1) para os dois casos: considerando apenas censura à direita e considerando simultaneamente censura à direita e truncatura à esquerda. Sendo que é visível que ignorando truncatura à esquerda leva a uma sobrestimação da sobrevivência.

Um total de 16 variáveis, apresentadas na Tabela 1, apresentaram diferenças significativas entre as curvas de sobrevivência das suas categorias (Tabela 1).

Cancro de mama triplo negativo (TN) é definidos pela ausência de es-

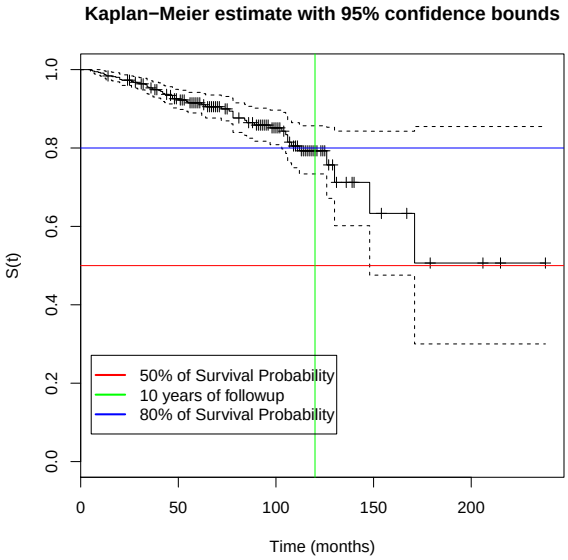


Figura 1: Curvas de estimativas de Kaplan-Meier dos pacientes de cancro da mama do Hospital de Braga.

trogénio, progesterona e expressão HER2 [7]. Embora o prognóstico de tumores triplo negativo permaneça incerto, é de conhecimento médico que os tumores TN têm usualmente pior prognóstico. No presente estudo contabilizaram-se 4.38% de casos TN, que representam 21,28% das mortes por câncer de mama. Como esperado, existe uma diferença significativa nas taxas de sobrevivência, para as categorias da variável triplo negativo (TN) sim contra não, como apresentado na Figura 2, e confirmado pelos resultados do teste log-rank (Tabela 1), sendo, como esperado os casos de cancro da mama TN aqueles com menor probabilidade de sobrevivência.

Embora a idade de diagnóstico tratada como uma variável contí-

Tabela 1: Resultados do teste de log-rank das variáveis com efeito significativo na sobrevivência dos pacientes

Resultados do teste Log-Rank	
Variável	**p-valor**
Estadio	< 0.001
Cancro Bilateral	0.002
Recidiva de cancro	< 0.001
Tratamento Neoadjuvante	< 0.001
Tipo de Cirurgia	< 0.001
Imagens de invasão vascular venosa	< 0.001
Biópsia de nódulos linfáticos sentinela	0.02
Expressão dos recetores de estrogénio	< 0.001
Expressão dos recetores de progesterona	< 0.001
Cancro triplo negativo	< 0.001
Grau de diferenciação	< 0.001
Tamanho do tumor primário	0.005
Envolvimento de nódulos linfáticos regional	< 0.001
Idade ao diagnóstico (categorizada)	0.017
Imagens de invasão linfática	0.04
Hormonoterapia	< 0.001

nua (como deveria ser) não apresentou um efeito estatisticamente significativo na sobrevivência dos pacientes, obteve-se um resultado diferente quando se categorizou esta mesma variável em dois grupos: pacientes com menos de 44 anos e pacientes com mais de 44 anos (ou idade igual). Verificou-se que casos que se incluam na categoria mais baixa (mulheres mais jovens) têm uma probabilidade de sobrevivência menor do que casos incluídos na categoria análoga (Figura 2), sendo essa diferença é estatisticamente significativa (Tabela 1).

É de salientar que, num primeiro momento, categorizou-se a idade ao diagnóstico de forma idêntica à realizada nos estudos de natureza semelhante reportados pelo RORENO. Subsequentemente agrupamos categorias que não diferiam estatisticamente em termos de probabilidade de sobrevivência.

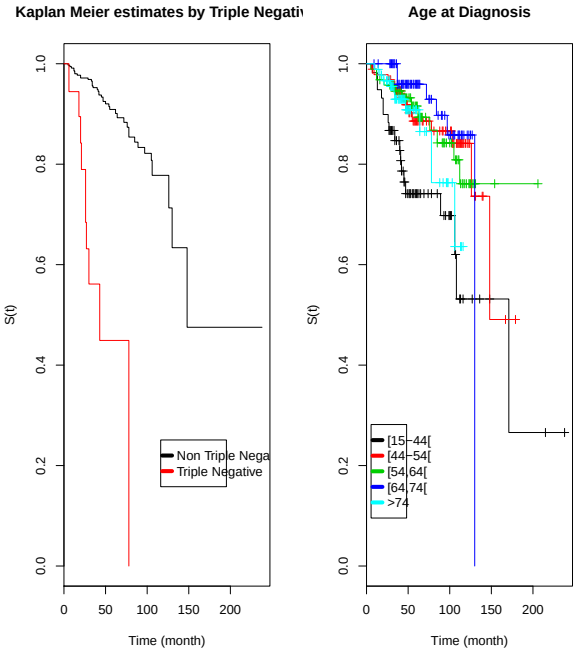


Figura 2: Curvas de KaplanMeier para as categorias das variáveis triplo negativo (triple negative) e idade ao diagnóstico (age at diagnosis).

Os resultados, como se pode observar na Figura 2, apontam para uma diferença significativa entre as curvas de sobrevivência das três categorias da variável grau de diferenciação (G1, G2 e G3), confirmados pelos valores do teste de logrank (Tabela 1). A probabilidade de sobrevivência de pacientes com tumor de grau G3 é menor em comparação com as outras duas categorias, e os casos com tumor diagnosticado com grau G1 tem uma maior probabilidade de sobrevivência.

Ainda, resultados confirmam que não há diferença significativa entre a taxa de sobrevivência entre as três primeiras categorias de Estágio do tumor (0, I e II), e também entre as duas últimas categorias desta covariável (III e IV). Assim sendo, foi possível agrupar as três primeiras categorias em uma única categoria (0/I/II) e ao grupo as duas últimas categorias numa só (III/IV). Os resultados sugerem (Figura 3) que os casos diagnosticados com estadio III ou IV têm menor probabilidade de sobrevivência em comparação com um tumor no estadio 0, I ou II, para estes pacientes.

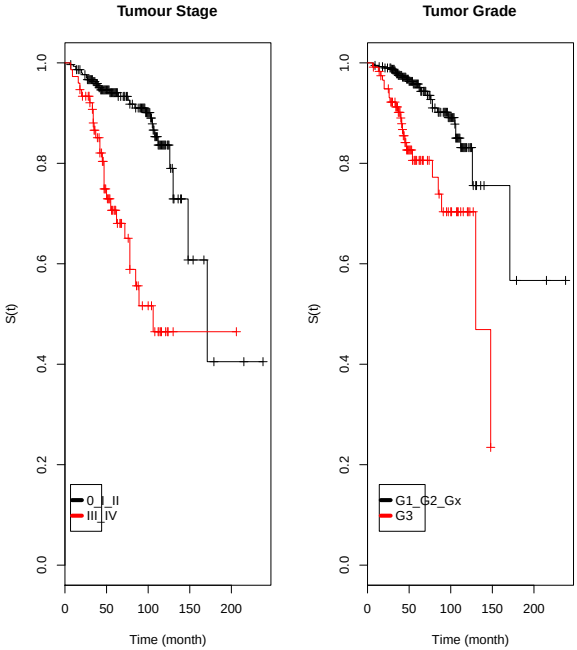


Figura 3: Curvas de KaplanMeier para as categorias das variáveis estadio (tumour stage) e grau de diferenciação (tumour grade).

Para a variável relacionada com o envolvimento regional de nódulos linfáticos, os nossos resultados sugerem que pacientes com tumor do tipo Nx, N0 ou N1 têm uma probabilidade de sobrevivência significativamente maior do que aqueles com maior grau (Figura 4). Foi possível determinar que não existem diferenças significativas entre as curvas de sobrevivência entre as categorias Nx, N0 e N1 e, também, entre as categorias N2 e N3.

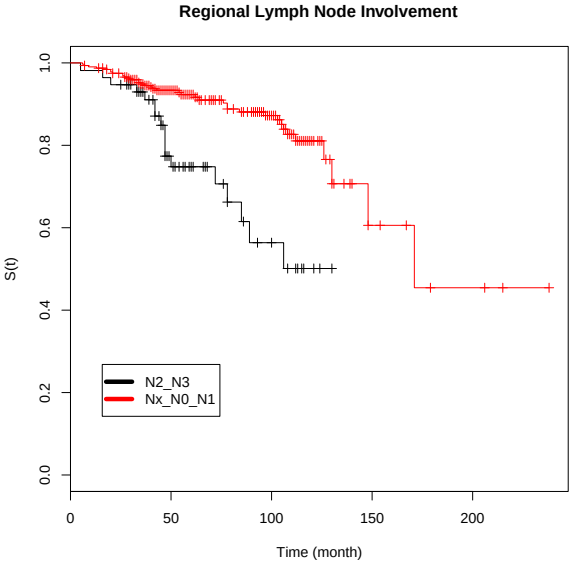


Figura 4: Curvas de Kaplan-Meier para as categorias da variável envolvimento regional de nódulos linfáticos (Regional lymph node involvement).

Depois de ajustar inúmeros modelos de sobrevivência com múltiplas covariáveis, começando, como já mencionado, com o modelo saturado, selecionou-se como modelo final o FRPM de zero nós com as

cinco covariáveis, descritas acima, com efeito significativo na sobrevivência destes pacientes. A Tabela 2 compara as estimativas obtidas (e respetivos intervalos de confiança de 95%) quando se considera o FRPM com as obtidas quando se ajusta ao CPHM. Apresentam-se, ainda, as estimativas para ambos os modelos considerando apenas censura à direita e considerando simultaneamente censura à direita e truncatura à esquerda. Como se pode constatar, os valores de estimativas são idênticos para ambos os modelos, e confirmam os resultados apontados acima. Constata-se, ainda, algumas diferenças nos resultados quando se considera ou ignora truncatura à esquerda. Por exemplo, a variável envolvimento regional de nódulos linfáticos não tem efeito significativo na sobrevivência dos pacientes quando se ignora truncatura à esquerda.

É de salientar que, nos modelos aqui apresentados, não se consideraram variáveis relacionadas com o tratamento, uma vez que as decisões do tratamento tomadas pela equipe médica recaem sobre as características do tumor, o que poderia traduzir uma forte correlação entre os diferentes tipos de variáveis relacionadas com o tratamento e as variáveis relacionadas com o tipo de tumor.

Como esperado, o risco de morrer de cancro da mama é significativamente maior (mantendo todos os outros valores iguais) em mulheres com cancro de mama triplo negativo (RR = 6.86). Analogamente, o risco de morte por cancro da mama para pacientes com 44 anos, ou mais, no momento do diagnóstico é mais baixo (RR = 0.38). O risco de morrer de cancro de mama é 7.00 vezes maior para os casos com tumor de estadio III ou IV em relação a casos com tumor de estadio 0, I ou II. Para além disso, o risco de morrer de um tumor com o tipo de grau G2 é 4.72 vezes maior do que de um tumor do tipo G1 grau, e aumenta o risco para pacientes com tumor do tipo G3 (RR = 6.99). Finalmente, aqueles que apresentam um envolvimento do tumor em nódulos linfáticos regionais de N2 ou N3 grau têm um risco mais elevado (RR = 2.24) de morrer de cancro da mama.

Com o intuito de avaliar graficamente a qualidade do ajuste dos modelos FRPM e CPHM selecionados, sobrepôs-se num mesmo gráfico (Figura 4) as curvas Kaplan-Meier, FRPM e CPHM para um paciente com seguintes características: tumor não triplo negativo, com idade ao diagnóstico ≤ 44, tumor no estadio III ou IV e tumor de grau G2.
É de salientar que, como Royston e Parmar [25] explicam, por convenção o modelo com zero nós significa que não foram especificados nenhum nó interno e nem de fronteira e, dessa forma, a distribuição de baseline trata-se da distribuição de Weibull.

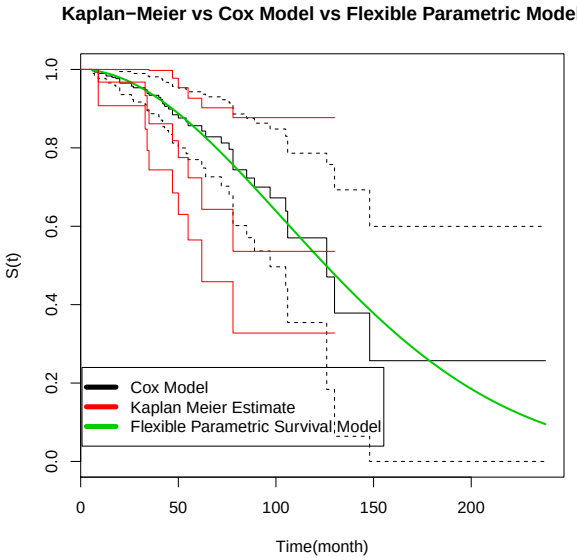


Figura 5: Curvas de Kaplan-Meier, FRPM e CPHM para a combinação de covariáveis: tumor não triplo negativo; idade ao diagnóstico ≥ 44; tumor no estadio 0, I ou II e grau G2, e grau de envolvimento de nódulos linfáticos Nx, N0 ou N1.

Tabela 2: Estimativas das razões de risco (RR) para os modelos de risco proporcional de Cox (CPHM) e para o modelo paramétrico flexível de Royston e Parmar (FRPM).

	Apenas Censura à Direita e Censura à Direita e Truncatura à Esquerda			
	CPHM	**FRPM (0 nós)**	**CPHM**	**FRPM (0 nós)**
Covariáveis	**RR IC 95%**	**RR IC 95%**	**RR IC 95%**	**RR IC 95%**
Triplo negativo (sim)	8.07 [3.27; 19.96]	7.10 [2.80; 17.99]	7.82 [2.81; 21.98]	6.86 [2.51; 18.7]
Idade ao diagnóstico (≥ 44)	0.39 [0.20; 0.73]	0.41 [0.22; 0.78]	0.37 [0.18; 0.71]	0.38 [0.20; 0.73]
Estadio (III ou IV)	3.89 [2.09; 7.25]	3.94 [2.13; 7.24]	7.58 [2.95; 19.45]	7.00 [2.80; 17.5]
Grau (G2)	4.51 [1.04; 19.47]	3.86 [0.96; 15.33]	5.05 [1.16; 21.94]	4.72 [1.09; 20.50]
Grau (G3)	6.14 [1.37; 27.65]	5.75 [1.38; 24.05]	6.39 [1.39; 29.47]	6.99 [1.53; 31.90]
Envolvimento nódulos linfáticos (N2 ou N3)	—	—	2.39 [0.86; 6.62]	2.24 [0.81; 6.21]

Tal como apresentado na Tabela 3, o pressuposto de riscos proporcionais para as covariáveis consideradas nos modelos finais não foi violado.

Tabela 3: P-valores obtidos no teste de riscos proporcionais

Covariáveis	**p-valor**
Triplo negativo (sim)	0.3885
Idade ao diagnóstico (≥ 44)	0.0621
Estadio (III ou IV)	0.5721
Grau (G2)	0.1727
Grau (G3)	0.4238
Envolvimento nódulos linfáticos (N2 or N3)	0.3312

A representação gráfica de resíduos de Cox-Snell (Figura 6), onde a linha preta sólida indica a estimativa de KaplanMeier da função de sobrevivência dos resíduos (com as linhas a tracejado correspondem os intervalos de 95% de confiança), e a linha a cinzento a função de sobrevivência a distribuição exponencial unitária, sugere um ajuste adequado do modelo de sobrevivência aos dados em estudo.

4.2 Análise longitudinal do marcador CA15-3

Apenas se registou informação disponível sobre valores marcadores tumorais CA15-3 para 534 dos 540 pacientes elegíveis para análise. Um total de 5166 medições de marcador tumoral CA15-3 representam todas as medições disponíveis desses pacientes desde o momento do diagnóstico de cancro da mama até ao final do estudo.
Uma vez que o pressuposto de normalidade da variável de resposta falhou, utilizou-se uma transformação logarítmica dos valores de CA15-3. Trata-se de uma transformação usual em marcadores biológicos. O gráfico de progressões individuais (Figura 7) apresenta a progressão dos valores de CA15-3, na escala logarítmica, para cada paciente, em relação ao valor de referência e a linha não paramétrica de *spline* suavizada que indica tendência média de progressão. A

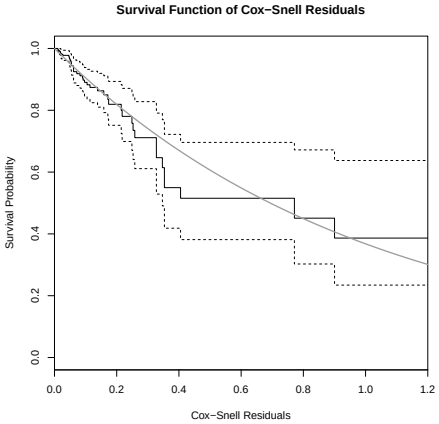


Figura 6: Sobreposição as estimativas de Kaplan-Meier da função de sobrevivência dos resíduos e da função de sobrevivência da distribuição exponencial unitária.

linha de *spline* sugere que, em média, a progressão do marcador aumenta num ritmo lento, permanecendo abaixo do valor de referência durante 10 anos (120 meses) após o cancro da mama ser diagnosticado. Após cerca de dez anos parece existir um momento onde a taxa de crescimento aumenta e o valor de referência é ultrapassado. Este fato pode apontar para a existência de um ponto de mudança na sua progressão no tempo. Assim, pareceu pertinente testar a existência de um ponto de mudança na progressão linear média do marcador.

No entanto, depois de se ajustar inúmeros modelos paramétricos saturados variando os valores do ponto de mudança (δ), a sua existência não foi significativa na progressão média do marcador.

A Tabela 4 apresenta os parâmetros estimados do modelo longitudinal final selecionado que explica a progressão do marcador no tempo, comparando as estimativas obtidas ajustando o modelo OLS

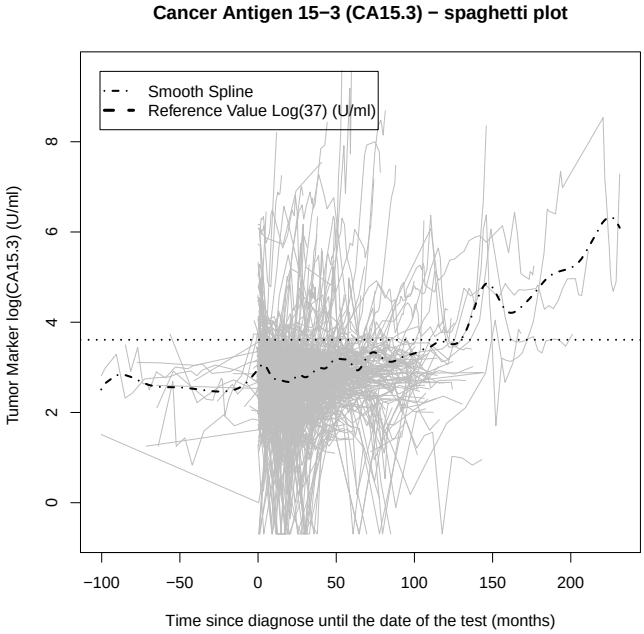


Figura 7: Gráfico de progressões individuais (spaghetti plot) para os valores do marcador tumoral CA15-3.

simples, e dos dois modelos longitudinais onde varia a estrutura de correlação e, ainda, os respetivos valores de log-verosimilhança. É de notar que, embora os valores das estimativas sejam similares para os três modelos, há variações na significância destes (p-valor).

A parte fixa do modelo longitudinal, que descreve a progressão média do marcador, é composta pelas seguintes covariáveis com efeito significativo na ordenada na origem do modelo linear: idade do diagnóstico, estadio (0 ou I ou II contra III ou IV), invasão vascular venosa (Sim contra Não), e expressão Ki.67 (baixo contra alto). A

Tabela 4: Estimativas dos parâmetros para o modelo linear ordinários e modelos longitudinais para o marcador tumoral CA15-3

	REE Modelo		**REG Modelo**		**OLS Modelo**	
	Est	**p-valor**	**Est**	**p-valor**	**Est**	**p-valor**
Ordenada na origem	2.226	< 0.001	2.215	< 0.001	2.287	< 0.001
Tempo (meses)	0.003	0.006	0.004	< 0.001	0.004	< 0.001
Idade ao diagnóstico	0.007	0.018	0.007	0.018	0.007	< 0.001
Estadio (III ou IV)	0.277	0.035	0.243	0.063	0.408	< 0.001
Invasão Vascular Venosa (Sim)	0.610	0.003	0.623	0.002	0.812	< 0.001
Ki.67 (baixo)	-0.168	0.067	-0.163	0.076	-0.226	< 0.001
ν^2	0.426		0.5011		.	
σ^2	0.031		0.066		.	
ϕ	10.552		6.119		.	
τ^2	0.403		0.284		.	
ξ^2	.		.		1.084	
Log Likelihood	-1953.611		-2472.825		-3447.758	

ordenada na origem do modelo, neste caso em particular, significa que um paciente com um tumor de estadio 0, I ou II, que não apresenta imagens de invasão vascular venosa e com uma expressão Ki.67 alta, em idade precoce de diagnóstico vai iniciar a progressão do marcador de tumor com um valor aproximado de 2.23, numa escala logarítmica.

Selecionando o modelo com o maior valor de logveromilhança, o modelo que incorpora efeitos aleatórios ao nível do indivíduo e estrutura de correlação serial exponencial para descrever a progressão média dos valores de CA15-3 pode-se inferir que a idade ao momento do diagnóstico afeta o valor inicial do marcador (numa escala logarítmica) a uma taxa de 0.0074 por ano de idade ao momento do diagnóstico. Um paciente com um tumor com estadio III ou IV implica um aumento do valor inicial do marcador de cerca de 0.2769, em

comparação com os casos com tumor no estadio 0, I ou II. Ainda, um tumor que apresente imagens da invasão vascular venoso tem um incremento valor inicial do marcador de 0.6097. Por outro lado, uma baixa expressão do biomarcador Ki.67 diminui o valor inicial da progressão do marcador em 0.1681.
A estrutura de correlação selecionada para representar a variabilidade dos dados é, como foi referido, a que incorpora efeitos aleatórios ao nível indivíduo com $\hat{\nu}^2 \approx 0.4359$, estrutura de correlação exponencial para descrever a variabilidade dentro do indivíduo com $\hat{\rho}(u) \approx \exp(-\frac{1}{10.5517}.|u|)$ e $\hat{\sigma}^2 \approx 0.0305$, e um erro de medição com variância $\hat{\tau}^2 \approx 0.4026$. A escolha da estrutura de correlação foi, também, sustentada pela sobreposição dos variogramas teóricos ajustados para os três modelos longitudinais contra o variograma empírico (Figura 8).
Comparando os modelos REE e REG com o modelo RE, a componente de correlação serial $W_i(t_{ij})$ demonstrou ser significativa. Este resultado reforça a necessidade de ter em conta correlação entre medições dentro de um mesmo indivíduo.
Finalmente, graficamente (Figuras 9 e 10) os pressupostos do modelo longitudinal relativamente à homogeneidade das variâncias dos erros de medição e na normalidade destes não parece ser passível de ser rejeitada.

4.3 Análise Conjunta dos dados longitudinais CA15-3 e de sobrevivência

A Tabela 5 apresenta os resultados da modelação conjunta do processo longitudinal dos valores do marcador tumoral CA15-3 (em escala logarítmica) e o processo de sobrevivência, onde o evento de interesse é a morte cancro da mama, obtidos fazendo uso do pacote *joineR* e do pacote *JM*. Estão indicadas as estimativas obtidas nas covariáveis consideradas para ambos os processos (longitudinal e de sobrevivência), bem como o erro padrão (SE) e os valores de prova respetivos, para os dois modelos conjuntos obtidos por cada pacote e, finalmente, o valor da log-verosimilhança de cada modelo.

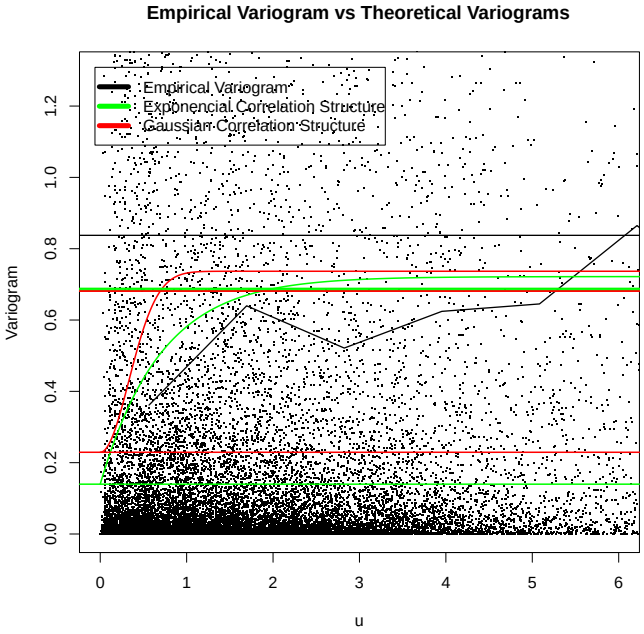


Figura 8: Sobreposição dos variogramas teóricos e do variograma empírico marcador tumoral CA15-3.

Analogamente à análise longitudinal, apresentada na secção anterior, uma vez que o pressuposto de normalidade da variável resposta longitudinal falhou, utilizou-se uma transformação logarítmica dos valores de CA15-3.
É importante ressalvar que, na presente análise conjunta apenas foram incluídos os pacientes com informação disponível dos valores de CA15-3. O que significa que apenas 534 pacientes foram elegíveis para a presente análise, ao passo que 540 pacientes foram incluídos na análise independente sobrevivência, apresentada na secção

Tabela 5: Estimativas dos parâmetros do modelo conjunto de dados longitudinais CA15-3 e de sobrevivência

	Pacote JoineR			**Pacote JM**		
Covariáveis	**Estimativa**	**SE**	**pvalor**	**Estimativa**	**SE**	**pvalor**
PROCESSO LONGITUDINAL						
Ordenada	2.212	0.170	< 0.0001	2.225	0.097	< 0.0001
Tempo (meses)	0.006	0.001	< 0.0001	0.006	0.001	< 0.0001
Idade ao Diagnóstico	0.007	0.002	0.003	0.005	0.002	< 0.001
Estadio (III ou IV)	0.054	0.099	0.584	0.134	0.078	0.088
Invasão Vascular Venosa (Sim)	0.753	0.269	0.005	0.642	0.112	< 0.0001
Ki.67 (baixo)	-0.136	0.062	0.090	-0.009	0.071	0.900
PROCESSO SOBREVIVÊNCIA						
Triplo Negativo (Sim)	1.878	0.408	< 0.0001	2.303	0.718	0.001
Idade ao Diagnóstico (≥ 44)	-0.070	0.371	0.850	-0.636	0.522	0.223
Estadio (III ou IV)	2.874	0.608	< 0.0001	3.557	0.706	< 0.0001
Grau						
(G2)	1.332	0.738	0.071	2.451	0.910	0.007
(G3)	1.513	0.738	0.040	2.804	0.859	0.001
Nódulos linfáticos (Nx ou N0 ou N1)	2.729	0.830	0.001	3.998	0.632	< 0.0001
ASSOCIAÇÃO LATENTE						
Parâmetro(s) de associação	γ_0 1.136	0.182	< 0.0001			
	γ_1 1.099	0.103	< 0.0001	α 1.194	0.143	< 0.0001
ν_1^2	0.2301			0.4970		
ν_2^2	1.6×10^{-7}			0.00038		
τ^2	0.0636			0.2566		
Loglikelihood	-2302.3			-2230.6		

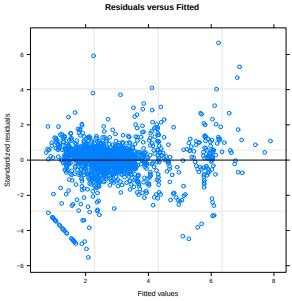


Figura 9: Gráfico de resíduos padronizados contra valores ajustados do marcador CA15-3.

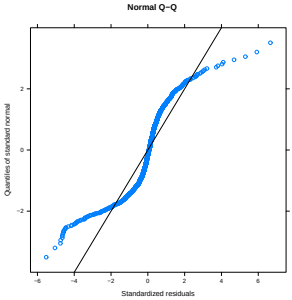


Figura 10: Q-Q plot of the CA15-3 longitudinal model.

anterior. Isso traduz-se em valores diferentes entre as estimativas dos parâmetros do modelo de sobrevivência independente e as obtidas para o modelo de sobrevivência que considera apenas os 534 pacientes.

No entanto, uma comparação entre as covariáveis que têm efeito significativo no processo de sobrevivência do modelo conjunto e as que têm efeito na análise de sobrevivência separada é pertinente. Covariáveis, tais como triplo negativo (sim contra não), estadio (0 ou I ou

II contra III ou IV) e envolvimento dos nódulos linfáticos regionais (Nx ou N0 ou N1 contra N2 ou N3) mantêm o seu efeito significativo na sobrevivência dos pacientes. Apesar de que o envolvimento dos nódulos linfáticos regionais apenas tinha mostrado efeito significativo quando se considera truncatura à esquerda no modelo de sobrevivência. A covariável grau de diferenciação do tumor (Gx ou G1 contra G2 contra G3) tem um efeito significativo marginal sobre a sobrevivência do paciente para o modelo ajustado com o pacote joineR, mas um efeito claramente significativo no modelo ajustado com o pacote *JM*. Ao contrário, a idade ao momento do diagnóstico (categorizada em ≤ 44 anos de idade contra ≥ 44 anos de idade) não apresenta um efeito significativo na sobrevivência dos pacientes para ambos os modelos conjuntos ajustados. No entanto, este resultado também pode estar associado pelo facto de que o efeito do tempo já é explicado pela progressão longitudinal do marcador tumoral no tempo.

É importante referir que as diferenças entre as estimativas obtidas para o processo de sobrevivência por ambos os pacotes é justificável pela forma como estes incorporam o efeito da variável longitudinal no processo de sobrevivência.

As diferenças na estrutura de correlação consideradas no modelo longitudinal ajustado na análise separada e o modelo longitudinal incorporado na modelação conjunta explicam as diferenças entre as estimativas dos parâmetros de regressão para o processo longitudinal. De facto, covariáveis como a idade no momento do diagnóstico, invasão vascular venosa (sim contra não) mantêm o seu efeito significativo nos valores iniciais da progressão linear do valor log CA15-3 dizer. No entanto, as covariáveis ki.67 (baixo contra alto) e estadio (0 ou I ou II contra III ou IV) não apresentam um efeito significativo sobre o processo longitudinal para ambos os modelos conjuntos considerados.

Para o modelo conjunto obtido com o pacote *JM*, a significância do parâmetro de associação α ($p - valor \leq 0.0001$) no processo de sobrevivência realça a necessidade de se considerar uma modelação conjunta dos dados. Esta ligação significativa confirma a relação en-

tre os processos longitudinal e sobrevivência. Analogamente, para o modelo conjunto ajustado com o pacote *joineR*, ambos os parâmetros de associação, γ_0 e γ_1, são significativos ($p - valor \leq 0.0001$) no processo de sobrevivência. O que confirma a necessidade de considerar uma modelação conjunta dos dados em estudo.
Relativamente aos resultados obtidos para o modelo conjunto ajustado com o pacote JM, através da determinação da razão de riscos (RR), é evidente que os indivíduos com valores elevados de CA15-3 (na escala logarítmica) tendem a ter uma pior probabilidade de sobrevivência ($RR = exp(1.1942) \approx 3.3$).
No âmbito das estimativas obtidas para o modelo conjunto ajustado com o pacote *joineR*, através do cálculo de razão de riscos, γ_0 indica que indivíduos com valores iniciais de CA15-3 que são mais elevados do que a média da população tendem a ter pior probabilidade de sobrevivência ($RR = exp(1.1356) \approx 3.11$). Analogamente, γ_1, indica que os indivíduos que têm um maior aumento na progressão média dos valores de CA15-3 tendem a ter pior probabilidade de sobrevivência ($RR = exp(1.0987) \approx 3.0$).
A representação gráfica de resíduos de CoxSnell (Figura 11), onde a linha preta sólida indica a estimativa de KaplanMeier da função de sobrevivência dos resíduos (com as linhas a tracejado correspondem os intervalos de 95% de confiança), e a linha a cinzento a função de sobrevivência a distribuição exponencial unitária, sugere um ajuste adequado do processo de sobrevivência no modelo conjunto.
Graficamente (Figuras 12 e 13) os pressupostos do modelo longitudinal relativamente à homogeneidade das variâncias dos erros de medição e na normalidade destes não parece ser passível de ser rejeitada.

5 Discussão

No presente trabalho propusemo-nos a realizar uma análise de dados de cancro da mama de pacientes acompanhados e diagnosticados na unidade de senologia do Hospital de Braga. Sendo que os dados re-

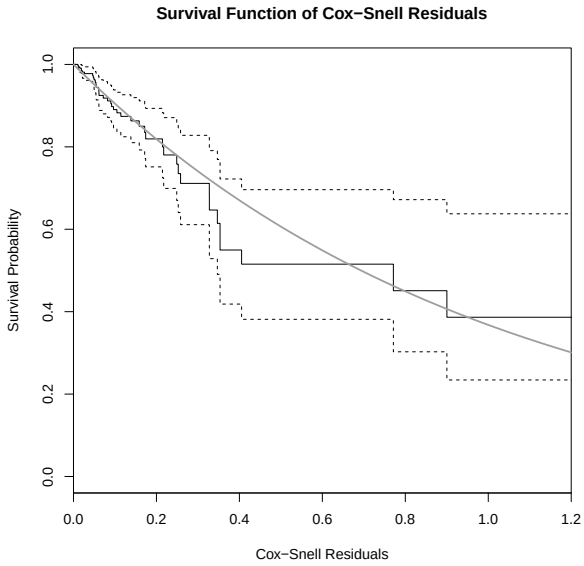


Figura 11: Sobreposição as estimativas de KaplanMeier da função de sobrevivência dos resíduos e da função de sobrevivência da distribuição exponencial unitária.

colhidos do sistema informático do Hospital de Braga se trataram de um conjunto de variáveis explicativas ao nível do paciente e do tumor, onde se incluíam medidas longitudinal de marcadores tumorais, fez-se uso de métodos estatísticos no âmbito da bioestatística para entender a progressão do cancro da mama nos pacientes deste hospital.

Nestes incluem-se métodos de análise de sobrevivência, em que o evento de interesse é a morte por cancro da mama e o tempo, em meses, é medido desde o diagnóstico de tumor maligno até ao momento em que ocorre o evento (ou até ao final do estudo, neste caso

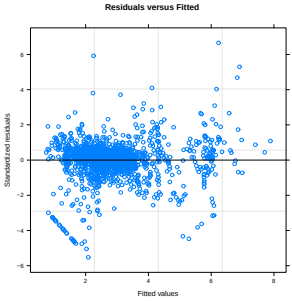


Figura 12: Gráfico de resíduos padronizados contra valores ajustados do marcador CA15-3.

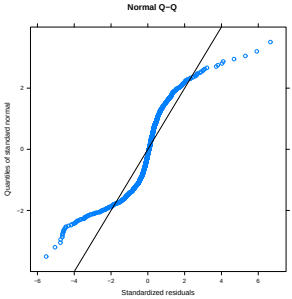


Figura 13: QQ plot of the CA15-3 longitudinal model.

particular foi considerada a data de 30/11/2014). E, ainda, métodos de análise longitudinal para compreender a progressão dos valores do marcador tumoral CA15-3. No entanto, como estes dois processos podem estar associados, uma vez que valores de CA15-3 acima do valor de referência de 37 U/ml são um sinal de alerta para uma possível recidiva do tumor (que poderá levar à morte por cancro da mama), é importante modelar conjuntamente estes dois processos. Ignorar uma possível associação poderá levar, como vários autores

apontam, a uma estimação enviesada dos parâmetros de interesse.
Após a realização de análises de sobrevivência e longitudinal separadas procedeu-se à modelação conjunta destes dois processos para inferir quanto à associação destes. Na análise de sobrevivência consideraram-se o modelo de riscos proporcionais de Cox [5] e o modelo paramétrico flexível de Royston-Parmar [25]. Para a análise longitudinal consideraram-se modelos longitudinais com várias estruturas de correlação definidos por Diggle, Heagerty, Liang e Zeger [9].
Da análise de sobrevivência individual resultaram como fatores de risco, estatisticamente significativos, na probabilidade de sobrevivência dos pacientes em estudo, as seguintes covariáveis: triplo negativo (sim contra não), idade ao momento do diagnóstico (categorizada em ≤ 44 anos de idade contra ≥ 44 anos de idade), estadio (0 ou I ou II contra III ou IV), grau de diferenciação do tumor (Gx ou G1 contra G2 contra G3) e envolvimento dos nódulos linfáticos regionais (Nx ou N0 ou N1 contra N2 ou N3). Constatou-se, ainda, que ignorar uma truncatura à esquerda no conjunto de dados em análise leva a uma sobrestimação das estimativas de sobrevivência.
Da análise longitudinal individual resultaram como fatores com efeito, estatisticamente significativo, na progressão média dos valores do marcador tumoral CA15-3, as seguintes covariáveis: idade ao diagnóstico, estadio (0 ou I ou II contra III ou IV), imagens de invasão vascular venosa (sim contra não) e expressão do biomarcador Ki.67 (baixo contra alto). Foi, ainda, possível constatar a existência de efeitos aleatórios ao nível dos indivíduos e, simultaneamente, uma correlação serial exponencial que representa a variabilidade das medições dentro de um mesmo indivíduo e não deve ser ignorada.
No entanto, modelando conjuntamente estes dois processos, adotando a metodologia de efeitos aleatórios [29], resulta que a associação entre os processos longitudinal e de sobrevivência é significativa sendo essencial o seu reconhecimento. Surgiram, tal como esperado, diferenças nas estimativas dos parâmetros relativos ao processo longitudinal e de sobrevivência comparando com as obtidas nas análises individuais respetivas. Levando, inclusive, que covariáveis com efei-

tos significativos em ambos os processos deixassem de o ter no modelo conjunto. No processo de sobrevivência, a idade ao diagnóstico deixou de ser um fator de risco na probabilidade de sobrevivência destes pacientes e no processo longitudinal as covariáveis estadio e ki.67 deixaram de ter efeito significativo na progressão média dos valores de CA15-3.

Dos modelos conjuntos considerados demonstra-se, ainda, que os indivíduos com valores elevados do marcador tumoral CA15-3 tendem a ter um risco mais elevado de morte por cancro da mama.

Da presente análise ressalta a necessidade de se modelar conjuntamente dados médicos como os considerados neste estudo, onde se tem acesso a dados de sobrevivência e longitudinais, de modo a obterem-se estimativas mais credíveis e uma perceção adequada do processo inerente a uma doença tão complexa como o cancro da mama. Ainda, uma análise desta natureza não seria possível sem o acesso a dados específicos a nível do paciente e do tumor o que realça a importância no investimento contínuo em estudos epidemiológicos implementados, por exemplo, nas unidades de senologia dos hospitais.

Como trabalho futuro tenciona-se incorporar modelos de cura na modelação conjunta, modelação conjunta de dados longitudinais multivariados (onde se consideram ambos os marcadores CEA e CA15-3 num mesmo modelo) e de sobrevivência e, ainda, o desenvolvimento de modelos de previsão individual.

Agradecimentos

Os autores agradecem o financiamento pela FCT-Fundação para a Ciência e Tecnologia. A autora Ana Borges desenvolveu o trabalho durante o doutoramento sendo bolseira da FCT com bolsa de referência SFRH/BD/74166/2010.

Referências

[1] Borges, A., Sousa, I., Castro, L.(2015). Longitudinal Analysis of Tumor Marker CEA of Breast Cancer Patients from Braga's Hospital. *REVSTAT Statistical Journal* 13(1), 63–78.

[2] Boyle, P., Ferlay, J. (2004). Cancer incidence and mortality in Europe. *Ann Oncol.* 16(3), 481–488.

[3] Cianfrocca, M., Goldstein, L.J. (2004). Prognostic and predictive factors in early-stage breast cancer. *The Oncologist* 9(6), 606–616.

[4] Cox, D.R., Snell, E.J. (1968). A general definition of residuals. *J. R. Stat. Soc., Ser. B* 30, 248–265, Discussion 265–275.

[5] Cox, DR. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, series B* 34, 87–220.

[6] David, H.A., Moeschberger, M.L. (1978). The Theory of Competing Risks. Griffin, London.

[7] Dawson, S.J., Provenzano, E., Caldas C. (2009). Triple negative breast cancers: clinical and prognostic implications. *Eur J Cancer* 45(1), 27–40.

[8] Diggle, P.J., Sousa, I., Chetwynd, A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine* 27, 2981–2998.

[9] Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S., (2002). Analysis of Longitudinal Data, 2nd edition. Oxford, England: Oxford University Press.

[10] dos Santos Silva, I., (1999). The role of cancer registries. In: dos Santos Silva I, ed. Cancer epidemiology. Principles and methods. Lyon: IARC: 385–403.

[11] Fitzgibbons, P.L., Page, D.L., Weaver, D., Thor, A.D., Allred, D.C. Clark, G.M. (2000). Prognostic factors in breast cancer, College of American Pathologists Consensus Statement 1999. *Archives of Pathology & Laboratory Medicine* 124(7), 966–978.

[12] Grambsch, P., Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–526.

[13] Harrington, D.P., Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika* 69, 553–566.

[14] Henderson, R., Diggle, P.J., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1(4), 465–480.

[15] Jackson, C.H. (2015). flexsurv: A Platform for Parametric Survival Modelling in R. version 0.6, http://CRAN.R-project.org/package=flexsurv.

[16] Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481.

[17] Klein, J.P., Moeschberger, M.L. (2003). *Survival Analysis Techniques for Censored and Truncated Data Second Edition.* Springer.

[18] Macià, F., Porta, M., Murta-Nascimento, C., Servitja, S., Guxens, M., Burón, A., Tusquets, I., Albanell, J., Castells, X. (2012). Factors affecting 5- and 10-year survival of women with breast cancer: an analysis based on a public general hospital in Barcelona. *Cancer Epidemiol.* 36(6), 554–559.

[19] McCrink, L.M., Marshall, A.H., Cairns, K.J. (2013). Advances in Joint Modelling: A Review of Recent Developments with Application to the Survival of End Stage Renal Disease Patients. *International Statistical Review*, 249–269.

[20] Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. and R Core Team (2012). JoineR: Joint modelling of repeated measurements and time-to-event data. R package version: 1.0-3, http://CRAN.R-project.org/package=joineR.

[21] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2015). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-121, http://CRAN.R-project.org/package=nlme.

[22] Pinheiro, P.S., Tyczynski, J.E., Bray, F., Amado, J., Matos, E., Parkin, D.M. (2003) Cancer incidence and mortality in Portugal. *Eur J Cancer* 39(17), 2507–2520.

[23] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[24] Rodrigues, V. (2011). Chapter 34. In *Manual de Ginecologia*, Permanyer, Portugal, 175–191.

[25] Royston, P., Parmar, M.K.B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21, 2175–2197.

[26] Sousa, I. (2011). A Review on Joint Modelling of Longitudinal Measurements and Time-to-event. *REVSTAT* 9(1), 57–81.

[27] Therneau, T. (2015). A Package for Survival Analysis in S. version 2.38, http://CRAN.R-project.org/package=survival.

[28] Trichopoulos, D., Adami, H.O., Ebkom, A., Hsieh, C.C., Lagiou, P. (2008). Early life events and conditions and breast cancer risk: from epidemioloy to etiology. *Int. J. Cancer* 122, 481–485.

[29] Wulfsohn, M.S., Tsiatis, A.A. (1997). A Joint model for survival and longitudinal data measured with error. *Biometrics* 53, 330–339.

[30] Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. *Chapman & Hall Book*.

Redução do viés do estimador de Hill: uma nova abordagem

Ivanilda Cabral
CMA, Universidade Nova de Lisboa
Universidade de Cabo Verde, *ivanilda.cabral@docente.unicv.edu.cv*

Frederico Caeiro
CMA e FCT, Universidade Nova de Lisboa, *fac@fct.unl.pt*

M. Ivette Gomes
CEAUL e DEIO, Universidade de Lisboa, *ivette.gomes@fc.ul.pt*

Palavras–chave: índice de valores extremos; estimação semi-paramétrica; redução do viés;

Resumo: Neste trabalho introduzimos um novo estimador do índice de valores extremos positivo, que resulta da redução de viés do clássico estimador de Hill. As propriedades assintóticas deste novo estimador são estudadas sob a validade duma condição de variação regular de terceira ordem e assumindo que os parâmetros de segunda ordem são conhecidos. Usamos também o método de simulação de Monte Carlo para analisar o comportamento do novo estimador para amostras de dimensão finita.

1 Introdução

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (*i.i.d*) de um modelo F, cuja ordenação ascendente resulta nas v.a.'s $(X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n})$, denominadas estatísticas ordinais crescentes. Vamos assumir que F pertence ao domínio de atração para máximos de uma função de distribuição G, isto é, que existem sequências de constantes normalizadoras a_n e

$b_n > 0$ tais que $P((X_{n:n} - a_n)/b_n \leq x) = F^n(a_n + b_n x) \underset{n\to\infty}{\longrightarrow} G(x)$ em todos os pontos de continuidade de G e escrevemos $F \in D(G)$. Caso G seja uma função de distribuição não degenerada, G é a distribuição de valores extremos $G_\xi(x) = \exp\{-(1+\xi x)_+^{-1/\xi}\}$, $\xi \in \mathbb{R}$, $x_+ = \max(x,0)$. O parâmetro de forma ξ é usualmente conhecido por *índice de valores extremos* e é o parâmetro que pretendemos estimar. Esta distribuição representa de modo unificado as três possíveis distribuições limite max-estáveis: Weibull ($\xi < 0$), Gumbel ($\xi = 0$) ou Fréchet ($\xi > 0$). Neste trabalho, admitimos ainda que F é um modelo de cauda pesada, ou seja, que $F \in D(G_\xi)$ com $\xi > 0$. Seja $U(t) = F^{\leftarrow}(1 - 1/t)$, $t \geqslant 1$, com $F^{\leftarrow}(t) = \inf\{x : F(x) \geqslant t\}$ a inversa generalizada de F. A condição necessária e suficiente que garante a convergência do máximo normalizado para G_ξ, $\xi > 0$, é que $U \in RV_\xi$, onde RV_α denota a classe das funções de variação regular em infinito de índice α, isto é, a classe das funções mensuráveis positivas $f(.)$ tais que $f(tx)/f(t) \underset{t\to\infty}{\longrightarrow} x^\alpha$, $\forall x > 0$, ([6]).

1.1 Estimação do índice de valores extremos

Nos modelos de cauda pesada, o estimador de Hill [12]

$$\hat{\xi}^H(k) = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{X_{n-i+1:n}}{X_{n-k:n}}, \quad k = 1, 2, \ldots, n-1, \tag{1}$$

é uma das principais referências sobre a estimação do parâmetro ξ. Este estimador é consistente caso $F \in D(G_\xi)$ com $\xi > 0$ e k represente uma sequência intermédia, isto é, uma sequência de valores inteiros ($1 \leqslant k \leqslant n-1$) tal que:

$$k = k_n \longrightarrow \infty \quad \text{e} \quad k/n \longrightarrow 0, \quad n \to \infty. \tag{2}$$

Para obtermos a distribuição limite de $\hat{\xi}^H(k)$, necessitamos de assumir a validade da seguinte condição de segunda ordem, relativa à velocidade de convergência de $U(tx)/U(t)$ para x^ξ,

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \frac{x^\rho - 1}{\rho}, \quad \forall x > 0 \tag{3}$$

onde $\rho \leq 0$ é um parâmetro de forma de segunda ordem que mede a velocidade de convergência. Quanto maior for o valor de $|\rho|$, maior será essa velocidade. A função $A(t)$ é de sinal constante para valores elevados de t, verificando $|A(t)| \in RV_\rho$, [5]. Se $\rho = 0$, consideramos o prolongamento por continuidade $\frac{x^\rho - 1}{\rho} \equiv \ln x$. Neste trabalho, vamos considerar a parametrização $A(t) = \xi\beta t^\rho$ com $\beta \neq 0$ parâmetro de "escala" de segunda ordem e $\rho < 0$.
Se as condições (2) e (3) forem satisfeitas, o estimador de Hill possui a seguinte representação assintótica em distribuição

$$\sqrt{k}(\hat{\xi}^H(k) - \xi) \stackrel{d}{=} \xi Z_k + \frac{\sqrt{k}A(n/k)}{1-\rho}(1 + o_p(1)), \qquad (4)$$

onde $Z_k = \sum_{i=1}^{k}(E_i - 1)/\sqrt{k}$ é assintoticamente normal padrão e E_i, $i = 1, 2, \ldots, k$ é uma sucessão de v.a.'s exponenciais unitárias independentes. Este estimador apresenta usualmente um viés assintótico acentuado quando k aumenta. A redução do viés de $\hat{\xi}^H(k)$, em (1), e de outros estimadores clássicos tem sido um tema relevante na literatura recente (ver [1] e suas referências). Os primeiros estimadores de viés reduzido tinham sempre o usual "trade-off" entre a variância e o viés. O "trade-off" anteriormente referido foi ultrapassado com a introdução de estimadores MVRB (do inglês "minimum variance reduced bias"). Nestes estimadores o termo dominante do viés do estimador de Hill, $A(n/k)/(1-\rho) = \xi\beta(n/k)^\rho/(1-\rho)$ é estimado e removido, sem alterar o valor da sua variância assintótica. O estimador MVRB mais simples foi introduzido em [2]. Este estimador, denotado CH (do inglês "Corrected Hill") é dado por

$$\hat{\xi}^{CH}_{\hat{\beta},\hat{\rho}}(k) = \hat{\xi}^H(k)\Big(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\Big), \quad k = 1, 2, \ldots, n-1. \qquad (5)$$

Os estimadores habitualmente usados para estimar ρ e β são aqueles que foram introduzidos em [4] e [7], respectivamente. Podemos encontrar algoritmos para a estimação de (ρ, β) em [8], entre outros. Apesar de $\hat{\xi}^{CH}_{\hat{\beta},\hat{\rho}}(k)$ resultar da redução do termo dominante do viés

assintotico do estimador de Hill, este estimador é muitas vezes enviesado para valores de k mais elevados (ver, por exemplo, a aplicação a dados reais em [9]). O viés é usualmente positivo. Para remover, eventualmente de modo mais adequado, o viés de $\hat{\xi}^H(k)$, introduzimos o denominado estimador NCH (do inglês "new corrected Hill") com a expressão funcional

$$\hat{\xi}^{NCH}_{\hat{\beta},\hat{\rho}}(k) = \hat{\xi}^H(k)\left(2 - e^{\frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}}\right), \quad k = 1, 2, \ldots, n-1. \tag{6}$$

Note que $d_1 = 1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}$ e $d_2 = 2 - e^{\frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}}$ são equivalentes até à segunda ordem e, para amostras de dimensão finita, temos $d_2 < d_1$ se $\hat{\beta} > 0$. Para avaliar a potencial redução de viés dos estimadores introduzidos em (5) e (6), vamos considerar neste trabalho os parâmetros ρ e β conhecidos e usar a notação

$$\hat{\xi}^{CH}(k) = \hat{\xi}^{CH}_{\beta,\rho}(k) \qquad \text{e} \qquad \hat{\xi}^{NCH}(k) = \hat{\xi}^{NCH}_{\beta,\rho}(k). \tag{7}$$

As propriedades do estimador NCH estão a ser estudadas pelos autores e serão apresentadas e discutidas em trabalho futuro.

1.2 Conteúdo do trabalho

Na seção 2 apresentamos as propriedades e distribuições assintóticas dos estimadores do índice positivo de valores extremos, ξ, apresentados em (7), sob uma condição de terceira ordem, dando ênfase ao estimador NCH. Seguidamente, na mesma seção, fazemos a comparação, em níveis ótimos, entre os estimadores $\hat{\xi}^{CH}(k)$ e $\hat{\xi}^{NCH}(k)$. Terminamos apresentando na seção 3 um estudo de simulação de Monte Carlo para obter o comportamento dos estimadores em estudo para amostras de dimensão finita dos modelos Fréchet e Burr.

2 Propriedades assintóticas

2.1 Resultados para níveis intermédios

Para obter mais informação acerca do viés dos estimadores em estudo, vamos impor a validade da seguinte condição de terceira ordem: existe uma função $B(t)$ que mede a velocidade de convergência de (3) tal que, para todo o $x > 0$,

$$\lim_{t\to\infty} \frac{\frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} - \frac{x^{\rho}-1}{\rho}}{B(t)} = \frac{x^{\rho+\rho'}-1}{\rho+\rho'}, \tag{8}$$

onde $\rho' \leq 0$ é um parâmetro de terceira ordem e $|B(t)| \in RV_{\rho'}$. Vamos ainda assumir que $A(t) = \xi\beta t^{\rho}$, $B(t) = \beta' t^{\rho'}$, com $\beta, \beta' \neq 0$, e $\rho, \rho' < 0$, todos estes parâmetros de valor conhecido.
Enunciamos primeiro a seguinte proposição que apresenta a representação assintótica dos estimadores $\hat{\xi}^H(k)$ e $\hat{\xi}^{CH}(k)$, em (1) e (7), respetivamente. A demonstração pode ser consultada em [10].

Proposição 2.1 *Consideremos que são válidas as condições* (2) *e* (8) *e que conhecemos os valores dos parâmetros de segunda ordem* ρ *e* β*. Então, podemos assegurar as seguintes representações em distribuição*

$$\sqrt{k}(\hat{\xi}^H(k) - \xi) \stackrel{d}{=} \xi Z_k + \frac{\sqrt{k}A(n/k)}{1-\rho} + \frac{\sqrt{k}A(n/k)B(n/k)}{1-\rho-\rho'}(1 + o_p(1))$$

e

$$\sqrt{k}(\hat{\xi}^{CH}(k) - \xi) \stackrel{d}{=} \xi Z_k + \left(\frac{\sqrt{k}A(n/k)B(n/k)}{1-\rho-\rho'} - \frac{\sqrt{k}A^2(n/k)}{\xi(1-\rho)^2}\right)(1 + o_p(1)),$$

onde Z_k *é a v.a. introduzida em* (4)*.*

Observação 2.2 *Podemos concluir que quando* $\sqrt{k}A(n/k) \to \lambda_1$*, finito, não necessariamente nulo,* $\sqrt{k}(\hat{\xi}^H(k) - \xi)$ *e* $\sqrt{k}(\hat{\xi}^{CH}(k) - \xi)$ *têm assintoticamente distribuição normal de valor médio* $\lambda_1/(1-\rho)$ *e* 0*, respetivamente, e variância* ξ^2*. Se adicionalmente considerarmos valores de* k *tais que* $\lambda_1 = \infty$*,* $\sqrt{k}A(n/k)B(n/k) \to \lambda_2$ *e*

$\sqrt{k}A^2(n/k) \to \lambda_3$, *ambos finitos,* $\sqrt{k}(\hat{\xi}^{CH}(k)-\xi)$ *é assintoticamente normal com valor médio* $\lambda_2/(1-\rho-\rho') - \lambda_3/(\xi(1-\rho)^2)$ *e variância* ξ^2.

O comportamento assintótico do novo estimador $\hat{\xi}^{NCH}(k)$, em (7), é enunciado num contexto de terceira ordem na seguinte proposição.

Proposição 2.3 *Nas condições da* Proposição 2.1*, o novo estimador* $\hat{\xi}^{NCH}(k)$ *possui a seguinte representação distribucional assintotica:*

$$\sqrt{k}(\hat{\xi}^{NCH}(k) - \xi) \stackrel{d}{=} \xi Z_k + \left(\frac{\sqrt{k}A(n/k)B(n/k)}{1-\rho-\rho'} - \frac{3\sqrt{k}A^2(n/k)}{2\xi(1-\rho)^2}\right)(1+o_p(1))$$

onde Z_k *é uma sucessão de v.a.'s assintoticamente normais padrão.*

Dem.: Usando a parameterização $A(n/k) = \xi\beta(n/k)^\rho$ e a aproximação: $e^x = 1 + x + \frac{x^2}{2} + o(x^2)$, quando $x \to 0$, temos

$$2 - e^{\frac{\beta(n/k)^\rho}{1-\rho}} = 2 - e^{\frac{A(n/k)}{\xi(1-\rho)}} = 1 - \frac{A(n/k)}{\xi(1-\rho)} - \frac{A^2(n/k)}{2\xi^2(1-\rho)^2}(1+o(1)).$$

Depois, usando o resultado da Proposição 2.1,

$$\begin{aligned}\hat{\xi}^{NCH}(k) &\stackrel{d}{=} \hat{\xi}^{H}(k)\left[1 - \tfrac{A(n/k}{\xi(1-\rho)} - \tfrac{1}{2}\left(\tfrac{A(n/k)}{\xi(1-\rho)}\right)^2(1+o_p(1))\right] \\ &\stackrel{d}{=} \xi + \tfrac{\xi}{\sqrt{k}}Z_k + \left(\tfrac{A(n/k)B(n/k)}{1-\rho-\rho'} - \tfrac{3A^2(n/k)}{2\xi(1-\rho)^2}\right)(1+o_p(1)).\end{aligned}$$

Consequentemente, podemos obter o resultado pretendido. ■

Como consequência das duas proposições anteriores, verificamos que os três estimadores do índice de valores extremos, definidos em (1) e (7), têm a mesma variância assintótica. Relativamente ao viés, verificamos que o termo dominante do viés assintótico dos estimadores NCH e CH é de ordem inferior a $A(n/k)$. Consequentemente, $\hat{\xi}^{NCH}(k)$ e $\hat{\xi}^{CH}(k)$ são estimadores MVRB de ξ. Temos também $\hat{\xi}^{NCH}(k) - \hat{\xi}^{CH}(k) \stackrel{d}{=} -A^2(n/k)(1+o_p(1))/(2\xi(1-\rho)^2)$, um resultado útil quando o viés dos dois estimadores tem o mesmo sinal.

2.2 Comparação assintótica em níveis ótimos

Nesta subseção, vamos comparar assintoticamente o estimador NCH com o estimador CH, nos respetivos níveis ótimos. A comparação será feita de modo similar à comparação em [11] para estimadores clássicos e em [3] para estimadores de viés reduzido. Aqui consideramos modelos de cauda pesada que verificam (8), com $\rho = \rho' < 0$, $A(t) = \xi\beta t^{\rho}$ e $B(t) = \beta' t^{\rho} = \zeta A(t)/\xi$, com $\zeta = \beta'/\beta$. Estas condições são verificadas por vários modelos de cauda pesada ([3]). Mencionamos, por exemplo, os modelos

- Fréchet, com função de distribuição $F(x) = \exp(-x^{-1/\xi})$, $x \geq 0$, $\xi > 0$ $(\rho' = \rho = -1,\ \beta = 0.5$ e $\beta' = 5/6)$;
- Burr com função de distribuição $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x \geq 0$, $\xi > 0$, $\rho < 0$ $(\rho' = \rho < 0$ e $\beta = \beta' = 1)$;
- t de Student, t_ν, com $\nu \in \mathbb{R}^+$ graus de liberdade. Os parâmetros ξ, ρ, ρ' e ζ são dados por $\xi = 1/\nu$, $\rho' = \rho = -2/\nu$ e $\zeta = \beta'/\beta = (\nu^2 + 4\nu + 2)/((\nu + 1)(\nu + 4)) \in (0.5, 1)$.

Vamos denotar por $\hat{\xi}^{\bullet}(k)$ um qualquer dos dois estimadores MVRB em (7). Então temos

$$\hat{\xi}^{\bullet}(k) \stackrel{d}{=} \xi + \frac{\sigma_{\bullet}}{\sqrt{k}} Z_k + b_{\bullet} A^2(n/k)(1 + o_p(1)), \tag{9}$$

onde Z_k é a sucessão de variáveis aleatórias introduzida em (4). A variância e o viés assintótico do estimador $\hat{\xi}^{\bullet}(k)$ são dados por $\sigma_{\bullet}^2/k$ e $b_{\bullet} A^2(n/k)$, respetivamente. O erro quadrático médio assintótico (AMSE) é então dado por $\mathrm{AMSE}[\hat{\xi}^{\bullet}(k)] = \sigma_{\bullet}^2/k + b_{\bullet}^2 A^4(n/k)$. Considerando dois estimadores $\hat{\xi}^{(1)}(k)$ e $\hat{\xi}^{(2)}(k)$ para os quais é válida a representação em (9), calculados nos respetivos níveis óptimos, $k_0^{(j)} := \arg\min_k \mathrm{AMSE}[\hat{\xi}^{(j)}(k)]$, e a notação $\hat{\xi}_0^{(j)} = \hat{\xi}^{(j)}(k_0^{(j)})$, j = 1,2, a eficiencia relativa assintótica, ARE (do inglês "Asymptotic Relative Efficiency"), de $\hat{\xi}_0^{(1)}$ relativamente a $\hat{\xi}_0^{(2)}$ é obtida através do

seguinte indicador ([3]):

$$ARE_{1|2} = ARE_{\hat{\xi}_0^{(1)}|\hat{\xi}_0^{(2)}} = \left[\left(\frac{\sigma_2}{\sigma_1}\right)^{-4\rho}\left|\frac{b_2}{b_1}\right|\right]^{\frac{1}{1-4\rho}}.$$

Quanto maior for o valor de $ARE_{1|2}$, melhor é o estimador $\hat{\xi}_0^{(1)}$. Para os estimadores em estudo, $\sigma_{CH} = \sigma_{NCH} = \xi$, $b_{CH} = \frac{\zeta}{\xi(1-2\rho)} - \frac{1}{\xi(1-\rho)^2}$ e $b_{NCH} = \frac{\zeta}{\xi(1-2\rho)} - \frac{3}{2\xi(1-\rho)^2}$. Consequentemente,

$$ARE_{NCH|CH} = \left[\left|\frac{2\zeta(1-\rho)^2 - 2(1-2\rho)}{2\zeta(1-\rho)^2 - 3(1-2\rho)}\right|\right]^{\frac{1}{1-4\rho}}.$$

Para o modelo Fréchet, a eficiência relativa assintótica é $ARE_{NCH|CH} = 1.11$, o que significa que, para este modelo, o estimador $\hat{\xi}^{NCH}(k)$ é assintoticamente mais eficiente do que o estimador $\hat{\xi}^{CH}(k)$, no respetivo nível ótimo. Relativamente ao modelo Burr, o valor do indicador $ARE_{NCH|CH}$ é superior a 1 se e só se $\rho < -0.809$. Para o modelo t de Student temos $ARE_{NCH|CH} > 1$ se e só se $0 < \nu < 1.02$.

3 Estudo de Simulação

3.1 Metodologia e resultados

Nesta seção apresentamos um estudo de simulação para analisar o comportamento dos estimadores, $\hat{\xi}^{H}$, $\hat{\xi}^{CH}$ e $\hat{\xi}^{NCH}$. Os resultados baseiam-se em 1000 amostras de dimensão n, para diferentes valores de n, dos modelos Fréchet com $\xi = 0.5$ e Burr com $\xi = 0.5$ e $\rho = -1$ e -0.5. Determinámos, para cada amostra de dimensão n, as estimativas $\hat{\xi}_i^{\bullet}(k)$, $k = 1, 2, \ldots, n-1$, $i = 1, 2, \ldots, 1000$, que nos permitiram obter estimativas do valor médio (E) e da raíz quadrada do erro quadrático médio (RMSE) dados por

$$E[\hat{\xi}^{\bullet}(k)] = \sum_{i=1}^{1000} \frac{\hat{\xi}_i^{\bullet}(k)}{1000} \quad \text{e} \quad RMSE[\hat{\xi}^{\bullet}(k)] = \sqrt{\sum_{i=1}^{1000} \frac{(\hat{\xi}_i^{\bullet}(k) - \xi)^2}{1000}}. \quad (10)$$

Apresentamos nas Figuras 1, 2 e 3 os valores simulados de $E[\hat{\xi}^{\bullet}(k)]$ e $RMSE[\hat{\xi}^{\bullet}(k)]$ em (10), para amostras de dimensão $n = 1000$ dos modelos em estudo. Com base nos valores dados por (10), determinámos $\hat{k}_0^{\bullet} = \arg\min_k RMSE[\hat{\xi}^{\bullet}(k))]$ com o qual obtivemos

$$E[\hat{\xi}_0^{\bullet}] = E[\hat{\xi}^{\bullet}(\hat{k}_0^{\bullet}))] \quad \text{e} \quad RMSE[\hat{\xi}_0^{\bullet}] = RMSE[\hat{\xi}^{\bullet}(\hat{k}_0^{\bullet})]. \tag{11}$$

Na Tabela 1 apresentamos os valores dos indicadores dados em (11), para várias dimensões de amostras e os modelos e valores de parâmetros usados nas figuras.

3.2 Conclusões

Analisando os gráficos e as tabelas, observamos que para os modelos Fréchet e Burr, em todas as dimensões de amostras consideradas, os valores médios simulados dos estimadores CH e NCH, no nível ótimo, estão mais próximas do verdadeiro valor de ξ do que os valores médios simulados do estimador de Hill no seu nível ótimo. No entanto, para os modelos Fréchet com $\xi = 0.5$ e Burr com $(\xi, \rho) = (0.5, -1)$, em todas as dimensões de amostras consideradas, os valores médios simulados do estimador NCH, no nível ótimo, são os que se encontram mais próximos do verdadeiro valor de ξ. O mesmo não acontece para o modelo Burr com $(\xi, \rho) = (0.5, -0.5)$ onde o estimador CH apresenta o melhor valor médio simulado.
Relativamente ao RMSE, no nível ótimo, o estimador NCH apresenta o menor valor, comparativamente com o RMSE dos estimadores de Hill e CH nos modelos Fréchet com $\xi = 0.5$ e Burr com $(\xi, \rho) = (0.5, -1)$. A exceção é no modelo Burr com $(\xi, \rho) = (0.5, -0.5)$.

Agradecimentos

À SPE e à FCT-UNL pelo apoio financeiro concedido a Ivanilda Cabral. Investigação parcialmente suportada por fundos nacionais através da FCT-Fundação para a Ciência e a Tecnologia, projectos PEst-OE/MAT/UI006/2014 (CEA/UL) e UID/MAT/00297/2013 (CMA/UNL).

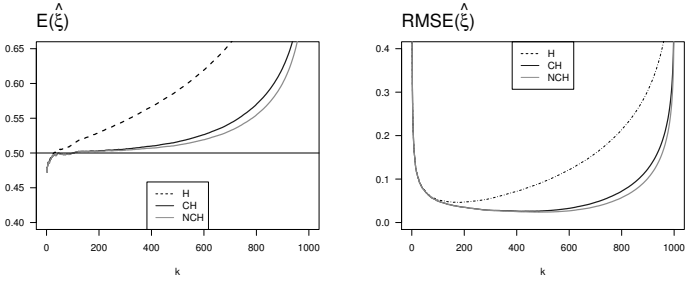


Figura 1: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 1000$ do modelo *Fréchet* com $\xi = 0.5$.

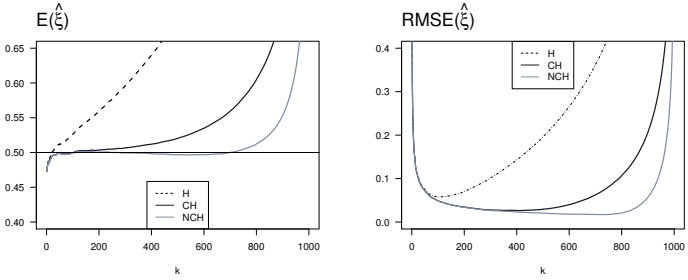


Figura 2: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 1000$ do modelo *Burr* com $(\xi, \rho) = (0.5, -1)$.

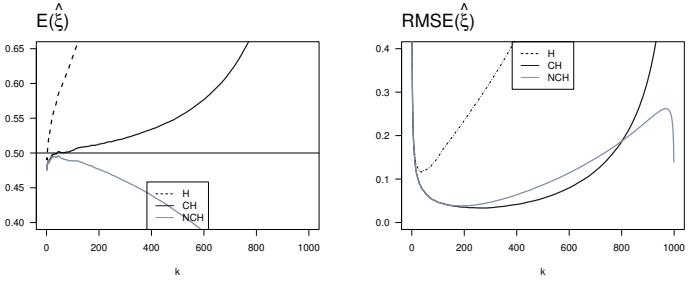


Figura 3: Valores médios (esquerda) e RMSE (direita) simulados para amostras de dimensão $n = 1000$ do modelo *Burr* com $(\xi, \rho) = (0.5, -0.5)$.

Tabela 1: Estimativas do valor esperado, no nível ótimo, dos estimadores do índice de valores extremos, $\hat{\xi}^{H}(k)$, $\hat{\xi}^{CH}(k)$ e $\hat{\xi}^{NCH}(k)$.

Fréchet $\xi = 0.5$ ($\rho = -1$, $\beta = 0.5$)							
n	100	200	500	1000	2000	5000	10000
$\hat{\xi}_0^H$	0.560	0.538	0.533	0.527	0.521	0.514	0.512
$\hat{\xi}_0^{CH}$	0.526	0.522	0.515	0.511	0.509	0.506	0.505
$\hat{\xi}_0^{NCH}$	**0.519**	**0.518**	**0.514**	**0.510**	**0.508**	**0.505**	**0.504**
Burr $\xi = 0.5$, $\rho = -1$ ($\beta = 1$)							
$\hat{\xi}_0^H$	0.570	0.543	0.542	0.527	0.525	0.519	0.513
$\hat{\xi}_0^{CH}$	0.528	0.519	0.516	0.513	0.508	0.506	0.505
$\hat{\xi}_0^{NCH}$	**0.514**	**0.503**	**0.502**	**0.504**	**0.501**	**0.500**	**0.500**
Burr $\xi = 0.5$, $\rho = -0.5$ ($\beta = 1$)							
$\hat{\xi}_0^H$	0.651	0.617	0.593	0.570	0.556	0.552	0.544
$\hat{\xi}_0^{CH}$	**0.534**	**0.530**	**0.519**	**0.519**	**0.513**	**0.509**	**0.507**
$\hat{\xi}_0^{NCH}$	0.456	0.466	0.470	0.481	0.482	0.488	0.490

Tabela 2: Estimativas da raíz quadrada do erro quadrático médio, no nível ótimo, dos estimadores do índice de valores extremos, $\hat{\xi}^{H}(k)$, $\hat{\xi}^{CH}(k)$ e $\hat{\xi}^{NCH}(k)$.

Fréchet $\xi = 0.5$ ($\rho = -1$, $\beta = 0.5$)							
n	100	200	500	1000	2000	5000	10000
$\hat{\xi}_0^H$	0.104	0.081	0.060	0.046	0.036	0.026	0.021
$\hat{\xi}_0^{CH}$	0.069	0.051	0.034	0.026	0.020	0.014	0.010
$\hat{\xi}_0^{NCH}$	**0.066**	**0.048**	**0.032**	**0.024**	**0.018**	**0.013**	**0.010**
Burr $\xi = 0.5$, $\rho = -1$ ($\beta = 1$)							
$\hat{\xi}_0^H$	0.131	0.103	0.074	0.057	0.045	0.033	0.026
$\hat{\xi}_0^{CH}$	0.071	0.053	0.035	0.026	0.020	0.013	0.010
$\hat{\xi}_0^{NCH}$	**0.056**	**0.038**	**0.024**	**0.017**	**0.012**	**0.008**	**0.005**
Burr $\xi = 0.5$, $\rho = -0.5$ ($\beta = 1$)							
$\hat{\xi}_0^H$	0.238	0.187	0.141	0.116	0.093	0.073	0.062
$\hat{\xi}_0^{CH}$	**0.080**	**0.063**	**0.043**	**0.033**	**0.025**	**0.019**	**0.014**
$\hat{\xi}_0^{NCH}$	**0.080**	0.066	0.049	0.038	0.030	0.022	0.018

Referências

[1] Beirlant, J., Caeiro, F., Gomes, M.I. (2012). An overview and open research topics in statistics of univariate extremes. *Revstat* 10(1), 1–31.

[2] Caeiro, F., Gomes, M.I., Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator. *Revstat* 3(2), 113–136.

[3] Caeiro, F., Gomes, M.I. (2011). Asymptotic comparison at optimal levels of reduced-bias extreme value index estimators. *Statistica Neerlandica* 65(4), 462–488.

[4] Fraga Alves, M.I., Gomes, M.I., de Haan, L. (2003). A new class of semiparametric estimators of the second order parameter. *Portugaliae Mathematica* 60(2), 193–213.

[5] Geluk, J., de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems.* Tech. Report CWI Tract 40, Centre for Mathematics and Computer Science, Amsterdam, Netherlands.

[6] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* 44, 423–453.

[7] 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.

[8] Gomes, M.I., Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal American Statistical Association* 102, 280–292.

[9] Gomes, M.I., Henriques-Rodrigues, L., Fraga Alves, M.I., Manjunath, B.G. (2013). Adaptative PORT-MVRB estimation: an empirical comparison of two heuristic algorithms. *J. Statist. Comput. Simul.* 83(6), 1129–1144.

[10] Gomes, M.I., Pestana, D., Caeiro, F. (2009). A note on the asymptotic variance at optimal Levels of bias-corrected Hill estimator. *Statistics Probability Letters* 79(3), 295–303.

[11] de Haan, L., Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* 52(1), 60–70.

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

Máximo de um modelo Ψ-INARMA

Sandra Dias
Pólo-CMAT e CEMAT, Dep. de Matemática, Universidade de Trás-os-Montes e Alto Douro, *sdias@utad.pt*

Maria da Graça Temido
CMUC, Dep. de Matemática, Fac. de Ciências e Tecnologia, Universidade de Coimbra, *mgtm@mat.uc.pt*

Palavras–chave: teoria de valores extremos, classe de Anderson, operador aleatório, modelos ARMA

Resumo: Estudamos um processo de variáveis inteiras que designamos $\Psi-$INARMA(1,1). Depois de provada a estacionaridade forte do processo, estabelecemos o comportamento de independência assintótica e de dependência local. Concluímos que a sucessão de máximos é atraída em distribuição para uma distribuição de Gumbel discreta, quando a sucessão residual pertence à classe de Anderson.

1 Introdução

Dada uma variável aleatória (v.a.) inteira X e $\eta \in]0,1[$, Aly e Bouzar [1] introduzem o operador aleatório $\odot_\Psi$, o qual faz corresponder ao par (η,X) a variável operada $\eta \odot_\Psi X \equiv Y_1 + Y_2 + \cdots + Y_X$, onde $\{Y_i\}$ é uma sucessão de v.a.'s i.i.d., independente de X, com função geradora de probabilidades (f.g.p.) $\Psi_t(z)$, $t = -\ln \eta$, pertencente a uma família específica. Nomeadamente, Ψ_t terá de verificar

$$\Psi_{t_1+t_2}(z) = \Psi_{t_1}(\Psi_{t_2}(z)), \quad |z| \leq r_Y$$

com $\Psi(0) \neq 0$, isto é, $P(Y = 0) > 0$. Neste trabalho r_Y representa o raio de convergência da f.g.p. da v.a. Y. A solução geral desta equação funcional é (para além da função identidade)

$\Psi_t(z) = g^{-1}(g(z) \pm t)$, onde g é uma função estritamente crescente. Em particular temos a família de f.g.p.

$$\Psi_t^{(\theta)}(z) = 1 - \frac{\bar{\theta}e^{-\bar{\theta}t}(1-z)}{\bar{\theta} + \frac{\theta}{2}(1-e^{-\bar{\theta}t})(1-z)}, \; |z-1| < \frac{2\bar{\theta}}{\theta(1-e^{-\bar{\theta}t})},$$

com $t \geq 0, \theta \in [0,1[$ e $\bar{\theta} = 1-\theta$. Se $\theta = 0$ então $\Psi_t^{(0)}(z) = 1-e^{-t}+e^{-t}z$, caso em que as v.a.'s Y_i possuem lei de Bernoulli. Neste caso particular $\odot_\Psi$ coincide com o operador aleatório binomial, introduzido por van Harn *et al.* [4], aqui denotado por $\star$, e bem conhecido nestes contextos. Aly e Bouzar [1] estudam o processo Ψ-INAR(1) descrito pela equação $X_n = \eta \odot_\Psi X_{n-1} + \epsilon_n$, onde $0 < \eta < 1$ e $\{\epsilon_n\}$ é uma sucessão de v.a.'s inteiras i.i.d., independentes das variáveis Y_i. Em McKenzie [6] é introduzido o processo ARMA(1,1) geométrico definido por $X_n = \beta \star Z_n + V_n W_{n-1}$, com $W_n = \eta \star W_{n-1} + U_n Z_n$, onde $\{Z_n\}$, $\{U_n\}$ e $\{V_n\}$ são sucessões de v.a.'s i.i.d., $\{U_n\}$ e $\{V_n\}$ têm ambas distribuição de Bernoulli (com parâmetros $1-\eta$ e $1-\beta$, respetivamente) e W_0 é independente de todas as outras v.a.'s.
Neste trabalho consideramos uma extensão do modelo de McKenzie na forma

$$X_n = \beta \odot_\Psi Z_n + V_n W_{n-1}, \text{ onde } W_n = \eta \odot_\Psi W_{n-1} + U_n Z_n,$$

com $\{Z_n\}$, $\{U_n\}$ e $\{V_n\}$ sob as mesmas hipóteses, a que chamamos processo Ψ-INARMA(1,1). Depois de provarmos a estacionaridade forte do processo, validamos o comportamento de independência assintótica e de dependência local induzido pelas condições $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$, introduzidas em Temido e Canto e Castro [8], onde $\{k_n\}$ é uma sucessão de inteiros não decrescente tal que $k_{n+1}/k_n \to r > 1, n \to +\infty$.
Recordemos tais condições. Seja $\{k_n\}$ uma sucessão crescente de inteiros positivos e $\{u_n\}$ uma sucessão de reais. A sucessão $\{X_n\}$ verifica $D_{k_n}(u_n)$ se, para quaisquer inteiros $1 \leq i_1 < ... < i_p < j_1 < ... < j_q \leq k_n$, com $j_1 - i_p > \ell_n$ e $A_j := \{X_j \leq u_n\}$, se tem

$$\left| P\big(\cap_{s=1}^p A_{i_s}, \cap_{m=1}^q A_{j_m}\big) - P\big(\cap_{s=1}^p A_{i_s}\big) P\big(\cap_{m=1}^q A_{j_m}\big) \right| \leq \alpha_{n,\ell_n},$$

onde $\lim_{n\to+\infty} \alpha_{n,\ell_n} = 0$, para alguma sucessão $\ell_n = o_n(k_n)$.
Temido e Canto e Castro [8] provam que sob $D_{k_n}(u_n)$ o limite em distribuição do máximo $M_{k_n} = \max(X_1,X_2,...,X_{k_n})$, se existir, é max-semiestável. Mais, a condição $D'_{k_n}(u_n)$, ocorre se existir uma sucessão de inteiros positivos $\{s_n\}$ tal que $k_n/s_n \to +\infty$, $s_n\alpha_{n,l_n} \to 0$ e

$$\lim_{n\to+\infty} k_n \sum_{j=2}^{[k_n/s_n]} P\left(X_1 > u_n, X_j > u_n\right) = 0.$$

Estas condições são adaptações das conhecidas condições de Leadbetter *et al.* [5] ao contexto de max-semiestabilidade.
Como se espera, sob $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$, as sucessões $\{P(M_{k_n} \leq x + b_n)\}$ e $\{F_X^{k_n}(x + b_n)\}$, quando convergentes, possuem o mesmo limite. Quanto à distribuição marginal do processo, assumimos que $\{Z_n\}$ pertence à classe de Anderson [2], isto é, à classe das f.d.'s F que verificam $(1 - F(n - 1))/(1 - F(n)) \to r > 1$, $n \to +\infty$. Em particular, perante a necessidade de recorrer a resultados de Hall [3], consideramos a subclasse da classe de Anderson constituída pelas f.d.'s que satisfazem

$$1 - F(z) \sim A[z]^{\xi} r^{-[z]},\ z \to +\infty,$$

onde $\xi \in \mathbb{R}$, $A > 0$ e $r > 1$, a qual será denotada por $\mathcal{C}_{\mathcal{A}}(r)$. Sob esta hipótese, provamos que o mesmo sucede com $\{W_n\}$ e $\{X_n\}$. Acrescente-se que em Temido [7] se prova que se $\{Z_n\}$ pertence à classe de Anderson, então $\{F_Z^{k_n}(x+b_n)\}$ converge para a f.d Gumbel discreta, $G(x) = \exp(-r_Z^{-[x]})$, $x \in \mathbb{R}$.
Estabelecidos estes resultados, concluímos que a sucessão de máximos do processo Ψ-INARMA(1,1) é atraída em distribuição para uma distribuição de Gumbel discreta.

2 Propriedades da variável operada

Nesta secção caracterizamos a f.g.p. $P_{\eta\odot_\Psi Z}$, assumindo que a f.g.p. Ψ_t, associada ao operador aleatório $\odot_\Psi$, possui raio de convergência

superior a 1. Consideremos a fórmula de Taylor de ordem 2 numa vizinhança do ponto $s = 1$, $\Psi_t(s) = \Psi_t(1) + \Psi_t^{'}(1)(s-1) + \frac{\Psi_t^{''}(\vartheta)}{2}(s-1)^2 = 1 + E(Y)(s-1) + \frac{\Psi_t^{''}(\vartheta)}{2}(s-1)^2$, onde $|\vartheta - 1| < |s-1|$ e $\vartheta := \vartheta(s)$. Considerando $h = s - 1$ e $\xi := \frac{\Psi_t^{''}(\vartheta)}{2}$, obtemos ainda $\Psi_t(1+h) = 1 + E(Y)h + \xi h^2 = 1 + E(Y)f(h)$, com $f(h) = h(1 + \xi h/E(Y))$. Em Aly e Bouzar [1] é estabelecido que $E(Y) = \eta^{\delta_\Psi}$, com $\delta_\Psi = -\ln \Psi_1^{'}(1)$, tendo-se $E(\eta \odot_\Psi Z) = \eta^{\delta_\Psi} E(Z)$.

Lema 2.1 *Consideremos a variável aleatória operada $X = \eta \odot_\Psi Z$ e admitamos que P_Z tem raio de convergência $r_Z > 1$.*

1. *Se $1 + E(Y)f(h) < r_Z$, então*

   (a) *$P_X(1+h) = P_Z(\Psi_t(1+h)) = 1 + E(Y)E(Z)h(1+o_h(1))$, $h \to 0$;*

   (b) *$P_X(1+h) \leq (1 + C_1 E(Y) f(h))^2$ onde C_1 é uma constante dependente de r_Z e de $E(Z)$ e $h \leq 1 - r_Z$.*

2. *Tem-se $E((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}) = P_Z(1 + (\eta_1\eta_2)^{\delta_\Psi} f_1(h) f_2(h) + \eta_1^{\delta_\Psi} f_1(h) + \eta_2^{\delta_\Psi} f_2(h))$, com $f_i(h) = h(1 + \xi h / E(Y_1^{(i)}))$, $i \in \{1,2\}$, e $1 + (\eta_1\eta_2)^{\delta_\Psi} f_1(h) f_2(h) + \eta_1^{\delta_\Psi} f_1(h) + \eta_2^{\delta_\Psi} f_2(h) < r_Z$.*

Dem.: 1. (a) Com efeito

$$\begin{aligned} P_X(1+h) &= E((1+h)^X) = E(E((1+h)^X | Z)) \\ &= \sum_{k \in S_Z} \prod_{i=1}^{k} E((1+h)^{Y_i}) P(Z = k) \\ &= \sum_{k \in S_Z} \left(\Psi_t(1+h)\right)^k P(Z = k) \qquad (1) \\ &= E\left[\left(\Psi_t(1+h)\right)^Z\right] = P_Z(\Psi_t(1+h)), \end{aligned}$$

onde

$$\sum_{k\in S_Z} \left(\Psi_t(1+h)\right)^k P(Z=k) = \sum_{k=1}^{+\infty} \left(1+E(Y)f(h)\right)^k P(Z=k)$$

$$= (1+E(Y)f(h))P(Z=1) + \sum_{k=2}^{+\infty}(1+kE(Y)f(h))P(Z=k)$$

$$+ \sum_{k=2}^{+\infty}\sum_{j=2}^{k} C_j^k \left(E(Y)f(h)\right)^j P(Z=k) \qquad (2)$$

$$= 1 + E(Y)f(h)E(Z) + \sum_{k=2}^{+\infty}\sum_{j=2}^{k} C_j^k \left(E(Y)f(h)\right)^j P(Z=k).$$

Por outro lado, como $C_{j+2}^k = \frac{k(k-1)}{(j+2)(j+1)} C_j^{k-2}$, obtemos

$$\sum_{j=2}^{k} C_j^k \left(E(Y)f(h)\right)^j = \left(E(Y)f(h)\right)^2 \sum_{j=0}^{k-2} C_{j+2}^k \left(E(Y)f(h)\right)^j$$

$$\leq \left(E(Y)f(h)\right)^2 k^2 \left(1+E(Y)f(h)\right)^{k-2},$$

pelo que a série em (2) não excede

$$\left(E(Y)f(h)\right)^2 \sum_{k=2}^{+\infty} k^2 \left(1+E(Y)f(h)\right)^{k-2} P(Z=k) \qquad (3)$$

o que, pelo critério de D'Alembert para séries numéricas positivas, representa uma série convergente pois $1 + E(Y)f(h) < r_Z$. Finalmente, devido a (1), (2) e (3), decorre

$$P_X(1+h) = 1 + \left(1 + \frac{\sum_{k=2}^{+\infty}\sum_{j=2}^{k} C_j^k \left(E(Y)f(h)\right)^j P(Z=k)}{E(Y)E(Z)f(h)} \right)$$

$$\times E(Y)E(Z)f(h),$$

com

$$0 \leq \frac{\sum_{k=2}^{+\infty} \sum_{j=2}^{k} C_j^k \left(E(Y)f(h)\right)^j P(Z=k)}{E(Y)E(Z)f(h)}$$

$$\leq \frac{\left(E(Y)f(h)\right)^2 \sum_{k=2}^{+\infty} k^2 \left(1+E(Y)f(h)\right)^{k-2} P(Z=k)}{E(Y)E(Z)f(h)}$$

$$\leq C_1 f(h) \to 0, \quad h \to 0^+,$$

onde C_1 representa uma constante positiva. Como $f(h) \sim h$, $h \to 0^+$, concluímos a prova.

(b) Como a série em (3) é convergente temos que $P_X(1+h) \leq 1+E(Y)E(Z)f(h)+C(E(Y)f(h))^2 \leq (1+C_1E(Y)f(h))^2$, onde $C_1 = \max\{\sqrt{C}/E(Z), 1\}$.

2. Assumido as duas sucessões de contagem $\{Y_j^{(1)}\}$ e $\{Y_j^{(2)}\}$ independentes, vem

$$\begin{aligned} E\left((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}\right) &= E\left(E\left((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}\right)|Z\right) \\ &= E\left(E\left((1+h)^{\eta_1 \odot_\Psi Z}\right) E\left((1+h)^{\eta_2 \odot_\Psi Z}\right)|Z\right) \\ &= \sum_{i=1}^{+\infty} (\Psi_{-\ln \eta_1}(1+h))^k (\Psi_{-\ln \eta_2}(1+h))^k P(Z=k) \\ &= \sum_{i=1}^{+\infty} (1+\eta_1^{\delta_\Psi} f_1(h))^k (1+\eta_2^{\delta_\Psi} f_2(h))^k P(Z=k). \end{aligned}$$

■

3 Estacionaridade forte do modelo

Uma vez que as f.g.p de W_n e de $U_n Z_n$, P_{W_n} e P_{UZ}, respectivamente, verificam $P_{W_n}(z) = P_{W_{n-1}}(\Psi_t(z)) P_{UZ}(z)$, $t = -\ln \eta$, $n \in \mathbb{N}$, admitindo que a f.g.p. $\Psi_t(z)$ verifica $\Psi_{t_1}(\Psi_{t_2}(z)) = \Psi_{t_1+t_2}(z)$, $|z| \leq r_Y$,

o processo $\{W_n\}$ admite a representação: $W_n =^d \eta^k \odot_\Psi W_{n-k} + \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-i}Z_{n-i}, \forall\, n \in \mathbb{N}, \forall\, k \geq 1$. Então, para $n \in \mathbb{N}$,

$$X_n \stackrel{d}{=} \beta \odot_\Psi Z_n + V_n \left(\eta^k \odot_\Psi W_{n-1-k} + \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i} \right).$$

Proposição 3.1 *Se $E(Z) < \infty$, então a sucessão $\{X_n\}$ é estritamente estacionária.*

Dem.: Recordemos que se $\{Q_n\}$ é uma sucessão de v.a.'s de média finita e $\sum_{n=1}^{+\infty} E(|Q_n|)$ é convergente então a série $\sum_{n=1}^{+\infty} Q_n$ é quase certamente absolutamente convergente. Como $\sum_{i=0}^{+\infty} E(\eta^i \odot_\Psi U_{n-i}Z_{n-i}) = E(UZ) \sum_{i=0}^{+\infty} \eta^{i\delta_\Psi} < +\infty$, obtemos que $W_n^{(k)} = \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-i}Z_{n-i} \xrightarrow[k]{q.c.} W'_n := \sum_{i=0}^{+\infty} \eta^i \odot_\Psi U_{n-i}Z_{n-i}$. Por outro lado, devido ao facto de a sucessão $\{U_nZ_n\}$ ser constituída por v.a.'s i.i.d., as f.g.p. dos vectores $(W_n^{(k)},W_{n+1}^{(k)},...,W_{n+t}^{(k)})$ e $(W_{n+\ell}^{(k)},W_{n+1+\ell}^{(k)},...,W_{n+t+\ell}^{(k)})$ são iguais, para qualquer $\ell > 1$, e assim $\{W_n^{(k)}\}$ é estritamente estacionária. Sendo a convergência q.c. de um vector equivalente à convergência q.c. das margens, provamos que estes dois vectores convergem q.c. para $(W_n^{'},W_{n+1}^{'},...,W_{n+t}^{'})$ e $(W_{n+\ell}^{'},W_{n+1+\ell}^{'},...,W_{n+t+\ell}^{'})$. Como a convergência q.c. implica a convergência em distribuição e o limite é único, concluímos que os vectores $(W'_n,W'_{n+1},...,W'_{n+t})$ e $(W'_{n+\ell},W'_{n+1+\ell},...,W'_{n+t+\ell})$ são i.d.. Atendendo ainda a que se tem $\eta^k \odot_\Psi W_{n-k} \xrightarrow[k]{q.c.} 0$, concluímos que $W_n \stackrel{d}{=} \sum_{i=0}^{+\infty} \eta^i \odot_\Psi U_{n-i}Z_{n-i}$ bem como $X_n \stackrel{d}{=} \beta \odot_\Psi Z_n + V_n \sum_{i=0}^{+\infty} \eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i}$, $n \in \mathbb{N}$, devido à

independência das sucessões $\{Z_n\}$ e $\{V_n\}$. Fica provado que $\{W_n\}$ e $\{X_n\}$ são estritamente estacionárias. ■

4 Distribuição limite do máximo

Comecemos por caracterizar a cauda das margens do processo em estudo. O seguinte lema, devido a Hall [3], é um resultado fundamental para este trabalho.

Proposição 4.1 *([3]) Sejam Y_1 e Y_2 duas v.a.'s independentes. Se $Y_1 \in \mathcal{C}_{\mathcal{A}}(r_{Y_1})$ e Y_2 tem f.g.p. finita para algum $r > r_{Y_1}$, então $Y_1 + Y_2 \in \mathcal{C}_{\mathcal{A}}(r_{Y_1})$, com A substituído por $AE((r_{Y_1})^{Y_2})$.*

Lema 4.2 *(Cauda da sucessão) Se as margens de $\{Z_n\}$ pertencem a $\mathcal{C}_{\mathcal{A}}(r_Z)$, então F_{X_n} pertence à mesma classe com*

$$P(X_n > z) \sim A^{'}[z]^{\xi} r_Z^{-[z]},\ z \to +\infty,$$

com $A^{'} = AE\left((\Psi_{-\ln\beta}(r_Z))^Z\right) E(r_Z^{\sum_{i=1}^{+\infty}\eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i}})$.

Dem.: Uma vez que se tem $X_n =^d V_n \sum_{i=1}^{+\infty}\eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i} + \beta \odot_\Psi Z_n + V_nU_{n-1}Z_{n-1}$ e $V_nU_{n-1}Z_{n-1} \in \mathcal{C}_{\mathcal{A}}(r_Z)$, há que provar que $\beta \odot_\Psi Z_n + V_n \sum_{i=1}^{+\infty}\eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i}$ tem f.g.p. finita, para algum $s > r_Z$, e aplicar o Lema 4.1. Escolha-se $M \geq 1$ tal que para $i > M$, $1 + \eta^{i\delta_\Psi} f(h) < r_Z < 1 + h$. Com efeito, usando o Lema 2.1, temos

$$\begin{aligned}
&\prod_{i=M}^{+\infty} E\left((1+h)^{\eta^i \odot_\Psi U_{n-1-i}Z_{n-1-i}}\right) \\
&< \prod_{i=M}^{+\infty}\left(1 + C_1\eta^{i\delta_\Psi} f(h)\right)^2 \leq \exp\left(\sum_{i=M}^{+\infty}\ln\left(1 + C_1\eta^{i\delta_\Psi} f(h)\right)^2\right) \\
&\leq \exp\left(\sum_{i=M}^{+\infty} 2C_1\eta^{i\delta_\Psi} f(h)\right) < +\infty.
\end{aligned}$$

Como $\beta \odot_\Psi Z_n$ também tem f.g.p. finita, o Lema 4.1 estabelece a conclusão. ■

Teorema 4.3 *Se $Z \in \mathcal{C}_\mathcal{A}(r_Z)$, então existem b_n e k_n tais que a sucessão estacionária $\{X_n\}$ verifica $D_{k_n}(x+b_n)$ e $D'_{k_n}(x+b_n)$, tendo-se*

$$P(M_{k_n} \leq x + b_n) \longrightarrow \exp(-r_Z^{-[x]}), n \longrightarrow +\infty, \forall\, x \in \mathbb{R}.$$

Dem.: Consideremos os inteiros i_1,...,i_p,j_1,...,j_q como na definição de $D_{k_n}(x+b_n)$. Seja $X_j^* = \beta \odot_\Psi Z_j + V_j \sum_{i=1}^{\ell_n - 1} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i}$ e usemos as notações $A_j := \{X_j \leq u_n\}$ e $A_j^* := \{X_j^* \leq u_n\}$. Como $A_j \subseteq A_j^*$ e $(X_{i_1}^*,...,X_{i_p}^*)$ e $(X_{j_1}^*,...,X_{j_q}^*)$ são independentes, temos, com $\varepsilon_n > \varepsilon > 0$,

$$\begin{aligned} P\left(\cap_{s=1}^p A_{i_s}, \cap_{t=1}^q A_{j_t}\right) &\leq P\left(\cap_{s=1}^p A_{i_s}^*\right) P\left(\cap_{t=1}^q A_{j_t}^*\right) \\ &\leq P\left(\cap_{s=1}^p \{X_{i_s} \leq u_n + \varepsilon_n\}\right) P\left(\cap_{t=1}^q \{X_{j_t} \leq u_n + \varepsilon_n\}\right) \\ &\quad + 3k_n P\left(V_1 \sum_{i=\ell_n}^{+\infty} \eta^i \odot_\Psi U_{-i} Z_{-i} > \varepsilon_n\right) \end{aligned} \tag{4}$$

onde, pela desigualdade de Markov, o último termo não excede

$$3(1-\beta)k_n \frac{E\left(\sum_{i=\ell_n}^{+\infty} \eta^i \odot_\Psi U_{-i} Z_{-i}\right)}{\varepsilon_n} = 3(1-\beta)E(UZ)\,\frac{k_n}{\varepsilon_n}\,\frac{\eta^{\delta_\Psi \ell_n}}{1-\eta^{\delta_\Psi}}.$$

Com $\ell_n = (k_n)^\alpha$, $\alpha \in]0{,}1[$, obtemos $\frac{k_n}{\delta_n}\,\eta^{\delta_\Psi \ell_n} \longrightarrow 0, n \longrightarrow +\infty$. A desigualdade contrária de (4) obtém-se similarmente.
Para provar que $D'_{k_n}(u_n)$ ocorre, comecemos por dividir o somatório da respectiva definição em duas parcelas de acordo com $j \leq \gamma_n - 1$ e $j \geq \gamma_n$. Uma vez que $X_j \leq X_j | \{V_j = 1\}$, $j \geq 1$, obtemos

$$\begin{aligned} &X_1 + X_j \leq T_j := \beta \odot_\Psi Z_1 + \beta \odot_\Psi Z_j + \eta^{j-2} \odot_\Psi U_1 Z_1 + \\ &+ \sum_{i=1}^{j-2} \eta^{i-1} \odot_\Psi U_{j-i} Z_{j-i} + \sum_{i=0}^{+\infty} (\eta^i \odot_\Psi U_{-i} Z_{-i} + \eta^{i+j-1} \odot_\Psi U_{-i} Z_{-i}). \end{aligned}$$

Então, pela desigualdade de Markov, decorre

$$
\begin{aligned}
& k_n \sum_{j=1}^{\gamma_n - 1} P(X_1 > u_n, X_j > u_n) \\
& \leq k_n \gamma_n \max_j P(X_1 + X_j > 2u_n) \\
& \leq k_n \gamma_n \max_j P(T_j > 2u_n) \\
& \leq k_n \gamma_n (1+h)^{-2u_n} \max_j E((1+h)^{T_j}), h > 0,
\end{aligned} \tag{5}
$$

onde

$$
\begin{aligned}
E((1+h)^{T_j}) \; = \; & P_Z(\Psi_{-\ln\beta}(1+h)) \times \\
& \times P_Z(\Psi_{-\ln\beta}(1+h)\Psi_{-\ln\eta^{j-2}}(1+h)) \prod_{i=1}^{j-2} P_{UZ}\left(\Psi_{-\ln\eta^{i-1}}(1+h)\right) \\
& \times \prod_{i=0}^{+\infty} P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right).
\end{aligned}
$$

Provamos apenas a convergência do último produtório, dada a similitude da convergência dos outros factores. Seja $\theta = \eta^{\delta_\Psi}$. Consideremos h tal que $\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h) = 1 + \theta^{2i+j-1} f^2(h) + \theta^i(1+\theta^{j-1})f(h) < r_Z$. Replicando os argumentos de Hall [3] (pag 372-373) e aplicando a propriedade 2 do Lema 2.1, obtemos

$$
\begin{aligned}
& P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right) \\
& = P_{UZ}\left(1 + \theta^{2i+j-1} f^2(h) + \theta^i(1+\theta^{j-1})f(h)\right) \\
& \leq P_{UZ}\left(1 + \theta^i f(h)\right)\left(1 + C_1\theta^i\theta^{j-1} f(h)\right)\left(1 + C_2\theta^{2i+j-1} f^2(h)\right) \\
& \leq \left(1 + C_3\theta^i f(h)\right)^2 \left(1 + C_1\theta^i\theta^{j-1} f(h)\right)\left(1 + C_2\theta^{2i+j-1} f^2(h)\right),
\end{aligned}
$$

pelo que

$$
\begin{aligned}
& \prod_{i=0}^{+\infty} P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right) \\
& \leq \exp\left((2C_3 + C_1\theta^{j-1})f(h) \sum_{i=0}^{+\infty} \theta^i + C_2\theta^{j-1} f^2(h) \sum_{i=0}^{+\infty} \theta^{2i}\right)
\end{aligned}
$$

o que é uniformemente limitado em j. Consideremos agora $b_n = n$, $k_n = [\frac{1}{A} n^{-\xi} r_Z^n]$, $s_n = [k_n^\alpha]$, com $\alpha \in]0{,}1[$, $\gamma_n = [(\frac{k_n}{s_n})^\mu]$, com $\mu \in]0{,}1[$, e $(1+h)^2 = r_Z^\phi$, com $\phi \in]1{,}2[$, de modo que $\mu(1-\alpha) < \phi - 1$. Atendendo a (5), fica provado que a parcela correspondente a $j \leq \gamma_n - 1$, tende para zero. Por outro lado, obtemos

$$\begin{aligned} k_n \sum_{j=\gamma_n}^{k_n/s_n} & P(X_1 > u_n, X_j > u_n) \\ & \leq \tfrac{k_n^2}{s_n} P(X_1 > u_n) P\left(\beta \odot_\Psi Z_j + U_{j-1} Z_{j-1}\right. \\ & \quad + \sum_{i=1}^{\gamma_n} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i} > u_n - \varepsilon) \\ & \quad + \tfrac{k_n^2}{s_n} P\left(\sum_{i=\gamma_n+1}^{+\infty} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i} > \varepsilon \right). \end{aligned} \tag{6}$$

Uma vez que $U_{j-1} Z_{j-1} \in \mathcal{C}_{\mathcal{A}}(r)$, atendendo mais uma vez ao Lema 4.1, também $\beta \odot_\Psi Z_j + U_{j-1} Z_{j-1} + \sum_{i=1}^{\gamma_n} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i}$ pertence à mesma classe. Então a primeira parcela do segundo membro de (6) é majorada por $r_Z^{-[x]} r_Z^{-[x-\varepsilon]} / s_n \longrightarrow 0$, $n \longrightarrow +\infty$. Por outro lado, a segunda parcela não excede

$$\frac{k_n^2}{s_n} \frac{E(\sum_{i=\gamma_n+1}^{+\infty} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i})}{\varepsilon} \leq \frac{k_n^2}{s_n} \frac{\theta^{\gamma_n}}{\varepsilon(1-\theta)} \longrightarrow 0,$$

quando $n \longrightarrow +\infty$. ■

Agradecimentos

O trabalho da primeira autora foi parcialmente financiado pela FCT - Fundação para a Ciência e a Tecnologia, pelos projectos UID/MAT/00013/2013 e UID/Multi/04621/2013. O trabalho da segunda autora foi parcialmente apoiado pelo Centro de Matemática da Universidade de Coimbra - UID/MAT/00324/2013, financiado pelo Governo Português através da FCT/MCTES e co-financiado pelo Fundo

Europeu de Desenvolvimento Regional através do Acordo de Parceria PT2020.

Referências

[1] Aly, E.A., Bouzar, N. (2005). Stationary solutions for integer-valued autoregressive processes. *International Journal of Mathematics and Mathematical Sciences* 1, 1–18.

[2] Anderson, C.W. (1970). Extreme value theory for a class of discrete distribution with applications to some stochastic processes. *Journal of Applied Probability* 7, 99–113.

[3] Hall, A. (2003). Extremes of integer-valued moving average models with exponential type tails. *Extremes* 6, 361-379.

[4] van Harn, K., Steutel, F.W., Vervaat, W. (1982). Self-decomposable discrete distributions and branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 61, 97–118.

[5] Leadbetter, M.R., Lindgren, G., Rootzén, H. (1983). Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, Berlin.

[6] McKenzie, E. (1986). Auto regressive-moving-average processes with negative binomial and geometric marginal distribution. *Advances in Applied Probability* 18, 679–705.

[7] Temido, M.G. (2002). Domínios de atracção de funções de distribuição discretas. In Carvalho, L. *et al.* (eds): *Novos Rumos em Estatística*, 415–426, Edições SPE, Lisboa.

[8] Temido, M.G., Canto e Castro, L. (2003). Max-semistable laws in extremes of stationary random sequences. *Theory of Probability and its Applications* 47, 365–374.

Duração média de períodos de ocupação contínua e probabilidade de bloqueio em sistemas oscilantes $M^X/G/1/(n,a,b)$

Fátima Ferreira
Universidade de Trás-os-Montes e Alto Douro, UTAD, Departamento de Matemática, CMAT e CEMAT, *mmferrei@utad.pt*

António Pacheco
Instituto Superior Técnico, Universidade de Lisboa, Departamento de Matemática e CEMAT, *apacheco@math.tecnico.ulisboa.pt*

Helena Ribeiro
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e CEMAT, *helena.ribeiro@ipleiria.pt*

Palavras–chave: Sistemas oscilantes, períodos de ocupação contínua, cadeias de Markov, probabilidade de bloqueio

Resumo: Neste trabalho, analisam-se características dos períodos de ocupação contínua de sistemas oscilantes $M^X/G/1/(n,a,b)$. Em particular, estendem-se os resultados de Pacheco e Ribeiro [6] a sistemas oscilantes não preemptivos, propondo um método para calcular a duração média de períodos de ocupação contínua. Estes resultados são combinados com os obtidos em Ferreira et al. [3] de forma a avaliar, para diferentes distribuições de serviço e de tamanho do grupo, a taxa de perdas de clientes em ciclos de ocupação e a probabilidade de bloqueio a longo prazo destes sistemas.

1 Introdução

Os sistemas oscilantes $M^X/G/1/(n,a,b)$ são filas de espera de capacidade finita, n, a que os clientes chegam em grupos de tamanho

aleatório segundo um processo de Poisson composto e são servidos, por ordem de chegada, por um único servidor. Os tamanhos dos grupos são variáveis independentes e identicamente distribuídas à variável X, com função de probabilidade $(f_l = P(X = l))_{l\in\mathbb{N}_+}$ e média finita $\bar{f}$. A sucessão dos tamanhos dos grupos e dos tempos entre chegadas são independentes. Contudo, contrariamente ao usual nos sistemas de filas de espera clássicos, nestes sistemas os tempos de serviço dos clientes não são independentes entre si, oscilando entre duas fases, 1 e 2, reagindo de uma forma dinâmica à congestão do sistema. Especificamente, a fase em que o sistema se encontra em cada instante é determinada pela evolução do número de clientes no sistema, de acordo com duas barreiras: a e b, $0 \leq a < b \leq n$. O sistema encontra-se na fase 1 quando está vazio e continua nesta fase até que o número de clientes no sistema atinja ou ultrapasse a barreira superior b. Após esse instante, o sistema muda para a fase 2, permanecendo nesta fase até ao instante subsequente em que o número de clientes no sistema passe a ser menor ou igual à barreira inferior a. Nesse instante o sistema passa de novo para a fase 1, e assim sucessivamente.

A duração de cada serviço é determinada pela fase em que o sistema se encontra no instante em que se inicia o serviço (sistema não preemptivo).

Os tempos de serviço iniciados com o sistema a operar na fase 1 têm duração aleatória S_1 com distribuição A_1 e média μ_1^{-1}, e os tempos de serviço iniciados com o sistema a operar na fase 2 têm duração aleatória S_2 com distribuição A_2 e média μ_2^{-1}, geralmente menor que μ_1^{-1}.

Assim, o estado do sistema em cada instante caracteriza-se a partir do processo em tempo contínuo, $Y(t) = (Y_1(t),Y_2(t))$, onde $Y_1(t)$ denota o número de clientes no sistema no instante t e $Y_2(t)$ a fase em que o sistema está a operar nesse mesmo instante. O processo $(Y(t))_{t\geq 0}$ tem espaço de estados

$$E^{(n,a,b)} = \{(i,1) : 0 \leq i \leq b-1\} \cup \{(i,2) : a+1 \leq i \leq n\}$$

e é um processo regenerativo markoviano (ver, e.g., [4]) associado à

sucessão dos instantes de saída de clientes do sistema.
Dada a relevância dos sistemas oscilantes no controlo da qualidade de serviço prestado com custos reduzidos, estes sistemas têm sido objeto de estudo nos últimos anos [1, 2, 3, 5]. Em particular, usando o método potencial, Chydzinski [1, 2] caracterizou a distribuição limite da ocupação de sistemas oscilantes com chegadas simples de Poisson e serviços com distribuição geral, de capacidade finita e infinita.
A análise destes sistemas em períodos de ocupação contínua, i.e., em períodos contínuos de utilização efetiva do servidor, é relevante do ponto de vista do operador, e fornece informação crucial para a sua gestão. Nesse âmbito, e no tocante a períodos de ocupação contínua de sistemas oscilantes $M^X/G/1/(n,a,b)$, tirando partido da estrutura regenerativa markoviana destes sistemas, Pacheco e Ribeiro [5] calcularam a distribuição do número de perdas consecutivas de clientes e Ferreira *et al.* [3] calcularam os momentos do número de perdas de clientes. Constatou-se que, em situações com elevada intensidade de tráfego, os sistemas com serviços de cauda pesada e distribuições do tamanho dos grupos de maior variabilidade apresentam menores número médios de perdas de clientes durante os períodos de ocupação contínua. Contudo, dependendo das distribuições de serviço e de tamanho do grupo consideradas, a duração média dos períodos de ocupação contínua pode variar consideravelmente, influenciando fortemente a taxa de perdas a longo prazo e a probabilidade de bloqueio destes sistemas, i.e., a fração a longo prazo de clientes que são rejeitados por não encontrarem lugar na fila quando da sua chegada.
Com vista a avaliar a taxa de perdas a longo prazo e a probabilidade de bloqueio de sistemas oscilantes $M^X/G/1/(n,a,b)$, propõe-se, na Secção 2, um método recursivo para o cálculo da duração média dos períodos de ocupação contínua dos sistemas em estudo. Estes resultados são combinados, na Secção 3, com os obtidos em Ferreira *et al.* [3] para o número médio de perdas em períodos de ocupação contínua, de forma a calcular a taxa de perdas de clientes em ciclos de ocupação e a probabilidade de bloqueio a longo prazo. Finalizamos este trabalho apresentando, na Secção 4, a aplicação dos resultados derivados a filas de espera com diferentes distribuições de serviço e

de tamanhos dos grupos, e, na Secção 5, algumas conclusões.

2 Duração média de períodos de ocupação contínua

Nesta secção propõe-se uma metodologia para o cálculo da duração média de períodos de ocupação contínua em sistemas oscilantes $M^X/G/1/(n,a,b)$. Consideram-se no estudo períodos de ocupação contínua iniciados com múltiplos clientes no sistema, denotando por $B_{(i,j)}^{(n,a,b)}$ a duração de um $(i,j)-$período de ocupação contínua da fila oscilante $M^X/G/1/(n,a,b)$. Especificamente, um (i,j)-período de ocupação contínua representa um período de ocupação contínua que se inicia com o sistema no estado (i,j), i.e., o período de tempo que se inicia com i clientes no sistema e o sistema a operar na fase j, com um cliente a iniciar serviço nesse instante, e termina no primeiro instante subsequente em que o sistema fica vazio.
Iniciamos com o estudo dos sistemas oscilantes com barreira inferior zero, $M^X/G/1/(n,0,b)$. Começamos por notar que, nestas filas, quando o período de ocupação contínua começa na fase 2, o sistema permanece nesta fase durante todo o período de ocupação contínua, independentemente do valor b da barreira superior. Deste modo, durante um $(i,2)-$período de ocupação contínua, o sistema oscilante $M^X/G/1/(n,0,b)$ comporta-se como um sistema regular $M^X/G/1/n$ com distribuição de serviço A_2 e, para $0 < b \leq i \leq n$, tem-se

$$B_{(i,2)}^{(n,0,b)} \stackrel{\mathrm{d}}{=} B_{(i,2)}^{(n,0,1)} \stackrel{\mathrm{d}}{=} B_i^{(n)} \tag{1}$$

com $\stackrel{\mathrm{d}}{=}$ a denotar a igualdade em distribuição e $B_i^{(n)}$ a duração de um período de ocupação contínua que se inicia com i clientes no sistema regular $M^X/G/1/n$ com distribuição de serviço A_2.
Assim, a duração média de um $(i,2)-$período de ocupação contínua no sistema oscilante $M^X/G/1/(n,0,b)$ pode obter-se de

$$E[B_{(i,2)}^{(n,0,b)}] = E[B_{(i,2)}^{(n,0,1)}] = E[B_i^{(n)}], \quad 0 < b \leq i \leq n \tag{2}$$

usando os procedimentos propostos em Pacheco e Ribeiro [6].
Por outro lado, se o período de ocupação contínua começa na fase 1, a duração média do $(i,1)-$período de ocupação contínua $(i < b)$ pode obter-se a partir do Teorema 2.1, tendo em conta a distribuição condicional do número de clientes que chegam ao sistema durante o serviço do cliente que inicia o período de ocupação contínua.

Teorema 2.1 *No sistema oscilante $M^X/G/1/(n,0,b)$, com $b > 1$, a duração média de um $(i,1)-$período de ocupação contínua é tal que*

$$E[B_{(i,1)}^{(n,0,b)}] = \xi_i - \tau_i \frac{\xi_0}{\tau_0} \tag{3}$$

para $0 < i < b$, com $(\xi_{b-1},\tau_{b-1}) = (0,1)$ e ξ_{j-1} e τ_{j-1}, com $j = b-1, b-2, \ldots, 1$, obtidos recursivamente por

$$\xi_{j-1} = \frac{\xi_j - \sum_{l=1}^{b-j-1} r_l \, \xi_{l+j-1} - \zeta_j}{r_0} \tag{4}$$

$$\tau_{j-1} = \frac{\tau_j - \sum_{l=1}^{b-j-1} r_l \, \tau_{l+j-1}}{r_0} \tag{5}$$

com r_l denotando a probabilidade de l clientes chegarem ao sistema durante um tempo de serviço com distribuição A_1[1] *e*

$$\zeta_j = E[S_1] + \sum_{l \geq b-j} r_l \, E[B_{(\min(l+j-1,n-1),2)}^{(n,0,b)}].$$

Dem.: Seja C a variável aleatória que denota o número de clientes que chegam ao sistema durante o serviço do primeiro cliente. Condicionando no número de clientes que chegam ao sistema durante esse serviço, a duração do $(i,1)-$período de ocupação contínua verifica

$$[B_{(i,1)}^{(n,0,b)} | C = l] \stackrel{\mathrm{d}}{=} \begin{cases} (S_1 | C = l) \oplus B_{(l+i-1,1)}^{(n,0,b)} & 0 \leq l < b - i \\ (S_1 | C = l) \oplus B_{\min((l+i-1,n-1),2)}^{(n,0,b)} & l \geq b - i \end{cases} \tag{6}$$

[1] Considerações sobre o cálculo das probabilidades r_l são apresentadas em [3]

para $i < b$, com $\oplus$ denotando a adição de variáveis aleatórias independentes, e tem-se

$$E[B_{(i,1)}^{(n,0,b)}] = \sum_{l=0}^{b-i-1} r_l \, E\left[(S_1|C=l) \oplus B_{(l+i-1,1)}^{(n,0,b)}\right] + \sum_{l \geq b-i} r_l \, E\left[(S_1|C=l) \oplus B_{(\min(l+i-1,n-1),2)}^{(n,0,b)}\right].$$

Separando, na expressão anterior, o termo $l = 0$,

$$r_0 \, E[B_{(i-1,1)}^{(n,0,b)}] = E\left[B_{(i,1)}^{(n,0,b)}\right] - \sum_{l=1}^{b-i-1} r_l \, E\left[B_{(l+i-1,1)}^{(n,0,b)}\right] - E[S_1] - \sum_{l \geq b-i} r_l \, E\left[B_{(\min(l+i-1,n-1),2)}^{(n,0,b)}\right]$$

pelo que, a duração média de um $(i-1,1)-$período de ocupação contínua pode ser escrita de uma forma recursiva, em função da duração média dos períodos de ocupação contínua iniciados com i ou mais clientes no sistema. Pode-se então concluir que

$$E[B_{(i,1)}^{(n,0,b)}] = \xi_i + \tau_i E[B_{(b-1,1)}^{(n,0,b)}] \tag{7}$$

com ξ_i e τ_i satisfazendo as equações (4) e (5), respetivamente. Fazendo $\xi_{b-1} = 0$ e $\tau_{b-1} = 1$, e usando o facto de $0 = E[B_{(0,1)}^{(n,0,b)}] = \xi_0 + \tau_0 E[B_{(b-1,1)}^{(n,0,b)}]$, obtém-se $E[B_{(b-1,1)}^{(n,0,b)}] = -\frac{\xi_0}{\tau_0}$. Finalmente, (3) decorre como consequência deste resultado e de (7). ■

Passamos agora ao estudo dos sistemas oscilantes com barreira inferior positiva, $M^X/G/1/(n,a,b)$ com $a > 0$. Da propriedade regenerativa markoviana nos instantes de saída de clientes decorre que, após o início dum $(i,j)-$período de ocupação contínua, com $(i,j) \in E^{(n,a,b)} \backslash \{(0,1),(1,1)\}$, o período de tempo que o sistema leva até ficar com um único cliente é independente do período de tempo

subsequente até o sistema ficar vazio. Em adição, fixando um cliente inicialmente presente no sistema e admitindo que este só será servido quando estiver sozinho no sistema, então: uma vez iniciado o $(i,j)-$período de ocupação contínua, o tempo que decorre até o sistema atingir o estado (1,1) – partindo do estado (i,j) – tem a mesma distribuição que a duração de um $(i-1,j)-$período de ocupação contínua num sistema oscilante $M^X/G/1/(n-1,a-1,b-1)$, com os mesmos parâmetros do sistema oscilante $M^X/G/1/(n,a,b)$ original, exceto a capacidade e barreiras.
Assim, para $(i,j) \in E^{(n,a,b)} \backslash \{(0,1),(1,1)\}$ e $a \geq 1$, tem-se

$$B_{(i,j)}^{(n,a,b)} \stackrel{\mathrm{d}}{=} B_{(i-1,j)}^{(n-1,a-1,b-1)} \oplus B_{(1,1)}^{(n,a,b)}. \tag{8}$$

Este resultado relaciona a duração de períodos de ocupação contínua iniciados com múltiplos clientes com a duração de períodos de ocupação contínua iniciados com um cliente e em sistemas similares com capacidade inferior. Por condicionamento no número de clientes que chegam durante o serviço do primeiro cliente servido no período de ocupação contínua, deduzimos no Teorema 2.2 um processo recursivo, na capacidade do sistema e barreiras, para o cálculo da duração média de um $(i,j)-$período de ocupação contínua num sistema oscilante $M^X/G/1/(n,a,b)$, com $0 < a < n-1$.

Teorema 2.2 *A duração média de um período de ocupação contínua num sistema oscilante $M^X/G/1/(n,a,b)$, com $0<a<n-1$, é tal que*

$$E\left[B_{(i,j)}^{(n,a,b)}\right] = E\left[B_{(i-1,j)}^{(n-1,a-1,b-1)}\right] + E\left[B_{(1,1)}^{(n,a,b)}\right] \tag{9}$$

para todo $(i,j) \in E^{(n,a,b)} \backslash \{(0,1),(1,1)\}$ e

$$\begin{aligned} r_0\, E\left[B_{(1,1)}^{(n,a,b)}\right] &= E\left[S_1\right] + \sum_{l=1}^{b-2} r_l\, E\left[B_{(l-1,1)}^{(n-1,a-1,b-1)}\right] \\ &\quad + r_{b-1}\, E\left[B_{(b-2,1+1_{\{a<b-1\}})}^{(n-1,a-1,b-1)}\right] + \sum_{l\geq b} r_l\, E\left[B_{(\min(l-1,n-2),2)}^{(n-1,a-1,b-1)}\right] \end{aligned} \tag{10}$$

com $1_{\{z\}}$ a denotar a função indicatriz da condição z.

Dem.: O resultado (9) decorre trivialmente da linearidade do valor esperado aplicada a (8). Condicionando no número de clientes que chegam ao sistema durante o serviço do cliente que inicia um (1,1)−período de ocupação contínua (C), tem-se

$$[B_{(1,1)}^{(n,a,b)}|C=l] \stackrel{\mathrm{d}}{=} \begin{cases} (S_1|C=0) & l=0 \\ (S_1|C=l) \oplus B_{(l,1)}^{(n,a,b)} & 1 \leq l \leq b-2 \\ (S_1|C=b-1) \oplus B_{(b-1,1+1_{\{a<b-1\}})}^{(n,a,b)} & l=b-1 \\ (S_1|C=l) \oplus B_{(\min(l,n-1),2)}^{(n,a,b)} & l \geq b \end{cases}$$

pelo que, tendo em conta a decomposição (9) e que $\sum_{l\geq 0} r_l = 1$, a duração média do (1,1)−período de ocupação contínua é tal que

$$\begin{aligned} E\left[B_{(1,1)}^{(n,a,b)}\right] =& E\left[S_1\right] + \sum_{l=1}^{b-2} r_l \, E\left[B_{(l-1,1)}^{(n-1,a-1,b-1)} \oplus B_{(1,1)}^{(n,a,b)}\right] \\ &+ r_{b-1} \, E\left[B_{(b-2,1+1_{\{a<b-1\}})}^{(n-1,a-1,b-1)} \oplus B_{(1,1)}^{(n,a,b)}\right] \\ &+ \sum_{l\geq b} r_l \, E\left[B_{(\min(l-1,n-2),2)}^{(n-1,a-1,b-1)} \oplus B_{(1,1)}^{(n,a,b)}\right] \end{aligned}$$

da qual resulta (10). ■

Finalmente, estudamos os sistemas oscilantes com barreira inferior $a = b-1 = n-1$, $M^X/G/1/(n,n-1,n)$. Nestas filas, quando o período de ocupação contínua começa na fase 1, o sistema opera na fase 1 durante todo o período de ocupação contínua. Assim $B_{(i,1)}^{(n,n-1,n)} \stackrel{\mathrm{d}}{=} B_i^{(n)}$ e $E[B_{(i,1)}^{(n,n-1,n)}] = E[B_i^{(n)}]$, onde $B_i^{(n)}$ denota a duração de um período de ocupação contínua que se inicia com i clientes no sistema regular $M^X/G/1/n$ com distribuição de serviço A_1.

Por outro lado, quando o período de ocupação contínua inicia na fase 2, o que ocorre apenas quando da chegada de $l \geq n$ clientes ao sistema vazio,

$$E[B_{(n,2)}^{(n,n-1,n)}] = E\left[S_2 \oplus B_{(n-1,1)}^{(n,n-1,n)}\right] = E\left[S_2\right] + E\left[B_{n-1}^{(n)}\right].$$

3 Probabilidade de bloqueio

Nesta secção, combinamos os resultados anteriores com os obtidos em Ferreira et al. [3] para o número médio de perdas de clientes em períodos de ocupação contínua, de forma a avaliar a probabilidade de bloqueio a longo prazo destes sistemas.
Atendendo à estrutura regenerativa markoviana dos sistemas oscilantes, num sistema oscilante $M^X/G/1/(n,a,b)$ a probabilidade de bloqueio a longo prazo é dada por

$$P_{bloqueio} = \frac{T^{(n,a,b)}}{\lambda \bar{f}}$$

onde λ denota a taxa de chegadas dos grupos, $\bar{f}$ o tamanho médio dos grupos, e $T^{(n,a,b)}$ a taxa de perdas de clientes a longo-prazo, cujo valor é igual à razão entre o número médio de perdas de clientes num ciclo de ocupação geral[2] e a duração média do mesmo, i.e.,

$$T^{(n,a,b)} = \frac{E[L^{(n,a,b)}]}{E[B^{(n,a,b)}] + \lambda^{-1}}$$

com $B^{(n,a,b)}$ denotando a duração de um período de ocupação contínua geral (i.e., iniciado com um número arbitrário de clientes no sistema) e $L^{(n,a,b)}$ o número de clientes perdidos no mesmo período. Capitalizando nos resultados derivados nas secções anteriores, por condicionamento no número de clientes que iniciam o período de ocupação contínua, obtém-se

$$E[B^{(n,a,b)}] = \sum_{i=1}^{b-1} f_i \, E[B^{(n,a,b)}_{(i,1)}] + \sum_{i=b}^{n} f_i \, E[B^{(n,a,b)}_{(i,2)}] + E[B^{(n,a,b)}_{(n,2)}] \sum_{i \geq n+1} f_i$$

e

$$E[L^{(n,a,b)}] = \sum_{i=1}^{b-1} f_i E[L^{(n,a,b)}_{(i,1)}] + \sum_{i=b}^{n} f_i E[L^{(n,a,b)}_{(i,2)}] + \sum_{i \geq n+1} f_i (i - n + E[L^{(n,a,b)}_{(n,2)}])$$

[2]Ciclo de ocupação geral: período de ocupação contínua adicionado do subsequente período em que o sistema se encontra vazio.

onde os $E[L_{(i,j)}^{(n,a,b)}]$ denotam o número médio de perdas de clientes em $(i,j)-$períodos de ocupação contínua, cujo cálculo é descrito em [3].

4 Ilustração numérica

Nesta secção, avalia-se a duração média de períodos de ocupação contínua e a probabilidade de bloqueio a longo prazo em ciclos de ocupação de diferentes sistemas oscilantes, ilustrando a sensibilidade destas medidas com respeito a diferentes intensidades de tráfego e a diferentes distribuições de serviço e de tamanho do grupo. Para o efeito, consideramos tamanhos dos grupos com distribuição Determinística com valor $\bar{f}$ ($Det(\bar{f})$) e com distribuição Geométrica de parâmetro $1/\bar{f}$ ($Geo(1/\bar{f})$) (de média comum $\bar{f}$). Consideramos ainda as seguintes distribuições dos tempos de serviço (de média comum μ^{-1}): Uniforme no intervalo $(0,2/\mu)$ ($U(0,2/\mu)$), Exponencial de taxa μ ($M(\mu)$), Pareto deslocada de parâmetros (κ,θ), com $\kappa > 1$ e $\theta = (\kappa-1)/\kappa\mu$ ($SP(\kappa,\theta)$) e Pareto Generalizada de parâmetros κ, θ e β, com $\theta = (\kappa-1)/(\mu\beta)$ ($GP(\kappa,\theta,\beta)$).
Os resultados derivados nas secções anteriores foram calculados com recurso a algoritmos implementados em MATLAB e usando as recursões propostas em [3] para o cálculo do número médio de perdas em períodos de ocupação contínua (p.o.c.). Para evidenciar as taxas de serviço consideradas em cada uma das fases nos exemplos numéricos, os sistemas oscilantes $M^X/G/1/(n,a,b)$ serão doravante denotados por $M^X/G(\mu_1)-G(\mu_2)/1/(n,a,b)$.
Na Figura 1 apresenta-se a evolução da duração média de um p.o.c. (geral) em função da intensidade de tráfego, para as diferentes distribuições de serviço e de tamanho de grupo consideradas. Como esperado, para cada sistema analisado, os resultados revelam um aumento da duração média dos p.o.c. com a intensidade de tráfego, sendo esse aumento pouco significativo para intensidades de tráfego pequenas e acentuando-se para taxas de tráfego mais elevadas. Em situações com baixa intensidade de tráfego, os sistemas

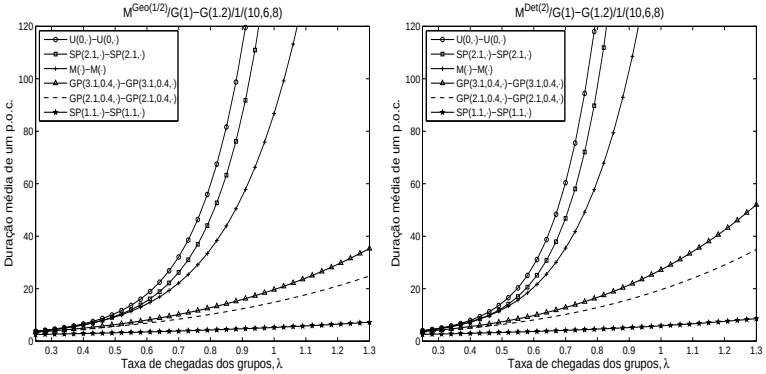


Figura 1: Duração média de um p.o.c. nos sistemas $M^X/G(1) - G(1.2)/1/(10,6,8)$ em função da taxa de chegadas de grupos.

apresentam p.o.c. de curtas durações médias, similares para as diferentes distribuições de serviço e de tamanho de grupo estudadas, apresentando-se bem mais elevadas e distintas (entre os diversos sistemas) para maiores intensidades de tráfego. Constatou-se ainda que os sistemas com serviços de cauda pesada ($SP(1.1,\cdot)$, $GP(2.1,0.4,\cdot)$ e $GP(3.1,0.4,\cdot)$) e distribuições do tamanho dos grupos de maior variabilidade ($Geo(1/2)$) apresentam menor duração média de p.o.c.. Contudo, conforme relatado em [3], a evolução com a taxa de chegadas dos grupos do número médio de perdas durante os p.o.c acompanha a tendência observada para a sua duração média, influenciando a taxa de perdas e a probabilidade de bloqueio a longo prazo.

De facto, como se observa na Figura 2, os sistemas com distribuições de serviço de cauda pesada, que apresentam menor número médio de perdas e menor duração média dos p.o.c., são precisamente os sistemas com maior probabilidade de bloqueio a longo prazo. Observa-se ainda que, sendo muito sensíveis a diferentes distribuições dos tempos de serviço, as probabilidades de bloqueio a longo prazo aparentam ser razoavelmente invariantes face a variações da distribuição do tamanho de grupos com a mesma média.

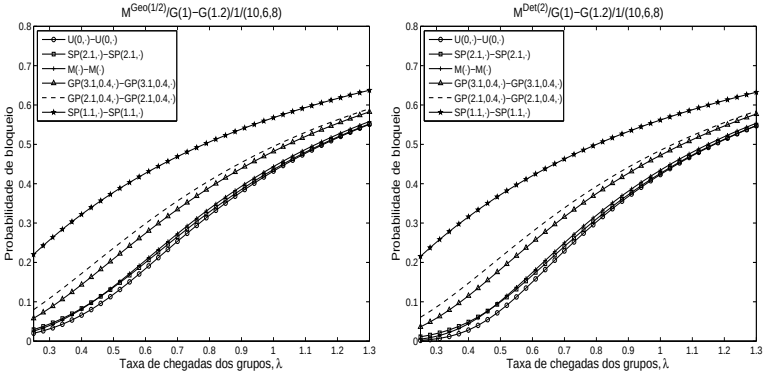


Figura 2: Probabilidade de bloqueio a longo prazo em ciclos de ocupação nos sistemas $M^X/G(1) - G(1.2)/1/(10,6,8)$ em função da taxa de chegadas de grupos.

5 Conclusões e trabalho futuro

Neste trabalho estudamos sistemas oscilantes $M^X/G/1/(n,a,b)$ não preemptivos, nos quais, no início de cada serviço, o servidor pode mudar o tipo de serviço (distribuição do tempo de serviço ou taxa de serviço), reagindo à evolução do número de clientes na fila. Tirando partido da estrutura regenerativa markoviana destes sistemas, derivou-se um procedimento recursivo para calcular a duração média de períodos de ocupação contínua e calculou-se a taxa de perda de clientes em ciclos de ocupação e a probabilidade de bloqueio a longo prazo. Os resultados derivados permitiram concluir que a duração média dos períodos de ocupação contínua e a probabilidade de bloqueio a longo prazo dependem fortemente das distribuições do tempo de serviço consideradas. Em particular, constatou-se que os sistemas oscilantes com distribuição de cauda pesada dos tempos de serviço (aqui ilustrada pela Pareto), tendo períodos de ocupação contínua de menor duração média, são os que apresentam maior pro-

babilidade de bloqueio a longo prazo, especialmente em cenários de elevada intensidade de tráfego.
A metodologia apresentada generaliza a aplicada em Pacheco e Ribeiro [6], para os sistemas regulares com chegadas simples, e será futuramente adaptada à análise de sistemas de filas de espera oscilantes com chegadas de clientes segundo um processo markoviano aditivo de chegadas [7].

Agradecimentos

Este trabalho foi elaborado com o apoio parcial da Fundação para a Ciência e a Tecnologia (FCT) pelo projeto UID/Multi/04621/2013.

Referências

[1] Chydzinski, A. (2002). The $M/G-G/1$ oscillating queueing system. *Queueing Systems* 42(3), 255–268.

[2] Chydzinski, A. (2004). The oscillating queue with finite buffer. *Performance Evaluation* 57(3), 341–355.

[3] Ferreira, F., Pacheco, A., Ribeiro, H. (2015). Moments of losses during busy-periods of regular and nonpreemptive oscillating $M^X/G/1/n$ systems. *Annals of Operations Research*. In press.

[4] Kulkarni, V.G. (1995). *Modeling and Analysis of Stochastic Systems*. Chapman and Hall, Londres.

[5] Pacheco, A., Ribeiro, H. (2008). Consecutive customer losses in regular and oscillating $M^X/G/1/n$ systems. *Queueing Systems* 58(2), 121–136.

[6] Pacheco, A., Ribeiro, H. (2008). Moments of the duration of busy periods of $M^X/G/1/n$ systems. *Probability in the Engineering and Informational Sciences* 22, 1–8.

[7] Pacheco, A., Prabhu, N.U., Tang, L.C. (2009). *Markov-Modulated Processes and Semiregenerative Phenomena*. World Scientific, Singapore.

O papel das estruturas geométricas na Estatística

Susana Ferreira
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CAMGSD — Centro de Análise Matemática, Geometria e Sistemas Dinâmicos, *susfer@ipleiria.pt*

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*

Palavras–chave: Inferência Estatística, variedade estatística, métrica Riemanniana.

Resumo: A Geometria da Informação recorre a conceitos e estruturas da geometria, tais como variedades diferenciáveis, espaço tangente, geodésicas e métrica Riemanniana, para analisar modelos estatísticos, nomeadamente famílias paramétricas de distribuições de probabilidade. Neste trabalho são exploradas algumas das potencialidades que este tipo de estruturas geométricas podem trazer para a Estatística.

1 Introdução

Uma família de distribuições de probabilidade, condicionada por um conjunto de parâmetros, pode ser modelada como uma variedade Riemanniana na qual cada distribuição corresponde a um ponto. Deste modo, o recurso às propriedades da geometria Riemanniana pode permitir obter mais informação acerca do modelo estatístico subjacente a um conjunto de dados. Esta ideia deu origem a resultados em diversas áreas da Estatística, nomeadamente em Inferência. Deste modo, estas estruturas geométricas (e outras semelhantes) podem ser

utilizadas com sucesso em diversas áreas da Estatística, como é comprovado em algumas referências clássicas deste tipo de construção, tais como [2, 3, 7, 8, 9, 17, 22, 23, 27]. Em particular, são utilizadas no estudo da informação (a aplicação de ferramentas da Geometria Diferencial na Estatística é frequentemente apelidada por Geometria da Informação) [5, 16], suficiência [25] e eficiência [19], quer em termos assintóticos [14] quer em termos de inferência baseada numa amostra finita [28], na estimação (pontual ou por intervalos) e em testes de hipóteses, na caracterização de distribuições específicas ou famílias de distribuições [1, 11, 12, 18], em problemas de estatística paramétrica [6, 26, 31], semi-paramétrica [4] e não paramétrica [10], abrangendo as diversas visões da Estatística, apesar de maior destaque na visão bayesiana objetiva [21, 30]. Por este motivo, neste trabalho começamos por apresentar resumidamente alguns conceitos da geometria Riemanniana que são depois interpretados em termos de variedades estatísticas, sendo apontadas algumas das suas aplicações.

2 Variedades Riemannianas

Variedades de dimensão n são estruturas que localmente se identificam com abertos de $\mathbb{R}^n$ através de aplicações designadas por cartas.

Definição 2.1 *Uma* ***variedade*** $\mathcal{M}$ *de dimensão n é um espaço topológico de Hausdorff que verifica a seguinte condição: para qualquer ponto $p \in \mathcal{M}$ existe uma vizinhança $\mathcal{U}$ homeomorfa a um conjunto de $\mathbb{R}^n$, ou seja, existe uma aplicação contínua com inversa contínua $\varphi : \mathcal{U} \to \mathbb{R}^n$.*

Uma coleção $(\mathcal{U}_i, \varphi_i)_{i \in I}$ de abertos $\mathcal{U}_i$ e homeomorfismos $\varphi_i : \mathcal{U}_i \to \mathbb{R}^n$ (**carta** ou **coordenada local**) tais que

- $\cup_{i \in I} \mathcal{U}_i = \mathcal{M}$,
- para qualquer $\varphi_j : \mathcal{U}_j \to \mathbb{R}^n$ com $\mathcal{U}_i \cap \mathcal{U}_j \neq \varnothing$, $\varphi_j \circ \varphi_i^{-1}$ é uma aplicação de classe $\mathcal{C}^k$,

designa-se por **atlas de classe** $\mathcal{C}^k$. As aplicações $\varphi_j \circ \varphi_i^{-1}$ designam-se mudança de coordenadas.

Definição 2.2 *A variedade $\mathcal{M}$ diz-se diferenciável (suave) se as aplicações mudança de coordenadas forem diferenciáveis ($\mathcal{C}^\infty$).*

Dada uma função $f : \mathcal{M} \to \mathbb{R}$ e um sistemas de atlas $(\mathcal{U}_i, \varphi_i)_{i\in I}$, podemos reescrever localmente esta função como sendo uma função $\tilde{f}$ de um subconjunto aberto de $\mathbb{R}^n$ tomando

$$\tilde{f}_i = f \circ \varphi_i^{-1}.$$

Dada uma variedade diferenciável $\mathcal{M}$ podemos considerar a união disjunta de todos os espaços tangente em cada ponto $\sqcup T_p\mathcal{M}$, este espaço é designado por **fibrado tangente**.
Dado um ponto $p \in \mathcal{M}$, fixando as coordenadas locais $\varphi(p) = \left(\theta^1, \cdots, \theta^n\right) \in \mathbb{R}^n$, consideremos uma curva $\gamma : [a,b] \to \mathcal{M}$ que passa em p e uma função $\mathcal{C}^\infty$ $f : \mathcal{M} \to \mathbb{R}$. A aplicação $\tilde{f} = f \circ \varphi^{-1}$ permite-nos definir a derivada da função f, ao longo da curva γ, da seguinte forma [5]:

$$\frac{d}{dt}\left(f\left(\gamma\left(t\right)\right)\right) = \left(\frac{\partial f}{\partial \theta^i}\right)_{\gamma(t)} \frac{d\gamma^i\left(t\right)}{dt}.$$

O espaço vetorial $T_p\mathcal{M}$ é gerado pelos vetores $\left\{\frac{\partial}{\partial \theta^i}\right\}$ para a escolha das coordenadas locais $\left(\theta^1, \cdots, \theta^n\right)$. Podemos definir uma aplicação $X : \mathcal{M} \to T_p\mathcal{M}$ que a cada $p \in \mathcal{M}$ faz corresponder um vetor tangente X_p. Esta aplicação é designada por **campo vetorial** e o conjunto de todas estas aplicações é denotado por $\mathcal{X}(\mathcal{M})$.
Em determinados espaços vetoriais $T_p\mathcal{M}$ pode ser definido um **produto interno** $\langle , \rangle_p$, ou seja, para cada par de vetores $X, Y \in T_p\mathcal{M}$, $\langle X, Y\rangle_p \in \mathbb{R}$. Se esta aplicação variar suavemente de ponto em ponto, definimos uma aplicação g, que a cada ponto $p \in \mathcal{M}$ faz corresponder o produto interno $g_p = \langle , \rangle_p$. A aplicação g, definida deste

modo, é um 2-tensor simétrico designado por **métrica Riemanniana**, ou seja,

$$g: \begin{array}{ccc} \mathcal{M} & \rightarrow & T_p\mathcal{M}^* \\ p & \mapsto & g_p \end{array},$$

onde $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \rightarrow \mathbb{R}$.
Em termos locais, se considerarmos uma carta local $\varphi_i : \mathcal{U}_i \rightarrow \mathbb{R}^n$ e as coordenadas locais $\varphi(p) = \left(\theta^1, \cdots, \theta^n\right)$, a aplicação g é definida pela matriz com entradas

$$g_{ij}(p) = \left\langle \frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j} \right\rangle_p .$$

Definição 2.3 *Uma variedade suave $\mathcal{M}$ munida de uma métrica Riemanniana é designada por uma* ***variedade Riemanniana.***

Exemplo 2.4 $(\mathbb{R}^n, \langle , \rangle)$ *onde a métrica é dada por* $\langle e_i, e_j \rangle = \delta_{ij}$ *(métrica euclidiana).*

Outra noção importante é de como relacionar os espaços tangente $T_p\mathcal{M}$ e $T_q\mathcal{M}$ onde p e q são pontos de $\mathcal{M}$. Se considerarmos uma curva suave $\gamma : [0,1] \rightarrow \mathcal{M}$ tal que $\gamma(0) = p$ e $\gamma(1) = q$, podemos definir uma aplicação que a cada ponto $\gamma(t)$ faz corresponder um vetor tangente $X(t) \in T_{\gamma(t)}\mathcal{M}$. Esta aplicação define um campo vetorial X ao longo da curva γ. Este conceito permite-nos, de certo modo, relacionar os diferentes espaços tangentes. O conceito mais geral deste raciocínio é dado pelo conceito de conexão afim.

Definição 2.5 *Uma* ***conexão afim*** ∇ *numa variedade $\mathcal{M}$ é uma aplicação $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \rightarrow \mathcal{X}(\mathcal{M})$ que a cada par de campos vetoriais (X,Y) faz corresponder um novo campo vetorial $\nabla_X Y$ que satisfaz as seguintes condições:*

1. $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$,
2. $\nabla_X (Y + Z) = \nabla_X Y + \nabla_X Z$,

3. $\nabla_X (fY) = f\nabla_X Y + X(f)\nabla_X Y$, onde $X(f)$ é a derivada direcional de f ao longo de X,

para qualquer campo vetorial $X, Y, Z \in \mathcal{X}(\mathcal{M})$ e quaisquer funções reais diferenciáveis f e g definidas em $\mathcal{M}$. Para uma visão mais detalhada pode ser consultado, por exemplo, [24].

3 Variedades Estatísticas

A Informação Geométrica surgiu do estudo geométrico da estimação em Estatística, considerando o espaço das distribuições de probabilidade, que constitui um modelo estatístico, uma variedade.
Suponhamos, então, que pretendemos estimar a distribuição de probabilidade (d.p.) que deu origem aos dados $x = (x_1, x_2, \ldots, x_N)$, os quais são caracterizadas por uma distribuição que representaremos por $p^\star$, a distribuição subjacente à origem dos dados. Consideremos, igualmente, uma família de distribuições de probabilidade definidas num qualquer conjunto $\mathcal{X}$ através da função densidade $p : \mathcal{X} \to \mathbb{R}$. Se $\mathcal{X} \subset \mathbb{R}^n$ então $\int_{\mathcal{X}} p(x)\,\mathrm{d}x = 1$ e $p(x) \geq 0$, $\forall x \in \mathcal{X}$ (onde $\int$ representa o integral múltiplo se $n \geq 2$ e, se $\mathcal{X}$ for discreto, p representa a derivada de Radon-Nikodym em ordem à medida de contagem). Contudo, esta construção pode ser efetuada considerando um espaço de probabilidade geral $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$, onde $\mathcal{X}$ representa o espaço-amostra, $\mathcal{B}(\mathcal{X})$ o espaço dos acontecimentos (σ-álgebra gerada por $\mathcal{X}$) e $\mathbb{P}$ a medida de probabilidade associada ao espaço mensurável $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.
Seja $\mathcal{M}$ um conjunto constituído por distribuições de probabilidade definidas em $\mathcal{X}$ e suponhamos que cada elemento de $\mathcal{M}$ pode ser parametrizado por n variáveis reais $\theta = \left(\theta^1, \cdots, \theta^n\right)$ tais que

$$\mathcal{M} = \left\{ p_\theta = p(x, \theta) \,\middle|\, \theta = \left(\theta^1, \cdots, \theta^n\right) \in \Theta \right\},$$

onde $\Theta \subset \mathbb{R}^n$ e a aplicação $\theta \mapsto p_\theta$ é injetiva. O conjunto $\mathcal{M}$ é designado por **modelo estatístico** n dimensional ou **modelo paramétrico** e Θ por **espaço dos parâmetros**. Deste modo, para a

estimação da d.p. $p^\star$, subjacente à origem das observações, começamos por estabelecer um modelo estatístico $\mathcal{M}$ que inclui as distribuições p candidatas a $p^\star$. De facto, apesar de $p^\star$ ser desconhecida, muitas vezes dispomos de informação prévia que nos transmite uma ideia sobre a forma de $p^\star$ permitindo, por vezes, restringir $\mathcal{M}$ a uma distribuição (ou família de distribuições) que depende dos valores do parâmetro $\theta \in \Theta$.

Para que seja possível trabalhar com o modelo $\mathcal{M} = \{p_\theta | \, \theta \in \Theta\}$ é usual assumirmos algumas condições de regularidade, nomeadamente que p_θ é diferenciável em relação aos parâmetros, Θ é um subconjunto aberto de $\mathbb{R}^n$ e a função $\theta \mapsto p_\theta$, de Θ para $\mathbb{R}$, é $\mathcal{C}^\infty$. Além disso, supomos que integração e diferenciação podem ser permutados (i.e., $\int \partial_i p(x,\theta) \, \mathrm{d}x = \partial_i \int p(x,\theta) \, \mathrm{d}x$ onde $\partial_i := \frac{\partial}{\partial \theta^i}$).

Se restringirmos $\mathcal{X}$ ao suporte da d.p. p, i.e., considerando que $p(x|\theta) > 0$, $\forall x \in \mathcal{X}$ e $\theta \in \Theta$ (supondo que o suporte de p_θ não depende de θ), então $\mathcal{M} \subset \mathcal{P}(\mathcal{X})$ onde

$$\mathcal{P}(\mathcal{X}) = \left\{ p : \mathcal{X} \to \mathbb{R} \,\middle|\, p(x) > 0, \forall x \in \mathcal{X}, \int p(x) \, \mathrm{d}x = 1 \right\}.$$

Deste modo, dado o modelo estatístico $\mathcal{M} = \{p_\theta | \, \theta \in \Theta\}$, a aplicação $\varphi : \mathcal{M} \to \mathbb{R}^n$ definida por $\varphi(p_\theta) = \theta$ permite considerar um sistema de coordenadas $\varphi = \left[\theta^i\right]$ de $\mathcal{M}$ (as coordenadas definem a distribuição). Se considerarmos ainda um difeomorfismo $\mathcal{C}^\infty$ ψ, de Θ para $\psi(\Theta)$ (uma bola aberta de $\mathbb{R}^n$), então na utilização dos parâmetros $\rho = \psi(\theta)$ em substituição de θ surge o modelo estatístico

$$\mathcal{M} = \left\{ p_{\psi^{-1}(\rho)} \,\middle|\, \rho \in \psi(\Theta) \right\}$$

que corresponde à mesma família de d.p. $\mathcal{M} = \{p_\theta | \, \theta \in \Theta\}$.

Por conseguinte, se considerarmos que duas parametrizações $\mathcal{C}^\infty$, difeomórficas entre elas, são equivalentes então $\mathcal{M}$ é uma variedade suave que denominamos por **variedade estatística** [5].

Rao [29] introduziu uma métrica Riemanniana no espaço das famílias de distribuições de probabilidade parametrizadas usando a matriz de informação de Fisher [20]. Com a métrica associada a essa

matriz é possível determinar a distância entre duas distribuições de probabilidade, bem como outras propriedades.
Seja $\mathcal{M}$ uma variedade estatística. Dado um ponto $\theta \in \Theta$, a **matriz de informação de Fisher** de $\mathcal{M}$ no ponto θ é a matrix $G(\theta) = [g_{ij}(\theta)]$,

$$g_{ij}(\theta) := \mathbb{E}_\theta \left[\partial_i \ell_\theta \, \partial_j \ell_\theta\right] = \int \partial_i \ell_\theta \, \partial_j \ell_\theta \, p(x,\theta) \, \mathrm{d}x, \;\; i,j \in \{1,\ldots,n\}$$

onde $\ell_\theta = \log\left(p(x,\theta)\right)$ e $\mathbb{E}_\theta$ representa o valor esperado em relação à distribuição p_θ. Deste modo, para o sistema de coordenadas $\left[\theta^i\right]$ a relação $g_{ij} = \langle \partial_i, \partial_j \rangle$ define uma métrica Riemanniana habitualmente denominada por **métrica de Fisher** [15]. Esta métrica não depende do sistemas de coordenadas, por conseguinte, podemos escrever

$$\langle X,Y \rangle_\theta = \mathbb{E}_\theta \left[(X\ell_\theta)(Y\ell_\theta)\right]$$

para todos os vetores tangentes $X,Y \in T_\theta(\mathcal{M})$.
Seja $F : \mathcal{X} \to \mathcal{Y}$ uma aplicação que transforma o valor da variável aleatória (v.a.) X em $Y = F(X)$. Deste modo, através da distribuição $p(x,\theta)$ de X podemos determinar a distribuição $q(y,\theta)$ que carateriza Y. Por outro lado, se a função $p(x,\theta)$ puder ser determinada através de $p(x,\theta) = q(F(x),\theta)\, r(x)$ então F é uma **estatística suficiente** (fatorização de Fisher-Neyman) uma vez que toda a dependência de $p(x,\theta)$ em relação a θ está incluída na distribuição $q(y,\theta)$ de $Y = F(X)$, i.e., será "suficiente" conhecer o valor de Y para estimar θ. Em geral, a perda de informação $\Delta G(\theta) = [\Delta g_{ij}(\theta)]$ causada ao resumir a informação dos dados x em $y = F(x)$ é dada por

$$\Delta g_{ij}(\theta) = \mathbb{E}_\theta \left[\partial_i \log\left(\tfrac{p(x,\theta)}{q(F(x),\theta)}\right) \; \partial_j \log\left(\tfrac{p(x,\theta)}{q(F(x),\theta)}\right)\right].$$

Por conseguinte, a perda é nula ($\Delta G(\theta) = [0]$) se

$$\partial_i \log\left(\tfrac{p(x,\theta)}{q(F(x),\theta)}\right) = 0,$$

para todos os valores de θ, x e i, sendo, neste caso, F um estimador suficiente para θ (não há perda de informação ao resumir a informação de x em $F(x)$).

Seja $\hat{\theta}$ a função das observações x que será utilizada para estimar os parâmetros desconhecidos θ, i.e., a aplicação $\hat{\theta} = [\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^n] : \mathcal{X} \to \mathbb{R}^n$ é um **estimador** de θ. Se $\mathbb{E}_\theta[\hat{\theta}(X)] = \theta$, $\forall \theta \in \Theta$, então $\hat{\theta}$ é um **estimador centrado** (EC). O **erro quadrático médio** (EQM) de um EC $\hat{\theta}$ pode ser expresso como a matriz de covariância $V_\theta[\hat{\theta}] = \left[v_\theta^{ij}\right]$ onde

$$v_\theta^{ij} := \mathbb{E}_\theta \left[\left(\hat{\theta}^i(X) - \theta^i \right) \left(\hat{\theta}^j(X) - \theta^j \right) \right].$$

Com este resultado podemos deduzir a **desigualdadde de Cramér-Rao** que garante que, num EC $\hat{\theta}$ verifica-se $V_\theta[\hat{\theta}] \geq G(\theta)^{-1}$ (no sentido em que $V_\theta[\hat{\theta}] - G(\theta)^{-1}$ é semidefinida positiva) [1]. O EC que atinja $V_\theta[\hat{\theta}] = G(\theta)^{-1}$, $\forall \theta$, é um **estimador eficiente** (com variância mínima entre os EC). Sublinhemos que nem sempre existem EC com variância $G(\theta)^{-1}$ e, além disso, é possível existirem estimadores enviesados com erro médio quadrático inferior. Todavia, há sempre uma sequência de estimadores $\hat{\theta}_N(x_1, \ldots, x_N)$ cujo EQM converge para $G(\theta)^{-1}$ quando $N \to \infty$ (estimador assintoticamente eficiente). Consideremos, agora, as n^3 funções $\Gamma^{(\alpha)}_{ij,k}$ que a cada ponto $\theta \in \Theta$ associa o valor (cf. Chentsov [13] e Amari [2])

$$\Gamma^{(\alpha)}_{ij,k}(\theta) := \mathbb{E}_\theta \left[\left(\partial_i \partial_j \ell_\theta + \frac{1-\alpha}{2} \, \partial_i \ell_\theta \, \partial_j \ell_\theta \right) (\partial_k \ell_\theta) \right], \quad \text{com} \quad \alpha \in \mathbb{R}.$$

Podemos definir uma conexão afim $\nabla^{(\alpha)}$, denominada por α-conexão, na variedade $\mathcal{M}$ através de

$$\left\langle \nabla^{(\alpha)}_{\partial_i} \partial_j, \partial_k \right\rangle_\theta = \Gamma^{(\alpha)}_{ij,k}$$

onde o produto interno é dado pela métrica de Fisher. A 0-conexão corresponde à conexão Riemanniana (ou conexão Levi-Civita) com respeito à métrica de Fisher. Por outro lado, as α-conexões verificam as seguintes duas igualdades

$$\nabla^{(\alpha)} = (1-\alpha)\nabla^{(0)} + \alpha \nabla^{(1)} = \frac{1+\alpha}{2} \nabla^{(1)} + \frac{1-\alpha}{2} \nabla^{(-1)},$$

razão pela qual é suficiente conhecer $\nabla^{(1)}$ (conexão exponencial) e $\nabla^{(-1)}$ (conexão mistura) [5].

Exemplo 3.1 (Família Exponencial) *Sejam* $\{F_1, \ldots, F_n\}$ *funções linearmente independentes (não constantes) definidas em* $\mathcal{X}$*,* K *uma função definida em* $\mathcal{X}$ *e* ψ *uma função definida em* Θ*, então*

$$p(x,\theta) = \exp\left\{K(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)\right\}$$

define a d.p. da família exponencial sendo $[\theta^i]$ *os seus parâmetros canónicos. Consequentemente,*

$$\partial_i \ell_\theta = F_i - \partial_i \psi(\theta)$$

e

$$\partial_i \partial_j \ell_\theta = -\partial_i \partial_j \psi(\theta),$$

pelo que

$$\Gamma^{(1)}_{ij,k} = -\partial_i \partial_j \psi(\theta) \mathbb{E}_\theta \left[\partial_k \ell_\theta\right] = 0$$

e, portanto, $\nabla^{(1)}$ *é uma conexão plana que é denominada por* ***conexão exponencial*** *ou* e-***conexão*** *(habitualmente representada por* $\nabla^{(e)}$*).*

Exemplo 3.2 (Mistura de distribuições) *Sejam* $\{F_1, \ldots, F_n\}$ *funções linearmente independentes (não constantes) definidas em* $\mathcal{X}$ *e* K *uma função definida em* $\mathcal{X}$*, então*

$$p(x,\theta) = K(x) + \sum_{i=1}^{n} \theta^i F_i(x),$$

define a d.p. da família de misturas de distribuições. Neste caso, a conexão $\nabla^{(-1)}$ *é plana sendo denominada* ***conexão mistura*** *ou* m-***conexão*** *e representada por* $\nabla^{(m)}$*.*

4 Conclusão

O recurso a conceitos da Geometria, tais como medidas, métricas e/ou distâncias entre objetos, está presente nas mais diversas áreas da Estatística e desempenha um papel crucial em todas as suas metodologias [32]. Atente-se, por exemplo, em conceitos elementares tais como a variância ou, em geral, os momentos (centrados ou não) de qualquer ordem, os quais têm um paralelismo óbvio com conceitos geométricos. Como tal, este paralelismo pode ser explorado para auxiliar na aplicação e na interpretação de conceitos da Estatística. Por outro lado, este recurso, sempre presente, a ferramentas da Geometria pode ser efetuado de forma despercebida, nomeadamente quando trabalhamos em espaços euclidianos nos quais estes conceitos são utilizados de forma mais intuitiva. Todavia, noutros casos, unicamente poderá ser efetuado com a utilização de estruturas mais complexas e abstratas, como ilustram as variedades estatísticas que neste trabalho foram apresentadas. Estas estruturas geométricas, praticamente desconhecidas para a maioria daqueles que fazem investigação em Estatística, assumem inequivocamente uma relevância notável no desenvolvimento da Estatística, razão pela qual consideramos pertinente a sua divulgação.

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito dos projetos PTDC/MATGEO/0675/2012 e UID/MAT/00006/2013.

Referências

[1] Amari S. (1982). Differential Geometry of curved exponential families – curvatures and information loss, *Ann. Stat.* 10, 357–385.

[2] Amari S. (1986). Differential Geometrical Methods in Statistics. *Lecture Notes in Statistics* 28, Springer-Verlag, Heidelberg.

[3] Amari S., Barndorff-Nielsen O.E., Kass R.E., Lauritzen S.L. and Rao C.R. (1987). *Differential Geometry in Statistical Inference*. Institute of Mathematical Statistics.

[4] Amari S., Kawanabe M. (1997). Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli* 3, 29–54.

[5] Amari S., Nagaoka H. (2000). Methods of Information Geometry. *Translations of Mathematical Monographs* 191, AMS.

[6] Barndorff-Nielsen O.E. (1986). Likelihood and observed geometries. *Ann Statist* 14, 856–873.

[7] Barndorff-Nielsen, O.E., Cox, D.R., Reid, N. (1986). The role of differential geometry in statistical theory. *Internat Statist Review* 54, 83–96.

[8] Barndorff-Nielsen O.E., Cox D.R. (1994). *Inference and asymptotics*. Chapman & Hall.

[9] Barndorff-Nielsen O.E., Cox D.R. (1989). *Asymptotic techniques for use in statistics*. Chapman & Hall.

[10] Bhattacharya A., Bhattacharya R. (2008). Nonparametric statistics on manifolds with applications to shape spaces, *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* 3, 282–301, Institute of Mathematical Statistics.

[11] Cena A., Pistone G. (2007). Exponential statistical manifold. *Ann Inst Stat Math* 59, 27–56.

[12] Chen W. (2014). A Note on Finding Geodesic Equation of Two Parameters Gamma Distribution. *Applied Mathematics* 5, 3511–3517.

[13] Chentsov N.N. (1972). Statistical Decision Rules and Optimal Inference, Nuaka, Moscow, 1972. Translated from Russian into English, American Mathematical Society, Rhode Island, 1982.

[14] Corcuera, J.M., Giummolè, F. (1999). On the relationship between α connections and the asymptotic properties of predictive distributions, *Bernoulli* 5(1), 163–176.

[15] Costa, S., Santos, S., and Strapasson, J. (2015). Fisher information distance: A geometrical reading. *Discrete Appl Math* 197, 59–69.

[16] Cover, T.M., Thomas, J.A. (1991). *Elements of Information Theory*, John Wiley & Sons, New York.

[17] Critchley, F., Marriott, P., Salmon, M. (2002). On preferred point geometry in statistics. *J. Stat. Plan. Inference* 102, 229–245.

[18] Efron B. (1978). The geometry of exponential families. *Ann. Statist.* 6, 362–376.

[19] Efron B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3, 1189–1217.

[20] Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. *Philos. T. Roy. Soc. A* 222, 309–368.

[21] Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *P. Roy. Soc. Lond. A Mat.* 196, 453–461.

[22] Kass, R.E., Vos, P.W. (1997). *Geometrical foundations of asymptotic inference*, John Wiley & Sons.

[23] Kass R.E. (1989). The Geometry of Asymptotic Inference. *Statistical Science*, 4(3), 188–234.

[24] Kobayashi, S. Nomizu, K. (1991). *Foundations of Differential Geometry*, Volume I. John Wiley & Sons.

[25] Kullback S., Leibler R.A. (1951). On Information and Sufficiency. *Ann. Math. Statist.* 22(1), 79–86.

[26] Marriott, P., Vos, P.W. (2004). On the global geometry of parametric models and information recovery. *Bernoulli* 10, 639–649.

[27] Murray, M.k., Rice, J.W. (1993). *Differential Geometry and Statistics*. Chapman & Hall, London.

[28] Pennec, X. (2006). Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *J. Math. Imaging Vis.* 25(1), 127–154.

[29] Rao, C.R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* 37, 81–91.

[30] Snoussi, H. (2005). Geometry of prior selection. *NeuroComputing* 67, 214–244.

[31] Uhler, C. (2012). Geometry of Maximum likelihood estimation in Gaussian graphical models, *The Annals of Statistics* 40(1), 238–261.

[32] Vos, P.W., Marriott, P. (2010). Geometry in statistics. *WIREs Comp Stat* 2(6), 686–694.

Aplicação do coeficiente RV em Controlo Estatístico da Qualidade

Adelaide Maria Figueiredo
Faculdade de Economia e LIAAD-INESC TEC Porto, Universidade do Porto, *adelaide@fep.up.pt*

Fernanda Otília Figueiredo
Faculdade de Economia da Universidade do Porto e CEAUL, Universidade de Lisboa, *otilia@fep.up.pt*

Palavras–chave: Carta de Controlo, Coeficiente RV, Controlo Estatístico da Qualidade, Monitorização de Processos, Simulação de Monte Carlo

Resumo: Em situações reais a avaliação da qualidade global de um produto ou de um serviço depende simultaneamente de várias caraterísticas de qualidade, pelo que o desenvolvimento de cartas de controlo para dados multivariados é crucial. Vamos considerar a carta de controlo proposta em Figueiredo e Figueiredo [6], baseada no coeficiente RV definido em [5], para monitorizar a estrutura de covariâncias de um processo multivariado. Para processos normais bivariados, estudaremos o desempenho da carta através do ARL (*average run length*), analisaremos também a distribuição do coeficiente RV quando o processo está sob controlo e estudaremos ainda várias caraterísticas da distribuição do RL (*run length*) quando o processo está fora de controlo.

1 Introdução

As cartas de controlo são as ferramentas usualmente utilizadas para a monitorização de processos em Controlo Estatístico da Qualidade. As cartas de controlo foram introduzidas por Shewhart em 1924 nos Bell Laboratories para a monitorização de processos industriais,

mas atualmente são aplicadas nas mais diversas áreas, entre elas, na Saúde e Medicina ([13]) e Genética, Ambiente e Finanças ([10]). As cartas de controlo são representações gráficas que têm por objetivo ajudar a tomar decisões sobre o estado do processo que está a ser monitorizado, isto é, ajudar a decidir se está sob controlo ou fora de controlo. Na literatura têm sido propostas várias cartas de controlo para monitorizar o vetor de valores esperados de um processo multivariado como por exemplo, a carta baseada na estatística T^2 de Hotelling [8], e variantes desta carta. Adicionalmente, diversas cartas para controlar a variabilidade de um processo multivariado têm sido propostas, tais como a carta baseada na variância generalizada $|S|$, variantes desta carta ([1] e outras), cartas baseadas no máximo das variâncias amostrais ou no máximo das amplitudes amostrais das p caraterísticas em estudo ([3], [4]), cartas propostas em [7], [11], [12], [14], entre outras. Existem também esquemas de controlo para monitorizar simultaneamente o vetor de valores esperados e a matriz de covariâncias do processo ([2], [15], etc.)

Neste trabalho iremos recorrer ao coeficiente RV (proposto em [5]) para desenvolver uma carta de controlo para monitorizar a estrutura de covariâncias associada a um conjunto de caraterísticas de um processo de controlo multivariado. Para o efeito, com base no coeficiente RV iremos comparar a matriz de covariâncias das p variáveis em estudo, associada a um conjunto de amostras de referência retiradas do processo quando o processo está sob controlo com a matriz de covariâncias dessas variáveis obtida num novo instante de tempo. Prosseguindo o trabalho [6], iremos neste estudo explorar a distribuição do coeficiente RV no contexto referido, de modo a podermos decidir devidamente sobre o estado do processo, i.e., se está sob controlo ou fora de controlo. Note-se que a distribuição exata do coeficiente RV não é conhecida e apenas existem aproximações a esta distribuição em determinados casos particulares. Em seguida, iremos avaliar o desempenho do procedimento para dois processos normais bivariados, com base em várias caraterísticas da distribuição do RL, a qual será previamente analisada para cada processo específico.

O trabalho está estruturado do seguinte modo: na secção 2 descrevemos a carta-RV; na secção 3 discutimos o desempenho da carta para processos normais bivariados.

2 Carta de controlo baseada no coeficiente RV

O coeficiente RV proposto por Escoufier [5] permite medir a semelhança entre duas matrizes semi-definidas positivas. Iremos utilizar este coeficiente como medida de semelhança entre duas matrizes de covariâncias na carta que vamos propor. O coeficiente RV entre as matrizes de covariâncias V_k e $V_{k'}$ é definido por

$$RV(V_k,V_{k'}) = \frac{\langle V_k,V_{k'}\rangle_{HS}}{\|V_k\|_{HS}\,\|V_{k'}\|_{HS}} = \frac{Tr(V_kV_{k'})}{\sqrt{Tr\left({V_k}^2\right)Tr\left({V_{k'}}^2\right)}},$$

onde $\langle V_k,V_{k'}\rangle_{HS} = Tr\left(V_kV_{k'}\right)$ representa o produto escalar de Hilbert-Schmidt entre V_k e $V_{k'}$, Tr representa o traço de uma matriz e $\|V_k\|_{HS} = \sqrt{\langle V_k,V_k\rangle_{HS}} = \sqrt{Tr\left(V_k\right)^2}$ é a norma Hilbert-Schmidt de V_k. O coeficiente RV varia entre 0 e 1 e quanto mais próximo de 1 for o coeficiente RV mais semelhantes são as matrizes V_k e $V_{k'}$.

A estrutura de covariâncias de um processo multivariado pode estimar-se através de uma matriz de covariâncias compromisso obtida a partir de um conjunto de amostras de referência retiradas do processo quando está sob controlo. Em seguida, definimos a matriz de covariâncias compromisso como na metodologia STATIS ([9]).

Consideremos: K amostras de referência de dimensão n em que cada observação da amostra é descrita por p variáveis, i.e., K quadros de dados $X_{n\times p}$ recolhidos em K instantes de tempo diferentes, quando o processo está sob controlo; as matrizes de covariâncias associadas a estes quadros de dados.

A matriz de covariâncias compromisso é definida como uma média

ponderada das K matrizes de covariâncias V_k:

$$V_{comp} = \sum_{k=1}^{K} \alpha_k V_k,$$

onde os pesos α_k são os elementos do vetor próprio associado ao maior valor próprio da matriz Z dos coeficientes RV entre os V_k's:

$$Z = \begin{pmatrix} 1 & RV(V_1,V_2) & \cdots & RV(V_1,V_K) \\ RV(V_2,V_1) & 1 & \cdots & RV(V_2,V_K) \\ \vdots & \vdots & \ddots & \vdots \\ RV(V_K,V_1) & RV(V_K,V_2) & \cdots & 1 \end{pmatrix}$$

A carta de controlo, que designamos por carta-RV, é implementada do seguinte modo.

- Para cada nova amostra j retirada do processo, representamos o valor do coeficiente RV entre a matriz de covariâncias associada a esta amostra e a matriz de covariâncias compromisso:

$$RV(V_j,V_{comp}) = \frac{Tr(V_j V_{comp})}{\sqrt{Tr\left({V_j}^2\right) Tr\left({V_{comp}}^2\right)}}.$$

- O Limite de Controlo da carta, LC, é calculado a partir da taxa de falsos alarmes:

$$\alpha = P(RV < LC | \text{processo está sob controlo}).$$

  Como a distribuição exata do coeficiente RV é desconhecida, fixamos o limite de controlo LC num percentil empírico de ordem α da distribuição por amostragem do coeficiente RV, quando o processo está sob controlo.

- Se $RV(V_j,V_{comp}) < LC$, considera-se que o processo está fora de controlo. Caso contrário, considera-se que o processo está sob controlo.

3 Desempenho da carta para um processo normal bivariado

Nesta secção, vamos analisar o desempenho da carta-RV para processos normais bivariados, usando como medida de eficiência o ARL - número esperado de amostras retiradas até à emissão de sinal, para uma taxa de falsos alarmes α de 0.005. Para processos normais multivariados de dimensão superior ($p = 3$ e $p = 4$), pode ver-se o desempenho da carta-RV em [6].
Nesta secção vamos também explorar a distribuição empírica do coeficiente RV quando o processo normal bivariado está sob controlo e estudar a distribuição do RL quando o processo está fora de controlo.

Gerámos processos normais bivariados $N_2(\boldsymbol{\mu},\Sigma)$, com $\boldsymbol{\mu} = (0,0)'$ e $\Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$. Considerámos diferentes estruturas da matriz de covariâncias quando o processo está sob controlo e quando está fora de controlo.

Observação 3.1 *Note-se que se considerássemos outro vetor de valores esperados, obteríamos os mesmos resultados, visto que o desempenho da carta é independente de tal vetor.*

Observação 3.2 *Note-se que a covariância entre as variáveis coincide com a correlação dado que as variâncias são unitárias.*

Para uma taxa de falsos alarmes $\alpha = 0.005$, estimamos o LC da carta; trata-se do quantil empírico de ordem 0.005 da distribuição sob controlo do coeficiente RV, obtido por simulação de Monte Carlo a partir de 100000 réplicas, recorrendo ao algoritmo descrito de seguida.

Algoritmo 3.3 *Para cada i=1,100000, repetir os passos:*

1. *Gerar 4 amostras de referência de dimensão n da distribuição $N_2(\boldsymbol{\mu}\Sigma)$,supondo o processo sob controlo.*

2. *Determinar a matriz de covariâncias compromisso,* V_{comp}.
3. *Gerar uma nova amostra de dimensão n da distribuição* $N_2(\boldsymbol{\mu},\Sigma)$, *supondo o processo sob controlo e determinar a respetiva matriz de covariâncias* V_i.
4. *Calcular* $RV(V_i,V_{comp})$.
   Determinar o quantil de ordem 0.005 dos valores RV obtidos.

A distribuição dos valores do coeficiente RV, quando o processo está sob controlo, encontra-se nas figuras 1 e 2, para a covariância nula e a covariância igual 0.75, respetivamente.

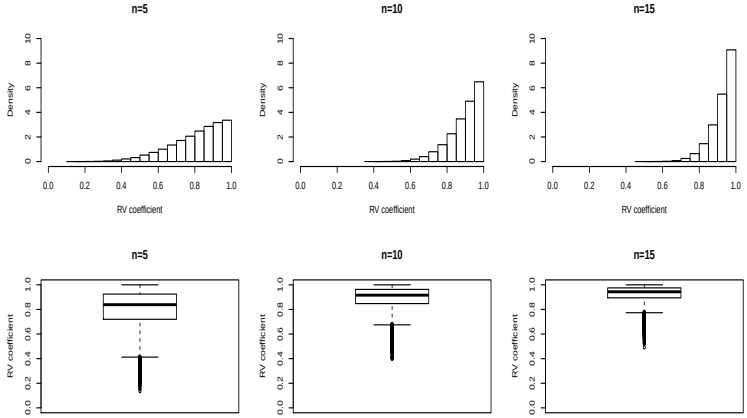


Figura 1: Distribuição do coeficiente RV quando o processo está sob controlo (covariância nula), para n = 5,10,15

Observa-se que a distribuição do coeficiente RV é assimétrica negativa, predominando os valores mais elevados do coeficiente RV, isto é, mais próximos de 1. Observa-se ainda, em qualquer dos casos, a existência de bastantes *outliers* inferiores na distribuição do coeficiente RV.

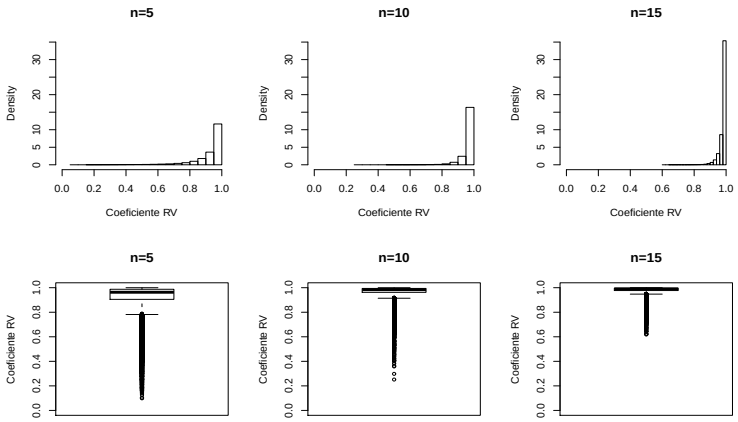


Figura 2: Distribuição do coeficiente RV quando o processo está sob controlo (covariância igual a 0.75), para $n = 5{,}10{,}15$

A distribuição do coeficiente RV está cada vez mais concentrada à direita (a dispersão diminui) à medida que aumenta a dimensão da amostra, isto é, o coeficiente RV toma valores cada vez mais próximos de 1 à medida que n aumenta. Assim, o Limite de Controlo da carta aumenta à medida que aumenta a dimensão da amostra.

Como medida de eficiência da carta, usamos o ARL. Para uma taxa de falsos alarmes $\alpha = 0.005$, estimámos o ARL a partir de 10000 réplicas para diferentes estruturas da matriz de covariâncias, recorrendo ao algoritmo descrito a seguir.

Algoritmo 3.4 *Para cada i=1,...,10000 (réplicas)*

1. *Repetir os passos abaixo até que a carta-RV emita sinal.*

   (a) *Gerar 4 amostras de referência de dimensão n da distribuição* $N_2(\boldsymbol{\mu},\Sigma)$ *com o processo sob controlo.*

(b) Calcular V_{comp}.

(c) Gerar uma amostra da distribuição $N_2(\boldsymbol{\mu},\Sigma)$ com o processo fora de controlo.

(d) Calcular $RV(V_i,V_{comp})$ e comparar com LC previamente estimado, usando o algoritmo 3.3.

2. Registar o número de amostras até à emissão de sinal, RL_i.

Calcular a média dos valores RL, i.e., a estimativa de ARL.

Na Tabela 1 apresentamos os resultados obtidos para n = 5,10,15 quando $\sigma_{12} = 0$ ou $\sigma_{12} = 0.75$ sob controlo. As estimativas de ARL sob controlo estão a negrito.

Tabela 1: Estimativas de LC e ARL para n = 5,10,15, sendo $\sigma_{12} = 0$ ou $\sigma_{12} = 0.75$ quando o processo está sob controlo

	σ_{12}= 0 sob controlo				σ_{12}= 0.75 sob controlo		
n	5	10	15	n	5	10	15
LC	0.360	0.593	0.698	LC	0.390	0.747	0.863
σ_{12}		ARL		σ_{12}		ARL	
-0.95	31.2	6.3	2.5	0.75	**178.4**	**204.9**	**180.3**
-0.5	122.3	48.2	22.4	0.5	39.7	16.6	9.4
-0.3	168.7	105.7	66.9	0.3	18.1	5.6	3.1
0	**198.6**	**188.3**	**193.4**	0.1	9.7	2.8	1.6
0.3	161.9	105.7	66.9	0	7.2	2.1	1.8
0.5	121.8	48.8	22.6	-0.3	3.5	1.8	1.1
0.75	68.0	15.8	6.0	-0.5	2.3	1.1	1.0
0.95	31.9	6.3	2.5	-0.75	1.4	1.0	1.0

A partir da Tabela 1, podem tirar-se as seguintes conclusões:

- O limite de controlo da carta e o ARL dependem da dimensão da amostra e da estrutura da matriz de correlações.

- O ARL quando o processo está sob controlo é elevado e aproximadamente igual ao valor esperado $\alpha^{-1} = 200$. Quando o processo está fora de controlo, o ARL diminui rapidamente à medida que a dimensão da amostra aumenta.

- Se a correlação é nula ou próxima de 0 sob controlo, a carta deteta a existência de correlações positivas ou negativas, sendo as maiores correlações em valor absoluto mais facilmente detetáveis.

- Se a correlação é elevada e positiva sob controlo, a carta-RV deteta diminuições na correlação, correlações negativas e também correlações nulas. No entanto, a carta-RV não deteta atempadamente aumentos de σ_{12} quando o seu valor alvo é igual a 0.75. A carta é mais sensível a detetar correlações elevadas negativas.

A distribuição do RL quando o processo está fora de controlo (covariância 0.3, 0.5 e 0.95) e a covariância é nula sob controlo é apresentada para $n = 5$ (Figura 3) e $n = 15$ (Figura 4). Considerámos outros valores da covariância quando o processo está fora de controlo, como por exemplo, 0.75, e os resultados foram semelhantes aos apresentados nas figuras 3 e 4. Resultados adicionais para o caso em que a covariância sob controlo é 0.75 permitiram tirar conclusões análogas às obtidas apartir das figuras 3 e 4.
Em qualquer um dos casos apresentados, a distribuição do RL é assimétrica positiva, pelo que há predominância de valores baixos do RL. Verifica-se que a mediana do RL é inferior ao valor médio do RL, i.e., ao ARL. Observa-se a existência de *outliers* superiores na distribuição do RL em qualquer uma das situações em que o processo está fora de controlo. À medida que nos afastamos da estrutura de correlações sob controlo, a dispersão da distribuição do RL diminui.

Observação 3.5 *Se a matriz de covariâncias compromisso V_{comp} é conhecida a priori, a distribuição do RL é geométrica, mas se estimarmos os parâmetros, a distribuição do RL não é geométrica.*

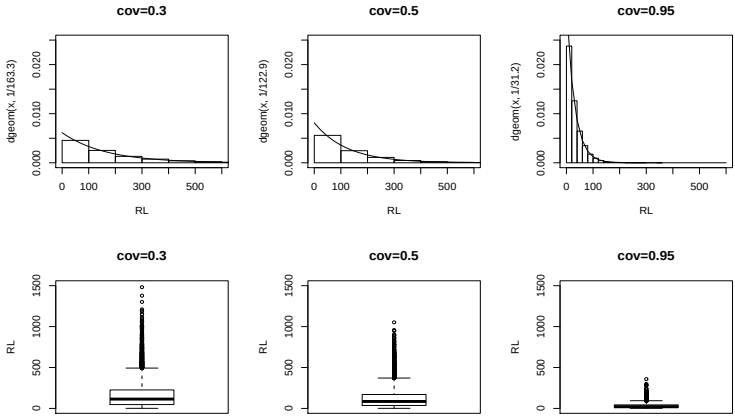


Figura 3: Distribuição do RL quando o processo está fora de controlo, para $n = 5$ e covariância sob controlo nula.

Neste caso, estimamos os parâmetros porque calculamos a matriz de covariâncias compromisso, mas a distribuição geométrica parece ajustar-se bem ao RL (ver figuras 3 e 4), pelo que se pode calcular a mediana e outros quantis da distribuição do RL a partir da distribuição geométrica.

Os resultados de simulação obtidos para processos normais bivariados sugerem que a carta-RV permite detetar facilmente alterações na correlação entre as variáveis, podendo constituir assim uma técnica de monitorização muito útil numa grande variedade de aplicações industriais.
Neste trabalho considerámos apenas dois valores da covariância sob controlo $\sigma_{12} = 0$ e $\sigma_{12} = 0.75$, mas em [6] foram considerados outros valores de σ_{12} e os resultados obtidos foram semelhantes aos apresentados aqui.
Para processos normais com $p = 3$ e $p = 4$ (ver [6]), os resultados

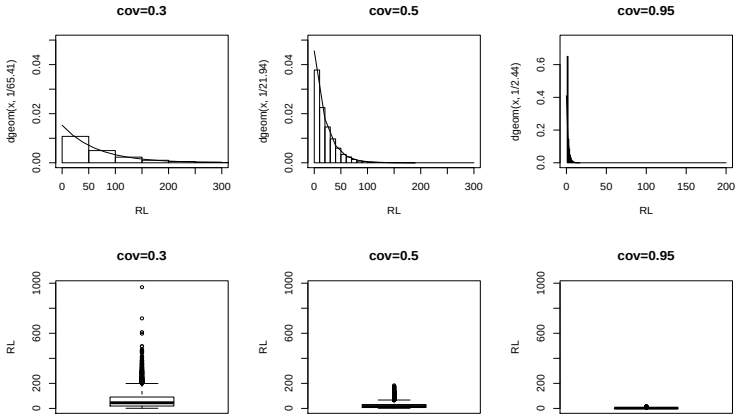


Figura 4: Distribuição do RL quando o processo está fora de controlo, para $n = 15$ e covariância sob controlo nula.

são análogos aos obtidos para $p = 2$ e permitem tirar conclusões que apoiam as referidas neste trabalho.
Referimos ainda que a carta-RV constitui uma contribuição útil para as cartas existentes na literatura para a monitorização da matriz de covariâncias.

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT (Fundação para a Ciência e a Tecnologia) no âmbito dos projetos UID/EEA/50014/2013 e PEst-OE/MAT/UI0006/2014.

Referências

[1] Alt, F.B. (1985). Multivariate quality control. In Kotz, S., Johnson, N.L. (eds.): *Encyclopedia of Statistical Sciences* (Vol. 6), 111-122. Wiley, New York.

[2] Chen, G., Cheng, S.W. and Xie, H. (2005). A new multivariate control chart for monitoring both location and dispersion, *Communications in Statistics - Simulation and Computation* 34, 203-217.

[3] Costa, A.F.B. and Machado, M.A.G. (2008a). A new chart based on sample variances for monitoring the covariance matrix of multivariate processes, *International Journal of Advanced Manufacturing Technology* 41, 770-779.

[4] Costa, A.F.B. and Machado, M.A.G. (2008b). A new multivariate control chart for monitoring the covariance matrix of bivariate processes, *Communications in Statistics – Simulation and Computation* 37, 1453-1465.

[5] Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 29, 751–760.

[6] Figueiredo, A. and Figueiredo, F. (2014). Monitoring the variability of a multivariate normal process using STATIS. In Gilli, M.,Gonzalez-Rodriguez, G., Neto-Reyes, A. (eds.): *Proceedings of COMPSTAT 2014*, 443-450.

[7] Hawkins, D. M. and Maboudou-Tchao, E. M. (2008). Multivariate exponentially weighted moving covariance matrix. *Technometrics* 50, 155-166.

[8] Hotelling, H. (1947). Multivariate quality control, illustrated by the air testing of sample bombsights. In Eisenhart, C., Hastay, M.W., Wallis, W.A. (eds.): *Techniques of Statistical Analysis*, 111-184. McGraw Hill, New York.

[9] Lavit, C., Escoufier, Y., Sabatier, R. and Traissac, P. (1994). The ACT (Statis method), *Computational Statistics and Data Analysis* 18, 97-119.

[10] Stoumbos, Z.G., Reynolds, Jr., M., Ryan, T.P. and Woodall, W.H. (2000). The state of statistical process control as we proceed into 21st century, *Journal of the American Statistical Association* 95, 992-998.

[11] Tang, P.F., Barnett, N.S. (1996a). Dispersion control for multivariate processes. *Australian Journal of Statistics* 38, 235-251.

[12] Tang, P.F., Barnett, N.S. (1996b). Dispersion control for multivariate processes - some comparisons. *Australian Journal of Statistics* 38, 253-273.

[13] Woodall, W.H. (2006). The use of control charts in health-care and public-health surveillance, *Journal of Quality Technology* 38, 89-104.

[14] Yeh, A.B., Lin, D.K.-J., McGrath, R.N. (2006). Multivariate control charts for monitoring covariance matrix: a review. *Quality Technology and Quantitative Management* 3, 415-436.

[15] Zhang, G. and Chang, S.I. (2008). Multivariate EWMA control charts using individual observations for process mean and variance monitoring and diagnosis, *International Journal of Production Research* 46, 6855-6881.

Distribuição de Pareto inflacionada em Controlo Estatístico da Qualidade

Fernanda Otília Figueiredo
Faculdade de Economia da Universidade do Porto, e CEAUL, Universidade de Lisboa, *otilia@fep.up.pt*

Adelaide Maria Figueiredo
Faculdade de Economia e LIAAD–INESC TEC Porto, Universidade do Porto, *adelaide@fep.up.pt*

M. Ivette Gomes
FCUL e CEAUL, Universidade de Lisboa, *ivette.gomes@fc.ul.pt*

Palavras–chave: controlo estatístico da qualidade, distribuição de Pareto inflacionada, planos de amostragem de aceitação por variáveis.

Resumo: Em medições efetuadas por cromatografia, devido a limitações frequentes da precisão dos cromatógrafos utilizados, justifica-se que os valores observados sejam truncados em determinados limiares (inferiores e/ou superiores), sendo por isso a distribuição subjacente aos dados inflacionada nestes valores. Neste trabalho mostramos a adequabilidade da distribuição de Pareto inflacionada para modelar este tipo de dados truncados e inflacionados, e definimos um plano de amostragem de aceitação por variáveis para inspecionar lotes de itens provenientes desta distribuição que será de grande utilidade prática.

1 Introdução

Em várias indústrias é importante controlar a presença de certas substâncias químicas que afetam a qualidade da matéria prima e

produtos finais. Em geral este tipo de controlo é feito através de análises de cromatografia realizadas em amostras retiradas de lotes de grande dimensão. A maior parte dos cromatógrafos utilizados não têm precisão suficiente para medir concentrações muito reduzidas ou demasiado elevadas destas substâncias, digamos abaixo ou acima de determinados limiares conhecidos, pelo que as medições resultantes são provenientes de distribuições contínuas inflacionadas, sendo muitos dos valores observados iguais ao valor destes limiares. Neste trabalho iremos dar ênfase à utilização de uma distribuição de Pareto inflacionada para modelar este tipo de dados truncados e inflacionados.

É importante referir que, mesmo controlando a qualidade da matéria prima e do produto final ao longo das várias etapas da sua produção, a existência destas substâncias, mesmo que em diminutas quantidades, pode violar as regras impostas pelas normas ISO 22000 Segurança Alimentar, por exemplo, e consequentemente acarretar elevados custos financeiros para as empresas. Assim para inspecionar os lotes de matéria prima e produto final será importante definir planos de amostragem de aceitação/rejeição de lotes que sejam eficientes, e por isso específicos para este tipo de dados.

Embora os planos de amostragem por atributos sejam os mais comuns, sendo as medições provenientes de um modelo contínuo, um plano de amostragem por variáveis é mais conveniente. É de referir que a dimensão da amostra que é necessária para garantir um determinado nível de proteção (por exemplo, em termos de risco do produtor/consumidor) é em geral menor no caso dos planos de amostragem por variáveis, sendo a principal desvantagem a apontar a estes planos a necessidade de conhecer a distribuição subjacente aos dados ou de a estimar.

Em Figueiredo, Figueiredo e Gomes [5] recorremos à metodologia *bootstrap* (Efron [3], Efron e Tibshirani [4], Davison e Hinkley [2]) e a simulações de Monte Carlo para comparar o desempenho de alguns planos de amostragem por variáveis para o mesmo tipo de dados, em que a representação gráfica dos mesmos sugere uma distribuição truncada e inflacionada. Para mais detalhes sobre planos de amos-

tragem ver, por exemplo, Montegomery [7], Gomes [6] e Carolino e Barão [1].
Neste trabalho, na Seção 2, iremos mostrar a adequabilidade da distribuição de Pareto inflacionada para modelar um conjunto de dados reais associados a análises de cromatografia, recorrendo ao método da máxima verosimilhança. Na Seção 3 definimos um plano de amostragem por variáveis para inspecionar lotes de grande dimensão num contexto de inspeção retificativa, admitindo que existe um limite de especificação superior para as medições. Analisamos as curvas *característica operacional* e da *qualidade média à saída* associadas a este plano, em termos do valor que decorre para os riscos do produtor e do consumidor face a algumas dimensões amostrais consideradas, e níveis de qualidade aceitável e rejeitável previamente fixados. Finalmente apresentamos algumas conclusões na Secção 4.

2 Distribuição de Pareto inflacionada

Seja X uma variável aleatória (v.a.) associada a um modelo de Pareto inflacionado, com *função distribuição* (f.d.) dada por

$$F(x;p,\xi,\delta) = p + (1-p)(1-(x/\delta)^{-1/\xi}), \quad x \geq \delta, \tag{1}$$

e *função densidade de probabilidade* (f.d.p.) dada por

$$f(x;p,\xi,\delta) = \begin{cases} (1-p)(\xi\delta)^{-1}(x/\delta)^{-1/\xi-1}, & x > \delta, \\ p, & x = \delta, \end{cases} \tag{2}$$

onde δ e ξ são, respetivamente, os parâmetros de escala e de forma, ambos positivos, e o parâmetro p representa a probabilidade associada ao ponto $x = \delta$, ou seja, ao ponto de truncatura do limite inferior do suporte da v.a. X, onde a distribuição subjacente aos dados aparece inflacionada. Note-se que se p e δ são fixos, quanto maior ξ, maior é o peso da cauda direita da distribuição subjacente aos dados, e consequentemente, maior a frequência de valores muito elevados.

2.1 Análise preliminar de um conjunto de dados de cromatografia

Para enfatizar a importância da distribuição de Pareto inflacionada em aplicações práticas, iremos considerar um conjunto de dados reais associados a medições (por cromatografia) da concentração de uma determinada substância química em itens de matéria prima retirados de um lote de grande dimensão. Para termos uma ideia do tipo de dados em estudo, apresentamos na Tabela 1 um quadro de frequências associado a estes dados. Por questões de confidencialidade não iremos referir qual o tipo de produto em análise nem a indústria que forneceu os dados. Apenas referimos que obtivemos o mesmo tipo de dados em diferentes tipos de matéria prima e de produto final analisados no estudo de consultadoria efetuado para esta empresa.

O objetivo da empresa é controlar os níveis de concentração desta substância química em lotes de grande dimensão, impondo o valor 4.0 como limite superior de especificação para o nível de concentração desta substância em cada item. Devido a alguma falta de precisão dos cromatógrafos que estão a ser utilizados nas medições, valores abaixo de 0.5 não oferecem garantia de estarem a ser bem quantificados, e por isso, com base em alguma experiência passada resultante de repetição das medições, todos estes valores são registados como sendo iguais a 0.5 (o que vai originar uma distribuição subjacente aos dados inflacionada neste valor).

A partir da Tabela 1 observamos que em 95% dos itens o nível de concentração da substância química é inferior ou igual a 4.0, e por isso aproximadamente 5% destes itens não satisfazem os requisitos da empresa. É de referir que 56.7% das medições são registadas com o valor 0.5, o que significa ou ausência da substância química ou quantificação não fiável devido a falta de precisão do equipamento. A distribuição subjacente aos dados além de ser inflacionada em 0.5, apresenta uma cauda direita pesada. Esta análise preliminar dos dados levou-nos a averiguar se o ajuste de uma distribuição de Pareto (distribuição de cauda direita pesada, muito utilizada em Teoria de Valores Extremos) inflacionada no valor 0.5 era adequado.

Tabela 1: Valores (agrupados em classes) da concentração da substância química em 1600 itens de matéria prima retirados de um lote para análise.

Classes	Número de itens (%)	Classes	Número de itens (%)
0.5	908 (56.7%)	]4.0,5.0]	20 (1.3%)
]0.5,1.0]	357 (22.3%)	]5.0,7.5]	17 (1.1%)
]1.0,2.0]	187 (11.7%)	]7.5,10.0]	16 (1.0%)
]2.0,3.0]	53 (3.3%)	]10.0,20.0]	10 (0.6%)
]3.0,4.0]	16 (1.0%)	>20.0	16 (1.0%)

2.2 Ajuste da distribuição de Pareto inflacionada ao conjunto de dados de cromatografia

Seja $(X_1,...,X_n)$ uma amostra aleatória de dimensão n de um modelo de Pareto inflacionado com f.d.p. $f(x;p,\xi,\delta)$ definida em (2). As estimativas de *máxima verosimilhança* (MV) dos parâmetros p, ξ e δ são os valores que maximizam o logaritmo da função de verosimilhança definida por

$$\ln L(p,\xi,\delta) = n_1 \ln p + n_2 \ln(1-p) - n_2 \ln(\xi\delta) - (1/\xi+1) \sum_{i=1}^{n_2} \ln(x_i/\delta), \tag{3}$$

onde n_1 e n_2 denotam, respetivamente, o número de observações iguais e maiores do que δ na amostra global de dimensão n. Assim, as estimativas de MV são definidas por:

$$\widehat{p} = \frac{n_1}{n}, \widehat{\xi} = \frac{1}{n_2} \sum_{i=1}^{n_2} \ln(x_i/\delta) \text{ e } \widehat{\delta} = \min x_i. \tag{4}$$

Para ajustar esta distribuição ao conjunto de dados de cromatografia em estudo, começamos por fixar $\delta = 0.5$, pois o equipamento não tem precisão suficiente para medir com rigor valores abaixo deste

limiar, e depois estimamos os outros parâmetros do modelo, ξ e p, pelo método da MV, tendo-se procedido do seguinte modo:

- Separamos os valores da amostra global de dimensão $n = 1600$ que são iguais a $\delta = 0.5$ (subamostra de dimensão n_1), dos restantes valores superiores a δ (subamostra de dimensão n_2);
- Para estimar p, consideramos a proporção de observações iguais a $\delta = 0.5$ na amostra global, e obtivemos $\hat{p} = 908/1600 = 0.5675$;
- Para estimar ξ, consideramos $\hat{\xi} = \sum_{i=1}^{n_2} \ln(x_i/\delta)/n_2$, e obtivemos $\hat{\xi} = 0.9286$.

Na Figura 1, apresentamos o histograma associado ao conjunto de dados e a curva da f.d.p. (estimada) correspondente à distribuição de Pareto inflacionada ajustada. Como se pode observar, o modelo é adequado para descrever este tipo de dados, o que se confirma também pelo valor que obtivemos para o valor-p do teste de ajustamento do qui-quadrado, que também efetuamos (valor-p=0.0655>0.05). De acordo com o modelo ajustado, a estimativa para a probabilidade de se obter uma medição superior ao valor 4.0 num item de matéria prima é de 4.61% (valor próximo da percentagem de observações superiores a 4.0 na amostra).

3 Plano de amostragem por variáveis

Suponhamos lotes de dimensão N bastante elevada, e que a característica de qualidade X a observar é proveniente de um processo com f.d. dada em (1), existindo um limite superior de especificação (LSE) para os valores a observar. Após a estimação *a priori* de p e de δ, vamos admitir que estes parâmetros são fixos e conhecidos.
Os planos de amostragem de aceitação mais comuns são delineados para controlar a fração de itens defeituosos, no nosso caso

$$\theta = \mathbb{P}(X > \text{LSE}) = (1-p)(\delta/\text{LSE})^{1/\xi}, \tag{5}$$

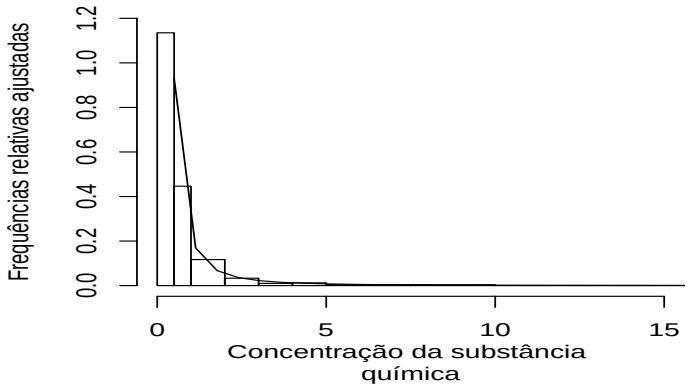


Figura 1: Histograma com os valores (agrupados em classes) da concentração da substância química nos 1600 itens analisados, construído com as frequências relativas ajustadas pelas amplitudes das classes, e f.d.p. da distribuição de Pareto inflacionada ajustada aos dados.

ou equivalentemente, um parâmetro do processo associado à produção de defeituosos, no nosso caso ξ, o qual pode ser expresso em função de θ e LSE através da expressão

$$\xi = \ln(\delta/\text{LSE})/\ln(\theta/(1-p)). \tag{6}$$

Note-se que para δ e p fixos, θ será pequeno se ξ o fôr.

3.1 Determinação de um plano de amostragem

De um modo geral, definir um plano de amostragem simples consiste em determinar a dimensão da amostra e a constante de aceitação que

para uma dada regra de decisão (função da estatística de controlo escolhida) permite obter um plano com um determinado desempenho, em geral, valores pré-determinados para os riscos do produtor e do consumidor. Por vezes, devido a restrições orçamentais e operacionais, é necessário definir um plano de amostragem para uma dimensão amostral fixa, escolhendo a constante de aceitação que permita obter um pré-determinado valor para o risco do produtor. Seja $\mathbb{P}(A|\theta)$ a probabilidade de aceitação de um lote com uma fração de defeituosos θ. Definindo *nível de qualidade aceitável* (NQA) como o pior nível de qualidade média para o processo que o produtor considera aceitável, esperando contudo que o processo opere com um nível de qualidade melhor do que este, e *nível de qualidade rejeitável* (NQR) como o pior nível de qualidade que o consumidor tolera aceitar num lote individual, os riscos do produtor e do consumidor, α e β, respectivamente, são definidos por

$$\alpha = \mathbb{P}(\overline{A}|\theta = \text{NQA}) \quad \text{e} \quad \beta = \mathbb{P}(A|\theta = \text{NQR}). \tag{7}$$

No nosso caso, sendo o estimador de máxima verosimilhança de ξ, $\hat{\xi} = \dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2} Y_i = \overline{Y}$, com $Y_i = \ln(X_i/\delta)$, consistente para ξ, obtido após dispormos de uma amostra retirada do processo onde n_2 das n observações da amostra são maiores do que δ, e atendendo a que a estatística $2n_2\overline{Y}/\xi$ segue uma distribuição qui-quadrado com $2n_2$ graus de liberdade, que denotaremos por $\chi^2_{2n_2}$, um plano de amostragem óbvio será baseado na seguinte regra de decisão:

$$\text{Aceitar o lote se } \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \leq k. \tag{8}$$

O valor n_2 e a constante de aceitação k do plano que permite obter riscos α e β fixos, para níveis de qualidade NQA e NQR também fixos, e para um limite superior de especificação LSE, devem satisfazer as condições

$$\begin{cases} \mathbb{P}\left(\overline{Y} \leq k | \xi = \dfrac{\ln(\delta/\text{LSE})}{\ln(\text{NQA}/(1-p))}\right) = 1-\alpha \\ \mathbb{P}\left(\overline{Y} \leq k | \xi = \dfrac{\ln(\delta/\text{LSE})}{\ln(\text{NQR}/(1-p))}\right) = \beta. \end{cases}$$

A partir destas equações determinamos o valor n_2 tal que

$$n_2 : \; F^{-1}_{\chi^2_{2n_2}}(1-\alpha) = \frac{\ln(\text{NQA}/(1-p))}{\ln(\text{NQR}/(1-p))} F^{-1}_{\chi^2_{2n_2}}(\beta) \tag{9}$$

e depois k é dado por

$$k = \frac{1}{2n_2} \frac{\ln(\delta/\text{LSE})}{\ln(\text{NQA}/(1-p))} F^{-1}_{\chi^2_{2n_2}}(1-\alpha), \tag{10}$$

onde $F^{-1}_{\chi^2_{2n_2}}$ denota a inversa da f.d. da distribuição $\chi^2_{2n_2}$.

3.2 Desempenho do plano de amostragem

No contexto do exemplo em estudo, considerando um limite superior de especificação LSE = 4.0, itens associados a medições de concentração da substância química acima de 4.0 são considerados defeituosos. Para ilustrar o desempenho do plano de amostragem definido na Subseção 3.1, assumimos δ conhecido igual a 0.5, $p = 0.5675$, fixo e igual ao valor da estimativa de MV, e admitimos que a deterioração da qualidade do lote se deve essencialmente a alterações no parâmetro ξ da distribuição.

Em muitas situações a análise do desempenho e a comparação de planos de amostragem pode ser efetuada, de forma satisfatória, com base na análise da curva *caraterística operacional* (CO), e da curva da *qualidade média à saída* (QMS). A curva CO, i.e., a curva que se ajusta aos pontos

$$(\theta, \mathbb{P}(A|\theta)), \text{para}\, \theta = 0, 1/N, \ldots, 1, \tag{11}$$

mostra o poder discriminatório do plano para aceitar/rejeitar lotes consoante a sua fração de defeituosos. A curva QMS, i.e., a curva

que se ajusta aos pontos

$$(\theta,\mathrm{QMS}(\theta) = \theta \times \mathbb{P}(A|\theta)) \,, \mathrm{para}\, \theta = 0{,}1/N, \ldots, 1, \qquad (12)$$

descreve aproximadamente a qualidade média de um lote de grande dimensão que resulta de um programa de retificação de lotes a 100% aplicado a uma sequência de lotes provenientes do mesmo processo. Na maior parte dos processos retificativos, retificar um lote a 100% consiste no seguinte: quando um lote é rejeitado pelo plano de amostragem, todos os itens do lote são inspecionados individualmente, e os itens defeituosos encontrados são substituídos por bons. Neste caso particular de estudo, e atendendo ao tipo de produto que se está a inspecionar, os lotes rejeitados são sujeitos ao seguinte tipo de retificação especial: todo o lote rejeitado irá ser alvo de um tratamento de limpeza especial, que na maior parte dos casos será suficiente para eliminar ou pelo menos reduzir a concentração da substância química existente em alguns dos itens; após este tratamento, os itens do lote irão ser repartidos e misturados com itens de outros lotes, que irão em seguida ser sujeitos a planos de amostragem de aceitação/rejeição, antes de entrarem em linha de produção ou serem enviados para venda. É de realçar que o QMS é um indicador de um nível médio de qualidade que resulta da inspeção de muitos lotes, e que por isso um lote particular pode ter uma qualidade pior. O valor máximo da curva QMS, denotado por LQMS, representa a pior qualidade média à saída que resulta da aplicação de um programa retificação de lotes a 100%. No nosso caso será aproximadamente igual à proporção de itens num lote com nível de concentração da substância química acima de 4.0 que passam o controlo, i.e., são enviados para a linha de produção (no caso de matéria prima) ou para venda (no caso de produto final). Finalmente será de referir que na avaliação do desempenho dos lotes submetidos a inspeção retificativa é usual ter-se também em consideração o número médio de itens inspecionados por lote. Atendendo ao tipo de retificação especial efetuada neste caso de estudo aos lotes rejeitados, não faz sentido calcular-se tal indicador, visto que o tratamento de limpeza

não é efetuado individualmente a cada item do lote mas sim ao lote na sua globalidade.

Na Figura 2 apresentamos as curvas CO e QMS associadas a riscos $\alpha = 15\%$ e $\beta = 30\%$ quando NQA = 5% e NQR = 10% (Plano I), e riscos $\alpha = 5\%$ e $\beta = 10\%$ quando NQA = 2.5% e NQR = 10% (Plano II). Os parâmetros n_2 e k que garantem planos com este nível de proteção foram obtidos através das equações (9) e (10). É de referir que a dimensão global da amostra que garante a obtenção deste valor n_2 é uma variável aleatória com distribuição binomial negativa, de valor médio $n_2/(1-p)$. Para a implementação prática dos planos de amostragem I e II pressupomos a existência de uma amostra de referência que nos forneça uma estimativa para p, e apenas podemos sugerir que se considere uma amostra de dimensão global $n_2/(1-p)$, sendo n_2 o valor (ideal) que desejaríamos ter para a dimensão da subamostra que permite implementar os planos com o nível de proteção referida. Obviamente que depois de observada tal amostra podemos ter um número de observações superior a δ mais ou menos próximo do valor n_2 ideal, e consequentemente o desempenho efetivo dos planos será mais ou menos semelhante aquele que é ilustrado nas figuras seguintes (apenas válido para a ordem de grandeza dos valores n_2 considerados nesta ilustração).

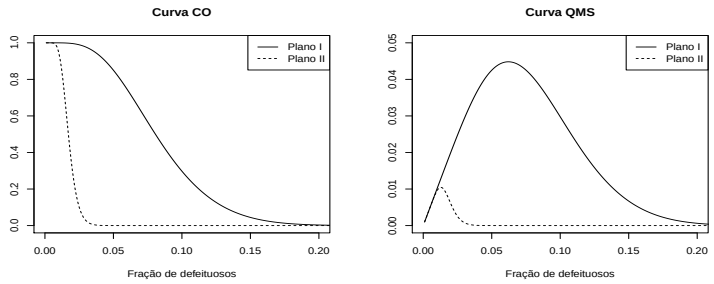


Figura 2: Curvas CO e QMS associadas ao plano I ($n_2 = 16$ e $k = 1.213$) e ao plano II ($n_2 = 109$ e $k = 0.642$).

A partir desta figura concluimos que o plano I aceita com probabilidade elevada lotes com uma percentagem de defeituosos significativa, e que o LQMS quando se usa este plano não é significativamente inferior ao valor de θ, se este fôr pequeno. Isto pode ser explicado pelo facto de termos fixado os riscos α e β em valores elevados, e o valor de n_2 ser muito pequeno. Se reduzirmos o valor dos riscos, aumentando consequentemente o valor de n_2, tal como no plano II, melhoramos significativamente o seu desempenho. O valor n_2 deste plano, e consequentemente o valor global da amostra a considerar, apesar de serem elevados, são aceitáveis para lotes de elevada dimensão, e em particular para os lotes e tipo de itens em análise neste estudo de consultoria.
Na Figura 3, os planos considerados foram desenhados fixando o valor n_2 e determinando a constante de aceitação k através da equação (10), de modo a obter um risco $\alpha = 5\%$ quando NQA $= 2.5\%$ (Plano I) e NQA $= 1\%$ (Plano II). Como era de esperar observamos melhorias significativas no desempenho dos planos quando n_2 aumenta.

4 Conclusão

Neste trabalho mostramos a importância da utilização de modelos inflacionados em aplicações, e apresentamos alguma motivação para a utilização da distribuição de Pareto inflacionada, uma vez que apresenta propriedades distribucionais simples, inclusive estimativas de máxima verosimilhança fáceis de calcular. Apresentamos um plano de amostragem de aceitação por variáveis para lotes de itens provenientes deste modelo, baseado numa regra de decisão bastante fácil de implementar, e fornecemos expressões analíticas para a determinação dos parâmetros do plano que permitem obter um determinado desempenho em função dos riscos do produtor e do consumidor. Ilustramos ainda o desempenho do plano através da representação das curvas CO e QMS para diferentes dimensões amostrais e/ou riscos do produtor e do consumidor.

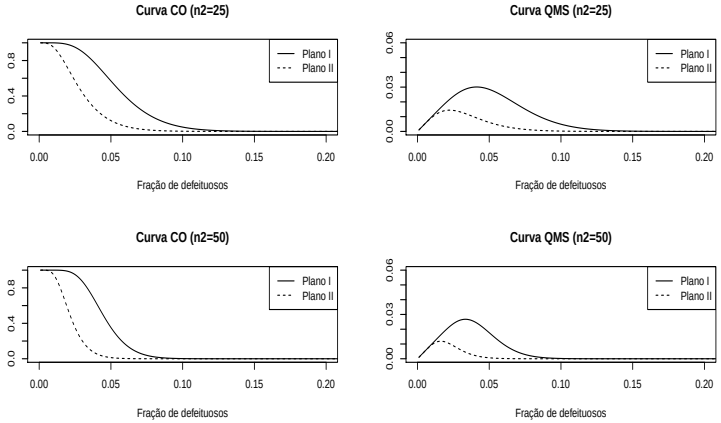


Figura 3: Curvas CO e QMS obtidas para um risco $\alpha = 5\%$ quando NQA = 2.5% (Plano I) e NQA = 1% (Plano II), e amostras de dimensão $n_2 = 25, 50$.

Agradecimentos

Este trabalho foi parcialmente financiado por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, Portugal, através dos projetos UID/MAT/00006/2013 e FCOMP-01-0124-FEDER-037281.

Referências

[1] Carolino, E., Barão, I. (2013). Robust methods in acceptance sampling. *Revstat* 11, 67–82.

[2] (2006). Davison, A., Hinkley, D.V. (2006). *Bootstrap Methods and their Application*. Cambridge University Press.

[3] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1–26.

[4] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

[5] Figueiredo, F., Figueiredo, A., Gomes, M.I. (2014). Comparison of sampling plans by variables using bootstrap and Monte Carlo simulations. *AIP Conference Proceedings* 1618, 535–538.

[6] Gomes, M.I. (2011). *Acceptance sampling.* In Lovric, M. (ed.), International Encyclopedia of Statistical Science, Part 1, 5-7, ISBN: 978-3-642-04897-5, Springer.

[7] Montgomery, D.C. (2009). *Introduction to Statistical Quality Control: a Modern Introduction*, 6th edition. John Wiley & Sons.

Matrizes de covariâncias para modelos lineares mistos aplicados ao estudo da variabilidade genética intravarietal de castas antigas de videira

Elsa Gonçalves
Secção de Matemática/DCEB e LEAF, Instituto Superior de Agronomia, Universidade de Lisboa; Associação Portuguesa para a Diversidade da Videira - PORVID, *elsagoncalves@isa.ulisboa.pt*
Antero Martins
LEAF, Instituto Superior de Agronomia, Universidade de Lisboa; PORVID, *anteromart@isa.ulisboa.pt*

Palavras–chave: modelos mistos, matrizes de covariâncias, variabilidade genética intravarietal, selecção da videira

Resumo: Neste trabalho propõem-se matrizes de covariâncias para modelos mistos usados no estudo da variabilidade genética intravarietal e analisam-se as consequências da sua utilização para fins de selecção. Faz-se a aplicação a uma variedade antiga de videira (Aragonez), tendo-se encontrado variabilidade genética intravarietal do rendimento e do grau brix mais elevadas em uma das suas principais regiões de cultura, indicando que provavelmente essa será a região de origem da casta.

1 Introdução

O ajustamento de modelos lineares mistos no contexto do melhoramento de plantas é uma prática corrente para a estimação de parametros genéticos importantes relativos às populacões em estudo para efeito de selecção. Uma determinada característica de um elemento da populacão experimental resulta da acção de um certo conjunto

de genes - o genótipo - mas modificada por desvios ambientais, originando o valor observado da característica, ou fenótipo. Isto é, para se conhecer o valor geneticamente transmissível (o único que interessa à selecção) são ajustados modelos mistos para decompor o valor fenotípico nas componentes genotípica e não genotípica. Cada vez mais, os recursos computacionais disponíveis têm permitido o ajustamento de modelos mistos de maior complexidade, nomeadamente no que se refere à estrutura das matrizes de covariâncias associadas aos efeitos aleatórios e aos erros aleatórios, sendo estas cada vez mais diversificadas e alternativas à utilização das clássicas matrizes de covariâncias diagonais [13]. Concretamente, essa diversificação compreende, por exemplo, o controlo da variabilidade espacial em ensaios de campo com elevado número de tratamentos, quando tal não foi eficientemente controlado pelos efeitos associados ao delineamento experimental. Nesses casos, a prática corrente passa pelo ajustamento de modelos com estrutura de covariância do erro que permita a modelação dessa variabilidade espacial, sendo o mais comum a modelação do erro como um processo auto-regressivo separável de primeira ordem (AR1×AR1) [13, 2, 3, 11]. Outra diversificação muito comum nas matrizes de covariâncias, surge no estudo de correlacões genéticas entre ambientes (interacção genótipo ×ambiente) e correlacões genéticas entre características, optando-se nestes casos geralmente por matrizes de covariâncias não estruturadas [4] ou baseadas na técnica multivariada de análise factorial [10, 13, 14]. Estes tipos de abordagem têm conduzido a um maior rigor nas práticas de selecção e melhoramento, actividades altamente responsáveis pelo aumento da produtividade e qualidade agrícolas. Porém, um tema raramente abordado teoricamente prende-se com o estudo da variabilidade genética intravarietal de variedades tradicionais no decurso de fases iniciais de selecção. No caso concreto de variedades antigas de videira, a amostra em estudo resulta da prospecção realizada nas suas principais regiões de cultura (subpopulações), pelo que a variabilidade genética existente em cada uma delas tenderá a ser diferente. Sob esta perspectiva, admitir uma variância genética comum a todas as regiões de cultura é, por vezes, irrealista. Na videira este tipo de

abordagem começou por ser tratado com o ajustamento de modelos individuais por característica e região de cultura da casta [7, 8]. Contudo, com os recursos computacionais actualmente disponíveis, o ajustamento de modelos lineares mistos com uma estrutura hierárquica dos efeitos aleatórios [6] e com matrizes de covariâncias não diagonais, torna-se um ponto crucial para o estudo da variabilidade genética intravarietal existente dentro da variedade antiga relativamente a várias características de interesse económico.
Neste trabalho faz-se uma abordagem raramente implementada no âmbito do melhoramento de plantas. Propõem-se matrizes de covariâncias para modelos mistos usados no estudo da variabilidade genética intravarietal e analisam-se as consequências da sua utilização para fins de selecção. É feita uma aplicação a uma variedade antiga de videira, uma das espécies agrícolas com maior importância económica e social em Portugal.

2 A metodologia proposta

Matricialmente, o modelo linear misto pode ser genericamente descrito como ([6])

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

em que $\mathbf{Y}$ é o vector $n \times 1$ das observações (valores fenotípicos), $\mathbf{X}$ 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 $\mathbf{e}$ é o vector $n \times 1$ de erros aleatórios. Os vectores $\mathbf{u}$ e $\mathbf{e}$ 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, $cov\,[\mathbf{u},\mathbf{e}] = \mathbf{0}$, $\mathbf{u} \cap \mathcal{N}_q\,(\mathbf{0},\mathbf{G})$, $\mathbf{e} \cap \mathcal{N}_n\,(\mathbf{0},\mathbf{R})$. A distribuição de $\mathbf{Y}$ admite-se assim normal multivariada, com vector de valores médios $\mathbf{X}\boldsymbol{\beta}$ e matriz de covariâncias $\mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$, $\mathbf{Y} \cap \mathcal{N}_n\,(\mathbf{X}\boldsymbol{\beta},\mathbf{V})$. Com o ajustamento deste tipo de modelo no contexto do melhoramento de plantas, os grandes objectivos são estimar as componentes de covariância (com base nas quais se avalia, por exemplo, a variabilidade

genética intravarietal) e fazer a selecção de um grupo de genótipos superiores com base no melhor preditor empírico linear não enviesado de $\mathbf{u}$ [5, 6],

$$\tilde{\mathbf{u}}_{EBLUP} = \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}(\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{EBLUE}), \tag{2}$$

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 [9], é actualmente o mais recomendado e utilizado para estimar componentes de covariância em grandes conjuntos de dados com estrutura complexa [6] e a inferência relativa às componentes de covariância é, em geral, baseada em testes de razão de verosimilhanças restritas. A comparação e selecção de modelos pode ser, quando possível, também feita com base, por exemplo, no critério de informação de Akaike [12]. No contexto biológico em que esta metodologia é proposta, as diversas variantes de modelos lineares mistos estão centradas na composição do vector $\mathbf{u}$ e na estrutura das matrizes $\mathbf{G}$ e $\mathbf{R}$. Vejamos alguns casos de aplicação.

Caso 1. O modelo linear misto mais simples é aplicável quando apenas uma característica (tipicamente o rendimento) é avaliada, os efeitos associados ao delineamento experimental admitem-se fixos, os efeitos genotípicos dos clones da casta admitem-se aleatórios ($\mathbf{u}$ é o vector $q \times 1$ dos efeitos genotípicos) e admite-se que as matrizes de covariâncias dos vectores $\mathbf{u}$ e $\mathbf{e}$ são, respectivamente, $\mathbf{G} = \sigma_g^2\mathbf{I}_q$, $\mathbf{R} = \sigma_e^2\mathbf{I}_n$, sendo $\mathbf{I}_q$ a matriz identidade $q \times q$ e $\mathbf{I}_n$ a matriz identidade $n \times n$. O ajustamento deste modelo permite obter as estimativas das variâncias genotípica e do erro para a característica avaliada $(\hat{\sigma}_g^2, \hat{\sigma}_e^2)$ e os melhores preditores empíricos lineares não enviesados (EBLUPs) dos efeitos genotípicos da característica avaliada. A sua aplicação faz sentido quando se estuda uma variedade antiga que é cultivada em apenas uma região, ou em várias, mas que tenham variabilidade genética similar, situação resultante da proximidade geográfica e de trocas frequentes de material de propagação entre as regiões.

Caso 2. O modelo linear misto mais realista numa fase inicial de selecção de uma variedade antiga de videira, quando se avalia apenas uma característica (tipicamente o rendimento), considera que $\mathbf{Y}$ é o vector $n \times 1$ dos valores fenotípicos de uma dada característica organizados por região de cultura, $\mathbf{Z}$ é a matriz de delineamento $n \times q$ dos efeitos genotípicos por região de cultura, $\mathbf{u}$ é o vector $q \times 1$ dos efeitos genotípicos por região de cultura ($q = \sum_{j=1}^{r} q_j$, sendo q_j o número de genótipos provenientes da região de cultura j e r o número de regiões de cultura/subpopulações estudadas) e $\mathbf{e}$ é o vector $n \times 1$ de erros aleatórios por região de cultura da casta ($n = \sum_{j=1}^{r} n_j$, sendo n_j o número de observações correspondentes aos genótipos provenientes da região de cultura j), ou seja, tem-se:

$$\mathbf{Zu} = \left[\begin{array}{cccc} \mathbf{Z}_{1(n\times q_1)} & \mathbf{Z}_{2(n\times q_2)} & \cdots & \mathbf{Z}_{r(n\times q_r)} \end{array} \right] \left[\begin{array}{c} \mathbf{u}_{1(q_1\times 1)} \\ \mathbf{u}_{2(q_2\times 1)} \\ \vdots \\ \mathbf{u}_{r(q_r\times 1)} \end{array} \right] = \sum_{j=1}^{r} \mathbf{Z}_j \mathbf{u}_j,$$

$$\mathbf{e} = \left[\begin{array}{cccc} \mathbf{e}_{1(n_1\times 1)}{}^T & \mathbf{e}_{2(n_2\times 1)}{}^T & \cdots & \mathbf{e}_{r(n_r\times 1)}{}^T \end{array} \right]^T.$$

Admite-se variâncias genéticas distintas por região de cultura, isto é, $\mathbf{G}_j = \sigma^2_{g_j} \mathbf{I}_{q_j}$, para $j = 1,...,r$, e $cov\,[\mathbf{u}_j, \mathbf{u}_{j'}] = \mathbf{0}$, para $\forall j \neq j'$, consequentemente, $\mathbf{G} = \oplus_{j=1}^{r} \mathbf{G}_j$, em que $\oplus$ representa a soma directa de matrizes. Quanto à estrutura da matriz de covariâncias do vector $\mathbf{e}$, quando necessário, admite-se igualmente variâncias dos erros aleatórios distintas por regiões de cultura da casta, isto é, $\mathbf{R}_j = \sigma^2_{e_j} \mathbf{I}_{n_j}$, para $j = 1,...,r$, e $cov\,[\mathbf{e}_j, \mathbf{e}_{j'}] = \mathbf{0}$, para $\forall j \neq j'$, consequentemente, $\mathbf{R} = \oplus_{j=1}^{r} \mathbf{R}_j$. Esta análise permite obter as componentes de variância genotípica e do erro para a característica avaliada em cada uma das regiões de cultura da casta ($\hat{\sigma}^2_{g_1}, \hat{\sigma}^2_{g_2}..., \hat{\sigma}^2_{g_r}, \hat{\sigma}^2_{e_1}, \hat{\sigma}^2_{e_2}, ..., \hat{\sigma}^2_{e_r}$) e, assim, quantificar a variabilidade genética intravarietal da casta por região. Usualmente admite-se que a subpopulação proveniente da região de cultura que apresenta maior variabilidade genética é a provável região de origem da casta. Também se obtêm os EBLUPs dos efeitos genotípicos para a característica avaliada. De notar que, sendo o EBLUP do efeito genotípico dependente da estrutura das

matrizes **G** e **R** (como expresso em 2), o ajustamento de modelos distintos conduz necessariamente a decisões de selecção diferentes. **Caso 3**. Modelo linear misto aplicável numa fase inicial de selecção de uma variedade antiga de videira quando se avaliam várias características (por exemplo, rendimento e características de qualidade do mosto). Neste caso, os vectores **Y**, **u** e **e** são dados por:

$$\mathbf{Y}=\begin{bmatrix} \mathbf{Y}_{1.1(n_1\times 1)} \\ \mathbf{Y}_{1.2(n_2\times 1)} \\ \vdots \\ \mathbf{Y}_{1.r(n_r\times 1)} \\ \mathbf{Y}_{2.1(n_1\times 1)} \\ \mathbf{Y}_{2.2(n_2\times 1)} \\ \vdots \\ \mathbf{Y}_{2.r(n_r\times 1)} \\ \vdots \\ \mathbf{Y}_{t.1(n_1\times 1)} \\ \mathbf{Y}_{t.2(n_2\times 1)} \\ \vdots \\ \mathbf{Y}_{t.r(n_r\times 1)} \end{bmatrix} \mathbf{u}=\begin{bmatrix} \mathbf{u}_{1.1(q_1\times 1)} \\ \mathbf{u}_{1.2(q_2\times 1)} \\ \vdots \\ \mathbf{u}_{1.r(q_r\times 1)} \\ \mathbf{u}_{2.1(q_1\times 1)} \\ \mathbf{u}_{2.2(q_2\times 1)} \\ \vdots \\ \mathbf{u}_{2.r(q_r\times 1)} \\ \vdots \\ \mathbf{u}_{t.1(q_1\times 1)} \\ \mathbf{u}_{t.2(q_2\times 1)} \\ \vdots \\ \mathbf{u}_{t.r(q_r\times 1)} \end{bmatrix} \mathbf{e}=\begin{bmatrix} \mathbf{e}_{1.1(n_1\times 1)} \\ \mathbf{e}_{1.2(n_2\times 1)} \\ \vdots \\ \mathbf{e}_{1.r(n_r\times 1)} \\ \mathbf{e}_{2.1(n_1\times 1)} \\ \mathbf{e}_{2.2(n_2\times 1)} \\ \vdots \\ \mathbf{e}_{2.r(n_r\times 1)} \\ \vdots \\ \mathbf{e}_{t.1(n_1\times 1)} \\ \mathbf{e}_{t.2(n_2\times 1)} \\ \vdots \\ \mathbf{e}_{t.r(n_r\times 1)} \end{bmatrix}.$$

Y é agora o vector $n \times 1$ dos valores fenotípicos para as várias características em análise, em cada uma das regiões de cultura (com $n = t\sum_{j=1}^{r} n_j$, sendo t o número de características avaliadas, n_j o número de observações correspondentes aos genótipos provenientes da região j, r o número de regiões), $\boldsymbol{\beta}$ é o vector $p\times 1$ de efeitos fixos (μ_i, para $i=1,...,t$; efeitos associados ao delineamento experimental por característica), **Z** é a matriz de delineamento $n \times q$ dos efeitos aleatórios (efeitos genotípicos por característica e região de cultura da casta), **u** é o vector $q\times 1$ dos efeitos genotípicos por característica e por região ($q = t\sum_{j=1}^{r} q_j$, sendo q_j o número de genótipos da região j e r o número de regiões) e **e** é o vector $n\times 1$ de erros aleatórios por característica e por região de cultura. A matriz de covariâncias do

vector $\mathbf{u}$, isto é, a matriz $\mathbf{G}$ é dada por:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{1.1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{1.1|2.1} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{G}_{1.1|t.1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{1.2} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{G}_{1.2|2.2} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{1.2|t.2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{1.r} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{1.r|2.r} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{1.r|t.r} \\ \mathbf{G}_{2.1|1.1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{2.1} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{G}_{2.1|t.1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{2.2|1.2} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{G}_{2.2} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{2.2|t.2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{2.r|1.r} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{2.r} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{2.r|t.r} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{t.1|1.1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{t.1|2.1} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{G}_{t.1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{t.2|1.2} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{G}_{t.2|2.2} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G}_{t.2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{t.r|1.r} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{t.r|2.r} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_{t.r}, \end{bmatrix} \quad (3)$$

em que $\mathbf{G}_{i.j} = \sigma^2_{g_{i.j}}\mathbf{I}_{q_j}$, para $i = 1,...,t$, $j = 1,...,r$, e $\mathbf{G}_{i.j|i'.j} = \sigma_{g_{i.j|i'.j}}\mathbf{I}_{q_j}$, $\forall i \neq i'$ e $j = 1,...,r$, sendo $\sigma^2_{g_{i.j}}$ a variância genética para a característica i na região j, $\mathbf{I}_{q_j}$ a matriz identidade $q_j \times q_j$, $\sigma_{g_{i.j|i'.j}}$ a covariância genética entre as características i e i' na região j, q_j o número de genótipos da região j. Esta análise permite quantificar a variabilidade genética por característica e por região, bem como obter os EBLUPs dos efeitos genotípicos para cada característica. Obtêm-se igualmente as estimativas das correlações genéticas entre características em cada região de cultura, sendo a correlação genética entre as características i e i' ($\forall i \neq i'$) na subpopulação proveniente da região de cultura j dada por $r_g = \frac{\sigma_{g_{i.j|i'.j}}}{\sigma_{g_{i.j}}\sigma_{g_{i'.j}}}$. Esta abordagem é importante, pois é útil compreender se ao seleccionar uma característica não se está a prejudicar outra com igual importância.

3 Uma aplicação

A aplicação é feita ao estudo da variabilidade intravarietal da casta Aragonez. Os dados são provenientes de um ensaio inicial da casta, contendo amostras de genótipos representativas da respectiva diversidade em distintas regiões de cultura de Portugal e de Espanha (60

genótipos do Alentejo (A), 66 do Douro (D), 67 de Rioja (R) e 62 de Valdepeñas (V)), num total de 255 genótipos. O ensaio foi instalado em Reguengos de Monsaraz, com um delineamento experimental em blocos completos casualizados (5 repetições, 3 plantas por parcela). Os dados utilizados referem-se às médias de anos de rendimento (kg/planta) e de grau brix obtidos por genótipo em cada parcela. Os modelos propostos na secção anterior foram ajustados no *Software R*, *package ASReml−R* [1] (método de estimação REML, algoritmo de informação média). Comecemos com o rendimento, de longe a característica mais correntemente avaliada. Foram ajustados vários modelos: o modelo descrito no caso 1 ($\mathbf{G} = \sigma^2_g \mathbf{I}_{255}$, $\mathbf{R} = \sigma^2_e \mathbf{I}_{1275}$), designado modelo M1; variantes do modelo descrito no caso 2, com matrizes $\mathbf{G} = \mathbf{G}_A \oplus \mathbf{G}_D \oplus \mathbf{G}_R \oplus \mathbf{G}_V$ ($\mathbf{G}_A = \sigma^2_{g_A} \mathbf{I}_{60}$, $\mathbf{G}_D = \sigma^2_{g_D} \mathbf{I}_{66}$, $\mathbf{G}_R = \sigma^2_{g_R} \mathbf{I}_{67}$, $\mathbf{G}_V = \sigma^2_{g_V} \mathbf{I}_{62}$) e $\mathbf{R} = \sigma^2_e \mathbf{I}_{1275}$, designado modelo M2A, e com a mesma matriz $\mathbf{G}$ mas com $\mathbf{R} = \mathbf{R}_A \oplus \mathbf{R}_D \oplus \mathbf{R}_R \oplus \mathbf{R}_V$ ($\mathbf{R}_A = \sigma^2_{e_A} \mathbf{I}_{300}$, $\mathbf{R}_D = \sigma^2_{e_D} \mathbf{I}_{330}$, $\mathbf{R}_R = \sigma^2_{e_R} \mathbf{I}_{335}$, $\mathbf{R}_V = \sigma^2_{e_V} \mathbf{I}_{310}$), designado modelo M2B. Os resultados obtidos com o ajustamento destes modelos (Tabela 1) sugerem heterogeneidade de variâncias entre regiões, particularmente associada a Valdepeñas. No entanto, as estimativas das componentes de variância obtidas para o Alentejo, Douro e Rioja admitem o ajustamento de uma outra variante do modelo descrito no caso 2, mais parcimoniosa, que considere que as regiões do Alentejo, Douro e Rioja (ADR) partilham a mesma variância genotípica do rendimento. Isto traduz-se no ajustamento de mais dois modelos: modelo com matrizes $\mathbf{G} = \mathbf{G}_{ADR} \oplus \mathbf{G}_V$ ($\mathbf{G}_{ADR} = \sigma^2_{g_{ADR}} \mathbf{I}_{193}$, $\mathbf{G}_V = \sigma^2_{g_V} \mathbf{I}_{62}$), e $\mathbf{R} = \sigma^2_e \mathbf{I}_{1275}$, designado modelo M2C; modelo com a mesma matriz $\mathbf{G}$, mas com $\mathbf{R} = \mathbf{R}_{ADR} \oplus \mathbf{R}_V$ (com $\mathbf{R}_{ADR} = \sigma^2_{e_{ADR}} \mathbf{I}_{965}$, $\mathbf{R}_V = \sigma^2_{e_V} \mathbf{I}_{310}$), designado modelo M2D. Todos os modelos atrás descritos partilham o mesmo termo $\mathbf{X}\boldsymbol{\beta}$, sendo $\boldsymbol{\beta}$ o vector de efeitos fixos (μ e os efeitos associados aos blocos completos, β_1, β_2, β_3, β_4 e β_5) e $\mathbf{X}$ a respectiva matriz de delineamento:

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{1}_{300} & \mathbf{I}_5 \otimes \mathbf{1}_{60} \\ \mathbf{1}_{330} & \mathbf{I}_5 \otimes \mathbf{1}_{66} \\ \mathbf{1}_{335} & \mathbf{I}_5 \otimes \mathbf{1}_{67} \\ \mathbf{1}_{310} & \mathbf{I}_5 \otimes \mathbf{1}_{62} \end{bmatrix},$$

onde **1** representa um vector de uns, **I** a matriz identidade e $\otimes$ o produto de Kronecker de matrizes.

Modelo	Var. genotípica	Var.erro	lr	par	AIC
Modelo M1	$\hat{\sigma}^2_g = 0{,}2577$	$\hat{\sigma}^2_e = 0{,}7701$	$-607{,}87$	2	1219,73
Modelo M2A	$\hat{\sigma}^2_{gA} = 0{,}2155$ $\hat{\sigma}^2_{gD} = 0{,}2241$ $\hat{\sigma}^2_{gR} = 0{,}1849$ $\hat{\sigma}^2_{gV} = 0{,}7671$	$\hat{\sigma}^2_e = 0{,}7701$	$-600{,}08$	5	1210,16
Modelo M2B	$\hat{\sigma}^2_{gA} = 0{,}1701$ $\hat{\sigma}^2_{gD} = 0{,}1594$ $\hat{\sigma}^2_{gR} = 0{,}1241$ $\hat{\sigma}^2_{gV} = 0{,}6205$	$\hat{\sigma}^2_{eA} = 0{,}7495$ $\hat{\sigma}^2_{eD} = 0{,}8357$ $\hat{\sigma}^2_{eR} = 0{,}8617$ $\hat{\sigma}^2_{eV} = 0{,}6217$	$-596{,}15$	8	1208,30
Modelo M2C	$\hat{\sigma}^2_{gADR} = 0{,}2072$ $\hat{\sigma}^2_{gV} = 0{,}7715$	$\hat{\sigma}^2_e = 0{,}7701$	$-600{,}16$	3	1206,33
Modelo M2D	$\hat{\sigma}^2_{gADR} = 0{,}1500$ $\hat{\sigma}^2_{gV} = 0{,}6238$	$\hat{\sigma}^2_{eADR} = 0{,}8180$ $\hat{\sigma}^2_{eV} = 0{,}6215$	$-596{,}89$	4	1201,78

Tabela 1: Estimativas das componentes de variância genotípica e do erro, log-verosimilhança restrita (lr), número de parâmetros de covariância (par) e critério de informação de Akaike (AIC) resultantes do ajustamento dos modelos M1, M2A, M2B, M2C e M2D.

Com base no AIC (Tabela 1) conclui-se que o modelo M2B, que admite variâncias genotípicas e do erro heterogéneas entre regiões, é preferível ao modelo M1, que considera variâncias genotípicas e do erro homogéneas entre regiões. Mas o modelo M2D, que admite variâncias homogéneas para Alentejo, Douro e Rioja e variâncias distintas para a região de Valdepeñas, é preferível ao modelo M2B. Em suma, de acordo com este critério, M2D revelou-se como sendo o melhor modelo. Conclui-se que a variabilidade genética do rendimento é idêntica no Alentejo, Douro e Rioja, sendo a região de Valdepeñas a que apresenta maior variabilidade genética intravarietal. Isto sugere que em futuros trabalhos de selecção valerá a pena focar a prospecção nesta região, assim como suporta a hipótese de esta ser a região de origem da casta, tendo-se posteriormente expandido para as outras regiões. Ao mesmo tipo de conclusão se chega através do teste de razão de verosimilhanças restritas. Comparando M2B e M2D, conclui-se que não diferem significativamente,

para qualquer nível de significância usual (os 2 modelos diferem em 4 parâmetros, o valor da estatística de razão de verosimilhanças restritas, $REMLRT$, é de 1,48, $valor - p = 0{,}8302$), optando-se, portanto, pelo modelo mais parcimonioso, M2D. Comparando M2D e M1, conclui-se que diferem significativamente para qualquer nível de significância usual (os 2 modelos diferem em 2 parâmetros, $REMLRT = 21{,}946$, $valor - p < 0{,}0001$). Isto é, rejeita-se a hipótese nula de igualdade das variâncias genotípicas e de igualdade das variâncias do erro entre as duas subpopulações (subpopulação conjunta de Alentejo, Douro e Rioja e subpopulação de Valdepeñas). Por fim, mostram-se na Tabela 2 alguns dos EBLUPs dos efeitos genotípicos do rendimento obtidos com o ajustamento dos modelos M1 e M2D que, como esperado, são diferentes de acordo com o modelo ajustado. Por exemplo, como o modelo M1 subestima a variância genotípica de Valdepeñas, os EBLUPs dos efeitos genotípicos para esta região são menores do que os obtidos com M2D. Por outro lado, como o modelo M1 sobreestima a variância genotípica associada ao Alentejo, Douro e Rioja, os EBLUPs dos efeitos genotípicos para estas regiões são maiores do que os obtidos com o modelo M2D. Neste caso concreto, as decisões de selecção devem ser tomadas de acordo com os EBLUPs obtidos com o ajustamento do modelo M2D.

Ord	EBLUPs, M1		EBLUPs, M2D	
1	Douro:cloneRZ1178	0,9200	Valdepenas:RZ4201	0,8821
2	Douro:cloneRZ0707	0,8755	Douro:RZ1178	0,6611
3	Rioja:cloneRZ7810	0,8675	Douro:RZ0707	0,6271
4	Rioja:cloneRZ8601	0,8483	Rioja:cloneRZ7810	0,6209
5	Douro:cloneRZ0136	0,7840	Rioja:cloneRZ8601	0,6063
6	Douro:cloneRZ6112	0,7822	Douro:cloneRZ0136	0,5571
7	Douro:cloneRZ6505	0,7435	Douro:cloneRZ6112	0,5558
8	Alentejo:cloneRZ1124	0,7296	Douro:cloneRZ6505	0,5262
9	Alentejo:cloneRZ1704	0,7229	Alentejo:cloneRZ1124	0,5156
10	Valdepenas:cloneRZ4201	0,7170	Alentejo:cloneRZ1704	0,5104

Tabela 2: EBLUPs dos efeitos genotípicos do rendimento (kg/planta) obtidos para os 10 melhores genótipos seleccionados com base nos modelos M1 e M2D.

Finalmente, vejamos uma aplicação com vista ao estudo da correlação genética entre rendimento e grau brix nas diversas regiões de cultura da casta. Para tal, foram ajustados vários modelos, variantes do modelo descrito no caso 3, partilhando todos eles o mesmo

termo $\mathbf{X}\boldsymbol{\beta}$, sendo $\boldsymbol{\beta}$ o vector de efeitos fixos (médias populacionais das características e os efeitos associados aos blocos completos por característica) e $\mathbf{X}$ a respectiva matriz de delineamento:

$$\boldsymbol{\beta}=\begin{bmatrix} \mu_{rend} \\ \beta 1_{rend} \\ \beta 2_{rend} \\ \beta 3_{rend} \\ \beta 4_{rend} \\ \beta 5_{rend} \\ \mu_{brix} \\ \beta 1_{brix} \\ \beta 2_{brix} \\ \beta 3_{brix} \\ \beta 4_{brix} \\ \beta 5_{brix} \end{bmatrix} \quad \mathbf{X}=\begin{bmatrix} \mathbf{1}_{300} & \mathbf{I}_5 \otimes \mathbf{1}_{60} & \mathbf{0}_{300} & \mathbf{0}_{300\times 5} \\ \mathbf{1}_{330} & \mathbf{I}_5 \otimes \mathbf{1}_{66} & \mathbf{0}_{330} & \mathbf{0}_{330\times 5} \\ \mathbf{1}_{335} & \mathbf{I}_5 \otimes \mathbf{1}_{67} & \mathbf{0}_{335} & \mathbf{0}_{335\times 5} \\ \mathbf{1}_{310} & \mathbf{I}_5 \otimes \mathbf{1}_{62} & \mathbf{0}_{310} & \mathbf{0}_{310\times 5} \\ \mathbf{0}_{300} & \mathbf{0}_{300\times 5} & \mathbf{1}_{300} & \mathbf{I}_5 \otimes \mathbf{1}_{60} \\ \mathbf{0}_{330} & \mathbf{0}_{330\times 5} & \mathbf{1}_{330} & \mathbf{I}_5 \otimes \mathbf{1}_{66} \\ \mathbf{0}_{335} & \mathbf{0}_{335\times 5} & \mathbf{1}_{335} & \mathbf{I}_5 \otimes \mathbf{1}_{67} \\ \mathbf{0}_{310} & \mathbf{0}_{310\times 5} & \mathbf{1}_{310} & \mathbf{I}_5 \otimes \mathbf{1}_{62} \end{bmatrix},$$

onde $\mathbf{1}$ representa um vector de uns, $\mathbf{I}$ a matriz identidade, $\mathbf{0}$ o vector nulo ou a matriz nula e $\otimes$ o produto de Kronecker de matrizes. Os modelos bivariados ajustados admitiram: homogeneidade de variâncias genéticas e do erro entre subpopulações e correlação entre rendimento e grau brix (modelo M3); heterogeneidade de variâncias genéticas entre subpopulações e correlações genéticas distintas entre rendimento e grau brix nas diferentes subpopulações (modelo M3A); heterogeneidade de variâncias genéticas e do erro entre subpopulações e correlações genéticas e do erro distintas entre rendimento e grau brix nas diferentes subpopulações (modelo M3B); heterogeneidade de variâncias genéticas entre duas subpopulações (ADR - Alentejo, Douro e Rioja e V - Valdepeñas) e correlações genéticas distintas entre rendimento e grau brix nas duas subpopulações (modelo M3C); heterogeneidade de variâncias genéticas e do erro entre duas subpopulações (ADR e V) e correlações genéticas e do erro distintas entre rendimento e grau brix nas duas subpopulações (modelo M3D). De acordo com os resultados obtidos (Tabela 3), verifica-se que, tal como observado para o rendimento, também para o grau brix a estimativa da componente de variância genotípica é superior na região de Valdepeñas. De entre os modelos ajustados, M3B e M3D revelaram um melhor ajustamento. Contudo, comparando formalmente estes dois modelos conclui-se que não diferem significativamente, para qualquer nível de significância usual (os 2 modelos diferem em 12 parâmetros, $REMLRT = 11{,}873$, $valor-p = 0{,}45593$),

optando-se, assim, pelo modelo mais parcimonioso que considera que as regiões do Alentejo, Douro e Rioja (ADR) partilham a mesma variância genotípica, quer para o rendimento quer para o grau brix (M3D). Verifica-se ainda que a subpopulação ADR não revela correlação genética entre as duas características analisadas, enquanto que essa correlação é moderadamente negativa na subpopulação de Valdepeñas. De facto, biológicamente é expectável que haja uma correlação negativa entre estas duas características e esta torna-se mais expressiva na região que apresenta maior variabilidade. É de salientar ainda que o resultado obtido relativo à correlação entre características obtido com M3D difere do resultado obtido com M3, o que ilustra bem a necessidade do ajustamento de um modelo com uma estrutura hierárquica dos efeitos aleatórios, pois este resultado terá consequências ao nível dos EBLUPS dos efeitos genotípicos do rendimento e do grau brix e, consequentemente, na selecção.

4 Considerações finais

Quando o ensaio de campo referente a uma variedade antiga contém genótipos de diferentes regiões de cultura da casta, o ajustamento de modelos lineares mistos deve assentar em estruturas de covariância que traduzam essa realidade. Este procedimento não só permite quantificar a variabilidade genética intravarietal de uma dada característica, como também conduz a um maior rigor nas decisões de selecção. Com a aplicação da metodologia proposta aos dados da casta Aragonez, concluiu-se que a variabilidade genética intravarietal do rendimento e do grau brix é idêntica nas regiões do Alentejo, Douro e Rioja, sendo a região espanhola de Valdepeñas a que apresenta maior variabilidade genética intravarietal para ambas as características, assim como a única região que revelou uma correlação genética moderada entre essas características (neste caso, moderadamente negativa).

Modelo M3	$\hat{\sigma}^2_{g_r} = 0{,}2577$ $\hat{\sigma}^2_{g_b} = 0{,}7090$ $\hat{\mathrm{r}}_{g_{r,b}} = -0{,}3135$	$\hat{\sigma}^2_{e_r} = 0{,}7701$ $\hat{\sigma}^2_{e_b} = 1{,}1106$ $\hat{\mathrm{r}}_{e_{r,b}} = -0{,}0285$	$lr = -1167{,}16$	$par = 6$	$AIC = 2346{,}31$
Modelo M3A	$\hat{\sigma}^2_{g_{A_r}} = 0{,}1646$ $\hat{\sigma}^2_{g_{D_r}} = 0{,}1721$ $\hat{\sigma}^2_{g_{R_r}} = 0{,}1425$ $\hat{\sigma}^2_{g_{V_r}} = 0{,}5943$ $\hat{\sigma}^2_{g_{A_b}} = 0{,}3338$ $\hat{\sigma}^2_{g_{D_b}} = 0{,}7826$ $\hat{\sigma}^2_{g_{R_b}} = 0{,}1949$ $\hat{\sigma}^2_{g_{V_b}} = 1{,}6092$ $\hat{\mathrm{r}}_{g_{A_{r,b}}} = 0{,}1362$ $\hat{\mathrm{r}}_{g_{D_{r,b}}} = 0{,}0933$ $\hat{\mathrm{r}}_{g_{R_{r,b}}} = -0{,}2544$ $\hat{\mathrm{r}}_{g_{V_{r,b}}} = -0{,}6315$	$\hat{\sigma}^2_{e_r} = 0{,}7701$ $\hat{\sigma}^2_{e_b} = 1{,}1105$ $\hat{\mathrm{r}}_{e_{r,b}} = -0{,}0269$	$lr = -1141{,}62$	$par = 15$	$AIC = 2313{,}24$
Modelo M3B	$\hat{\sigma}^2_{g_{A_r}} = 0{,}1678$ $\hat{\sigma}^2_{g_{D_r}} = 0{,}1584$ $\hat{\sigma}^2_{g_{R_r}} = 0{,}1245$ $\hat{\sigma}^2_{g_{V_r}} = 0{,}6265$ $\hat{\sigma}^2_{g_{A_b}} = 0{,}3312$ $\hat{\sigma}^2_{g_{D_b}} = 0{,}6754$ $\hat{\sigma}^2_{g_{R_b}} = 0{,}2463$ $\hat{\sigma}^2_{g_{V_b}} = 1{,}6383$ $\hat{\mathrm{r}}_{g_{A_{r,b}}} = -0{,}0154$ $\hat{\mathrm{r}}_{g_{D_{r,b}}} = 0{,}2721$ $\hat{\mathrm{r}}_{g_{R_{r,b}}} = -0{,}2805$ $\hat{\mathrm{r}}_{g_{V_{r,b}}} = -0{,}6107$	$\hat{\sigma}^2_{e_{A_r}} = 0{,}7500$ $\hat{\sigma}^2_{e_{D_r}} = 0{,}8371$ $\hat{\sigma}^2_{e_{R_r}} = 0{,}8606$ $\hat{\sigma}^2_{e_{V_r}} = 0{,}6210$ $\hat{\sigma}^2_{e_{A_b}} = 1{,}1124$ $\hat{\sigma}^2_{e_{D_b}} = 1{,}3710$ $\hat{\sigma}^2_{e_{R_b}} = 0{,}9556$ $\hat{\sigma}^2_{e_{V_b}} = 1{,}0135$ $\hat{\mathrm{r}}_{e_{A_{r,b}}} = 0{,}1243$ $\hat{\mathrm{r}}_{e_{D_{r,b}}} = -0{,}2042$ $\hat{\mathrm{r}}_{e_{R_{r,b}}} = 0{,}0105$ $\hat{\mathrm{r}}_{e_{V_{r,b}}} = -0{,}0223$	$lr = -1131{,}34$	$par = 24$	$AIC = 2310{,}67$
Modelo M3C	$\hat{\sigma}^2_{g_{ADR_r}} = 0{,}1585$ $\hat{\sigma}^2_{g_{V_r}} = 0{,}6022$ $\hat{\sigma}^2_{g_{ADR_b}} = 0{,}4382$ $\hat{\sigma}^2_{g_{V_b}} = 1{,}6300$ $\hat{\mathrm{r}}_{g_{ADR_{r,b}}} = 0{,}0352$ $\hat{\mathrm{r}}_{g_{V_{r,b}}} = -0{,}6368$	$\hat{\sigma}^2_{e_r} = 0{,}7701$ $\hat{\sigma}^2_{e_b} = 1{,}1108$ $\hat{\mathrm{r}}_{e_{r,b}} = -0{,}0305$	$lr = -1146{,}81$	$par = 9$	$AIC = 2311{,}61$
Modelo M3D	$\hat{\sigma}^2_{g_{ADR_r}} = 0{,}1489$ $\hat{\sigma}^2_{g_{V_r}} = 0{,}6319$ $\hat{\sigma}^2_{g_{ADR_b}} = 0{,}4278$ $\hat{\sigma}^2_{g_{V_b}} = 1{,}6627$ $\hat{\mathrm{r}}_{g_{ADR_{r,b}}} = 0{,}0403$ $\hat{\mathrm{r}}_{g_{V_{r,b}}} = -0{,}6184$	$\hat{\sigma}^2_{e_{ADR_r}} = 0{,}8180$ $\hat{\sigma}^2_{e_{V_r}} = 0{,}6215$ $\hat{\sigma}^2_{e_{ADR_b}} = 1{,}1420$ $\hat{\sigma}^2_{e_{V_b}} = 1{,}0142$ $\hat{\mathrm{r}}_{e_{ADR_{r,b}}} = -0{,}0334$ $\hat{\mathrm{r}}_{e_{V_{r,b}}} = -0{,}0211$	$lr = -1143{,}21$	$par = 12$	$AIC = 2310{,}42$

Tabela 3: Estimativas das componentes de variância genotípica e do erro, das correlações genotípicas e do erro entre rendimento (r) e grau brix (b), log-verosimilhança restrita (lr), número de parâmetros de covariância (par) e critério de informação de Akaike (AIC) resultantes do ajustamento dos modelos M3, M3A, M3B, M3C e M3D.

Referências

[1] Butler, D., Cullis, B.R., Gilmour, A.R, Gogel, B.J. (2007). ASReml-R reference manual. NSW Department of Primary Industries, Queensland.

[2] Gonçalves, E., St.Aubyn, A., Martins, A. (2007). Mixed spatial models for data analysis of yield on large grapevine selection field trials. *Theoretical and Applied Genetics*, 115, 653–663.

[3] Gonçalves, E., Carrasquinho, I., St.Aubyn, A., Martins, A.(2013). Broad-sense heritability in the context of mixed models for grapevine initial selection trials. *Euphytica*, 189, 379–391.

[4] Gonçalves, E., Carrasquinho, I., Almeida, R., Pedroso, V., Martins, A. (2016). Genetic correlations in grapevine and their effects on selection. *Australian Journal of Grape and Wine Research*, 22, 52–63.

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

[6] McCulloch, C.E., Searle, S.R., Neuhaus, J.M. (2008). Generalized, linear and mixed models. John Wiley & Sons, New York.

[7] Martins, A., Carneiro L.C., Gonçalves, E., Eiras-Dias, J.E. (2006). Méthodologie pour lanalyse et conservation de la variabilité génétique des cépages. Proc.29th World Congress Vine and Wine, 25-30Junho, Logroño, Espanha.

[8] Martins, A., Gonçalves, E. (2015). Grapevine breeding programmes in Portugal. In Grapevine Breeding Programs for the Wine Industry. A. G. Reynolds ed., Woodhead Publishing, Elsevier, UK, 159–182.

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

[10] 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.

[11] Piepho, H., Mohring, J., Pflufelder, M., Hermann, W., Williams E. (2015). Problems in parameter estimation for power and AR(1) models of spatial correlation in designed field experiments. *Communications in Biometry and Crop Science*, 10, 3–16.

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

[13] 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.

[14] Smith, A., Ganesalingam, A., Kuchel, H., Cullis, B. (2015). Factor analytic mixed models for the provision grower information from national crop variety testing programs. *Theoretical and Applied Genetics*, 118, 55–72.

Dados, ética e investigação em saúde tropical: constrangimentos e desafios

Luzia Gonçalves
Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa e Centro de Estatística e Aplicações da Universidade de Lisboa, *luziag@ihmt.unl.pt*

Palavras–chave: Dados, ética, protocolo de investigação, desafios

Resumo: Ultimamente, a intervenção da estatística antes da recolha de dados está mais presente na investigação em saúde, quer pela pressão do financiamento e dos conselhos de ética, quer pela pressão das publicações científicas. A recolha de dados de qualidade em saúde tropical é fundamental para actuar, por vezes, de forma rápida e em tempo real (e.g., Ebola ou Zika). Sendo os pressupostos teóricos iguais a qualquer outra área, na prática existem especificidades no terreno que por vezes tornam esta recolha mais difícil e desafiante. Este trabalho tem como objectivo apresentar e discutir alguns aspectos que emergem da acção crescente da bioestatística em contextos africanos, no âmbito de projectos de investigação, que tanto pode potencializar como restringir as ligações à estatística teórica estabelecida.

1 Introdução

Numa era de recolha e de armazenamento de grandes quantidades de dados, co-existem preocupações com a qualidade e a quantidade de dados obtidos em contextos rurais e urbanos de zonas tropicais que podem incluir populações de difícil acesso e/ou simplesmente terem dificuldades acrescidas por questões culturais ou éticas (entre outras). A elaboração do protocolo de investigação – documento que antecede a realização de uma investigação, sendo fundamental

para as candidaturas a financiamento, comissões ou conselhos de ética e mesmo para a obtenção de algumas pós-graduações ou graus académicos – tem contribuído para a afirmação da estatística mais atempadamente nos projectos de investigação. A literatura sobre a elaboração de protocolos de investigação é extensa [1, 2, 5, 3, 4]. Na investigação em saúde, em geral, é necessária a aprovação do protocolo de investigação por parte dos comités de ética, perante os quais o investigador assume compromissos sobre a recolha, a informatização, o armazenamento e, eventual, destruição dos dados após a investigação, podendo condicionar os tratamentos estatísticos mais morosos e complexos. Neste aspecto, a partilha dos dados com os estatísticos teóricos seria uma mais valia, porém, perante o(s) comité(s) de ética o investigador pode ter assumido o compromisso de os dados ficaram apenas afectos à equipa de investigação original. Por outro lado, na fase de publicação dos artigos científicos, existe uma tendência crescente para disponibilizar as bases de dados de forma a confirmar os resultados. Nesta experiência de terreno pode haver constrangimentos, mas existem desafios interessantes que emergem dos diversos problemas em saúde tropical.

2 A diversidade de dados

A multidimensionalidade dos problemas em saúde pode exigir a recolha de dados em diferentes camadas que vão do nível macro ao micro e podem envolver diferentes populações interrelacionadas que idealmente deveriam ser tratados de forma conjunta e não fragmentada. Hoje em dia, a epidemiologia clássica dá lugar a outras correntes, como a eco-epidemiologia [5] de forma a contemplar os aspectos multifactoriais associados à doença. Por exemplo, num estudo de uma doença tropical, além do agente da doença, dos vectores e dos hospedeiros, poderemos ter dados ligados ao ambiente, ao ordenamento do território, às infraestruturas, etc. No projecto UPHI-STAT: *O planeamento urbano e as desigualdades em saúde: passando das estatísticas macro para as micro*, que decorreu na cidade da Praia, na

ilha de Santiago, em Cabo Verde, essa diversidade de dados também aparece num contexto de doenças cardiovasculares [6]. Além dos dados de caracterização do planeamento urbano e das infraestruturas existentes naquela cidade, recolheram-se dados através da aplicação de um questionário, aplicado por entrevistadores locais, e efectuaram-se medições antropométricas por uma equipa de nutricionistas, obtendo dados sobre a caracterização sociodemográfica, opiniões e percepções sobre infraestruturas e a segurança na cidade, aspectos ligados à percepção sobre o estado de saúde, hábitos alimentares, actividade física no lazer, no trabalho e nas deslocações, etc.

3 Protocolo de Investigação, aspectos éticos e a estatística antes da recolha de dados

Hoje em dia, a par da investigação também no ensino, os cursos de saúde e ciências biomédicas têm algumas unidades curriculares sobre métodos de investigação, comunicação e escrita científica, ou mesmo para os alunos prosseguirem para a elaboração das suas dissertações ou teses de doutoramento podem ser obrigatórias a elaboração e a defesa de um protocolo de investigação. Assim, os investigadores de diversas áreas reconhecem cada vez mais a importância da definição da população em estudo, a dimensão da amostra, dos métodos de amostragem e a descrição do plano de análise estatística dos dados, sem ainda os ter recolhido, na secção de Material e Métodos. Pensar na estatística ainda sem dados, é um desfio importante que tem fomentado a intervenção dos estatísticos nas equipas de investigação desde o início do planeamento e do delineamento do projecto de investigação, o que nem sempre acontecia há uns anos atrás. A secção de Material e Métodos tem um elevado peso nas avaliações dos projectos de investigação por parte dos financiadores. Por exemplo, na apoio à investigação de doenças tropicais, no programa TDR, *the*

Special Programme for Research and Training in Tropical Diseases, podemos encontrar o guião de avaliação dos projectos submetidos (ver `http://www.who.int/tdr/grants/application_reporting_forms/application_assessment_form_sample.pdf?ua=1`[3]) sendo apresentados diversos critérios ligados à estatística, à epidemiologia e à ética. Certamente, no guião de avaliação de uma candidatura, e posteriormente nas publicações, constam as seguintes questões: (i) a população e a amostra estão correctamente definidos? (ii) o tamanho da amostra é apropriado? (iii) os métodos estão articulados com os objectivos da investigação? (iii) os métodos estatísticos são apropriados e descritos adequadamente? Por outro lado, o protocolo de investigação é geralmente submetido à apreciação de um ou mais conselhos ou comités de ética que podem ter normas diferentes. Por exemplo, pode haver necessidade de pedir aprovação aos comités de ética do país (ou instituições) de onde é oriundo o financiamento, do comité de ética (se aplicável) das instituições proponentes ou parceiras, e ainda do país onde se realiza a investigação (se for diferente). Por exemplo, no projecto UPHI-STAT pediu-se aprovação ao Conselho de Ética do IHMT (Doc. n.24-2013-PI) e ao Comité Nacional de Ética para a Pesquisa em Saúde (Doc. n.52/2013), em Cabo Verde. Enquanto no primeiro caso, o pedido era obrigatório por envolver o estudo de seres humanos, no segundo caso era facultativo pois não implicava a colheita de amostras biológicas. Porém, é sempre aconselhável que o país de acolhimento tenha conhecimento da investigação e, neste caso, o armazenamento dos questionários e dos consentimentos informados ficou a cargo do actual Instituto Nacional de Saúde de Cabo Verde, como instituição de acolhimento por indicação do Ministério da Saúde daquele país.

Na submissão aos conselhos de ética é essencial a descrição dos riscos e benefícios da recolha de dados e os seus procedimentos, o anonimato e a confidencialidade dos dados, o acesso, o armazenamento e a conservação dos dados e também a destruição da base de dados. Não havendo regras pré-definidas para a destrui-

[3] Último acesso a 25/03/2016

ção da base de dados, é frequente fazer-se referência a 5 anos, após o final do projecto ou após a última publicação dos resultados. A referência ao acesso à base de dados também pode limitar a partilha de dados com os estatísticos que não estejam afectos à equipa de investigação do projecto. É comum, o investigador principal do projecto declarar que a utilização dos dados fica restrita à equipa de investigação. Recorde-se que frequentemente os objectivos da investigação propostos no protocolo, salvo raras excepções, não são de natureza estatística, isto é, não implicam que se desenvolvam novos métodos estatísticos. Assim, podem surgir algumas questões sobre a utilização dos dados recolhidos para desenvolvimentos estatísticos teóricos não previstos inicialmente. Note-se que a Lei da Protecção de Dados Pessoais (Lei nº 67/98 e Lei n.º 103/2015, de 24/08 – `http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=156&tabela=leis&so_miolo=`), para qualquer área, no artigo 5º reforça a necessidade de apenas recolher dados (...) *para finalidades determinadas, explícitas e legítimas, não podendo ser posteriormente tratados de forma incompatível com essas finalidades* (alínea b); *Adequados, pertinentes e não excessivos relativamente às finalidades para que são recolhidos e posteriormente tratados.* Embora, no ponto 2, do mesmo artigo, haja a resalva de possíveis alterações *Mediante requerimento do responsável pelo tratamento, e caso haja interesse legítimo, a CNPD pode autorizar a conservação de dados para fins históricos, estatísticos ou científicos por período superior (...).* Relativamente, aos conselhos de ética também existe essas possibilidade de solicitar alterações.

4 Recolha de dados em diferentes contextos

Na prática, uma das maiores dificuldades para recolher uma amostra aleatória relaciona-se com a disponibilidade de bases de amostragem credíveis e acessíveis. O movimento dos refugiados e o estudo de populações em guerra tem colocado desafios aos investigadores que no

terreno constatam a inviabilidade dos métodos de amostragem clássicos. Assim, têm sido desenvolvidos métodos para populações de difícil acesso [7, 8, 9]. A Organização Mundial da Saúde tem vindo a alertar para a necessidade de estudar as populações que vivem em bairros informais das grandes cidades que albergam um número elevado de residentes e que frequentemente não entram nas estatísticas oficiais [10, 11]. Porém, nomeadamente em África, devido ao desenvolvimento dos sistemas de informação geográficos por vezes torna-se possível utilizar métodos baseados na utilização de coordenadas geográficas das residências de forma a chegar aos indivíduos, por exemplo, para aplicar um questionário porta-a-porta [6]. Mesmo que o protocolo de investigação possa descrever a população convenientemente, tendo prevista uma base de amostragem fiável, que o cálculo do tamanho da amostra possa ter sido efectuado de uma forma credível atendendo aos parâmetros em estudo, ao transpor a teoria para a prática pode haver uma discrepância considerável. Em determinadas situações, os indivíduos seleccionados para integrar a amostra pode não dar o seu consentimento (oral ou mais frequentemente escrito) para participar no estudo. O consentimento informado deve apresentar uma linguagem acessível aos indivíduos da população em estudo. Este documento deve fornecer aos indivíduos participantes, ou seus representantes legais, informação adequada sobre: (i) os objectivos da investigação; (ii) investigadores e os seus contactos; (iii) a explicação dos métodos e a utilização dos resultados; (iv) definir a forma de transmissão dos resultados aos participantes; (v) explicar os benefícios esperados, potenciais riscos do estudo e incómodos que lhe possam estar associados, bem como outros aspetos relevantes do estudo. Os indivíduos devem ter a opção de recusar a participação ou de a interromper a qualquer momento. Neste caso, existem omissões e vazios que levantam questões sobre até quando o participante pode contactar o investigador e desistir da sua participação do estudo. Num estudo transversal, supostamente após a informatização já não haverá ligação por entre o consentimento informado e o questionário, sendo este um ponto de paragem. No entanto, num estudo longitudinal, como a ligação aos dados pessoais está assegurada até

ao fim, em qualquer momento poderá fazer sentido que o sujeito seja retirado mesmo em fases tardias do estudo. A taxa de não respostas pode ser elevada em determinados contextos e mais frequentes em determinados sub-grupos podendo por em causa a representatividade das amostras. No projecto UPHI-STAT, na cidade da Praia, usaram-se balanças de bioimpedância, para recolher dados da massa gorda, massa muscular e óssea, exigindo a pesagem do indivíduo descalço. Se este estudo tivesse sido realizado em algumas cidades de Angola, seria de esperar uma reduzida participação dos homens, pelo facto de terem que se descalçar. Mesmo quando não existem aspectos culturais tão marcantes, a maior participação das mulheres em estudos de saúde é relatada com frequência em diversos contextos [12]. Em Cabo Verde, em dois estudos transversais, um em contexto urbano [6], e outro na ilha de Santiago [13] registam uma maior participação das mulheres que dos homens (64.4% vs 35.6% e 68.7% vs 31.3%, respectivamente). Apesar de haver inquéritos demográficos e de saúde de 5 em 5 anos e os censos serem cada vez mais regulares nos países africanos, continua a não ser fácil ter dados populacionais que permitam introduzir ponderadores na análise que exprimam da melhor forma a realidade. Porém, ao recolher dados por questionários a taxa de resposta em determinados contextos em África até é melhor que a verificada em países europeus, não sendo um questão meramente da saúde tropical.

5 Notas finais

A intervenção da estatística, antes da recolha de dados, está mais presente na investigação em saúde, quer pela pressão do financiamento e dos conselhos de ética, quer pela pressão das publicações científicas. Relativamente às publicações científicas tendem a exigir que as bases de dados sejam fornecidas, o que por vezes pode entrar em "conflito" com o assumido perante os conselhos de ética. Actualmente, o espaço lusófono está em expansão, em termos de ensino, de investigação e de consultoria, devendo haver cada vez mais atenção

à recolha de dados nestes contextos. Apesar dos constrangimentos no terreno, a investigação em saúde tropical está a oferecer inúmeras oportunidades à estatística. Os contextos rurais e urbanos de zonas tropicais, que podem incluir populações de difícil acesso e/ou simplesmente terem dificuldades acrescidas por questões culturais ou éticas (entre outras), oferecem novos desafios para que as estratégias de amostragem sejam inovadas. A supervisão e a monotorização da recolha de dados no terreno devem ter também uma atenção máxima, pois não adianta sofisticar a modelação estatística se os dados não apresentam a qualidade desejada. Por outro lado, operar em terrenos tão ricos, pode potenciar o desenvolvimento de novas metodologias estatísticas, fazendo com que a teoria seja desenvolvida em função das necessidades da prática.

Agradecimentos

Trabalho parcialmente financiado pela Fundação para a Ciência e Tecnologia (FCT) – Portugal – projectos PTDC/ATP/EUR/5074/2012 e PEst/OE/MAT/UI0006/2014. Um agradecimento especial à equipa do projecto UPHI-STAT.

Referências

[1] Brooks, N. (1996). *Writing a grant application.* In G.Parry & F.N.Watts (eds.), Behavioural and Mental Health Research: A Handbook of Skills and Methods (2nd edition). Hove: Erlbaum.

[2] Robson, C. (2002). *Real world research.* Oxford: Blackwell. (2nd edition).

[3] Robson, C. (2007). *How to do a research project - A guide for undergraduate students.* Blackwell Publishing.

[4] Médicos de Medicina Geral e Familiar(2008). *Investigação Passo a Passo – Perguntas e Respostas Essenciais para a Investigação Clínica.* Núcleo de Investigação da APMCG. Focom XXI, Lda.

[5] March, M., Susser, E. (2006). The eco- in eco-epidemiology. *Int. J. Epidemiol* 35, 1379–1383.

[6] Gonçalves, L., Santos, Z., Amado, M., Alves, D., Simões, R., Delgado, A., Correia, A., Velez Lapão, L., Cabral, J., Craveiro, I. (2015) Urban Planning and Health Inequities: looking in a small-scale in a City of Cape Verde. *PLoS ONE* 10(11): e0142955.

[7] Platt, L., Wall, M., Rhodes, T., Judd, A., Hickman, M., Johnston, L.G., et al. (2006). Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injecting drug users across nine studies in Russia and Estonia. *J Urban Health* 83(1):39–53.

[8] Southern, D.A., Lewis, S., Maxwell, C.J., Dunn, D.R., Noseworthy, T.W., Corbett, G., et al. (2008). Sampling hard-to-reach populations in health research: yield from a study targeting Americans living in Canada. *BMC Med Res Methodol.* 8:57–57.

[9] Kral, A.H., Malekinejad, M., Vaudrey, J., Martinez, A.N., Lorvick, J., McFarland, W., et al. (2010). Comparing respondent-driven sampling and targeted sampling methods of recruiting injection drug users in San Francisco. *J Urban Health* 87(5):839–850.

[10] WHO (2008). *Our cities, our health, our future. Acting on social determinants for health equity in urban settings.* Report to the WHO Commission on Social Determinants of Health from the Knowledge Network on Urban Settings.

[11] Unger, A., Riley, L. (2007) Slum health: From understanding to action. *PLoS Med.* 4(10):e295.

[12] Galea, S., Tracy, M. (2007) Participation Rates in Epidemiologic Studies. *Ann Epidemiol* 17:643–653.

[13] Rodrigues, L., Reis, P.D. (2013). *Conhecimentos, Atitudes e Práticas sobre o Paludismo em Cabo Verde.* The Global Fund to Fight AIDS, Tuberculosis and Malaria, Ministério de Saúde de Cabo Verde.

Uma aplicação da distribuição *a priori* árvore de Pólya no estudo da adequabilidade do modelo exponencial

Maria João Polidoro
ESTGF|CIICESI, Instituto Politécnico do Porto e CEAUL, *mjp@estgf.ipp.pt*

Fernando Magalhães
ISCAP, Instituto Politécnico do Porto e CEAUL, *fjmm@iscap.ipp.pt*

Palavras–chave: processo de Dirichlet, árvore de Pólya, teste de ajustamento bayesiano

Resumo: Nas últimas duas décadas tem-se assistido a um grande desenvolvimento de novas técnicas de inferência bayesiana associadas a uma abordagem não paramétrica. Surge, assim, o conceito de modelo bayesiano não paramétrico. O modelo bayesiano é designado de paramétrico se a distribuição de probabilidade utilizada para modelar os dados, $\{F_\theta : \theta \in \Theta\}$, tem uma forma conhecida e está indexada por um vetor de parâmetros de dimensão finita, usualmente desconhecido, o que requer a especificação de uma distribuição *a priori* sobre Θ. Segundo a abordagem bayesiana não paramétrica, procura-se uma classe mais geral de modelos $\{F : F \in \mathcal{F}\}$, o que requer a especificação de uma distribuição *a priori* sobre $\mathcal{F}$, o espaço de todas as medidas de probabilidade, denominada de modelo bayesiano não paramétrico. Na literatura estatística a distribuição *a priori* processo de Dirichlet (DP) é a mais referenciada e estudada. No entanto, a natureza discreta das distribuições obtidas a partir de um DP, limita a sua aplicação, nomeadamente nos problemas de modelação de dados contínuos. Em contraste, a distribuição *a priori* árvore de Pólya (PT), que permite a modelação de dados contínuos e é uma generalização do DP, não tem sido tão amplamente utilizada. Neste trabalho, pretende-se dar a conhecer uma aplicação da distribuição *a priori* árvore de Pólya no estudo da adequabilidade

do modelo exponencial, algumas das suas limitações e formas de as contornar.

1 Introdução

O processo de Dirichlet (DP) foi introduzido por Ferguson [1] como uma distribuição *a priori*, designada por F, sobre o espaço de todas as medidas de probabilidade $\mathcal{F}$. É, provavelmente, a distribuição mais utilizada na inferência bayesina não paramétrica pelas suas interessantes propriedades, como a da conjugação e a da existência de métodos que geram realizações de DP. No entanto, o DP gera quase certamente distribuições discretas, limitando a sua aplicação a muitos problemas estatísticos como, por exemplo, a estimação de densidades. Para contornar esta limitação, Lo [2] e Escobar e West [3] propõem um modelo de mistura por processos de Dirichlet (DPM), isto é, a distribuição desconhecida é definida utilizando um modelo de mistura de distribuições contínuas onde os pesos da mistura passam a ser uma medida de probabilidade aleatória que segue um DP.

Alternativamente, a distribuição *a priori* árvore de Pólya (PT), que é provavelmente o modelo bayesiano não paramétrico mais simples (Nieto-Barajas e Mueller [4]), tem como principal vantagem a modelação direta de distribuições discretas, contínuas ou absolutamente contínuas, para determinados valores dos seus parâmetros. Por outro lado, goza de imensas propriedades interessantes, em particular, goza da propriedade de conjugação para dados censurados à direita, o que não acontece no DP. A ideia base da distribuição PT encontra-se em Ferguson [5], mas foi Lavine [6, 7] e Mauldin [8] que desenvolveram e catalogaram detalhadamente a base teórica desta distribuição. Posteriormente, apareceram generalizações da distribuição PT, tais como, misturas de PT (Hanson e Johnson [9] e Hanson [10]), PT multivariadas (Paddock, Ruggeri, Lavine e West [11]), PT opcionais (Wong e Ma [12]) e Rubbery PT (Nieto-Barajas e Mueller [4]), conduzindo a um grande leque de aplicações de métodos não

paramétricos em problemas de inferência estatística e em diferentes áreas de investigação, nomeadamente, em problemas de estimação de densidades, regressão, curvas ROC, análise de sobrevivência, medições repetidas e validação de modelos, só para mencionar algumas.

Uma das soluções proposta pela abordagem bayesiana, para o estudo da adequabilidade de um modelo probabilístico paramétrico a um conjunto de dados observados, consiste em definir um modelo bayesiano não paramétrico alternativo que incorpore o modelo paramétrico em estudo. Seguidamente, a averiguação da adequabilidade do modelo é feita através de métodos de comparação de modelos, destacando-se o factor de Bayes como método de eleição para a comparação.

Neste trabalho, apresenta-se uma aplicação da utilização da distribuição PT no estudo da adequabilidade do modelo exponencial. É, ainda, apresentado um estudo de simulação para comparar o desempenho do teste de ajustamento bayesiano não paramétrico com alguns testes clássicos.

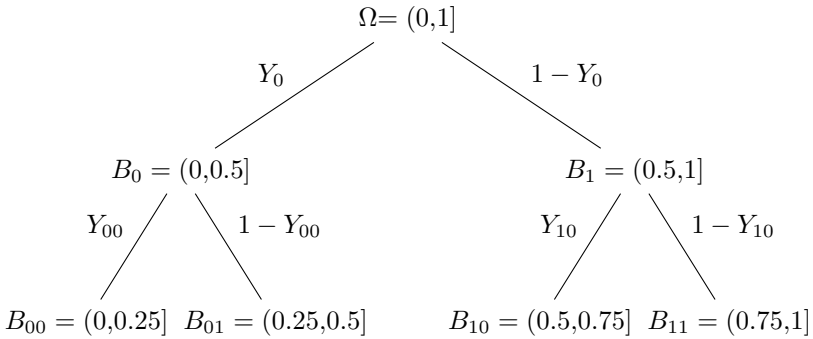
2 Distribuição Árvore de Pólya Finita

Uma distribuição PT com M níveis para G [4] é construída dividindo o espaço amostral Ω em intervalos disjuntos, utilizando o particionamento binário em árvore (ver Figura 1) e atribuindo probabilidades aleatórias a cada um dos ramos dessa árvore, ou seja, é definida por:

1. Uma sequência finita de partições binárias $\Pi = \{B_{\varepsilon_{1:m}}\}$, onde $\varepsilon_{1:m} = \varepsilon_1\varepsilon_2\cdots\varepsilon_m$ com $\varepsilon_j \in \{0,1\}$ para $j = 1,2,\ldots,m$ e $m = 1,2,\ldots,M$.

2. Um conjunto de variáveis aleatórias independentes com distribuição beta, $Y_{\varepsilon_{1:m}0} \sim \text{Beta}(\alpha_{\varepsilon_{1:m}0},\alpha_{\varepsilon_{1:m}1})$, representando a

[4] Utiliza-se G em vez de F para distinguir os dois modelos, não paramétrico e paramétrico, respetivamente.



$Y_0 \sim \text{Beta}(\alpha_0,\alpha_1), Y_{00} \sim \text{Beta}(\alpha_{00},\alpha_{01})$ e $Y_{10} \sim \text{Beta}(\alpha_{10},\alpha_{11})$
$Y_0 = \Pr(X_i \in B_0) = G(B_0)$
$Y_{00} = \Pr(X_i \in B_{00} | X_i \in B_0) \Rightarrow G(B_{00}) = Y_0 Y_{00}$
$Y_{10} = \Pr(X_i \in B_{10} | X_i \in B_1) \Rightarrow G(B_{10}) = (1 - Y_0) Y_{10}$.

Figura 1: Ilustração de uma distribuição PT com dois níveis, $M = 2$, para uma partição binária do espaço amostral, $\Omega = (0,1]$.

probabilidade de cada observação pertencer a cada um dos ramos da árvore.

3. Um conjunto de parâmetros não negativos $\mathcal{A} = \{\alpha_{\varepsilon_{1:m}}, m = 1,2,\ldots,M\}$.

A distribuição marginal de um qualquer conjunto $B_{\varepsilon_{1:m}}$, no m-ésimo nível, é dada por

$$G(B_{\varepsilon_{1:m}}) = \prod_{j=1,\varepsilon_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} \prod_{j=1,\varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1} 0}).$$

Uma distribuição PT com M níveis é determinada pelas partições Π e pelos parâmetros da distribuição beta em $\mathcal{A}$ e representa-se por $G \sim \text{PT}_M(\Pi,\mathcal{A})$.

Propriedades

1. Uma propriedade interessante da distribuição PT é a de que ela pode gerar distribuições de probabilidade contínuas. Para isso, basta que os parâmetros da distribuição beta em $\mathcal{A}$ aumentem rapidamente, por exemplo, considerando $\alpha_{\varepsilon_{1:m}} = cm^2$, com a constante $c > 0$. Por outro lado, se os parâmetros da distribuição beta diminuem rapidamente, por exemplo, se $\alpha_{\varepsilon_{1:m}} = c/2^m$, então a distribuição PT reduz-se ao caso particular de uma distribuição DP.

2. Outra propriedade atrativa, do ponto de vista prático, é a da facilidade de centrar a distribuição PT em torno de uma qualquer distribuição G_0, $E[G(B)] = G_0(B)$. Uma forma de o fazer é considerar que os limites (inferior e superior) dos intervalos que formam a partição coincidam com quantis de G_0 e supondo que $\alpha_{\varepsilon_{1:m}0} = \alpha_{\varepsilon_{1:m}1}$. Ou seja, no m-ésimo nível, os intervalos são definidos por $B_{\varepsilon_{1:m}} = \left(G_0^{-1}\left(\frac{k-1}{2^m}\right), G_0^{-1}\left(\frac{k}{2^m}\right)\right]$, para $m = 1,2,\ldots,M$, e $k = 1,2,\ldots,2^m$, onde $G_0^{-1}(0) = -\infty$ e $G_0^{-1}(1) = +\infty$, se $X_i \in \mathbb{R}$. Alternativamente, pode-se centrar a distribuição PT considerando uma partição arbitária fixa Π e fazendo com que $\alpha_{\varepsilon_{1:m}} = cG_0(B_{\varepsilon_{1:m}})$, $c > 0$.

3. A distribuição PT tem uma expressão fechada para a distribuição preditiva *a priori*.

Embora a distribuição PT possua propriedades interessantes, apresenta algumas limitações práticas, tais como: (i) é dependente da partição considerada; (ii) a densidade preditiva *a posteriori* é descontínua nos pontos extremos dos intervalos das partições; e (iii) a inerente dificuldade na escolha de F_0.

Para contornar estas limitações (Lavine [6], Hanson e Johnson [9] e Hanson [10]) substituem G_0 por uma distribuição paramétrica F_θ e consideram distribuições *a priori* para os hiperparâmetros, $h(\theta)$.

O modelo resultante é designado por mistura finita de árvores de Pólya (MPT) e permite, em particular, suavizar as descontinuidades nos pontos extremos dos intervalos das partições. Mais pormenores sobre distribuições *a priori* árvores de Pólya podem ser encontrados em Polidoro [14].

3 Aplicação

A distribuição exponencial é uma das mais simples e importante distribuições utilizada na modelação de dados que representam o tempo até à ocorrência de um determinado acontecimento. O estudo da adequabilidade da distribuição exponencial é fundamental para validar as inferências realizadas.

Neste trabalho, seguindo o procedimento proposto por Berger e Guglielmi [13] e como exemplo de aplicação, propõe-se um teste de ajustamento bayesiano não paramétrico para o estudo da adequabilidade da distribuição exponencial (H_0) considerando como modelo bayesiano não paramétrico alternativo (H_1) a distribuição MPT. A averiguação da adequabilidade do modelo proposto na hipótese nula é realizada utilizando o factor de Bayes. **Teste de Ajustamento Bayesiano**

O modelo bayesiano paramétrico (H_0 ou F_θ) é dado por

$$\begin{gathered} X_i|\theta \overset{\text{iid}}{\sim} \text{Exp}(\theta), \text{ for } i = 1,2,\ldots,n, \\ \theta \sim h(\theta). \end{gathered}$$

e o modelo bayesiano não paramétrico (H_1 ou G) é dado por

$$\begin{gathered} X_1,X_2,\ldots,X_n|G \overset{\text{iid}}{\sim} G \\ G|\Pi,\mathcal{A}_\theta \sim \text{MPT}_M(\Pi,\mathcal{A}_\theta) \\ \theta \sim h(\theta), \end{gathered}$$

onde $\text{MPT}_M(\Pi,\mathcal{A}_\theta)$ é a distribuição mistura finita PT, com parâmetros $(\Pi,\mathcal{A}_\theta)$ e $h(\theta)$ é a distribuição *a priori* para θ.

O factor de Bayes (BF) a favor do modelo paramétrico (H_0) e contra o modelo não paramétrico (H_1) é dado por

$$\mathrm{BF}_{01}(x) = \frac{p_0(x)}{p_1(x)}.$$

As distribuições preditivas *a priori* (sob H_0 e H_1) são, respetivamente, $p_0(x) = \int_\Theta f(x|\theta)h(\theta)d\theta$ e $p_1(x) = \int_\Theta p(x|\theta)h(\theta)d\theta$, onde $p(x|\theta) = f(x|\theta)\psi(\theta)$, com

$$\psi(\theta) = \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha'_{\varepsilon_{1:m}(x_j)}(\theta)\left(\alpha_{\varepsilon_{1:m-1}0(x_j)}(\theta) + \alpha_{\varepsilon_{1:m-1}1(x_j)}(\theta)\right)}{\alpha_{\varepsilon_{1:m}(x_j)}(\theta)\left(\alpha'_{\varepsilon_{1:m-1}0(x_j)}(\theta) + \alpha'_{\varepsilon_{1:m-1}1(x_j)}(\theta)\right)},$$

onde $\varepsilon_{1:m}(x_j)$ é o índice $\varepsilon_1\varepsilon_2\cdots\varepsilon_m$ que identifica o subconjunto da partição $B_{\varepsilon_1\cdots\varepsilon_m}$, para cada nível m, que contém x_j, $\alpha'_{\varepsilon_{1:m}(x_j)}(\theta)$ é igual a $\alpha_{\varepsilon_{1:m}(x_j)}(\theta)$ mais o número de observações entre $\{x_1, \ldots, x_{j-1}\}$ que pertencem a $B_{\varepsilon_1\cdots\varepsilon_m}(x_j)$. Para cada x_j, o limite superior $m^*(x_j)$, no produto, representa o menor nível m tal que nenhum x_i, $i < j$, pertence a $B_{\varepsilon_1\cdots\varepsilon_m}(x_j)$.

O cálculo do BF é simplificado porque pode ser escrito como

$$\mathrm{BF}_{01}(x) = \left[\int_\Theta \psi(\theta)h(\theta|x)d\theta\right]^{-1} = \left\{E[\psi(\theta)|x]\right\}^{-1},$$

onde $h(\theta|x) = f(x|\theta)h(\theta)/p_0(x)$, isto é, pode ser escrito como o inverso de uma média *a posteriori*, sob H_0. Caso se possa simular uma amostra aleatória θ_1,θ_2,...,θ_L da densidade *a posteriori* $h(\theta|x)$, o método de Monte Carlo direto aproxima o BF pelo inverso da média empírica

$$\widehat{\mathrm{BF}}_{01}(x) = \left[\frac{1}{L}\sum_{l=1}^{L}\psi(\theta_l)\right]^{-1}.$$

O BF é uma medida da evidência provida pelos dados a favor de uma das hipóteses (modelos) em confronto. Intuitivamente, o melhor modelo corresponde àquele que apresente o maior valor da distribuição

preditiva *a priori* para x. Um BF muito grande ou muito pequeno relativamente ao valor numérico um representa uma evidência muito forte nos dados a favor de uma das hipóteses contra a outra hipótese.

Nesta aplicação, para centrar a distribuição MPT em torno da distribuição exponencial, considerou-se a situação particular das partições (Π) fixas, substituindo o parâmetro θ pelo seu estimador de máxima verosimilhança: $B_{\varepsilon_{1:m}} = \left(F_{\widehat{\theta}}^{-1}\left(\frac{k-1}{2^m}\right), F_{\widehat{\theta}}^{-1}\left(\frac{k}{2^m}\right)\right]$, para $m = 1,2,\ldots,M$, e $k = 1,2,\ldots,2^m$, onde $F_{\widehat{\theta}}^{-1}(0) = 0$ e $F_{\widehat{\theta}}^{-1}(1) = +\infty$, uma vez que $X_i \in \mathbb{R}^+$. Para $\varepsilon_{1:m-1} = \varepsilon_1\varepsilon_2\cdots\varepsilon_{m-1}$, os parâmetros da distribuição beta, $\mathcal{A}_\theta$, são definidos por

$$\alpha_{\varepsilon_{1:m-1}0}(\theta) = c_m\left(\frac{F_\theta(B_{\varepsilon_{1:m-1}0})}{F_\theta(B_{\varepsilon_{1:m-1}1})}\right)^{1/2}$$

e

$$\alpha_{\varepsilon_{1:m-1}1}(\theta) = c_m\left(\frac{F_\theta(B_{\varepsilon_{1:m-1}1})}{F_\theta(B_{\varepsilon_{1:m-1}0})}\right)^{1/2},$$

onde $c_m \propto \eta^{-1}\rho(m)$, $\eta > 0$.

A função $\rho(m)$ é definida por forma a que a distribuição MPT se adapte a distribuições amostrais contínuas, por exempo, considerando $\rho(m) = m^2, m^3, 2^m, 4^m$ e 8^m. O parâmetro η controla a variância da distribuição MPT em torno da sua média, isto é, a variação dos valores de η determina quão concentrada está a distribuição MPT da distribuição exponencial. Estudos de simulação sugerem que: para valores de $\eta \to 0$ a distribuição MPT está mais concentrada em torno da distribuição paramétrica e o BF irá convergir para um; para valores de $\eta \to \infty$ a distribuição MPT está mais afastada da distribuição paramétrica e o BF será muito grande; entre estes dois extremos, o BF, por vezes, aumenta com η, mas também pode, inicialmente, diminuir para depois aumentar. Por conseguinte, opta-se por uma análise de robustez, calculando o BF

para vários valores de η e, seguidamente, escolhe-se o valor mínimo obtido, $\min(\widehat{\mathrm{BF}}_{01}(x))$, como uma escolha conservativa.

4 Estudo de Simulação

Com o objetivo de comparar o desempenho do teste bayesiano proposto com algumas das estatísticas de teste clássicas mais potentes: CO_n - Cox e Oakes; EP_n - Epps e Pulley; $\overline{\mathrm{CM}}_n$ - Cramér-von Mises modificada; AD_n - Anderson Darling; $\mathrm{BH}_{n,a=1}$ - Baringhaus e Henze e $\mathrm{T}_{n,a=2.5}$ - Henze e Meintanis, realizou-se um estudo de simulação Monte Carlo. Na primeira parte do estudo, são simuladas amostras da distribuição exponencial padrão, isto é, supondo H_0 verdadeira. Caso seja rejeitada H_0 então é cometido um erro do tipo I. Na segunda parte, são simuladas amostras supondo H_0 falsa, neste caso, utilizaram-se as distribuições frequentemente consideradas em outros estudos: Ga=Gama; Wei=Weibull; χ^2=Qui-Quadrado; HCa=Half-Cauchy; LN=Log-Normal e HN=Half-Normal e com diferentes taxas de falha. Caso seja rejeitada H_0, toma-se uma decisão correta. As estimativas empíricas, para a taxa de erro tipo I e para a potência, são calculadas através da proporção de vezes que a hipótese nula é rejeitada, com base em 500 amostras simuladas para três dimensões diferentes: $n = 25$; 50 e 100 (utilizou-se para nível de significância $\alpha = 5\%$).

No teste bayesiano definiu-se $M = 6$ níveis, para a construção das partições da àrvore de Pólya, considerou-se $\rho(m) = 4^m$ e $\eta = 2^s$, com s a tomar todos os valores inteiros no intervalo [-6,6]. Foram gerados $L = 2000$ valores da distribuição *a posteriori* Gama($a + n\bar{x}$,$b + n$), com $a = b = 0.001$ (utilizou-se para o parâmetro θ a distribuição *a priori* não informativa da família conjugada natural, Gama(a,b)) para calcular, para cada η e l, os valores dos parâmetros da distribuição beta e a respetiva estimativa do BF. Finalmente, é escolhido o valor mínimo das 13 estimativas calculadas. Mais pormenores sobre estes parâmetros, podem ser encontrados em Polidoro [14].

Tabela 1: Média (e desvio padrão) da estimativa empírica para a proporção de rejeições corretas para cada um dos testes.

		Teste						
Distr.	n	$\widehat{\mathrm{BF}}_{01}(x)$	EP_n	CO_n	$\overline{\mathrm{CM}}_n$	AD_n	$\mathrm{BH}_{n,a=1}$	$\mathrm{T}_{n,a=2.5}$
Exp(1)	25	0.05	0.038	0.034	0.048	0.046	0.042	0.046
		(0.028)	(0.022)	(0.021)	(0.030)	(0.023)	(0.024)	(0.036)
	50	0.050	0.034	0.042	0.044	0.052	0.044	0.052
		(0.030)	(0.017)	(0.031)	(0.023)	(0.027)	(0.026)	(0.023)
	100	0.050	0.056	0.050	0.050	0.048	0.054	–
		(0.031)	(0.032)	(0.033)	(0.027)	(0.037)	(0.031)	–
Ga(2,1)	25	0.696	0.588	0.578	0.592	0.554	0.650	0.676
		(0.060)	(0.088)	(0.066)	(0.088)	(0.076)	(0.077)	(0.092)
	50	0.989	0.924	0.984	0.918	0.928	0.956	0.912
		(0.019)	(0.025)	(0.025)	(0.030)	(0.021)	(0.033)	(0.045)
	100	1	0.996	0.998	0.996	0.998	0.998	–
			(0.008)	(0.006)	(0.008)	(0.006)	(0.006)	–
Ga(0.8,1)	25	0.110	0.114	0.148	0.110	0.146	0.136	0.026
		(0.046)	(0.049)	(0.052)	(0.044)	(0.052)	(0.047)	(0.019)
	50	0.214	0.212	0.266	0.202	0.248	0.232	0.092
		(0.071)	(0.067)	(0.068)	(0.068)	(0.078)	(0.077)	(0.037)
	100	0.318	0.348	0.452	0.328	0.392	0.380	–
		(0.083)	(0.063)	(0.095)	(0.079)	(0.093)	(0.074)	–
Wei(0.5,1)	25	0.990	0.938	0.976	0.940	0.974	0.966	0.706
		(0.024)	(0.037)	(0.025)	(0.035)	(0.019)	(0.023)	(0.068)
	50	1	0.998	1	1	1	1	0.998
			(0.006)					(0.010)
	100	1	1	1	1	1	1	–
								–
Wei(1.2,1)	25	0.208	0.150	0.124	0.164	0.144	0.164	0.214
		(0.040)	(0.041)	(0.048)	(0.041)	(0.047)	(0.042)	(0.072)
	50	0.294	0.270	0.260	0.276	0.224	0.242	0.318
		(0.038)	(0.039)	(0.057)	(0.056)	(0.053)	(0.071)	(0.058)
	100	0.635	0.598	0.594	0.574	0.530	0.612	–
		(0.029)	(0.033)	(0.057)	(0.042)	(0.024)	(0.044)	–
$\chi^2(1)$	25	0.576	0.626	0.810	0.614	0.782	0.704	0.250
		(0.068)	(0.067)	(0.043)	(0.075)	(0.049)	(0.079)	(0.054)
	50	0.880	0.882	0.982	0.872	0.962	0.926	0.738
		(0.034)	(0.033)	(0.020)	(0.025)	(0.028)	(0.028)	(0.031)
	100	1	0.992	1	0.988	1	1	–
			(0.014)		(0.017)			–
HCa(0,1)	25	0.766	0.764	0.742	0.770	0.754	0.758	0.250
		(0.030)	(0.039)	(0.033)	(0.036)	(0.038)	(0.035)	(0.058)
	50	0.968	0.950	0.928	0.956	0.936	0.946	0.564
		(0.020)	(0.025)	(0.030)	(0.023)	(0.025)	(0.021)	(0.059)
	100	0.998	0.998	0.998	0.998	0.998	0.998	–
		(0.004)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	–
LN(0,1)	25	0.270	0.134	0.090	0.170	0.170	0.122	0.056
		(0.038)	(0.049)	(0.037)	(0.040)	(0.052)	(0.050)	(0.021)
	50	0.352	0.168	0.140	0.266	0.342	0.206	0.072
		(0.042)	(0.067)	(0.057)	(0.075)	(0.078)	(0.057)	(0.039)
	100	0.740	0.234	0.184	0.446	0.704	0.296	–
		(0.038)	(0.048)	(0.063)	(0.074)	(0.047)	(0.062)	–
HN(0,1)	25	0.267	0.232	0.160	0.246	0.198	0.246	0.322
		(0.036)	(0.042)	(0.056)	(0.041)	(0.044)	(0.037)	(0.046)
	50	0.586	0.500	0.372	0.514	0.408	0.462	0.586
		(0.040)	(0.085)	(0.079)	(0.071)	(0.073)	(0.087)	(0.060)
	100	0.866	0.848	0.722	0.864	0.792	0.810	–
		(0.038)	(0.048)	(0.068)	(0.044)	(0.057)	(0.061)	–

5 Resultados e Conclusões

Na Tabela 1 apresentam-se a média (e o desvio-padrão) das estimativas empíricas da taxa de erro tipo I e da potência dos diferentes testes. Para as distribuições alternativas com função taxa de falha crescente (Ga(2,1), Wei(1.2,1) e HN(0,1)), notou-se que a potência empírica do teste de ajustamento bayesiano é quase sempre superior à dos testes clássicos. Por outro lado, quando as amostras simuladas são obtidas a partir de distribuições alternativas com taxa de falha decrescente, como é o caso da distribuição Wei(0.5,1), o teste de ajustamento bayesiano é, pelo menos, tão potente quanto os clássicos. No entanto, é ligeiramente menos potente do que alguns testes clássicos para as restantes distribuições com função taxa de falha decrescente (Ga(0.8,1) e $\chi^2(1)$), talvez por estas duas distribuições estarem mais próximas de uma distribuição rexponencial padrão. Para a distribuição Half-Cauchy, a potência empirica do teste bayesiano é comparável com a dos testes clássicos e para a distribuição LogNormal, particularmente quando as amostras são de pequena dimensão, o teste bayesiano é o que apresenta melhor desempenho. Assim, pode afirmar-se que o estudo de simulação efetuado, permite concluir que o teste bayesiano não paramétrico proposto para o estudo da adequabilidade da distribuição exponencial tem, de uma forma geral, um excelente desempenho.

Como trabalho futuro, pretende-se investigar a possibilidade de generalizar o teste de ajustamento bayesiano para outras distribuições. Além disso e simultaneamente, pretende-se analisar qual o impacto nos resultados do teste de ajustamento, utilizando a ideia de Nieto-Barajas e Mueller [4] que consiste em introduzir algum tipo de dependência entre as variáveis Y_ϵ, dentro do mesmo nível da partição, para ultrapassar o problema da descontinuidade nos extremos das partições.

Agradecimentos
Trabalho financiado pela FCT - Fundação para a Ciência e a Tec-

nologia, através do projeto UID/MAT/00006/2013.

Referências

[1] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1, 209–230.

[2] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* 12, 351–357.

[3] Escobar, M. D. e West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* 90, 577–588.

[4] Nieto-Barajas, L. E., e Mueller, P. (2012). Rubbery Polya Tree. *Scand. J. Stat.* 39(1), 166-184.

[5] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* 2, 615–629.

[6] Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* 20, 1222–1235.

[7] Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* 22, 1161–1176.

[8] Mauldin, R. D., Sudderth, W. D. e Williams, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* 20, 1203–1221.

[9] Hanson, T., Johnson, W. (2002). Modeling regression errors with a mixture of Polya trees. *J. Amer. Statist. Assoc.* 97, 1020–1033.

[10] Hanson, T. (2006). Inference for mixture of finite Polya tree models. *J. Amer. Statist. Assoc.* 101, 1548–1565.

[11] Paddock, S., Ruggeri, F., Lavine, M., e West, M. (2003). Randomised Polya tree models for nonparametric Bayesian inference. *Statist. Sinica* 13, 443–460

[12] Wong, W. e Ma, L. (2010). Optional Polya tree and Bayesian inference. *Ann. Statist.* 38, 1433–1459.

[13] Berger, J. O., e Guglielmi, A. (2001). Bayesian Testing of a Parametric Model versus Nonparametric Alternatives. *J. Amer. Statist. Assoc.* 96, 174–184.

[14] Polidoro, M. J. (2014). Metodologia Bayesiana e Adequação de Modelos. Tese de doutoramente, Universidade de Lisboa.

Previsão multi-passos: comparação de três abordagens com aplicação ao consumo de energia elétrica em Cabinda

António Casimiro Puindi
Faculdade de Economia da Universidade do Porto & CIDMA, *up201301154@fc.up.pt*

Geslie Fernandes
Faculdade de Economia da Universidade do Porto, *geslief@gmail.com*

Maria Eduarda Silva
Faculdade de Economia da Universidade do Porto & CIDMA, *mesilva@fep.up.pt*

Palavras–chave: energia elétrica, previsão multi-passos, sazonalidade.

Resumo:

Este trabalho constitui um estudo comparativo de três estratégias de previsão multi-passos do consumo de energia elétrica em Cabinda. As estratégias consideradas são: uma estratégia recursiva baseada no modelo de inovações em espaço de estados que comporta sazonalidades múltiplas, uma estratégia direta que usa redes neuronais artificiais e uma estratégia de retificação que combina previsões obtidas recursivamente a partir de um modelo linear com retificações obtidas com uma estratégia direta usando gradiente *boosting*. Considera-se como conjunto de treino a série temporal do consumo horário de energia elétrica (em mega-Watt) na cidade de Cabinda entre 1 de Janeiro de 2011 e 30 de Setembro de 2014 e preveem-se as 24h do dia 1 de Outubro. A estratégia recursiva mostra-se a mais adequada para captar as principais características da dinâmica do consumo horário de energia elétrica em Cabinda.

1 Introdução

Atualmente, a previsão de valores futuros de uma variável de interesse usando apenas os valores observados no passado, constitui um desafio em muitos cenários da vida real. Este problema pertence ao domínio da análise de séries temporais. Distingue-se previsão a um-passo quando se pretende prever apenas a próxima observação e previsão multi-passos quando o objetivo é obter previsões para vários momentos no futuro. Tradicionalmente, a previsão multi-passos é obtida recorrendo a uma estratégia recursiva, na qual é estimado um único modelo para a série temporal, geralmente com base na minimização do erro a um-passo. A previsão para o passo h obtém-se à custa das previsões anteriores, iterando o modelo. Mais recentemente, tem sido proposto o cálculo direto das previsões multi-passos, recorrendo à estimação de um modelo de séries temporais para cada horizonte de previsão, de modo a que as previsões sejam calculadas com base apenas nas observações. Se à estratégia recursiva está associado o problema do aumento da incerteza (variância) com o aumento do horizonte de revisão, à estratégia directa está associado o problema de uma função de previsão irregular, uma vez que são calculados diferentes modelos. Estas e outras considerações importantes sobre este tema podem ser consultadas com detalhe em [2]. Em particular, este autor refere que a escolha entre as duas estratégias corresponde a um *trade-off* entre enviesamento e variância. Com o objectivo de colmatar as deficiências das abordagens mencionadas, [2] propôs uma estratégia que denominou como estatégia de retificação e que combina previsões obtidas recursivamente a partir de um modelo linear com retificações obtidas com uma estratégia direta usando gradiente *boosting* e que tem um desempenho no mínimo comparável ao melhor das duas, direta e recursiva. Este trabalho apresenta um estudo comparativo da precisão das três estratégias de previsão multi-passos aplicadas ao consumo de energia elétrica (em mega-Watt) na cidade de Cabinda, Figura 1(a). De facto, o conhecimento da evolução do consumo de energia elétrica é fundamental quando se pretende dimensionar os sistemas de geração

de energia de modo a garantir uma oferta suficiente. Mais, a gestão eficiente dos sistemas de geração de energia elétrica requer previsões precisas para horizontes que podem ser curtos, por exemplo a próximas 24 horas, como podem ser muito longos, por exemplo a próxima década,[8]. Assim é de todo o interesse estudar qual a melhor estratégia de previsão a usar em cada contexto. O restante do trabalho está organizado da seguinte forma: Secção 2 descreve as estratégias de previsão consideradas; a Secção 3 apresenta e discute os resultados da aplicação das estratégias à série do consumo horário de energia elétrica em Cabinda. As considerações finais estão apresentadas na Secção 4.

2 Estratégias de Previsão

Considere-se uma série temporal $(y_1, y_2, \ldots, y_N)$ para a qual se pretende obter previsões (pontuais) para os $h = 1, \ldots, H$ valores seguintes. É sabido que o preditor óptimo no sentido do erro médio quadrático é a média condicional $\mu_{t+h|t} = \mathrm{E}(y_{t+h}|y_N)$. Neste trabalho a qualidade de previsão é aferida pela raiz quadrada do erro médio quadrático (RMSE) e pelo erro absoluto percentual médio (MAPE). Denote-se $\mathbf{y}_t = (y_1, y_2, \ldots, y_{t-p+1})$, onde p designa um desfasamento (*lag*).

2.1 Estratégia recursiva

A estratégia recursiva consiste em estimar um modelo para a série temporal $y_t = m(\mathbf{y}_t; \boldsymbol{\theta}) + e_t$, onde $\mathrm{E}(e_t) = 0$ e $\boldsymbol{\theta}$ é um vetor de parâmetros. Neste trabalho considera-se o seguinte modelo proposto por [1], designado por TBATS, adequado para séries temporais com M sazonalidades - sazonalidade múltipla.

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{M} S_{t-m_i}^{(i)} + d_t \tag{1}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t$$

$$S_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \mathrm{sen}\lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{2}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \mathrm{sen}\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{3}$$

$$d_t = \sum_{i=1}^{p} \phi_i d_{t-1} + \sum_{j=1}^{q} \theta_i \epsilon_{t-i}$$

onde $y_t^{(\omega)}$ denota uma transformação Box-Cox com parâmetro ω da série temporal y_t; l_t é o nível local; b_t denota a tendência; m_1,...,m_M representam os períodos sazonais; $S_t^{(i)}$ representa a i-ésima componente sazonal com representação trignométrica dada por (2) e (3), $\lambda_j^{(i)} = 2\pi j/m_i$, $s_{j,t}^{(i)}$ e $s_{j,t}^{*(i)}$ são o nível e o crescimento estocástico da i-ésima componente sazonal; k_i, é o número de harmônicas necessárias para a i-ésima componente sazonal; ϵ_t é um ruído branco. O modeloTBATS pode escrever-se como um modelo de inovações em espaço de estados

$$y_t^{(\omega)} = \mathbf{w}^{'} \mathbf{x}_{t-1} + \varepsilon_t$$

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\varepsilon_t$$

onde $\mathbf{w}^{'}$ é uma matriz linha, $\mathbf{g}$ é uma matriz coluna, $\mathbf{F}$ é uma matriz e $\mathbf{x}_t = (l_t, b_t, s_t^{(1)}, ..., s_t^M, d_t, d_{t-1}, d_{t-p+1}, \varepsilon_t, \varepsilon_{t-1}, ..., \varepsilon_{t-q+1})^{'}$ é o vetor de estados no tempo t e que é não observado.

A estimação dos parâmetros do modelo, baseada em suavização exponencial, está descrita em [1] e [5]. A distribuição preditiva da

variável transformada $y_t^{(\omega)}$ para o período futuro $N+h$, isto é a distribuição de $y_{N+h|N}^{(\omega)}$, dado o vetor de estado final $\mathbf{x}_N$ e os parâmetros, é Gaussiana com $\mathrm{E}(y_{N+h|N}^{(\omega)}) = \mathbf{w}^{'}\mathbf{F}^{h-1}\mathbf{x}_N$ e $\mathrm{var}(y_{N+1|N}^{(\omega)}) = \sigma^2$, $\mathrm{var}(y_{N+h|N}^{(\omega)}) = \sigma^2[1 + \sum_{j=1}^{h-1} c_j^2],\ h \geq 2$, onde $c_j = \mathbf{w}^{'}\mathbf{F}^{j-1}\mathbf{g}$. A distribuição preditiva de $y_{N+h|N}$ é não Gaussiana mas podem obter-se previsões pontuais e intervalares usando a transformação Box-Cox inversa de quantis apropriados da da distribuição de $y_{N+h|N}^{(\omega)}$, mais detalhes em [1] e [5]. Neste trabalho, foi usada a *package* `forecast` [4] para estimação do modelo e calculo das previsões. As previsões pontuais de passo h designam-se por $\hat{\mu}_{n+h|n} = m^{(h)}(\mathbf{y}_n; \boldsymbol{\theta})$.

2.2 Estratégia direta

Na estratégia direta considera-se um modelo $y_t = m_h(\mathbf{y}_{t-h}; \boldsymbol{\theta}_h) + e_{t,h}$, para cada horizonte h que se estima minimizando

$$\hat{\boldsymbol{\theta}}_h = \underbrace{argmin}_{\boldsymbol{\theta}_h \in \boldsymbol{\theta}_h} \sum_t [y_t - m_h(\mathbf{y}_{t-h}; \boldsymbol{\theta}_h)]^2$$

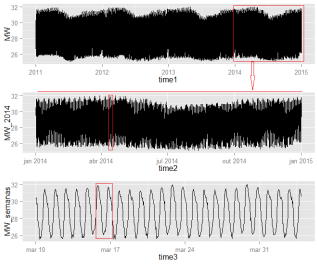
num conjunto de treino D_{train} para o horizonte h, [2]. As previsões para o passo h são obtidas do modelo correspondente, $\hat{\mu}_{n+h|n} = m_h(\mathbf{y}_{t-h}; \boldsymbol{\theta}_h)$. Neste trabalho os modelos $m_h(\cdot)$ são redes neuronais do tipo *multilayer perceptron*, MLP que permitem modelar relações não-lineares complexas entre um conjunto de variáveis de entrada e uma variável de saída. Uma rede neuronal é uma rede de nós organizadas em camadas, incluíndo: uma camada de entrada, produzida com as variáveis de entrada; uma ou mais camadas intermediárias, chamadas camadas ocultas que contêm nós escondidos; e um camada de saída com uma variável de saída. Redes MPL utilizam nas camadas ocultas funções de ativação não-lineares, como a função sigmoide. Considera-se também o caso particular de redes lineares, LIN, nas quais o número de nós escondidos é zero, [3].

2.3 Estratégia de retificação

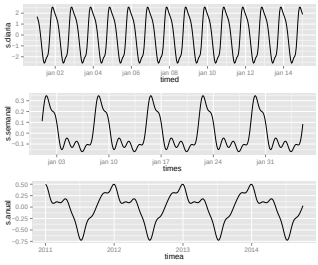
A estratégia de retificação é uma estratégia de previsão com 2 fases: primeiro modela-se a série temporal com um modelo autoregressivo $y_t = m(\mathbf{y}_{t-p}; \boldsymbol{\phi}) = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t$ e produzem-se previsões recursivas a partir do modelo estimado, $\hat{\mu}_{t+h|t} = m^{(h)}(\mathbf{y}_{t-p}; \hat{\boldsymbol{\phi}})$; em seguida corrigem-se as previsões aplicando a estratégia direta aos erros de previsão do modelo linear. Por outras palavras ajusta-se o modelo $y_t - m^{(h)}(\mathbf{y}_{t-p}; \hat{\boldsymbol{\phi}}) = m_h(\mathbf{y}_{t-h}; \boldsymbol{\theta}_h) + \epsilon_{t,h}$. As previsões são obtidas para cada horizonte h adicionando as retificações às previsões de base: $\hat{\mu}_{n+h|n} = m^{(h)}(\mathbf{y}_{t-p}; \hat{\boldsymbol{\phi}}) + m_h(\mathbf{y}_{t-h}; \hat{\boldsymbol{\theta}}_h)$. Esta estratégia reduz as irregularidades associadas aos diferentes modelos de previsão na estratégia directa uma vez que os modelos de retificação estão de algum modo associados pela facto de operarem sobre os erros de previsão do mesmo modelo linear. Claro que para os diferentes horizontes os modelos de retificação podem diferir mas essas diferenças serão menores ao modelar os resíduos da estratégia recursiva, comparativamente aos modelos resultantes de uma estratégia direta pura. Para mais detalhes consultar [2].

Neste trabalho, considera-se uma estratégia de *boosting* na fase de retificação. O *boosting* é uma técnica cujo objetivo é estimar a função de regressão ótima $\hat{f}(*) = \underbrace{arg\ \ min}_{f(*)}\{E_{Y,X}[\rho(Y, f(X))]\}$, onde $\rho(y, f(*)) = (y - f)^2$ representa a função de perda definida por L_2 (erro quadrático), que conduz à regressão clássica dos mínimos quadrados com $f(x) = E(Y|X = x)$. Em alternativa, considera-se a função de perda L_1 (erro absoluto), $\rho(y, f(*)) = |y - f|$, que conduz à regressão na mediana, [3]. Assim, os modelos de retificação são estimados usando o algoritmo de gradiente *boosting* sobre *P-splines* com as função de perda: L_1 e L_2, [2]. A ordem do modelos AR a usar na primeira fase é escolhida minimizando o AIC.

3 Aplicação à série de consumo horário de energia elétrica em Cabinda



(a) *Cronograma em diferentes escalas temporais:* 4 *anos (painel de topo),* 2014 *(painel do meio),* 4 *semanas em Março/Abril* 2014 *(painel do fundo)*



(b) *Padrões sazonais: intra–diária (painel de topo), intra–semanal (painel do meio); intra–anual (painel do fundo).*

Figura 1: Consumo horário de energia elétrica (em mega-Watt) em Cabinda

A Figura 1(a) apresenta o consumo horário de energia elétrica em Cabinda entre 1 de Janeiro de 2011 e 31 de Dezembro de 2014. O primeiro painel que representa os 4 anos de observações, indica a presença de sazonalidade mas não de tendência. As observações para o ano 2014, segundo painel, indicam que nos meses de verão, especialmente em Fevereiro, Março e Abril, se verifica um maior consumo. O terceiro painel mostra, não só, o ciclo intra-diário mas também o efeito fim de semana, com aumento de consumo ao sábado e domingo. No ciclo intra-diário o máximo do consumo ocorre entre as 19:00 e as 22:00 horas, com o pico às 21:00h. O consumo mínimo ocorre às 9:00. Verificam-se ainda efeitos de calendário, em especial o Natal com um comportamento similar ao do fim de semana (sá-

bado e domingo). Portanto, a série apresenta sazonalidade múltipla com frequências $m_1 = 24h$, $m_2 = 168h$ e $m_3 = 8766h$. O objetivo é usar a série do consumo de energia elétrica do período entre 1 de Janeiro de 2011 e 30 de Setembro de 2014 para prever o consumo do dia 1 de Outubro de 2014, recorrendo às 3 estratégias descritas na secção anterior. As previsões obtidas recorrendo ao modelo descrito na secção 2.1 são designadas por *REC*. Na estratégia direta foram construídas redes neuras artificiais cujas variáveis de entrada, ditadas pela função de autocorrelação amostral, são a procura desfasada em 24, 134, 168 e 169 horas. As previsões resultantes são designadas por $DirMLPeDirLIN$. As previsões resultantes da estratégia de retificação, são designadas por *RTY-L1* e *RTY-L2*.

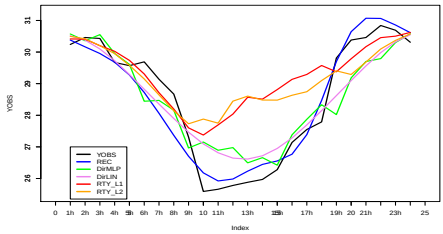


Figura 2: *Previsão do consumo das primeiras* 24 *horas de Outubro de* 2014*: comparação das três estratégias*

As medidas de precisão de previsão, Figura 2, mostram o desempenho favorável da estratégia recursiva em comparação com as outras estratégias. As distribuições preditivas das três estratégias estão representadas nas Figuras 3, 4 e 5. A função densidade de probabilidade da procura horária foi estimada utilizando a estimativa da densidade *kernel*, [7]. Os pontos na marcados no eixo das abcissas representam o consumo real observado às 12:00 e às 21:00. A seleção desses dois períodos horários foi feita por corresponderem a um período de carga de base e um período de carga de ponta, respetivamente.
Os valores do consumo real para os dois períodos horários escolhidos são valores típicos da distribuição preditiva obtida pela estratégia recursiva,

Previsão de 24 horas	Precisão	
	RMSE	MAPE
REC	0.468	1.284
$DirMLP$	0.813	2.315
$DirLIN$	0.739	2.189
$RTY - L1$	1.362	3.731
$RTY - L2$	1.362	3.833

Tabela 1: *Medidas de precisão da previsão observada na Figura 2*

Figuras 3. No entanto, a distribuição preditiva obtida pela estratégia directa parece sobre-estimar e sub-estimar o consumo nas horas base e ponta, respectivamente. Este efeito é mais notória nos resultados obtidos pela estratégia de retificação Figura 5.

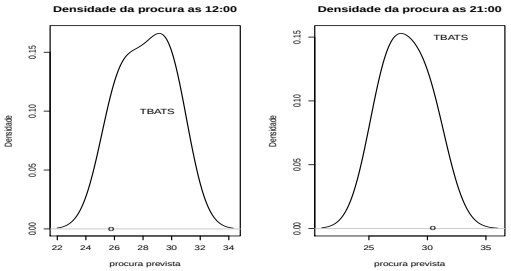
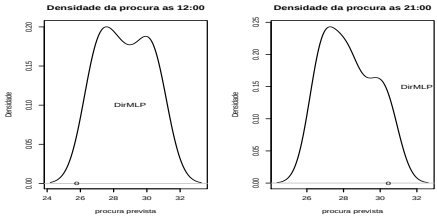


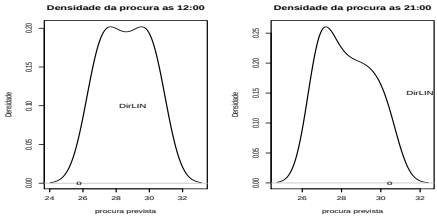
Figura 3: *Estratégia recursiva: avaliação da distribuição da previsão do consumo versus consumo real marcado por um ponto na abcissa.*

4 Considerações finais

Os resultados obtidos apontam para a necessidade de aprofundar o estudo comparativo das diversas abordagens à previsão uma vez que não seria de esperar que a estartégia de retificação tivessse pior desempenho. No prosseguimento deste trabalho, estuda-se a possibilidade da extensão



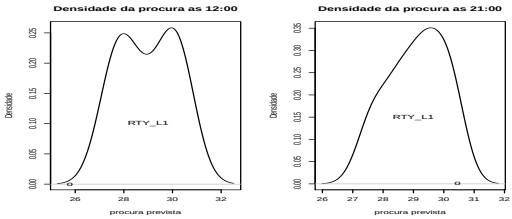
(a) Estratégia DirMLP



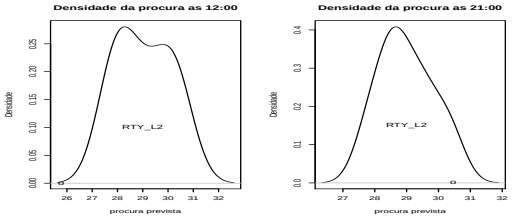
(b) Estratégia DirLIN

Figura 4: *Estratégia direta: avaliação da distribuição da previsão de consumo versus consumo real marcado por um ponto na abcissa.*

do modelo de inovação em espaço de estados que comporta sazonalidades múltiplas, com a incorporação de variáveis externas, que podem melhorar as previsões. O aumento sistemático do consumo de eletricidade aos sábados e domingos, é também uma característica crucial no estudo da variabilidade do consumo de energia elétrica, dando lugar ao estudo dos efeitos de calendário. Finalmente, os resultados foram obtidos, em geral, usando as funções *tbats*, *nnet* e *mboost* dos pacotes *forecast* e *caret* do *R*, [4] e [6]. Embora o pacote *forecast* permita aplicar diversos métodos de previsão, não está ainda disponível um pacote para aplicação de diferentes estratégias de previsão multi-passos à frente, especialmente para os algoritmos de aprendizagem automática, tal como a estratégia direta associada. Assim sendo, procedeu-se à implementação de funções para obter previsões recursivas e diretas geradas por algoritmos lineares e não lineares de aprendizagem automática.



(a) Estratégia RTY-L1



(b) Estratégia RTY-L2

Figura 5: *Estratégia de retificação: avaliação da distribuição da previsão do consumo versus consumo real marcado por um ponto na abcissa.*

Agradecimentos

Os autores agradecem a Empresa Nacional de Distribuição de Eletricidade (ENDE) de Angola/Cabinda pela disponibilização dos dados utilizados nesse estudo. Este trabalho foi parcialmente suportado pelo CIDMA através do projecto UID/MAT/04106/2013 financiado pela FCT-Fundação para a Ciência e Tecnologia de Portugal.

Referências

[1] Alysha, M.De Livera, Rob J. Hyndman & Ralph D.S. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential

Smoothing. *Journal of the American Statistical Association*, 106:496, 1513–1527, DOI:10.1198/jasa.2001.tm09771.

[2] Ben Taieb, S. (2014). Machine learning strategies for multi-step-ahead time series forecasting. PhD Thesis, Départment d'Informatique, Université Libre de Bruxelles, Belgium.

[3] Fernandes, G. (2015). Previsão multi-passos em séries temporais: estratégias clássicas e de aprendizagem automática. Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão, Faculdade de Economia da Universidade do Porto.

[4] Hyndman, R.J. (2015). forecast: Forecasting Functions for Time Series and Linear Models. R package version 6.2, `https://cran.r-project.org/web/packages/forecast/forecast.pdf`

[5] Hyndman, R.J., A.B.Koehler, J.K.Ord and R.D.Snyder (2008). *Forecasting with Exponential Smoothing: the state space approach.* Springer-Verlang, 137–143.

[6] R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, url = https://www.R-project.org/.

[7] Silverman, B. W. (1986). *Density Estimation.* London: Chapman and Hall.

[8] Taylor, J.W. (2010). Triple Seasonal Methods for Short–term Electricity Demand Forecasting. *European Journal of Operational Research* 204, pp.139–152.

Deteção de *outliers* no modelo de equações simultâneas usando o estimador GMM robusto

Anabela Rocha
ISCA e CIDMA, Universidade de Aveiro, *anabela.rocha@ua.pt*

Manuela Souto de Miranda
DMat e CIDMA, Universidade de Aveiro, *manuela.souto@ua.pt*

João Branco
IST e CEMAT, Universidade de Lisboa, *jbranco@math.ist.utl.pt*

Palavras–chave: SEM (*Simultaneous Equation Model*), SUR (*Seemingly Unrelated Regressions*), robustez, *outliers*, GMM (*Generalized Method of Moments*).

Resumo: O modelo SEM é uma generalização do modelo de regressão multivariado que assume dependência entre equações. Esta característica do SEM cria dificuldades adicionais às que já existem na deteção de *outliers* em modelos multivariados. Neste trabalho, propõe-se um novo método para detetar *outliers* em SEM. A proposta baseia-se numa versão robusta do estimador GMM e adapta ao SEM uma metodologia que foi recentemente utilizada para o modelo SUR, uma vez que este modelo também pressupõe dependência entre equações. As técnicas aplicadas mostraram-se adequadas para a deteção de *outliers*; o desempenho deste método foi comparado com o dos métodos convencionais, com base num estudo de simulação e num conjunto de dados reais. Os resultados mostraram vantagens na utilização da metodologia robusta que aqui se propõe, o que resulta numa mais valia do uso destes modelos na resolução de uma grande variedade de problemas que surgem na prática.

1 Introdução

Os modelos SEM e SUR são frequentemente usados em Econometria e generalizam o modelo de regressão multivariado. O SEM apresenta algumas caraterísticas específicas exigindo processos de estimação mais elaborados do que os que se usam no modelo de regressão ou no SUR.
De entre os estimadores tradicionais para o SEM destacam-se o estimador $3SLS$ (*Three Stages Least Squares*), que é o mais popular, e o estimador GMM. Estes estimadores apresentam boas propriedades, mas não são robustos, sendo muito sensíveis a desvios em relação ao modelo especificado ou à presença de *outliers*. Uma versão robusta para o estimador GMM foi apresentada em Rocha [8].
No presente trabalho foram consideradas sugestões de estimação robusta desenvolvidas para o modelo SUR por Bilodeau e Duchesne [2] e em Hubert *et al* [3]. Estudou-se o desempenho do estimador GMM robusto com base num estudo de simulação, no qual se mantiveram os cenários e os critérios contemplados em Hubert *et al* [3] para o modelo SUR. Este estudo evidenciou a vantagem da estimação robusta quando se verificam desvios dos pressupostos assumidos para o modelo, quer ao nível da localização, quer ao nível da dispersão. Por outro lado, estudou-se um conjunto de dados reais com o objetivo de proceder à deteção de *outliers* univariados e multivariados, adaptando ao SEM os procedimentos robustos usados em Bilodeau e Duchesne [2] e em Hubert *et al.* [3] para o modelo SUR. Este estudo mostrou vantagem nesta metodologia robusta para a deteção de observações atípicas, tanto a nível univariado como multivariado.

Todos os cálculos foram realizados com o programa *R-3.2.1.*

2 Modelo de equações simultâneas

O SEM é caraterizado por um sistema de equações interdependentes que inclui variáveis endógenas e variáveis exógenas. O SEM gene-

raliza o modelo de regressão multivariado, no sentido em que admite erros correlacionados com regressores e erros heterocedásticos.

Exemplo 2.1 *Um exemplo clássico de SEM é o Modelo Keynesiano simples, definido por:*

$$\begin{cases} y_t &= c_t + x_t \\ c_t &= \beta + \gamma y_t + \varepsilon_t \ , \end{cases}$$

onde, para um momento t, c_t representa o consumo (variável endógena), y_t representa o rendimento (variável endógena), x_t representa o investimento (variável exógena), ε_t é o erro aleatório, γ e β são os parâmetros estruturais.
Como se pode observar na 1^aequação, o rendimento depende do consumo, mas o consumo também é influenciado pelo rendimento, de acordo com a 2^aequação, mostrando a interdependência existente entre as equações do modelo.

Uma forma muito usada para escrever o SEM é a forma estrutural:

$$\mathbf{Y\Gamma} + \mathbf{XB} + \mathbf{E} = \mathbf{0},$$

onde $\mathbf{Y}$ e $\mathbf{X}$ são as matrizes de observações das variáveis endógenas e exógenas, respetivamente, $\mathbf{E}$ é a matriz dos erros aleatórios, $\mathbf{\Gamma}$ e $\mathbf{B}$ são as matrizes dos parâmetros estruturais.
Outra representação do SEM que é conveniente para a estimação dos parâmetros é dada pela equação:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{e}, \tag{1}$$

onde $\mathbf{Z} = diag\left[\begin{array}{ccc} \mathbf{Z}_1 & \cdots & \mathbf{Z}_M \end{array}\right]$, com $\mathbf{Z}_i = \left[\begin{array}{cc} \mathbf{Y}_i & \mathbf{X}_i \end{array}\right]$.
Note-se que entre as variáveis $\mathbf{Z}_i$, que são as variáveis explicativas do SEM, há variáveis endógenas que são correlacionadas com os erros, fazendo com que a estimação por GLS (*Generalized Least Squares*) conduza a um estimador não consistente. Este problema pode ser resolvido utilizando variáveis instrumentais e aplicando a seguir a estimação por GLS. Este processo é designado por estimador $3SLS$.

O SEM escrito na forma (1) é semelhante em termos formais ao modelo SUR. No entanto importa distinguir os dois tipos de modelos: enquanto que no SEM há variáveis endógenas entre as variáveis explicativas em (1), no SUR tal não acontece e a correlação entre equações é devida a fatores externos ao modelo, que se refletem apenas na correlação não nula entre erros de diferentes equações.

Exemplo 2.2 *Um exemplo de SUR, publicado em Judge et al. [4], refere-se a duas empresas americanas do mesmo ramo (General Electric e Westinghouse), onde cada equação traduz a relação entre o investimento bruto anual dessa empresa (Y_1 e Y_2) e as ações emitidas (X_1) e o capital social (X_2) da empresa. O modelo é constituído por um sistema de duas equações da forma:*

$$\begin{cases} Y_{1t} &= \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + u_{1t} \\ Y_{2t} &= \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t} \end{cases},$$

A presença de fatores que influenciam ambas as empresas vai provocar a existência de correlação entre os erros das duas equações. Ao contrário do que acontecia no SEM, as variáveis explicativas do SUR não são correlacionadas com os erros, pelo que a estimação por GLS permite obter um estimador consistente, ao contrário do que acontece na estimação do SEM.

3 Estimação do SEM

De entre os estimadores tradicionais do SEM, destacam-se o estimador $3SLS$ e o estimador GMM. Estes estimadores têm boas propriedades sob um conjunto de pressupostos do modelo, nomeadamente no modelo normal, mas podem sofrer grandes perturbações quando há desvios em relação ao modelo e, em particular, na presença de observações atípicas na amostra.
A estimação robusta surge como uma alternativa conveniente pois é pouco sensível a ligeiros afastamentos dos pressupostos assumidos

para o modelo. De entre os principais trabalhos sobre estimação robusta em *SEM*, destacam-se as propostas de Amemiya [1], Maronna e Yohai [6], Krishnakumar e Ronchetti [5] e Rocha [8].
No seguimento, vai usar-se a versão robusta do estimador GMM proposta em Rocha [8], a qual será designada por estimador $GMMR$. Resumidamente, o algoritmo que permite obter esse estimador, consiste no seguinte procedimento:
P.1. Obter estimativas iniciais dos resíduos, aplicando regressão robusta por equação com base no estimador LTS (*Least Trimmed Squares*), proposto em Rousseeuw [9].
P.2. Estimar a matriz de covariâncias dos erros, usando o estimador OGK (*Orthogonalized Gnanadesikan-Kettenring*), publicado em Maronna e Zamar [7], aplicado aos resíduos obtidos no passo P.1.
P.3. Resolver o problema de minimização de uma função de Huber com resíduos ponderados pelas estimativas das covariâncias obtidas no passo P.2.
Neste trabalho foram adaptados ao SEM os procedimentos robustos sugeridos em Rousseeuw e Van Zomeren [10]. No estudo de simulação desenvolveu-se uma adaptação ao SEM dos cenários e critérios de avaliação do desempenho de estimadores, também usados em Hubert *et al.* [3] para o modelo SUR.
Para a deteção de *outliers* univariados e multivariados procedeu-se à adaptação ao SEM dos princípios usados em Bilodeau e Duchesne [2] e em Hubert *et al.* [3], os quais foram originalmente propostos por Rousseeuw e Van Zomeren [10] para estimadores LMS (*Least Median of Squares*) e MVE (*Minimum Volume Ellipsoid*).

4 Estudo de simulação

Para estudar o desempenho do estimador $GMMR$, efetuou-se um estudo de simulação, gerando as observações de acordo com um SEM particular, já trabalhado por outros autores.
O SEM considerado foi proposto por Judge *et al.* [4], com forma

estrutural definida pelo sistema:

$$\left\{\begin{array}{rrrrrl} -\mathbf{Y}_1 & +\mathbf{Y}_2\gamma_{21}+\mathbf{Y}_3\gamma_{31}+\mathbf{X}_1\beta_{11} & & & +\mathbf{e}_1 & =\mathbf{0} \\ \mathbf{Y}_1\gamma_{12}-\mathbf{Y}_2 & & +\mathbf{X}_1\beta_{12}+\mathbf{X}_2\beta_{22}+\mathbf{X}_3\beta_{32}+\mathbf{X}_4\beta_{42} & & +\mathbf{e}_2 & =\mathbf{0} \\ & \mathbf{Y}_2\gamma_{23}-\mathbf{Y}_3 & +\mathbf{X}_1\beta_{13}+\mathbf{X}_2\beta_{23} & & +\mathbf{X}_5\beta_{53}+\mathbf{e}_3 & =\mathbf{0} \end{array}\right. .$$

Na simulação, mantiveram-se os valores dos parâmetros e das variáveis exógenas tal como em Judge *et al.* [4]. Para comparar o desempenho dos estimadores em diferentes condições, simularam-se amostras adaptando ao SEM os cenários usados por Hubert *et al.* [3]: consideraram-se várias distribuições dos erros, nomeadamente, distribuição Normal 3D, com percentagens de contaminação 0, 5, 10 e 30%. Contaminaram-se os valores da variável $\mathbf{Y}_2$, por esta variável ser explicativa nas primeira e terceira equações e por ser variável dependente na segunda equação.
Para cada distribuição anteriormente referida, geraram-se 100 amostras de dimensões 30 e 100, calcularam-se as estimativas dos parâmetros e os resíduos a partir dos estimadores $GMMR$ e $3SLS$.
Com o objetivo de avaliar o desempenho dos estimadores, utilizaram-se os indicadores usados por Hubert *et al.* [3] para o modelo SUR, com base em N amostras:

$$\textbf{Viés} : \left\| 1/N \sum_{k=1}^{N} \hat{\delta}^{(k)} - \delta \right\|, \tag{2}$$

$$\textbf{Erro Quadrático Médio (EQM)} : 1/N \sum_{k=1}^{N} \left\| \hat{\delta}^{(k)} - \delta \right\|^2. \tag{3}$$

Na Tabela 1 encontram-se os resultados relativos ao viés dos estimadores $GMMR$ e $3SLS$, no caso da dimensão amostral n=30 e para diferentes graus de contaminação, de acordo com (2). Os valores mostram que o estimador $GMMR$ tem melhor desempenho nos cenários de contaminação.
Na Tabela 2 encontram-se os resultados relativos ao erro quadrático médio obtido para os estimadores $GMMR$ e $3SLS$, no caso da dimensão amostral n=30 e para os mesmos cenários de contaminação, de acordo com (3). Os valores mostram que, tal como aconteceu em relação ao viés, também relativamente ao critério do erro quadrático médio, os melhores resultados são encontrados com o estimador $GMMR$, desde que a distribuição esteja contaminada. Os resultados obtidos para a dimensão amostral n=100 conduzem às mesmas conclusões, pelo que não são aqui apresentados; ainda

Viés	*3SLS*	*GMMR*
Normal	0.789	2.638
Normal-ct5	80.326	29.071
Normal-ct30	101.744	8.748

Tabela 1: Valores do viés dos estimadores $3SLS$ e $GMMR$, para amostras de dimensão n=30 e diferentes graus de contaminação.

EQM	*3SLS*	*GMMR*
Normal	142.979	208.191
Normal-ct5	6 481.576	6 037.276
Normal-ct30	10 419.45	6 432.463

Tabela 2: Valores do EQM para os estimadores $GMMR$ e $3SLS$, para dimensão amostral n=30 e diferentes graus de contaminação.

assim, é de notar que o estimador $GMMR$ apresentou menor variabilidade. Por motivos idênticos, os resultados para a contaminação 10% não são relatados nas tabelas 1 e 2, uma vez que conduzem a conclusões análogas às dos restantes graus de contaminação.
Em face dos resultados e para as situações simuladas, podemos concluir que o estimador $3SLS$ apenas produz melhores resultados no modelo Normal sem contaminação. Desde que exista contaminação, e para qualquer dos graus considerados, o estimador $GMMR$ mostra-se superior.

5 Deteção de *outliers* no SEM

Como já referimos, a deteção de *outliers* é uma tarefa difícil neste tipo de modelos, não só por estarem presentes as dificuldades conhecidas com observações multivariadas, mas também porque a dependência entre equações mascara ainda mais as observações realmente atípicas.

Motivados pela necessidade de dispor de um meio de diagnóstico de deteção de *outliers* em SEM, e na ausência de outras propostas na bibliografia sobre o assunto, decidiu-se seguir de perto a metodologia aplicada por outros autores para o modelo SUR, nomeadamente em Bilodeau e Duchesne [2] e em Hubert *et al.* [3]. Esses autores sugerem que se investigue, separadamente, a deteção de *outliers* univariados e multivariados.
A deteção de *outliers* univariados baseia-se numa representação gráfica, para cada equação. Propomos que os valores dos resíduos obtidos para cada estimador sejam representados contra os valores da distância de Mahalanobis robusta das observações das variáveis explicativas de cada equação. Os limites a considerar para o eixo dos resíduos são as retas horizontais definidas pelos valores $+2.5$ e -2.5, subjacentes à hipótese de que os erros têm distribuição Normal; no eixo horizontal, onde se registam as distâncias de Mahalanobis, sugere-se a reta vertical definida pelo valor da raiz quadrada de um quantil elevado da distribuição qui-quadrado com $k_i - 1$ graus de liberdade, onde k_i é o número de variáveis explicativas da i-ésima equação, incluindo o termo constante. Este tipo de representação gráfica de resíduos permite simultaneamente avaliar a qualidade do ajustamento (através do eixo dos resíduos) e identificar pontos de alavanca (através do eixo da distância de Mahalanobis).
Importa também e sobretudo detetar *outliers* multivariados no modelo. Para a determinação de *outliers* multivariados propõe-se um outro tipo de gráfico, representando nas ordenadas as distâncias de Mahalanobis (clássicas ou robustas) dos resíduos multivariados do ajustamento robusto e nas abcissas a sequência (ou os índices) das observações. Relativamente aos limites para detetar *outliers* multivariados, os princípios foram os já referidos para o caso univariado, isto é, no eixo das ordenadas usar a reta horizontal definida pelo valor da raiz quadrada de um quantil elevado da distribuição qui-quadrado com $k-1$ graus de liberdade, onde k é o número de variáveis explicativas do modelo, incluindo o termo constante.
Para ilustrar o método proposto, apresenta-se um exemplo de um SEM com dados reais, já trabalhado por outros autores, permitindo deste modo a comparação de resultados.

Exemplo 5.1 *Em Maronna e Yohai [6] é modelado por um SEM um conjunto de dados reais da economia da Argentina, relativos ao período entre 1956 e 1984, com a seguinte forma estrutural:*

$$\begin{cases} -\mathbf{Y}_1 & +\mathbf{Y}_3\gamma_{31} & & & +\mathbf{X}_5\beta_{51}+\mathbf{e}_1 & = \mathbf{0} \\ \mathbf{Y}_1\gamma_{12}-\mathbf{Y}_2 & & +\mathbf{X}_1\beta_{12} & & +\mathbf{X}_4\beta_{42}+\mathbf{X}_5\beta_{52}+\mathbf{e}_2 & = \mathbf{0} \\ \mathbf{Y}_1 & -\mathbf{Y}_2-\mathbf{Y}_3 & +\mathbf{X}_1 & +\mathbf{X}_2-\mathbf{X}_3+\mathbf{X}_4 & & = \mathbf{0} \end{cases},$$

onde as variáveis exógenas são X_1, o investimento bruto, X_2, o volume de exportações, X_3, os custos de impostos, X_4, o consumo público e X_5, a variável unitária; as variáveis endógenas são Y_1, o consumo privado, Y_2, o volume de importações e Y_3, o rendimento. Os dados correspondem a registos anuais das respetivas variáveis do sistema anterior.

Como se referiu, procede-se separadamente à deteção de *outliers* univariados e multivariados. Para identificação das observações atípicas, optou-se pelo quantil 0.975 da distribuição qui-quadrado, para a distância de Mahalanobis, e marcaram-se os anos no eixo das abcissas.

5.1 Análise do Exemplo 5.1 no caso univariado:

Observando a Figura 1, não se deteta a existência de *outliers*, nem de grandes valores de resíduos para a equação 1, uma vez que em nenhum dos dois gráficos aparecem pontos para além das retas limites. Logo, seja na perspetiva clássica com o estimador $3SLS$ e a distância de Mahalanobis convencional, seja do ponto de vista robusto com o estimador $GMMR$ e a distância de Mahalanobis robusta, não há indícios de observações discordantes na 1ª equação.

Relativamente à 2ª equação, a que se refere a Figura 2, as conclusões são diferentes. De facto, ambas as imagens mostram que não há resíduos a destacar (seja por $3SLS$, seja por $GMMR$), uma vez que não existem pontos para além dos limites considerados no eixo vertical (dos resíduos). Ao analisar o limite relativo à distância de Mahalanobis (no eixo horizontal), verifica-se que na imagem da direita da Figura 2 são destacados diversos pontos, não realçados na imagem da esquerda. Isto significa que, com a metodologia robusta baseada na estimação por $GMMR$ e na distância de Mahalanobis robusta, foi possível detetar *outliers* que passavam despercebidos na metodologia clássica, baseada em estimação por $3SLS$ e na distância de Mahalanobis clássica.

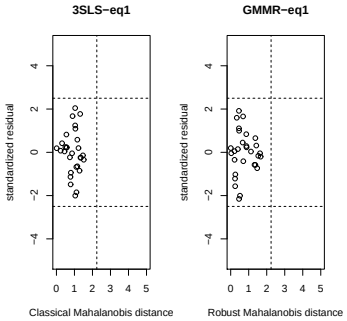


Figura 1: Deteção gráfica de *outliers* da 1ªequação: resíduos com os estimadores $3SLS$ e $GMMR$, contra a distância de Mahalanobis clássica e robusta.

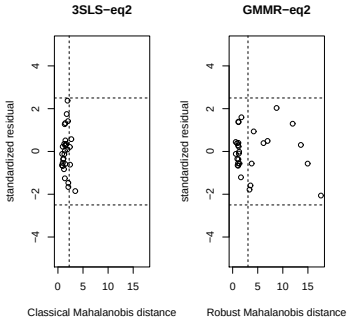


Figura 2: Deteção gráfica de *outliers* da 2ªequação: resíduos com os estimadores $3SLS$ e $GMMR$, contra a distância de Mahalanobis clássica e robusta.

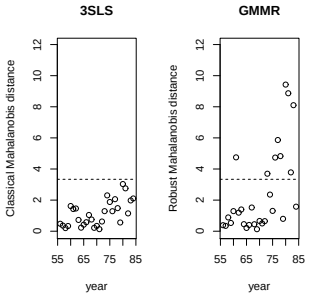


Figura 3: Deteção gráfica de *outliers* do sistema: distância de Mahalanobis clássica e robusta dos resíduos multivariados, com o estimador $GMMR$, contra os anos.

5.2 Análise do Exemplo 5.1 no caso multivariado:

Na Figura 3, observando o limite relativo à distância de Mahalanobis (no eixo vertical), destacam-se diversos pontos na imagem da direita, os quais não aparecem na imagem da esquerda. Isto traduz que a metodologia robusta, que combina a estimação por $GMMR$ com a distância de Mahalanobis robusta, permitiu detetar *outliers* multivariados que não eram notados com a metodologia clássica.

6 Comentários finais

Realizou-se um estudo de simulação que evidenciou a vantagem da estimação robusta ($GMMR$), quando se verificam desvios dos pressupostos assumidos para o modelo, quer ao nível da localização, quer ao nível da dispersão. Estudou-se a deteção de *outliers* univariados e multivariados no SEM procedendo à adaptação de metodologias propostas anteriormente para outros modelos. Os novos procedimentos para a deteção de *outliers* mostraram-se mais eficazes. Os métodos robustos que se propõem neste trabalho mostraram-se preferíveis na deteção de observações atípicas no modelo SEM, quer na perspetiva univariada, quer na multivariada.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação Portuguesa para a Ciência e Tecnologia (FCT-Fundação para a Ciência e a Tecnologia), por meio do CIDMA - Centro de Investigação e Desenvolvimento em Matemática e Aplicações, dentro do projeto UID / MAT / 04106/2013.

Referências

[1] Amemiya, T. (1982). Two stage least absolute deviation estimators, *Econometrica*,50, 689–711.

[2] Bilodeau, M. and Duchesne, P. (2000). Robust estimation of the *SUR* model. *The Canadian Journal of Statistics*, Vol. 28, 2, 277–288.

[3] Hubert, M., Verdonk, T. and Yorulmaz, O. (2014). Fast robust *SUR* with applications to the multivariate chain ladder method. *ROBUST@Leuven, Publications, Technical reports.*

[4] Judge, G., Griffiths, W., Lutkepohl, Hill, R. and Lee, T. (1988). *Introduction to the theory and practice of econometrics, second edition*, John Wiley & Sons, New York.

[5] Krishnakumar, J. e Ronchetti, E. (1997). Robust estimators for simultaneous equations models, *Journal of Econometrics*, 78, 295–314.

[6] Maronna, R. e Yohai, V. (1997). Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference*, 57, 233–244.

[7] Maronna, R. e Zamar, R. (2002). Robust estimates of location and disersion for high-dimensional datasets, *Technometrics*, 44, 307–317.

[8] Rocha, A. (2010). *Estimação robusta em Modelos Lineares de Equações Simultâneas, Tese de Doutoramento*, Universidade de Aveiro.

[9] Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.

[10] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, 85, 633–639.

Estimação em misturas pseudo-convexas

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*

Miguel Felgueiras
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, CIGS-IPL — Centro de Investigação em Gestão para a Sustentabilidade, *mfelg@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*

Palavras–chave: misturas pseudo-convexas, distribuições estáveis para extremos, estimação paramétrica, simulação.

Resumo: Neste trabalho são comparados, via simulação, os desempenhos de estimadores paramétricos para misturas pseudo-convexas geradas pela distribuição exponencial e pela distribuição função potência, duas distribuições estáveis para extremos quando esta definição é estendida para permitir alterações do parâmetro de forma.

1 Introdução

Na teoria dos valores extremos uma distribuição fechada para mínimo (máximo) é referida como min-estável (max-estável) e desempenha um papel fundamental na caraterização do mínimo (máximo) observado num fenómeno aleatório (cf., por exemplo, [1] e [2]). Com o objetivo de obter distribuições mais flexíveis, [3] generaliza esta de-

finição de modo a permitir alterações no parâmetro de forma. Com base nesta nova família de distribuições desenvolve um trabalho preliminar sobre misturas pseudo-convexas, deduzindo algumas das suas propriedades mais relevantes. [4] clarifica esta definição e coloca o enfoque na estimação dos parâmetros. Em particular, analisa estimadores para misturas pseudo-convexas geradas pela distribuição exponencial. Neste artigo, dando continuidade ao trabalho desenvolvido em [4], é analisada a performance de estimadores baseados no método dos momentos e no método da máxima verosimilhança em misturas pseudo-convexas geradas pela distribuição exponencial e pela distribuição função potência.

2 Distribuições estáveis para extremos

Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias (v.a.) absolutamente contínuas, independentes e identicamente distribuídas (i.i.d.) com função de distribuição (f.d.) F e função de sobrevivência (f.s.) $\overline{F}$ (i.e., $\overline{F}(x) := 1 - F(x)$), e seja $X_{i:n}$ a sua i-ésima estatística ordinal. A v.a. X com f.d. F é estável para mínimo ou min-estável (minE) se existirem sequências de constantes $\{\alpha_n \in \mathbb{R}^+\}$ e $\{\beta_n \in \mathbb{R}\}$ tais que se verifica a igualdade em distribuição: $X_{1:n} \overset{d}{=} \alpha_n X + \beta_n$, $\forall n \in \mathbb{N}$, com $X \sim F$. Esta igualdade equivale a que $\overline{F}$ verifique

$$\overline{F}_{X_{1:n}}(x) = \overline{F}^{\,n}(x) = \overline{F}\left(\frac{x - \beta_n}{\alpha_n}\right), \forall x \in \mathbb{R},$$

onde $\overline{F}_{X_{1:n}}$ representa a f.s. de $X_{1:n}$. Assim, o mínimo de n v.a. i.i.d. a $X \sim F$ também é caraterizado pela distribuição F (com eventual alteração de escala e/ou localização) se F é minE. A distribuição geral de valores extremos para mínimos (GEVm_γ), com f.s.

$$\overline{F}_{\mathrm{GEVm}_\gamma}(x) = \begin{cases} \exp\left\{-\left[1-\gamma x\right]^{-1/\gamma}\right\},\ 1+\gamma x > 0 & \text{se} \quad \gamma \neq 0 \\ \exp\left\{-\exp(x)\right\},\ x \in \mathbb{R} & \text{se} \quad \gamma = 0, \end{cases}$$

inclui todas as distribuições minE, considerando $\alpha_n = n^\gamma$ e $\beta_n = \gamma^{-1}\left(1 - n^\gamma\right)$ se $\gamma \neq 0$ ou $\alpha_n = 1$ e $\beta_n = -\ln(n)$ se $\gamma = 0$. A GEVm_γ

contém as distribuições min-Gumbel ($\gamma = 0$), min-Fréchet ($\gamma > 0$) e Weibull ($\gamma < 0$). O parâmetro γ é o índice de valores extremos que mede o peso da cauda-esquerda de F. A distribuição GEVm_γ pode ser generalizada de forma a incorporar parâmetros de localização (μ) e escala (σ) através de $F_{\text{GEVm}_\gamma}(x;\mu,\sigma) = F_{\text{GEVm}_\gamma}((x-\mu)/\sigma)$. Esta distribuição é fundamental na Teoria dos Valores Extremos, uma vez que se o mínimo de n v.a. converge para uma distribuição não degenerada quando $n \to \infty$, então terá de convergir para a distribuição GEVm_γ — Teorema do Valor Extremo (Fisher-Tippett--Gnedenko).

Todos os resultados para o mínimo de uma sequência de v.a. contínuas i.i.d. podem ser adaptados para o máximo uma vez que $Y_{1:n} = -X_{n:n}$ (e $Y_{n:n} = -X_{1:n}$) se $Y = -X$. Deste modo, em vez de ser analisada a cauda-esquerda de F será investigado o peso da cauda-direita de $\overline{F}$. Por conseguinte, a v.a. X com f.d. F é max-estável (maxE) ou estável para máximo se existirem as sequências de constantes reais $\{\alpha_n \in \mathbb{R}^+\}$ e $\{\beta_n \in \mathbb{R}\}$ tais que $X_{n:n} \stackrel{d}{=} \alpha_n X + \beta_n$, $\forall n \in \mathbb{N}$, i.e., a f.d. F verifica

$$F_{X_{n:n}}(x) = F^n(x) = F\left(\frac{x-\beta_n}{\alpha_n}\right), \forall x \in \mathbb{R},$$

onde $F_{X_{n:n}}$ representa a f.d. de $X_{n:n}$. Assim, nas distribuições maxE o máximo de n v.a. i.i.d. a X tem a mesma distribuição (com uma eventual alteração de escala e/ou localização) que X. As únicas distribuições max-estáveis estão contidas na distribuição geral de valores extremos para máximos (GEVM_ξ) com f.d. dada por

$$F_{\text{GEVM}_\xi}(x) = \begin{cases} \exp\left\{-\left[1+\xi x\right]^{-1/\xi}\right\},\ 1+\xi x > 0 & \text{se} \quad \xi \neq 0 \\ \exp\left\{-\exp(-x)\right\},\ x \in \mathbb{R} & \text{se} \quad \xi = 0, \end{cases}$$

onde $F_{\text{GEVM}_\xi}(x) = \overline{F}_{\text{GEVm}_\xi}(-x)$. A GEVM_ξ inclui as distribuições de Gumbel ($\xi = 0$), Fréchet ($\xi > 0$) e max-Weibull ($\xi < 0$) e pode incluir parâmetros de localização e escala ($F_{\text{GEVM}_\xi}(x;\mu,\sigma) = F_{\text{GEVM}_\xi}((x-\mu)/\sigma)$). O parâmetro ξ é o índice de valores extremos que mede o peso da cauda-direita de $\overline{F}$.

2.1 Extensão da definição para permitir alterações do parâmetro de forma

A classe de distribuições estáveis pode ser estendida de forma a incluir alterações do parâmetro de forma (cf. [3] e [4]). Assim, F é uma distribuição min-estável estendida para alterações de forma (minE_f) se existirem as sequências de constantes normalizadoras $\{\alpha_n \in \mathbb{R}^+\}$, $\{\beta_n \in \mathbb{R}\}$ e $\{\gamma_n \in \mathbb{R}\}$ tais que $X_{1:n} \stackrel{d}{=} \alpha_n X + \beta_n$, $\forall n \in \mathbb{N}$, com $X \sim F_{\gamma_n}$, onde F_{γ_n} representa o mesmo tipo de distribuição que F com uma (eventual) alteração no valor do parâmetro de forma (γ_n representa o novo valor). Por conseguinte, é equivalente a

$$F_{X_{1:n}}(x) = 1 - \overline{F}^{\,n}(x) = F_{\gamma_n}\left(\frac{x-\beta_n}{\alpha_n}\right), \forall x \in \mathbb{R}.$$

Além da distribuição GEVm_γ (sem alteração do parâmetro de forma), outros exemplos de distribuições minE_f são a Logística Generalizada (tipo II), GL2(μ,σ,γ), com f.s. dada por

$$\overline{F}\,(x) = \left[\frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)}\right]^{\gamma},\ x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma,\gamma \in \mathbb{R}^+,$$

verificando-se $X_{1:n} \sim \text{GL2}\,(\mu,\sigma,n\gamma)$; e a distribuição Pareto Generalizada, GP(μ,σ,γ), com f.s. dada por

$$\overline{F}\,(x) = \left[1+\frac{x-\mu}{\gamma\sigma}\right]^{-\gamma},\ x > \mu,\ \mu \in \mathbb{R},\ \sigma,\gamma \in \mathbb{R}^+,$$

que verifica $nX_{1:n} + (1-n)\mu \sim \text{GP}\,(\mu,\sigma,n\gamma)$.
De forma análoga, F é uma distribuição max-estável estendida para alterações de forma (maxE_f) se existirem as sequências de constantes normalizadoras $\{\alpha_n \in \mathbb{R}^+\}$, $\{\beta_n \in \mathbb{R}\}$ e $\{\gamma_n \in \mathbb{R}\}$ tais que $X_{n:n} \stackrel{d}{=} \alpha_n X + \beta_n$, com $X \sim F_{\gamma_n}$, $\forall n \in \mathbb{N}$, i.e.,

$$F_{X_{n:n}}(x) = F^{\,n}(x) = F_{\gamma_n}\left(\frac{x-\beta_n}{\alpha_n}\right), \forall x \in \mathbb{R}.$$

Outros exemplos de distribuições maxE_f (além da GEVM_ξ) são a Logística Generalizada (Tipo I) com f.d. $F(x) = \left[1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right]^{-\gamma}$,

$x, \mu \in \mathbb{R}$, $\sigma, \gamma \in \mathbb{R}^+$; e a Função Potência com f.d. $F(x) = x^\gamma$, $x \in (0,1)$ e $\gamma \in \mathbb{R}^+$.
Esta extensão (inclusão de alterações do valor do parâmetro de forma) permite generalizar a classe de distribuições estáveis para extremos, mas não garante as mesmas propriedades. Por outro lado, depende ainda do conceito parâmetro de forma (que não tem, que tenhamos conhecimento, uma definição precisa como acontece com os parâmetros de localização e escala). Porém, esta generalização origina uma família mais vasta de distribuições que permitirá gerar misturas pseudo-convexas.

3 Distribuições pseudo-convexas geradas por distribuições estáveis

As misturas pseudo-convexas geradas por distribuições estáveis para extremos foram definidas em [3], tendo a sua definição sido aprofundada e diversas das suas propriedades deduzidas em [4]. Seja X uma v.a. $\min E_f$ com f.d. F, então a v.a. X_m com f.d. F_{X_m} definida por

$$F_{X_m}(x) = (1+\omega) F(x) - \omega F_{X_{1:2}}(x),\ \omega \in [-1,1],$$

é uma mistura pseudo-convexa (MPC) gerada pela distribuição $\min E_f$ F. Notemos que X_m é uma mistura convexa para $\omega \in [-1,0)$ e não convexa para $\omega \in (0,1]$ de F e $F_{X_{1:2}}$.
O mesmo raciocínio pode ser aplicado para o máximo. Seja X uma v.a. $\max E_f$ com f.d. F, então a v.a. X_M com f.d. F_{X_M} definida por

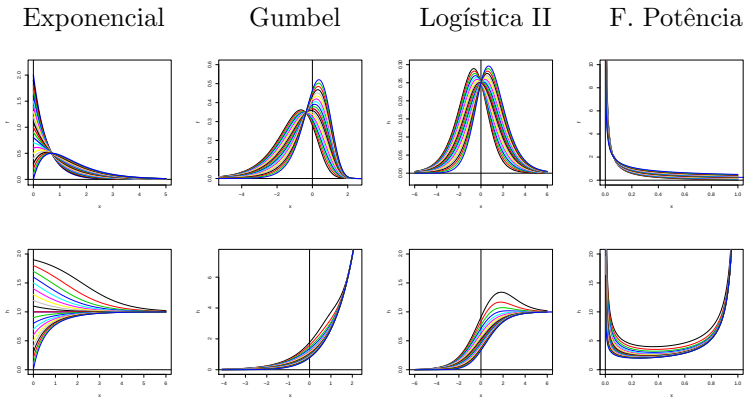
$$F_{X_M}(x) = (1-\omega) F(x) + \omega F_{X_{2:2}}(x),\ \omega \in [-1,1],$$

é uma MPC gerada pela distribuição $\max E_f$ F. Assim, X_M é uma mistura convexa para $\omega \in (0,1]$ e não convexa para $\omega \in [-1,0)$ de F e $F_{X_{2:2}}$. As fórmulas de F_{X_m} e F_{X_M} podem ser reescritas na forma

$$F_{X_m}(x) = F_{X_M}(x) = F(x)\left[1 - \omega \overline{F}(x)\right],\ \omega \in [-1,1],$$

que só depende de $F(x)$ e ω. Deste modo, as MPC têm os mesmos parâmetros que F mais o parâmetro ω da mistura. Os gráficos representados na Figura 1 ilustram a forma da função densidade de MPC

Figura 1: Densidades e *hazard rates* de MPC para $\omega = -1 + 0.1k$, com $k = 0{,}1, \ldots, 20$



para diferentes valores de ω (linha superior) e correspondentes distribuições da taxa de risco (*hazard rates* — $r_X(x) := f_X(x)\, \overline{F}_X^{-1}(x)$) que estão representadas na linha inferior.

Outras propriedades das MPC podem ser deduzidas. Em particular, o k-ésimo momento, com $k \in \mathbb{N}$, é dado por

$$\mathbb{E}\left(X_m^k\right) = \mathbb{E}\left(X^k\right) + \omega\left[\mathbb{E}\left(X^k\right) - \mathbb{E}\left(X_{1:2}^k\right)\right],$$

se $\mathbb{E}\left(X^k\right)$ e $\mathbb{E}\left(X_{1:2}^k\right)$ existirem e

$$\mathbb{E}\left(X_M^k\right) = \mathbb{E}\left(X^k\right) + \omega\left[\mathbb{E}\left(X_{2:2}^k\right) - \mathbb{E}\left(X^k\right)\right],$$

se $\mathbb{E}\left(X^k\right)$ e $\mathbb{E}\left(X_{2:2}^k\right)$ existirem. Simulações destas distribuições podem ser realizadas recorrendo a (cf. [3])

$$\xi_{X_m}(p) = \xi_{X_M}(p) = F^{-1}\left[\frac{\omega - 1 + \sqrt{(1-\omega)^2 + 4\omega p}}{2\omega}\right],$$

onde $\xi_{X_m}(p)$ e $\xi_{X_M}(p)$ representam o p-ésimo quantil, com $p \in (0{,}1)$, de X_m e X_M respetivamente, sendo p caraterizado pela distribuição uniforme padrão. Outras propriedades gerais destas distribuições podem ser consultadas em [4].

3.1 MPC gerada pela distribuição Exponencial

Seja X uma v.a. com distribuição exponencial de parâmetro $\lambda \in \mathbb{R}^+$ e f.d. dada por $F(x) = 1 - e^{-\lambda x}$, $x > 0$. Esta distribuição é minE_f pois $X_{1:n} \sim \text{Exp}(n\lambda)$ (por outro lado, sendo uma distribuição minE, uma vez que é uma Weibull, também será minE_f). A f.d. e a função densidade da MPC gerada pela distribuição Exponencial (MPC_Exp) X_m são dadas por

$$\begin{aligned} F_{X_m}(x) &= 1 - \left[1 + \omega\left(1 - e^{-\lambda x}\right)\right] e^{-\lambda x} \quad \text{e} \\ f_{X_m}(x) &= (1+\omega)\,\lambda e^{-\lambda x} - 2\omega\lambda e^{-2\lambda x}. \end{aligned}$$

Os momentos de ordem k de X_m, com $k \in \mathbb{N}$, são dados por

$$\mathbb{E}\left(X_m^k\right) = \tfrac{k!}{\lambda^k}\left[1 + \omega\left(1 - \tfrac{1}{2^k}\right)\right],$$

e um possível estimador pelo método dos momentos (EMM) será

$$\widetilde{w} = 2\left(\lambda\overline{X} - 1\right) \qquad \text{e} \qquad \widetilde{\lambda} = \tfrac{3\overline{X} + \sqrt{9\overline{X}^2 - 4m_2}}{2m_2},$$

com $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ e $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$. A função de verosimilhança de λ e ω dada a amostra aleatória $X = (X_1, \cdots, X_n)$ é

$$\mathcal{L}(\lambda, w|X) = \prod_{i=1}^n \lambda \exp(-\lambda X_i)\left(1 + w - 2w\exp(-\lambda X_i)\right),$$

e a função log-verosimilhança $\ell(\lambda, \omega|X) = \ln \mathcal{L}(\lambda, w|X)$ será

$$\ell(\lambda, \omega|X) = n\ln(\lambda) - n\lambda\overline{X} + \sum_{i=1}^n \ln\left(1 + w - 2w\exp(-\lambda X_i)\right).$$

Deste modo, as derivadas parciais são

$$\begin{aligned} \tfrac{\partial \ell(\lambda,\omega|X)}{\partial \lambda} &= \tfrac{n}{\lambda} - n\overline{X} + \sum_{i=1}^n \tfrac{2\omega X_i \exp(-\lambda X_i)}{1+\omega-2\omega\exp(-\lambda X_i)}, \\ \tfrac{\partial \ell(\lambda,\omega|X)}{\partial \omega} &= \sum_{i=1}^n \tfrac{1-2\exp(-\lambda X_i)}{1+\omega-2\omega\exp(-\lambda X_i)}, \end{aligned}$$

pelo que, para determinar um valor aproximado do vetor $(\lambda_{\text{EMV}}, \omega_{\text{EMV}})$ que maximiza a função de verosimilhança, teremos de recorrer a métodos numéricos.

3.2 MPC gerada pela distribuição função potência

Seja X uma v.a. com distribuição função potência (FP) com parâmetro $\gamma \in \mathbb{R}^+$ e f.d. dada por $F(x) = x^\gamma$, $x \in (0,1)$ e $\gamma \in \mathbb{R}^+$. Esta distribuição é maxE_f pois $X_{n:n} \sim \text{FP}(n\gamma)$. A f.d. e a função densidade da MPC gerada pela distribuição função potência (MPC_FP) X_M são

$$F_{X_M}(x) = x^\gamma(1-\omega)+\omega x^{2\gamma} \quad \text{e} \quad f_{X_M}(x) = \gamma x^{\gamma-1}(1-\omega+2\omega x^\gamma),$$

sendo os primeiros momentos dados por

$$\mathbb{E}(X_M) = \frac{\gamma(2\gamma+1+\omega)}{(\gamma+1)(2\gamma+1)} \quad \text{e} \quad \mathbb{E}\left(X_M^2\right) = \frac{\gamma(\gamma+1+\omega)}{(\gamma+2)(\gamma+1)}.$$

Consequentemente, possíveis estimadores para γ e ω obtidos pelo método dos momentos (EMM) são

$$\begin{aligned} \widetilde{\gamma} &= \frac{3\overline{X}-3m_2+\sqrt{(3\overline{X}-3m_2)^2-4(m_2-2\overline{X}+1)(2m_2-\overline{X})}}{2(m_2-2\overline{X}+1)}, \\ \widetilde{\omega} &= \frac{\overline{X}(\gamma+1)(2\gamma+1)}{\gamma} - 2\gamma - 1. \end{aligned}$$

Por outro lado, a função de verosimilhança de λ e ω dada a amostra aleatória $X = (X_1, \cdots, X_n)$ é

$$\mathcal{L}(\gamma,w|X) = \gamma^n \prod_{i=1}^{n} x_i^{\gamma-1}\left(1-\omega+2\omega x_i^\gamma\right),$$

sendo a respetiva função log-verosimilhança $\ell(\gamma,\omega|X)$ dada por

$$\ln \mathcal{L}(\gamma,w|X) = n\ln(\gamma) + (\gamma-1)\sum_{i=1}^{n}\ln(x_i) + \sum_{i=1}^{n}\left(1-\omega+2\omega x_i^\gamma\right).$$

Por conseguinte, as derivadas parciais são dadas por

$$\begin{aligned} \frac{\partial \ell(\gamma,\omega|X)}{\partial \gamma} &= \frac{n}{\gamma} + \sum_{i=1}^{n}\ln(x_i) + \sum_{i=1}^{n}\frac{2\omega x_i^\gamma \ln(x_i)}{1-\omega+2\omega x_i^\gamma}, \\ \frac{\partial \ell(\gamma,\omega|X)}{\partial \omega} &= \sum_{i=1}^{n}\frac{-1+2x_i^\gamma}{1-\omega+2\omega x_i^\gamma}. \end{aligned}$$

Deste modo, tal como no caso anterior, teremos de recorrer a métodos numéricos para obter uma aproximação do vetor $(\gamma_{\text{EMV}},\omega_{\text{EMV}})$ que maximiza a função de verosimilhança.

4 Simulações: resultados e comentários

Nesta secção analisamos o desempenho de estimadores paramétricos para MPC através de simulação de Monte Carlo (10^3 réplicas) recorrendo ao *software* R. Para tal, foram geradas misturas pseudo-convexas com base nas duas distribuições referidas na secção anterior (MPC_{Exp} e MPC_{FP}) tendo os seus parâmetros (os associados à distribuição e o $\omega \in [-1,1]$ associado à mistura) sido estimados recorrendo ao método dos momentos bem como a estimadores baseados em métodos iterativos numéricos que visam obter (de forma aproximada) uma estimativa de máxima verosimilhança (EMV). Nesta análise foi aplicado o algoritmo de Newton-Raphson sendo o valor inicial igual à estimativa obtida pelo EMM (como regra geral). Como medidas de comparação do desempenho dos estimadores, foram utilizadas o enviesamento (Viés), o enviesamento absoluto relativo (EAR) e o desvio quadrático médio (DQM). Os estimadores foram analisados sob diversos cenários, nomeadamente utilizando duas dimensões da amostra $n \in \{100,1000\}$ e diferentes valores para os parâmetros bem como para o valor inicial do EMV.

Os resultados patentes na Tabela 1 mostram que, em relação à MPC_{Exp}, os resultados melhoram com o aumento de ω, enquanto que na MPC_{FP} ocorre o oposto (obtendo-se bons resultados quando as misturas são não convexas e resultados menos satisfatórios quando as misturas são convexas). Como era expectável, o aumento da dimensão da amostra melhora a qualidade das estimativas e o EMV obtém (quase sempre) melhor performance que o EMM. Por outro lado, a alteração do valor do parâmetro associado à distribuição (λ ou γ) não parece ter (em termos relativos) grande impacto na qualidade das estimativas obtidas. Por fim, de referir que as estimativas obtidas pelo EMV utilizando como valor inicial $(\lambda_0, \omega_0) = (\overline{x}, 0)$ na

Tabela 1: Estimação do parâmetro associado à distribuição — λ, γ

	EMM							EMV						
ω	$-.75$	$-.50$	$-.25$	.00	.25	.50	.75	$-.75$	$-.50$	$-.25$	.00	.25	.50	.75
	MPC_{Exp}, $\lambda = 1$, $n = 100$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.6159	.3124	.153	.0699	.0418	.0227	.0203	.6147	.3123	.1518	.0677	.0380	.0182	.0078
EAR	.6162	.3316	.2061	.1599	.1379	.1263	.1046	.6149	.3307	.2042	.1578	.1349	.1221	.0916
DQM	.4777	.165	.0708	.0401	.0301	.0259	.0178	.4743	.1635	.0692	.0388	.0286	.0240	.0136
	MPC_{Exp}, $\lambda = 1$, $n = 1000$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.323	.0713	$-.003$	$-.004$	$-.007$	.0092	.0037	.3215	.0717	$-.003$	$-.004$	$-.007$	.0085	.0038
EAR	.323	.1058	.0728	.0514	.0463	.0442	.0303	.3215	.1057	.0729	.0515	.0457	.0428	.0291
DQM	.1177	.0162	.0093	.0044	.0035	.0028	.0014	.1169	.0162	.0093	.0044	.0034	.0027	.0013
	MPC_{Exp}, $\lambda = 10$, $n = 1000$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	3.126	1.023	.0948	-9.00	$-.009$	$-.011$	.0328	3.108	1.025	.0972	-9.00	-012	$-.015$	.0347
EAR	.3126	.1175	.0721	.9004	.0448	.0406	.035	.3108	.1176	.0724	.9004	.0444	.0395	.0338
DQM	11.53	2.054	.833	81.07	.3344	.2577	.1927	11.43	2.053	.8367	81.07	.3283	.2456	.1796
	MPC_{Exp}, $\lambda = 1$, $n = 1000$ e $(\lambda_0, \omega_0) = (\overline{x}, 0)$													
Viés	.3213	.0984	.0149	.0064	.0011	.0008	.0027	.3193	.0982	.014	.0047	-001	$-.002$	$-.001$
EAR	.3213	.1148	.0722	.0572	.0453	.0394	.0353	.3193	.1145	.0717	.0548	.0438	.0371	.0316
DQM	.12	.0201	.0081	.0053	.0033	.0024	.002	.1188	.02	.0079	.005	.0031	.0021	.0016
	MPC_{FP}, $\gamma = 3$, $n = 100$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.0113	.0963	.0176	$-.024$	.1895	.4608	1160	.0193	.0933	.0226	$-.012$	.168	.5518	1133
EAR	.0952	.1087	.1259	.1622	.2067	.2716	.3985	.0919	.1101	.1285	.1574	.2088	.27	.4019
DQM	.1168	.1952	.2355	.3598	.6119	1129	2359	.1106	.1834	.2564	.3393	.6475	1167	2374
	MPC_{FP}, $\gamma = 3$, $n = 1000$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.0058	.0057	.0057	.0188	$-.060$	.1629	.8089	.0061	.0056	.0075	.0226	-089	.0119	.5007
EAR	.0257	.0288	.0369	.0562	.0936	.1137	.2696	.025	.0257	.0362	.0552	.0996	.1313	.1912
DQM	.0102	.0113	.0204	.0417	.1151	.1804	.8578	.0088	.0095	.0195	.0403	.1512	.2164	.5895
	MPC_{FP}, $\gamma = 10$, $n = 1000$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.0005	$-.051$	.0237	.082	$-.092$	$-.042$	2.253	$-.022$	$-.038$	.0082	.0365	-073	$-.034$	1.600
EAR	.0271	.0339	.0391	.0538	.086	.1126	.2257	.0249	.03	.0355	.0533	.0864	.1116	.1817
DQM	.1088	.184	.2351	.4241	1.245	1.704	7.556	.0913	.1416	.2057	.5285	1.304	1.864	6.078
	MPC_{FP}, $\gamma = 3$, $n = 1000$ e $(\gamma_0, \omega_0) = (\overline{x}/(1-\overline{x}), 0)$													
Viés	$-.008$	$-.001$	.0225	$-.015$	$-.018$	.0804	.8509	$-.006$	$-.007$	.0214	$-.012$	-086	.03	.4251
EAR	.0274	.0291	.0403	.0571	.0835	.103	.2836	.0268	.0282	.0388	.0544	.1027	.1043	.195
DQM	.0107	.0118	.0227	.0507	.0954	.1341	.939	.0098	.011	.0216	.0453	.1866	.1493	.6511

MPC_{Exp} e $(\gamma_0, \omega_0) = (\overline{x}/(1-\overline{x}), 0)$ na MPC_{FP} apresentam apenas ligeiras diferenças relativamente à utilização das estimativas obtidas pelo EMM como valor inicial, o que parece mostrar que o EMV é robusto, convergindo para o mesmo máximo local independentemente do valor inicial utilizado.

Em relação à estimação do parâmetro ω da mistura, na Tabela 2 podemos constatar que a performance dos estimadores do parâmetro ω é análoga à dos estimadores do parâmetro associado à distribuição (quando um obtém bons resultados o outro também obtém). Por

Tabela 2: Estimação do parâmetro associado à mistura — ω

	EMM							EMV						
ω	$-.75$	$-.50$	$-.25$	.00	.25	.50	.75	$-.75$	$-.50$	$-.25$	.00	.25	.50	.75
	MPC_{Exp}, $\lambda = 1$, $n = 100$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.7036	.4195	.2358	.1264	.0752	.0284	.0108	.6885	.4035	.2132	.0975	.0366	$-.003$	.0017
EAR	.9382	.8555	1193	–	1011	.4939	.2523	.918	.8231	1112	–	.8671	.4018	.202
DQM	.5838	.2601	.1368	.1103	.098	.092	.0535	.5543	.2391	.1163	.0894	.0731	.0647	.0391
	MPC_{Exp}, $\lambda = 1$, $n = 1000$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.3934	.1121	$-.012$	$-.005$	$-.014$	.0231	.0107	.394	.1111	$-.013$	$-.007$	$-.014$	.0151	.0075
EAR	.5245	.2978	.4865	–	.3587	.1632	.0978	.5253	.295	.4776	–	.3067	.127	.0634
DQM	.1739	.033	.023	.0145	.012	.0104	.0076	.1746	.0322	.0222	.0126	.009	.006	.0033
	MPC_{Exp}, $\lambda = 10$, $n = 1000$ e $(\lambda_0, \omega_0) = (\lambda_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.3814	.1456	.0121	$-.006$	$-.008$	$-.006$	.0062	.3829	.1448	.0081	$-.009$	$-.010$	$-.010$	.0045
EAR	.5085	.3191	.4676	–	.3617	.1701	.1029	.5106	.3172	.4535	–	.3275	.1378	.0671
DQM	.1651	.0371	.0217	.0158	.0136	.0116	.0098	.1668	.0367	.0206	.0144	.0115	.0075	.0041
	MPC_{Exp}, $\lambda = 1$, $n = 1000$ e $(\lambda_0, \omega_0) = (\overline{x}, 0)$													
Viés	.3906	.1451	.0224	.007	.0034	.0023	.0032	.3919	.1439	.02	.0027	.0004	$-.000$	$-.003$
EAR	.5208	.3289	.4785	–	.344	.1627	.1012	.5225	.3263	.4701	–	.3043	.1293	.0623
DQM	.1734	.0402	.0216	.0159	.012	.0103	.0091	.1747	.0396	.0208	.0142	.0095	.0065	.0036
	MPC_{FP}, $\gamma = 3$, $n = 100$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.0185	$-.060$	.0158	.0766	$-.046$	$-.172$	$-.449$	.0109	$-.059$	.0156	.0702	$-.033$	$-.216$	$-.4368$
EAR	.2394	.3497	.9407	–	1199	.6775	.6132	.2097	.3458	.9442	–	1222	.6376	.6012
DQM	.0479	.0492	.0931	.1413	.1213	.1757	.3407	.0445	.0471	.1025	.1404	.1332	.1723	.3382
	MPC_{FP}, $\gamma = 3$, $n = 1000$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	.0037	.0053	.0055	$-.007$	.0337	$-.069$	$-.317$	.0033	.0055	.0051	$-.009$	.0509	.0043	$-.191$
EAR	.0639	.1007	.2787	–	.598	.32	.422	.057	.0945	.2695	–	.6514	.3825	.2928
DQM	.0042	.0043	.0076	.0128	.0331	.0368	.1276	.0031	.0038	.0072	.0123	.0492	.0503	.0845
	MPC_{FP}, $\gamma = 10$, $n = 1000$ e $(\gamma_0, \omega_0) = (\gamma_{\text{EMM}}, \omega_{\text{EMM}})$													
Viés	$-.003$	.0029	$-.001$	$-.002$	.0208	.0035	$-.266$	.0029	$-.001$	.0024	.007	.0181	.0046	$-.184$
EAR	.0788	.124	.2983	–	.5871	.3205	.3549	.0608	.1182	.2712	–	.5902	.3173	.2894
DQM	.0057	.0064	.009	.0152	.0353	.0336	.1021	.003	.0053	.0073	.0197	.0384	.0372	.0841
	MPC_{FP}, $\gamma = 3$, $n = 1000$ e $(\gamma_0, \omega_0) = (\overline{x}/(1-\overline{x}), 0)$													
Viés	.0069	$-.017$	$-.004$	.0166	.0223	$-.035$	$-.351$	.0041	$-.011$	$-.003$	.0145	.0628	$-.011$	$-.140$
EAR	.0659	.1135	.2574	–	.5305	.292	.4682	.0579	.1086	.2491	–	.689	.3045	.2476
DQM	.0044	.0053	.0065	.0146	.0271	.0284	.1603	.0033	.0047	.0061	.0132	.0578	.0344	.0764

outro lado, alterações dos valores dos parâmetros λ e γ não parecem ter influência significativa na estimação de ω.

5 Conclusão

As MPC geradas por distribuições estáveis para extremos (estendida para alterações de forma) dão origem a uma família rica de distribuições, que assume distintas formas, que poderão ser utilizadas para modelar dados reais. Neste artigo foram utilizados estimado-

res paramétricos (EMM e EMV) em MPC geradas pela distribuição exponencial e pela distribuição função potência. Nas simulações realizadas foram obtidos bons resultados quando as misturas são não convexas (principalmente quando se utiliza os EMV), mas resultados insatisfatórios quando as misturas são convexas. Deste modo, com o objetivo de fundamentar a utilidade das MPC na modelação de fenómenos aleatórios, além da extensão a outras MPC geradas por distribuições $\min E_f$ e $\max E_f$, iremos brevemente investigar o desempenho de outras metodologias de estimação de forma a obter estimativas fiáveis quando as misturas são convexas.

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto UID/MAT/00006/2013.

Referências

[1] Beirlant, J., Caeiro, F., Gomes, M.I. (2012). An overview and open research topics in statistics of univariate extremes, *Revstat* 10, 1–31.

[2] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*, Wiley, England.

[3] Felgueiras, M., Martins, J.P., and Santos, R. (2012). Pseudo-convex Mixtures, *Numerical Analysis and Applied Mathematics ICNAAM 2012, AIP Conf. Proc.* 1479, 1125–1128.

[4] Santos, R., Felgueiras, M., and Martins, J.P. (2016). Pseudo-convex Mixtures Generated by Shape-extended Stable Distributions for Extremes, *Journal of Statistical Theory and Practice*.

O Operador *Thinning* na Modelação de Séries Temporais de Valores Inteiros

Manuel G. Scotto
CEMAT e Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, *manuel.scotto@tecnico.ulisboa.pt*

Palavras–chave: séries temporais, valores inteiros, operador *thinning*

Resumo: Este artigo visa proporcionar uma pesquisa abrangente sobre os operadores *thinning* propostos na literatura para modelar séries temporais de valores inteiros. Em seguida serão introduzidos os modelos homólogos discretos do processo autorregressivo convencional de primeira ordem, assim como uma extensão para séries de contagem com suporte finito.

1 Introdução

Uma caraterística comum e predominante em muitas séries temporais que se estudam na prática, é serem constituidas por valores inteiros. Este tipo de dados surge naturalmente associado, por exemplo, a processos de contagem de acontecimentos, objetos ou indivíduos, sendo, portanto, de todo o interesse o estudo de métodos de modelação e análise adequados. Exemplos deste tipo particular de séries temporais podem ser encontrados num vasto leque de áreas de investigação, da biologia e medicina às finanças e economia [21].

Tradicionalmente, as séries temporais de valores inteiros têm vindo a ser analisadas como se o seu suporte fosse o conjunto dos números reais. Nos casos em que as séries representam contagens de valores elevados este procedimento poderá, eventualmente, funcionar pela aplicação do teorema limite central, no entanto, em certas situações, nomeadamente quando as observações apresentam valores reduzidos,

ignorar a natureza dos dados pode conduzir a resultados sem grande significado. Na tentativa de ultrapassar estas (e outras) limitações, nas duas últimas décadas foram propostas várias classes de modelos para descrever e caraterizar adequadamente séries de valores inteiros. A maior parte dos modelos que têm surgido na literatura podem ser classificados em duas classes: a classe de modelos `INARMA` (do inglês *INteger-valued AutoRegressive Moving Average*, ver e.g., [21]) e a classe de modelos `GARCH` de valores inteiros (do inglês *Generalized AutoRegressive Conditional Heteroscedastic*) com distribuição condicional na classe das leis discretas infinitamente divisíveis [5]. Neste artigo a atenção centrar-se-á no estudo de modelos pertencentes à primeira classe. Assim, serão apresentados os modelos `INAR` de primeira ordem baseados no operador *thinning* binomial e alguns modelos para séries de contagem com suporte finito.

2 Operadores *thinning*

As diferentes famílias de modelos que fazem parte da classe de processos `INARMA` partilham o mesmo princípio: construir modelos análogos aos modelos `ARMA`(p,q) convencionais do tipo

$$X_t = \sum_{i=1}^{p} \alpha_i \cdot X_{t-i} + \sum_{j=1}^{q} \beta_j \cdot \epsilon_{t-j} + \epsilon_t,\ t \in \mathbb{Z}, \tag{1}$$

em que α_i $(i = 1, \ldots, p)$ e β_j $(j = 1, \ldots, q)$ são constantes reais e (ϵ_t) é uma sucessão de variáveis aleatórias (v.a's) contínuas, independentes e identicamente distribuídas (i.i.d.), para dados de contagem. É importante salientar que os modelos `ARMA` convencionais não são, em princípio, de grande utilidade na modelação de séries de valores inteiros visto que o simples processo de multiplicação de um escalar real por um valor real ou inteiro conduz à obtenção de um valor real. Uma forma de ultrapassar esta dificuldade é substituir a operação multiplicação em (1) por uma outra operação cujo resultado seja sempre um valor inteiro. Por outro lado, torna-se também necessário

a adoção de uma distribuição discreta para a sucessão (ϵ_t). De entre as diversas operações propostas na literatura destaca-se a família de operações baseadas no operador *thinning*. O conceito de operador *thinning* surge naturalmente em variáveis de contagem, sempre que num conjunto de elementos cada um for selecionado (ou eliminado) com uma certa probabilidade. O operador *thinning* mais popular é a operador *thinning* binomial sugerido por Steutel e Van Harn [20], e definido do modo seguinte:

Definição 2.1 *Seja Z uma variável aleatória discreta com suporte no conjunto $\{0,1,\ldots,n\}$ ou $\mathbb{N}_0$, e $\alpha \in [0;1]$. Define-se o operador thinning binomial entre α e Z como*

$$\alpha \circ Z := \begin{cases} \sum_{j=1}^{Z} \xi_j(\alpha) & Z > 0 \\ 0 & Z = 0 \end{cases},$$

sendo os ξ_j's uma sucessão i.i.d. de v.a's de Bernoulli com parâmetro α ($\xi_j \sim Be(\alpha)$), independentes de Z. A sucessão $\{\xi_j(\alpha) : j \in \mathbb{N}\}$ denomina-se sucessão de contagem.

De acordo com a definição anterior, a variável aleatória $\alpha \circ Z|Z \sim Bi(Z,\alpha)$, isto é, segue uma distribuição binomial com parâmetros Z e α. A interpretação deste operador é a seguinte: seja uma população com Z elementos, sendo a probabilidade de qualquer um dos elementos possuir uma determinada caraterística igual a α. Se os indivíduos dessa população possuem essa caraterística de forma independente uns dos outros, então o número de elementos da população que possui essa caraterística é dado por $\alpha \circ Z$.

As principais propriedades do operador *thinning* binomial são apresentadas na Tabela 1. Como se pode ver, este operador partilha algumas propriedades com a multiplicação usual, nomeadamente a associatividade entre parâmetros *thinning* em termos de igualdade em distribuição e também das propriedades relativas a momentos de primeira ordem. No entanto, a multiplicação usual goza da propriedade distributiva da soma de escalares relativamente à multiplicação

1.	$0 \circ Z = 0$; $1 \circ Z = Z$;
2.	$\alpha_1 \circ Z + \alpha_2 \circ Z \stackrel{d}{\neq} (\alpha_1 + \alpha_2) \circ Z$, α_1,$\alpha_2 \in [0;1]$;
3.	$\alpha_1 \circ (\alpha_2 \circ Z) \stackrel{d}{=} (\alpha_1\alpha_2) \circ Z$, α_1,$\alpha_2 \in [0;1]$;
4.	$E[\alpha \circ Z] = \alpha E[Z]$; $E[(\alpha \circ Z)Y] = \alpha E[ZY]$;
5.	$V[\alpha \circ Z] = \alpha^2 V[Z] + \alpha(1-\alpha)E[Z]$;
6.	$\Phi_{\alpha \circ Z}(s) = \Phi_Z(1-\alpha(1-s))$ com $\Phi_X(s) = E(s^X)$.

Tabela 1: Algumas propriedades do operador *thinning* binomial.

com uma variável aleatória Z (em termos de igualdade em distribuição), propriedade esta que deixa de ser válida quando a multiplicação é substituída pelo operador *thinning* binomial. Note-se também que o operador *thinning* binomial introduz um termo acrescido na variância, dado por $\alpha(1-\alpha)E[Z]$. Este termo corresponde à variância de uma variável aleatória $Bi(E[Z],\alpha)$. É também importante salientar que, em geral, os momentos de ordem superior a um que envolvem o operador *thinning* binomial, não são iguais aos respetivos momentos quando se usa a multiplicação usual em vez do referido operador. Por outro lado, uma questão que habitualmente se coloca em relação à distribuição de $\alpha \circ Z$, é saber em que casos as distribuições de $\alpha \circ Z$ e Z pertencem à mesma família. Puig e Valero [15] mostraram que a condição necessária e suficiente para isto acontecer é que $\Phi_Z(s) = g(\mu_Z(s-1))$, sendo $g(\cdot)$ uma função real analítica e $\mu_Z = E[Z]$.

Várias modificações do operador *thinning* binomial têm sido propostas nos últimos anos para torná-los mais flexíveis na modelação de sucessões de contagem. Latour [10] introduziu o operador *thinning* generalizado cuja definição é idêntica à definição do operador *thinning* binomial, mas com a diferença das variáveis ξ_j's não serem necessariamente do tipo 0-1. Um caso particular do operador

thinning generalizado é o operador *thinning* estendido proposto por Zhu e Joe [24] em que os ξ_j's formam uma sucessão i.i.d. de v.a's com a mesma distribuição que uma variável aleatória ξ com função geradora de probabilidade

$$\Phi_\xi(s) = \frac{(1-\alpha)+(\alpha-\gamma)s}{(1-\alpha\gamma)-(1-\alpha)\gamma s},\ \gamma \in (0;1],$$

de média $E[\xi] = \alpha$ e variância $V[\xi] = \alpha(1-\alpha)(1+\gamma)/(1-\gamma)$. Através deste operador, os autores introduziram o conceito de distribuição autodecomponível para inteiros generalizada. O operador *thinning* binomial obtém-se fixando $\gamma = 0$. Mais recentemente, Zhu e Joe [25] propuseram um novo operador *thinning*, chamado operador *thinning* esperado, que inclui como casos particulares os operadores *thinning* biniomial, generalizado e estendido. A definição deste operador é a seguinte:

Definição 2.2 *Seja* $\{\xi(\alpha) : 0 \leq \alpha \leq 1\}$ *uma família de v.a's autogeneralizadas*[5] *com suporte no conjunto* $\mathbb{N}_0$ *e* $E[\xi(\alpha)] < \infty$. *Define-se o operador thinning esperado entre* α *e* Z *como*

$$\alpha \otimes Z := \begin{cases} \sum_{j=1}^{Z} \xi_j(\alpha) & Z > 0 \\ 0 & Z = 0 \end{cases},$$

sendo os ξ_j*'s uma sucessão i.i.d. de v.a's com a mesma distribuição que* $\xi(\alpha)$, *independentes de* Z, *com* $E[\xi(\alpha)] < 1$ *para todo* $\alpha \in (0;1)$.

É de salientar que em todos os operadores apresentados anteriormente assume-se que os ξ_j's são independentes. No entanto, em

[5] Uma variável aleatória $Y(\alpha)$ diz-se autogeneralizada, em relação ao parâmetro α, se $\Phi_{Y(\alpha)}(\Phi_{Y(\alpha)}(s;\alpha);\alpha') = \Phi_{Y(\alpha)}(s;\alpha\alpha')$, para todo α,$\alpha' \in [0,1]$. É importante salientar que a variável $Y(\alpha)$ satisfaz a propriedade $Y(\alpha) \otimes Y(\alpha') \stackrel{d}{=} Y(\alpha\alpha')$, para $0 \leq \alpha$,$\alpha' \leq 1$ (*closure property*), sendo " $\otimes$ " o operador *thinning* esperado introduzido na definição 2.2.

muitas situações práticas tal imposição é muito restritiva. Para ultrapassar esta limitação, Ristić et al. [16] propuseram a seguinte representação para as variáveis da sucessão de contagem:

$$\xi_i = (1 - V_i)W_i + V_i Y, \quad i \in \mathbb{N}, \tag{2}$$

sendo (W_i) e (V_i) sucessões i.i.d. de v.a's de Bernoulli de parâmetros $\phi \in [0;1]$ e $\theta \in [0;1]$ independentes entre si, e independentes da variável aleatória $Y \sim Be(\phi)$. A representação (2) implica que (ξ_j) é uma sucessão de v.a's de Bernoulli dependentes com parâmetro $\phi \in [0;1]$ e $Corr(\xi_i,\xi_j) = \theta^2 \neq 0$ para $\theta \neq 0$ e $i \neq j$. O caso $\theta = 0$ corresponde ao operador *thinning* binomial.

Uma outra generalização foi proposta por Gomes e Canto e Castro [6] e Zheng et al. [23]. Estes autores consideraram o caso em que α é também uma variável aleatória com suporte no conjunto $[0;1)$.

Uma das limitações do operador *thinning* binomial e das suas várias modificações é o facto de poderem ser utilizados, unicamente, na modelação de séries de contagem de valores não negativos. No caso de ter que lidar com séries de contagem que apresentem valores inteiros negativos, Kim e Park [9] propuseram a seguinte extensão do operador *thinning* binomial.

Definição 2.3 *Seja Z uma variável aleatória discreta com suporte no conjunto $\mathbb{Z}$ e $|\alpha| \in [0;1]$. Define-se o operador thinning binomial sinalizado entre α e Z como*

$$\alpha \odot Z := sgn(\alpha) \cdot sgn(Z) \cdot \sum_{j=1}^{|Z|} \xi_j(\alpha),$$

com sgn(x)=1 se $x \geq 0$ e -1 se $x < 0$, e os ξ_j's uma sucessão i.i.d. de v.a's de Bernoulli com parâmetro $|\alpha|$.

Generalizações deste estimador têm sido sugeridas por Alzaid and Omair [2], Kachour e Truquet [7] e Zhang et al. [22]. Propriedades

destes (e outros) operadores podem ser consultadas em Scotto et al. [18].

Embora no caso univariado já exista, como se viu, um vasto leque de operadores *thinning*, a literatura sobre extensões para o caso bivariado e multivariado é escassa. A generalização do operador *thinning* binomial para o caso bivariado foi proposta por Scotto et al. [19]. Estes autores introduziram o operador *thinning* binomial bivariado cuja definição é a seguinte:

Definição 2.4 *Seja* $\boldsymbol{X} = [X_1, X_2]'$ *um vetor aleatório e* $\boldsymbol{\alpha}$ *o vetor de parâmetros* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \phi_\alpha)$ *with* $0 < \alpha_1, \alpha_2 < 1$, *e* $|\phi_\alpha| \leq 1$. *Define-se o operador thinning binomial bivariado entre* $\boldsymbol{X}$ *e* $\boldsymbol{\alpha}$ *como*

$$\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X} \sim \ BVB_{\mathrm{II}}\left(X_1, X_2, \min\{X_1, X_2\}; \alpha_1, \alpha_2, \phi_\alpha\right),$$

isto é, $\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X}$ *segue uma distribuição binomial bivariada de tipo II.*

Este operador apresenta um conjunto de caraterísticas importantes, nomeadamente o facto de as distribuições condicionais marginais serem binomiais, e a dependência entre as duas componentes de $\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X}$ poder ser positiva ($\phi_\alpha > 0$) ou negativa ($\phi_\alpha < 0$). As propriedades de este operador foram analisadas por Scotto et al. [19]. Outros operadores para o caso bivariado e multivariado podem ser consultados em Karlis [8] e Scotto et al. [18].

3 Modelos para séries de contagem com suporte infinito

McKenzie [11] e Al-Osh e Alzaid [1] introduziram, independentemente, o modelo autorregressivo de primeira ordem para valores inteiros `INAR`(1), substituindo a operação multiplicação em (1) pelo operador *thinning* binomial. O modelo `INAR`(1) é definido pela equação recursiva

$$X_t = \alpha \circ X_{t-1} + \epsilon_t,\ t \in \mathbb{Z}, \tag{3}$$

em que $\alpha \in [0;1)$,[6] sendo (ϵ_t) uma sucessão de v.a's i.i.d. de valores inteiros não negativos com $E[\epsilon_t] < \infty$ e $V[\epsilon_t] < \infty$, onde para cada instante $t \in \mathbb{Z}$, ϵ_t é independente de X_{t-1} e de $\alpha \circ X_{t-1}$.

O processo INAR(1) partilha várias propriedades com o modelo AR(1), nomeadamente o facto da função de autocorrelação (FAC) tender exponencialmente para zero. No entanto para o processo INAR(1) os valores da FAC são sempre positivos. Uma outra caraterística importante do processo INAR(1) é que qualquer distribuição autodecomponível[7] para inteiros (DSD[8]) pode ser usada como distribuição marginal para X_t. Fazem parte desta classe, por exemplo, a distribuição de Poisson e a distribuição binomial negativa.

O modelo INAR(1) tem sido generalizado de várias formas, nomeadamente considerando que o parâmetro α varia ao longo do tempo de forma (a) determinística ou; (b) estocástica. Fazem parte da primeira categoria o modelo INAR(1) proposto por Monteiro et al. [13] em que o $\alpha = \alpha_t$ varia periodicamente ao longo do tempo, o modelo SETINAR (do inglês *Self-Exiciting Threshold INteger AutoRegressive*) proposto por Monteiro et al. [14] em que

$$\alpha = \alpha_1 I(X_{t-1} \leq r) + \alpha_2 I(X_{t-1} > r),\ r \in \mathbb{N},$$

e o modelo introduzido por Brännäs [3] quem propôs um modelo INAR(1) em que o parâmetro α varia ao longo do tempo através de covariáveis fixas. Em particular, este autor adotou a seguinte especificação para $\alpha \equiv \alpha_t = 1/[1 + \exp\{\boldsymbol{y}_t\boldsymbol{\omega}\}]$, sendo $\boldsymbol{y}_t$ o vetor

[6] Neste caso, o modelo INAR(1) em (3) diz-se ser estável. O caso $\alpha = 1$ é usualmente referido como instável.

[7] Uma função distribuição em $\mathbb{N}_0$ com função geradora de probabilidades P diz-se autodecomponível para inteiros se $P(s) = P(1 - \alpha + \alpha s)P_\alpha(s)$, sendo $|s| < 1$ e $\alpha \in (0;1)$, em que P_α é uma função geradora de probabilidades. Em termos de v.a's isto significa que X é autodecomponível se $X \stackrel{d}{=} \alpha \circ X + X_\alpha$, em que as variáveis $\alpha \circ X$ e X_α são independentes.

[8] Do inglês *Discrete Self-Decomposable*. A família de distribuições DSD é uma subclasse da classe de distribuições infinitamente divisíveis discretas.

de covariáveis fixas e $\boldsymbol{\omega}$ o correspondente vetor de parâmetros desconhecidos. Finalmente, modelos com coeficientes a variar ao longo do tempo de forma aleatória têm sido propostos por Roitershtein e Zhong [17] e Gomes e Canto e Castro [6].

4 Modelos para séries de contagem com suporte finito

Quando são analisadas séries de contagem em que o suporte é finito os modelos apresentados na secção anterior não são de grande utilidade. Uma forma de ultrapassar esta situação é considerar como distribuição marginal do processo a distribuição binomial[9]. Neste caso, [11] propôs a seguinte representação para X_t:

$$X_t = \alpha \circ X_{t-1} + \beta \circ (N - X_{t-1}),\ t \in \mathbb{Z}, \tag{4}$$

em que os operadores *thinning* são independentes e, para cada t fixo, independentes de $(X_s)_{s<t}$, sendo $\beta := \pi(1-\nu)$, $\alpha := \beta + \nu$, e

$$\nu \in [\max\{-\pi/(1-\pi), -(1-\pi)/\pi\}, 1]. \tag{5}$$

Se $X_0 \sim \text{Bi}(N,\pi)$, então a equação (4) gera um processo `AR`(1) binomial estacionário com distribuição marginal $\text{Bi}(N,\pi)$. A condição (5) garante que $\alpha, \beta \in [0;1]$. A interpretação deste modelo é a seguinte: assumindo que no instante $t-1$ existem no sistema N elementos, cada um deles no estado 0 ou 1, X_{t-1} representa o número de elementos cujo estado é 1, pelo que $\alpha \circ X_{t-1}$ representa o número de unidades no estado 1 no instante t. O termo $\beta \circ (N - X_{t-1})$ representa o número de elementos que transitam do estado 0 para o estado 1 no instante t. Os elementos transitam de estado independentemente uns dos outros com probabilidade β.

[9] A distribuição binomial não pertence à classe de distribuições DSD pelo que não é adequada como distribuição marginal para o modelo `INAR` em (3).

Este tipo de modelos foram usado por Brännäs e Nordström [4] para analisar a taxa de ocupação hoteleira. Ou autores consideram o caso $N = N_t$ em que N_t representa o número total de quartos (isto é, a soma dos disponíveis e dos ocupados) num dado hotel, no dia t. Assim, X_t representa o número de quartos ocupados no dia t. O termo $\alpha \circ X_{t-1}$ representa o número de quartos que permanecem ocupados entre os dias $t-1$ e t. Isto implica que no dia t o hotel tem $N_t - \alpha \circ X_{t-1}$ quartos disponíveis, pelo que $\beta \circ (N_t - X_{t-1})$ representa o número de quartos vagos que são ocupados no dia t.

Uma extensão do modelo em (4) foi recentemente proposta por Möller et al. [12] na qual os autores introduziram um modelo AR(1) binomial por limiares, sendo X_t representado da forma seguinte:

$$X_t \;=\; \phi_t \circ X_{t-1} \;+\; \eta_t \circ (N - X_{t-1}),\, t \in \mathbb{Z},$$

em que

$$\phi_t := \alpha_1 I(X_{t-1} \leq r) + \alpha_2 I(X_{t-1} > r)$$

e

$$\eta_t := \beta_1 I(X_{t-1} \leq r) + \beta_2 I(X_{t-1} > r),$$

com $r \in \mathbb{N}$.

Agradecimentos

O autor quer agradecer às colegas Conceição Ribeiro e Clara Cordeiro pelo convite para apresentar este trabalho no congresso.

Referências

[1] Al-Osh, M.A., Alzaid, A.A. (1987). First order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis* 8, 261–275.

[2] Alzaid, A.A., Omair, M.A. (2014). Poisson difference integer valued autoregressive model of order one. *Bulletin of the Malaysian Mathematical Sciences Society* 37, 465–485.

[3] Brännäs, K. (1995). Explanatory variables in the AR(1) count data model. *Umeå Economic Studies* 381.

[4] Brännäs, K., Nordström, J. (2006). Tourist accommodation effects of festivals. *Tourism Economics* 12, 291–302.

[5] Gonçalves, E., Mendes-Lopes, N., Silva, F. (2015). Infinitely divisible distributions in integer-valued GARCH models. *Journal of Time Series Analysis* 36, 503–527.

[6] Gomes, D., Canto e Castro, L. (2009). Generalized integer-valued random coefficient for a first order structure autoregressive (RCINAR) process. *Journal of Statistical Planning and Inference* 139, 4088–4097.

[7] Kachour, M., Truquet, L. (2011). A p-order signed integer-valued autoregressive (SINAR(p)) model. *Journal of Time Series Analysis* 32, 223–236.

[8] Karlis, D. (2015). Models for multivariate count time series. In Davis, R.A., Holan, S.H., Lund, R., Ravishanker, N. (eds.): *Handbook of Discrete-Valued Time Series*. Chapman and Hall/CRC, 404–424.

[9] Kim, H.Y., Park, Y. (2008). A non-stationary integer-valued autoregressive model. *Statistical Papers* 49, 485–502.

[10] Latour, A. (1998). Existence and stochastic structure of a non-negative integer-valued autoregressive processes. *Journal of Time Series Analysis* 4, 439–455.

[11] McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin* 21, 645–650.

[12] Möller, T.A., Silva, M.E., Weiß, C.H., Scotto, M.G., Pereira, I. (2015). Self-exciting threshold binomial autoregressive processes. *AStA Advances in Statistical Analysis* (no prelo).

[13] Monteiro, M. Scotto, M.G., Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference* 140, 1529–1541.

[14] Monteiro, M., Scotto, M.G., Pereira, I. (2012). Integer-valued self-exciting threshold autoregressive processes. *Communications in Statistics - Theory and Methods* 41, 2717–2737.

[15] Puig, P., Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli* 13, 544–555.

[16] Ristić, M.M., Nastić, A.S., Miletić Ilić, A.V. (2013). A geometric time series model with dependent Bernoulli counting series. *Journal of Time Series Analysis* 34, 466–476.

[17] Roitershtein, A., Zhong, Z. (2013). On random coefficient INAR(1) processes. *Science China Mathematics* 56, 177–200.

[18] Scotto, M.G., Weiß, C.H., Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling* 15, 590-618.

[19] Scotto, M.G., Weiß, C.H., Silva, M.E., Pereira, I. (2014). Bivariate binomial autoregressive models. *Journal of Multivariate Analysis* 125, 233–251.

[20] Steutel, F.W., van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability* 7, 893–899.

[21] Turkman, K.F., Scotto, M.G., de Zea Bermudez, P. (2014). *Non-Linear Time Series: Extreme Events and Integer Value Problems.* Springer & Verlag, Switzerland.

[22] Zhang, H., Wang, D., Zhu, F. (2010). Inference for INAR(p) processes with signed generalized power series thinning operator. *Journal of Statistical Planning and Inference* 140, 667–683.

[23] Zheng, H., Basawa, I.V., Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference* 173, 212–229.

[24] Zhu, R., Joe, H. (2003). A new type of discrete self-decomposability and its applications to continuous-time Markov processes for modeling count data time series. *Stochastic Models* 19, 235–254.

[25] Zhu, R., Joe, H. (2010). Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference* 140, 1874–1888.

Pseudo-valores-p e meta análise

Paulo Semblano
CEAUL, Universidade de Lisboa e CGD, *paulo.semblano@cgd.pt*

M. Fátima Brilhante
CEAUL, Universidade de Lisboa e Departamento de Matemática da Universidade dos Açores, *maria.fa.brilhante@uac.pt*

Dinis Pestana
Universidade de Lisboa, CEAUL e Instituto de Investigação Científica Bento da Rocha Cabral, *dinis.pestana@fc.ul.pt*

Fernando Sequeira
CEAUL e DEIO, Universidade de Lisboa, *fjsequeira@fc.ul.pt*

Palavras–chave: meta análise; valores de prova-p problemáticos; controvérsia de Mendel-Fisher; combinações pseudo-convexas.

Resumo: No contexto de combinação de valores de prova-p em meta análise, na perspetiva clássica, sob validade da hipótese nula H_0, o valor de prova-p é considerado uma observação de uma variável aleatória uniforme padrão. No entanto, mesmo sob validade de H_0, nem sempre será possível assumir que os valores de prova-p são observações de uma uniforme, se aceitarmos que há a possibilidade de alguns investigadores repetirem experiências, quando o resultado da primeira não se coaduna com as suas expectativas ou, por exemplo, por desconfiarem dos resultados obtidos. Neste caso propõe-se que seja utilizado como modelo uma mistura convexa entre uma variável aleatória uniforme padrão, o máximo e o mínimo de r variáveis aleatórias independentes e identicamente distribuídas uniformes padrão. Descreveremos este modelo apresentando algumas particularidades e daremos especial atenção à estimação do parâmetro de mistura quando $r = 2$.

1 Introdução

Seja p_j o valor de prova-p decorrente do teste T_j : H_{0j} *vs.* H_{1j}, onde os T_j, $j = 1,2,...,n$ são testes independentes, assumindo-se em geral que são trabalho de equipas diferentes sobre um mesmo problema, que interessa harmonizar. A questão foi inicialmente abordada por Tippett [4] — que considerou que a hipótese nula composta $\boldsymbol{H_0}$: H_{0j} verdadeira, $j = 1,...,n$ (*versus* $\boldsymbol{H_1}$: $\exists j \in \{1,...,n\}$ para o qual H_{1j} verdadeira) deve ser rejeitada ao nível α se $p_{1:n} = \min_{1\leq j\leq n} p_j < 1-(1-\alpha)^{1/n}$ —, e por Fisher [2], que propôs como critério de rejeição $-2\sum_{j=1}^{n} \ln p_j > \chi^2_{2n,1-\alpha}$.
Por outras palavras, quer Tippett quer Fisher assumiram que os valores de prova-p são observações de uniformes padrão independentes. Tsui e Weerahandi [5] criticaram esse pressuposto, dando início à investigação de *valores de prova-p generalizados*, observações de variáveis aleatórias cuja distribuição não é necessariamente uniforme, porque existirão valores de j para os quais H_{1j} é verdadeira.
Admitir que existem j tais que H_{1j} é verdadeira não é a única razão para abordar a combinação de valores de prova-p de forma diversa das soluções tradicionais. No interessante trabalho de Pires e Branco [3] sobre a controvérsia Mendel–Fisher é analisada a possibilidade de Mendel, ou algum dos seus colaboradores, quando insatisfeito com um resultado experimental, replicar a experiência e reportar o que considerou ser o resultado mais fiável — leia-se "mais consentâneo com os seus interesses".
Neste caso sob validade de H_0 seria reportado um *pseudo-valor-p* que seria o máximo (respetivamente o mínimo) de valores de prova-p uniformes padrão, e que sob validade de H_1 seria o máximo (respetivamente o mínimo) de valores de prova-p generalizados, não uniformes. Brilhante *et al.* [1] investigaram o modelo

$$X_m = \begin{cases} U & U_{\lfloor 2+\frac{|m|}{2^m}\rfloor:2} \\ 1-\frac{|m|}{2} & \frac{|m|}{2} \end{cases}, \; m \in [-2,2] \qquad (1)$$

onde $\lfloor x \rfloor$ denota o maior inteiro não superior a x, isto é de misturas

convexas de uniforme com o máximo ($m \in (0,2]$), ou com o mínimo ($m \in [-2,0)$) de duas uniformes independentes, que corresponde a haver uma proporção $\frac{|m|}{2}$, $m \in [-2,2]$, de extremos de uniforme na sequência de valores de prova-p reportados, evidenciando as dificuldades de estimação do parâmetro de mistura (a proporção ingénua ou intencional de pseudo-valores-p), ao usar o modelo (1).
Neste trabalho analisamos questões de não-identificabilidade mais complexas quando se admite que o reporte de valores pelos diversos investigadores é uma mistura de erros diversos, no sentido em que uns reportam o máximo e outros reportam o mínimo de um certo número, r, de replicações da experiência,

$$X_{r;p,q} = \left\{ \begin{array}{ccc} U & U_{r:r} & U_{1:r} \\ 1-p-q & p & q \end{array} \right. , \; 0 \leq \min\{p,q,1-p-q\}. \quad (2)$$

2 A família $X_{r;p,q}$ de misturas de Uniforme com Beta(r,1) ou Beta(1,r)

Repare-se que a família $X_{r;p,q}$ (2) resulta da mistura de uma variável aleatória uniforme padrão com uma variável aleatória Beta(r,1) e uma Beta(1,r), assumindo que é reportado o máximo de r réplicas da experiência com probabilidade $0 \leq p \leq 1$, o mínimo com probabilidade $0 \leq q \leq 1$ e que $0 \leq p+q \leq 1$.
É curioso verificar que se pode reescrever $X_{r;p,q}$ como uma mistura pseudo-convexa de máximos $\{U_{1:1} \equiv U,\, U_{2:2},\, U_{3:3} \ldots,\, U_{r:r}\}$, com $0 \leq \min\{p,q\},\; p+q \leq 1$. Note-se que,

$$\begin{aligned} F_{X_{r;p,q}}(x) &= (1-p-q)F_U(x) + pF_{U_{r:r}}(x) + qF_{U_{1:r}}(x) \\ &= [1-p+q(r-1)]F_U(x) + \sum_{j=2}^{r-1} \binom{r}{j} q(-1)^{j+1} F_{U_{j:j}}(x) \\ &\quad + [p-(-1)^r q]F_{U_{r:r}}(x). \end{aligned}$$

De forma análoga, $X_{r;p,q}$ pode também ser representada como uma mistura pseudo-convexa de mínimos.

2.1 Caso particular de $r = 2$

Quando $r = 2$, utilizando o resultado anterior, a função de distribuição poderá ser escrita como uma mistura pseudo-convexa de máximos, da seguinte forma:

$$F_{X_{2;p,q}}(x) = (1 - p + q)F_U(x) + (p - q)F_{U_{2:2}}(x). \tag{3}$$

O peso q associado ao mínimo é incorporado no peso do máximo, ocorrendo os necessários ajustamentos no peso de U. Neste caso, o parâmetro que interessará estudar será $k = p - q$.
Se $p = 0$, ou $q = 0$, obtém-se uma variável aleatória da família X_m (1). Por exemplo, se $q = 0$, resultará $m = 2p$ (com $p > 0$) e a função de distribuição será dada por:

$$F_{X_{2;p,0}}(x) = (1 - p)F_U(x) + pF_{U_{2:2}}(x).$$

Os problemas de identificabilidade tornam-se ainda mais evidentes no caso de $r = 2$, $p = q$ — um modelo natural quando se pensa num esquema similar: o investigador ou obtém um único valor-p que reporta, ou replica a experiência e reporta o segundo valor que observa, que então com igual probabilidade será um máximo ou um mínimo. Neste caso, $X_{2,p,p} \overset{\mathrm{d}}{=} U$,

$$F_{X_{2;p,p}}(x) \overset{\mathrm{d}}{=} F_U(x).$$

Assim, se $p = q$ a distribuição de $X_{2;p,p}$ será uniforme, qualquer que seja o valor de $p \in [0, \frac{1}{2}]$. Consequentemente, se com igual probabilidade os experimentadores repetirem a experiência e reportarem o máximo ou o mínimo, dessas duas experiências, manter-se-á válida a asserção inicial de que os valores reportados serão provenientes de uma distribuição uniforme.
Repare-se que $X_{2;p,p} \overset{\mathrm{d}}{=} U$ poderá ser visto como um caso particular de uma mistura de uma variável aleatória W com função de distribuição contínua com o máximo $W_{2:2}$ de duas réplicas de W com probabilidade p e o mínimo $W_{1:2}$ de duas réplicas de W com a mesma probabilidade p ($0 \leq p \leq 0.5$). Ou seja,

$$W_{2;p,p} = \left\{ \begin{array}{ccc} W & W_{2:2} & W_{1:2} \\ 1-2p & p & p \end{array} \right. \stackrel{\mathrm{d}}{=} W.$$

3 Estimação do parâmetro k

No caso em que $r = 2$, $X_{2;p,q} = X_{2;k}$ é uma variável aleatória da família X_m e, consequentemente, a estimação do parâmetro k, que incorpora a informação da diferença entre p e q (com $k = p - q$ e $-1 \leq k \leq 1$), herda todas as dificuldades da estimação de m na referida família. Assim, a função de distribuição (3) em função do parâmetro k será,

$$F_{X_{2;k}}(x) = (1-k)F_U(x) + kF_{U_{2:2}}(x). \tag{4}$$

Brilhante *et al.* [1] investigaram diversos métodos para a estimação do parâmetro m, aqui optamos por estimar o parâmetro k efetuando em simultâneo um teste de ajustamento do modelo e a estimação do parâmetro, "invertendo" os testes de ajustamento de forma a identificar a região do parâmetro onde os dados apresentam maior "concordância" com o modelo ajustado, obtendo o que por vezes se designa por intervalo de confiança de *score* .

Assim, para obter as estimativas de k utiliza-se o teste de ajustamento de Kolmogorov-Smirnov (K-S), supondo que $F^*_{n;k}$ é a função de distribuição empírica de uma amostra proveniente do modelo $F_{X_{2;k}}$, com k desconhecido e n a dimensão da amostra. Uma estimativa para k pode ser obtida identificando o valor que minimiza a distância de K-S: o que é equivalente a maximizar o valor de prova-p associado ao teste de ajustamento de K-S.

Procede-se de forma idêntica utilizando o teste de ajustamento de Anderson-Darling (A-D), um teste que habitualmente apresenta melhores resultados do que o de K-S quando a distribuição em causa tem "caudas" mais pesadas (não sendo o caso aqui).

3.1 Esquema de simulação

Com o intuito de analisar os resultados das estimativas para k, obtidas através dos testes de ajustamento de K-S e de A-D, fez-se um estudo de simulação para valores de $k = -1(0.05)1$ e amostras de dimensão $n = 10$, 25, 50 e 100.
Para cada valor de k e para cada valor de n, o procedimento de implementação da simulação pode ser resumido em cinco etapas:

1. gerar uma amostra de n números pseudo-aleatórios com distribuição uniforme padrão, $\boldsymbol{u} = u_1, u_2, \cdots, u_n$;

2. obter a amostra $\boldsymbol{x} = x_1, x_2, \cdots, x_n$ com a distribuição de mistura $F_{X_{2;k}}$, através do método da transformação inversa: quando $k \neq 0$, $x = \frac{k-1+\sqrt{1-2k+k^2+4ku}}{2k}$, e $x = u$, quando $k = 0$;

3. determinar os valores k que maximizam os valores de prova-p associados aos dois testes de ajustamento;

4. repetir 20 000 vezes os passos 1, 2 e 3;

5. para todas as estimativas obtidas determinar os intervalos de confiança de *score* e calcular também as estatísticas de interesse, por exemplo a média, o viés e o erro quadrático médio.

Procedendo deste modo os valores de prova-p são gerados aleatoriamente através do modelo (4) recorrendo ao Teorema da Transformação Uniformizante. Desta forma, o verdadeiro valor do parâmetro é conhecido e está fixo para cada caso de simulação. As amostras geradas em cada passo da simulação são utilizadas para efetuar os testes de ajustamento de K-S e de A-D para os diversos valores admissíveis do parâmetro. Sendo assim possível identificar o valor que maximiza o valor de prova-p associado a cada um dos testes de ajustamento e determinar os intervalos de confiança de *score*, comparando os resultados obtidos através dos dois testes de ajustamento estudados.

3.2 Resultados da estimação de k

Repare-se que as estimativas obtidas através do teste de ajustamento de A-D são melhores do que as fornecidas pelo teste de ajustamento de K-S, sendo a diferença entre os dois métodos mais evidente nas amostras de menor dimensão.

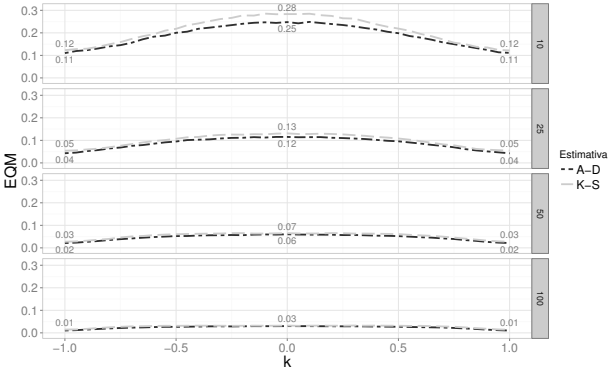


Figura 1: Erro quadrático médio das estimativas para k obtidas através dos testes de ajustamento de A-D e de K-S, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

Através da análise da Figura 1 pode confirmar-se que os valores do erro quadrático médio das estimativas obtidas pelo teste de ajustamento de A-D são sempre mais baixos do que o das estimativas obtidas através do teste de ajustamento de K-S. Note-se que os valores mais altos são observados quando $k = 0$ e vão diminuindo à medida que $|k|$ se aproxima de 1, isto é, quando a componente de mistura tem um sinal mais forte (componente dominante Beta não uniforme).

Observe-se o viés das estimativas de k na Figura 2 e repare-se na tendência de diminuição do enviesamento à medida que a dimensão da amostra aumenta, denotando-se um aumento do enviesamento quando a componente de mistura tem um sinal mais forte ($|k|$ está

próximo de 1).

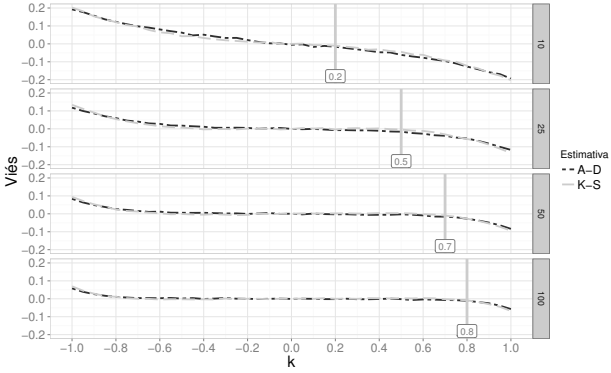


Figura 2: Viés das estimativas para k obtidas através dos testes de ajustamento de A-D e de K-S, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

Nas Figuras 3 e 4 estão representados os valores médios das estimativas obtidas utilizando os testes de ajustamento de A-D e de K-S, podendo uma vez mais verificar-se que os valores obtidos estão muito próximos do verdadeiro valor de k, denotando-se valores de viés mais elevados à medida que diminui a dimensão da amostra. Um olhar mais atento permite confirmar que o valor médio das estimativas obtidas com o teste de ajustamento de A-D se encontra, de um modo geral, mais próximo do verdadeiro valor de k.

Os resultados até agora analisados parecem ser interessantes. No entanto, observe-se a amplitude média dos intervalos de confiança de *score* com um nível de significância de 95% na Figura 5: a amplitude média dos intervalos de confiança de *score* é sempre superior a 1 para amostras de dimensão 10, chegando a ser 1.58 e 1.66 quando $k = 0$; Quando $k = 0$, o valor mais baixo (0.67) é observado em amostras de dimensão 100, sendo todos os outros valores próximos ou superiores a 1; valores de amplitude média dos intervalos de confiança inferiores a

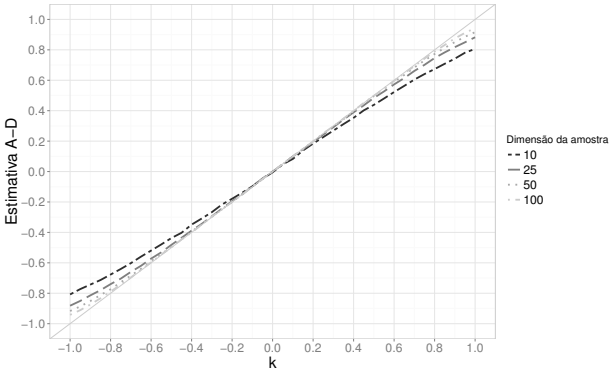


Figura 3: Valor médio das estimativas para k obtidas através do teste de ajustamento de A-D, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

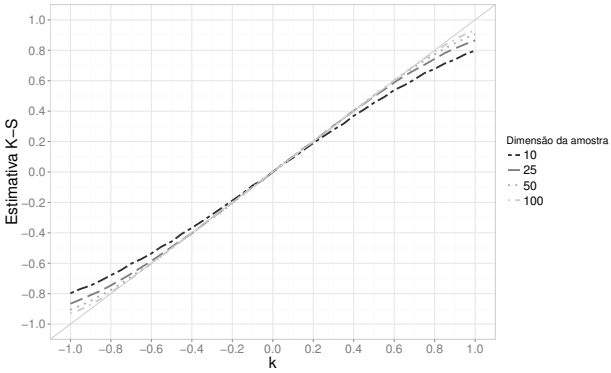


Figura 4: Valor médio das estimativas para k obtidas através do teste de ajustamento de K-S, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

0.5 só se encontram no teste de ajustamento de A-D em amostras de dimensão 50 para $|k|$ próximo de 1 e, nos dois testes de ajustamento em amostras de dimensão 100 para $|k|$ próximo de 1.

Parece denotar-se que, em praticamente todos os casos simulados, os intervalos de confiança obtidos cobrem grande parte do espaço do parâmetro.

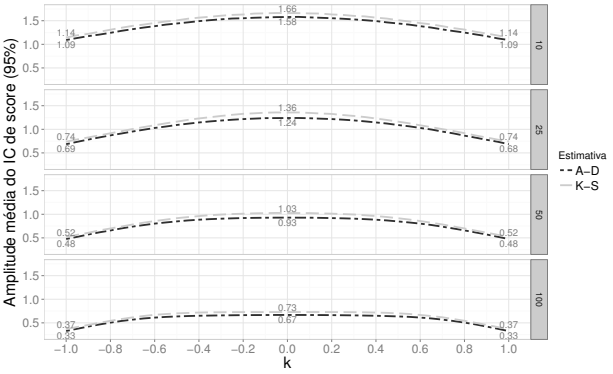


Figura 5: Amplitude média dos intervalos de confiança de *score* utilizando os testes de ajustamento de A-D e de K-S, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

Uma outra forma de utilizar esta informação será considerar a proporção de intervalos de confiança de *score* que incluem $k = 0$, ou seja, os casos em que não seria excluída a hipótese de $X_{2;k}$ ter distribuição uniforme padrão. Na Figura 6 encontra-se representada esta informação, podendo-se constatar que, para amostras de dimensão 10 e 25, não se rejeita a hipótese de ajustamento da uniforme padrão ($k = 0$) para qualquer valor de k e que, para amostras de dimensão 50 e 100, só se rejeita a hipótese de ajustamento da uniforme padrão para valores de $|k|$ próximos de 1.

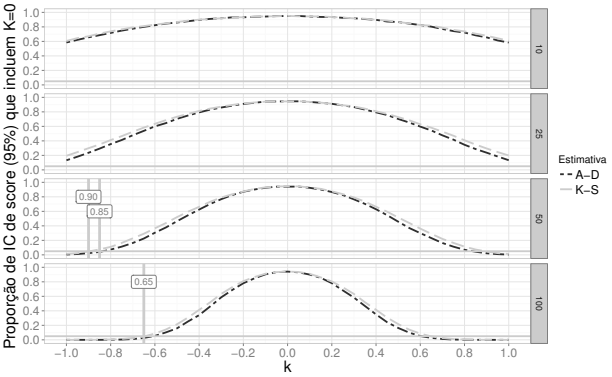


Figura 6: Proporção dos intervalos de confiança de *score* que incluem $k = 0$. Intervalos de confiança de *score* obtidos utilizando os testes de ajustamento de A-D e de K-S, para os diversos valores de k e para amostras de dimensão $n = 10$, 25, 50 e 100.

4 Resumo dos resultados

Entre os métodos de estudados destaca-se o teste de ajustamento de A-D com resultados melhores do que os obtidos através do teste de ajustamento de K-S, uma diferença que tende a esbater-se à medida que a dimensão da amostra aumenta. Salienta-se também a tendência generalizada de melhoria dos resultados à medida que a dimensão da amostra aumenta:

- O enviesamento tende para zero à medida que a amostra aumenta. Para amostras de pequena dimensão observa-se um maior enviesamento que se acentua quando $|k|$ está perto de 1.

- Intervalos de confiança de *score* com nível de significância de 95% com elevada amplitude. Amplitude que vai diminuindo à medida que $|k|$ se aproxima de 1 (componente dominante Beta) e que aumenta a dimensão da amostra.

- Elevada proporção de intervalos de confiança de *score* com nível de significância de 95% que incluem $k = 0$: para amostras de dimensão 10 e 25 não se rejeita a hipótese de ajustamento da uniforme padrão (k=0) para qualquer valor de k; para amostras de dimensão 50 e 100 só se rejeita a hipótese de ajustamento da uniforme padrão para $|k| \approx 1$.

Os resultados ilustram a dificuldade deste problema de estimação. Consequentemente, as implicações a nível de meta análise ganham novo relevo, mostrando a necessidade de reaquacionar a teoria de valores de prova combinados tendo em linha de conta não só a possibilidade de H_{1j} ser verdadeira para alguns j como também a possibilidade de haver "batota" no reporte dos valores de prova-p originais.

Agradecimentos

Os autores agradecem o financiamento pela Fundação para a Ciência e a Tecnologia, através do projeto UID/MAT/00006/2013.

Referências

[1] Brilhante, M.F., Pestana. D., Semblano, P., Sequeira, F. (2014). On the Proportion of Non Uniform Reported p-values. In T.E. Tsimos, G. Psihoyios, Ch. Tsitouras and Z. Anatassi (eds.): *ICNAAM 2014, AIP Conference Proceedings.* (in press)

[2] Fisher, R.A. (1932). *Statistical Methods for Research Workers.* (4ª ed.) London, Oliver and Boyed.

[3] Pires, A.M., Branco, J.A. (2010). A statistical model to explain the Mendel-Fisher controversy. *Statistical Science*, **25**, 545–565.

[4] Tippett, L.H.C. (1931). *The Methods of Statistics.* London, Williams & Norgate.

[5] Tsui, K., Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602-607.

A few notes on using prevalence of infection in malaria elimination settings

Nuno Sepúlveda
LSHTM and CEAUL, *nuno.sepulveda@lshtm.ac.uk*

Chris Drakeley
LSHTM, *chris.drakeley@lshtm.ac.uk*

Keywords: epidemiology, proportion, confidence interval, posterior estimation.

Abstract: Last decade has witnessed a global effort to decrease malaria mortality and incidence rates. This effort led to strong decreases in disease transmission to the point that elimination and eradication might be achieved in several countries soon. In this context, there is a research interest in understanding the statistical power of current data analysis approaches to handle with such targets and whether this power can be improved in some way. The goal of the paper is to discuss a few statistical problems of using prevalence of infection in real data from Swaziland and Iran, two countries on the cusp of malaria elimination.

1 Introduction

Malaria is a parasitic disease affecting millions of people worldwide with the highest death toll in subsaharan Africa. Latest official statistics revealed a steadily decrease of malaria-related deaths and infection rates in the last decade [1]. This decrease in disease burden reached infectivity levels where malaria elimination and eradication might be envisioned up to 2030 in countries like Cape Verde or Sri Lanka. However, the path to a disease free setting has several hurdles [2], specifically, on how to obtain sufficient economical, logistic and scientific support to design, execute and study the impact of future

malaria elimination campaigns.
According to the World Health Organization (WHO), a given region can be classified as in a pre-elimination, elimination or eradication stage if the number of cases per 1,000 people at risk is 5, < 1, or 0, respectively. From a statistical standpoint, these epidemiological settings bring into the horizon interesting statistical problems in the frontier of stochastic phenomena. In general, malaria burden is measured by the number of cases officially reported or the number of infected individuals detected in a surveillance study. As a desirable outcome, the path towards malaria elimination gradually reduces the number of infected cases in the population and, therefore, surveillance or impact studies are likely to deal with true prevalences of infection close to 0. In this situation, point estimates for the prevalence might be in the borderline of the parameter space whereas the respective confidence intervals might not have the expected theoretical properties or not even calculable depending on the method used. Therefore, it is important to understand which statistical inference methods correctly quantify the underlying uncertainty.
This paper aims to discuss the statistical problems of using prevalence of infection in a context of malaria elimination. Section 2 focuses on the problems associated with the estimation via a statistical interval while section 3 deals with the problem of classifying a given population in the respective malaria elimination stage. Finally, Section 4 concludes with a few remarks and avenues for future research.

2 Estimation of prevalence of infection

One of the major epidemiological problems in malaria elimination studies is to accurately detect the presence of infection in asymptomatic individuals. In practice, there are three classes of diagnostic tests: (i) visual inspection of blood slides on the microscope; (ii) the rapid diagnostic tests that detect the expression of specific parasite genes that are activated upon infection; (iii) molecular assays where

specific DNA fragments of the malaria parasite are amplified when present, allowing their detection in a gel. As expected, each diagnostic test has its own sensitivity (probability of detecting a truly infected individual) and specificity (probability of detecting a truly non-infected individual) and scope of applicability in practice. For example, microscopy testing requires specific training and the availability of lab experts that can identify malaria parasites at their different stages of infection. Rapid diagnostic tests are typically easy and cheap to use but tend to fail in asymptomatic infections where the malaria parasite count is low. The molecular assays are by far the most reliable diagnostic tests. However, they require specific lab conditions and qualified teams of lab experts that are not broadly available in malaria-endemic countries. Concerning sensitivity and specificity, recent research performing an extensive meta-analysis showed that the performance of the above diagnostic tools vary with the underlying transmission intensity with important implications for malaria elimination strategies [3]. Combining the results from different diagnostic tools tends to increase the chance of detecting all infected individuals in a sample. This was attempted in large national cross-sectional study in Swaziland [4]. Of the 4,330 participants tested, three were malaria positive for rapid diagnostic testing. Additional 2 infected individuals were detected using molecular assay testing. However, the simultaneous use of different diagnostic tools is not common practice in malaria endemic countries mostly due to economic, logistic and technical constraints. Therefore, the estimation of the underlying sensitivity and specificity becomes an important aspect of any malaria epidemiological study. In absence of prior information, estimating sensitivity is highly problematic in samples from regions on the cusp of malaria elimination due to the high chance of not sampling infected individuals as demonstrated in a cross-sectional study of 1,500 individuals from Iran [5]. In that case, one assumes a given value for sensitivity and specificity and carries on with the analysis accordingly.

After detecting the presence of infection in the study participants, the main objective is then to estimate the prevalence of infection in

the population. For the subsequent discussion, let's assume perfect sensitivity or specificity of the diagnostic tools or at least the use of many different diagnostic tools that ensure the detection of all infected individuals. To estimate the prevalence of infection (hereafter denoted by π), the simplest statistical framework is to assume a large population size and a sampling with replacement in order to bring the popular Binomial distribution into data analysis. As a basic knowledge of statistical inference, the maximum likelihood estimate of π is given by the sample proportion. To estimate the uncertainty associated with that estimate, there are several methods to calculate the respective confidence interval. Recent research studied the statistical properties of these confidence intervals and the respective implications for sample size calculation in wide range of the value for the true population [6]. In a disease elimination setting, π is expected to be very close to 0, a situation where the confidence intervals are prone for overshooting (i.e., lower bound might be negative), degeneracy (i.e., the confidence interval is a single point) or incorrect coverage. The most well-known example is the popular Wald confidence interval at $(1-\alpha)\times 100\%$,

$$\hat{\pi} \pm q_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \tag{1}$$

where $\hat{\pi}$ is an estimate of π, $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution, and n is the sample size. For $\hat{\pi}$ close to zero, it is often observed that the lower bound might be negative, which is not an admissible value for π. When $\hat{\pi}=0$, the resulting confidence interval is degenerate as the lower and upper bound coincide and equal to 0. The degeneracy and overshooting of the Wald confidence interval can be avoided by using the Clopper-Pearson confidence interval. However, this exact confidence interval tends to overestimate the expected coverage.

In practice, the popular R software provides different packages for estimating a proportion. For disease elimination purposes, the RSurveillance package brings into the community a suite of statistical tools to help monitoring populations on the cusp of being free from

a given infection. This package includes frequentist and Bayesian methods for estimating a proportion via an interval: Clopper-Pearson method (hereafter referred to as the exact confidence interval), Wilson score that is based on the inversion of the score test for a proportion, the method proposed by Agresti and Coull, and Bayesian credible intervals based on a $Beta(\alpha, \beta)$ prior distribution for π. With respect to Bayesian methods, it is common to use non-informative prior distributions such as the Jeffrey's and Uniform distributions ($\alpha = \beta = 1/2$ and $\alpha = \beta = 1$, respectively). These methods are then the starting point to understand the uncertainty underlying the data.

As examples of application, Table 1 shows the respective results for estimating the prevalence of infection in the above-mentioned malaria elimination studies from Swaziland (5 malaria positive cases out of 4,330 individuals) and Iran (no malaria cases detected in a sample of 1,500 individuals). These statistical intervals imply different degrees of uncertainty associated with the prevalence estimation. On the one hand, the Jeffrey's credible interval, although using a non-informative prior, is the method providing the highest precision (i.e., lowest range) for π in both datasets. This result might be explained in part by the fact that the Jeffrey's prior distribution implies a high prior probability for very low values of π in relation to a Uniform prior distribution. For example, the prior probability of $\pi < 0.01$ is 0.064 and 0.010 using the Jeffrey's and Uniform prior distribution, respectively. On the other side of the spectrum, the Agresti-Coull confidence intervals are the longest ones in both data sets. More importantly, the respective confidence interval for the Iran data set shows a problem of overshooting (negative lower bound), thus, illustrating the difficulty of obtaining coherent estimates for π in a malaria elimination setting.

After calculating different intervals for π, the next step of the analysis is to understand which one provides the most accurate and reliable quantification of the underlying uncertainty. For that matter one can assess the performance of a given interval in terms of the frequentist concepts of coverage and expected length, or the costs

Table 1: Estimation of π in malaria studies from Swaziland and Iran using 95% confidence intervals (Exact, Wilson, and Agresti-Coull) and 95% central credible intervals (Jeffreys and Uniform).

Study	Method	Lower bound	Upper bound
Swaziland	Exact	0.00038	0.00269
	Wilson	0.00049	0.00270
	Agresti-Coull	0.00041	0.00279
	Jeffreys	0.00044	0.00253
	Uniform	0.00051	0.00269
Iran	Exact	0.00000	0.00246
	Wilson	0.00000	0.00247
	Agresti-Coull	-0.00005	0.00298
	Jeffreys	0.00000	0.00128
	Uniform	0.00000	0.00199

associated with future sample size calculations [8]. As a preliminary step towards a more comprehensive comparison, this paper focus on coverage using the sample sizes adopted in the above cross-sectional studies as case studies. The comparison is carried out using $\pi = 0.005$ and 0.001, two thresholds used by WHO to differentiate populations in pre-elimination and elimination stages, respectively. To estimate coverage of each interval, asymptotic expansions were used as available in the R package `binom` [7]. Data simulation was also performed because asymptotic expansions might not be so accurate when the true prevalence is in the borderline of the parameter space, as demonstrated elsewhere [8]. Data simulation consisted of 10,000 binomial samples per pair of sample size and prevalence. Coverage was estimated as the proportion of times each interval included the true prevalence.

The simulated and approximated coverage are in good agreement with each other, thus, indicating they are reliable estimates of the

true coverage (Table 2). The three general comments can be made to the results. The first one is that none of the interval agreed with the nominal value of 95% for the coverage even for the large sample size of 4,330 individuals. This observation is very important in practice because one expects to obtain correct coverage by increasing the sample size, which might not be true in the present setting. The second comment refers to the inflated coverage of exact confidence intervals irrespective of the sample size and true prevalence. This confidence interval could be regarded as the default choice if one wishes to adopt a very conservative approach for data analysis. A less conservative choice is to use the Uniform-based credibility interval that can be seen as the a shrinkage version of the exact confidence interval as suggested by Thulin [8]. However, this interval should be used with caution because it might lead to an under coverage situation for relatively large sample sizes ($n = 1{,}500$). The third comment relates to the tendency of obtaining intervals closer to the correct coverage for the malaria pre-elimination threshold than for the malaria elimination one. This result can be easily seen by the range of the estimated coverages. For example, using a sample size of 1,500 individuals, the range of coverages is from 0.9377 to 0.9741 for $\pi = 0.005$ whereas the respective range for $\pi = 0.001$ is from 0.9345 to 0.9815. Interestingly, when the sample size is 1,500 individuals, coverage is somewhat unstable because it can be lower and higher than expected depending on the interval used. In the increased sample size of 4,300, all the intervals with the exception of the exact confidence intervals are in agreement with undercoverage and overcoverage for $\pi = 0.005$ and 0.001, respectively. Overcoverage and undercoverage have important practical implications since the former situation might be prone to overestimate uncertainty while the latter the opposite. Combining all these above comments together, there is no simple answer for choosing the best interval using a sample size of 1,500 individuals. For the sample size of 4,330 individuals where coverage seems to be more consistent across intervals, the Uniform-based credible intervals are the closest ones to the nominal value of 95%, thus, being the recommended approach in what

Table 2: Approximate coverage (simulated estimates in brackets) of the different 95% confidence and credible intervals for π.

π	Interval	$n = 1{,}500$	$n = 4{,}330$
0.005	Exact	0.9587 (0.9590)	0.9606 (0.9602)
	Wilson	0.9377 (0.9370)	0.9486 (0.9485)
	Agresti-Coull	0.9741 (0.9743)	0.9486 (0.9484)
	Jeffreys	0.9587 (0.9590)	0.9461 (0.9455)
	Uniform	0.9377 (0.9370)	0.9487 (0.9485)
0.001	Exact	0.9815 (0.9828)	0.9734 (0.9728)
	Wilson	0.9345 (0.9335)	0.9540 (0.9549)
	Agresti-Coull	0.9815 (0.9828)	0.9672 (0.9711)
	Jeffreys	0.9815 (0.9828)	0.9540 (0.9549)
	Uniform	0.9345 (0.9335)	0.9540 (0.9549)

this criterium is concerned.

3 Classification of the malaria elimination stage

The classification of a given population into the respective malaria elimination stage is an important task for public health authorities because each stage implies different epidemiological strategies (see guidelines in 2015 WHO report [1]). There are four broad classification stages: not yet in a pre-elimination stage (NPE, $\pi \geq 0.005$), pre-elimination (PE, $0.001 \leq \pi < 0.005$), elimination (E, $\pi < 0.001$) and eradication (no malaria cases in at least three consecutive years, not in the scope of this paper). The statistical investigation consists of determining the underlying malaria elimination stage of the population given the data. A simple way to do it is to adopt the worst case scenario by comparing the upper bound of the intervals with

the WHO thresholds. Coming back to Table 1, both Swaziland and Iran would appear to be in a pre-elimination stage. A more informative approach is to apply a Bayesian perspective to the problem. It is then common to use $Beta(\alpha, \beta)$ for the prior distribution of π under a Binomial sampling. In that case, the posterior distribution for π is given by $Beta(\alpha + x, \beta + n - x)$ where x is the number of malaria cases detected in a sample of size n. The above classification problem is then easily solved by calculating the following posterior probabilities:

$$\begin{aligned} \theta_{NPE} &= P(\pi \geq 0.005|n, x, \alpha, \beta), \\ \theta_{PE} &= P(0.001 \leq \pi < 0.005|n, x, \alpha, \beta), \\ \theta_{E} &= P(\pi < 0.001|n, x, \alpha, \beta). \end{aligned} \tag{2}$$

These probabilities are simply calculated using the cumulative probability distribution of the Beta posterior distribution. As usual, one can adopt a conservative approach to data analysis by considering Jeffrey's or Uniform prior distributions for π. Under this assumption (Table 3), Swaziland is most likely to be in a pre-elimination stage ($\theta_{PE} = 0.653$ and $\theta_{PE} = 0.732$ for Jeffrey's and Uniform prior distribution, respectively) while Iran is further down in its elimination stage ($\theta_E = 0.917$ and $\theta_E = 0.777$ for Jeffrey's and Uniform prior distributions, respectively). It is worth noting the strong effect of the prior distribution on the posterior classification probabilities. Like alluded above, the Jeffrey's distribution provides high prior probability to π close to either 0 or 1. Therefore, this distribution, although seen as non-informative in most statistical applications, might be considered as informative in malaria elimination settings. This issue will be investigated in more detailed elsewhere.

4 Concluding remarks

This paper described a few statistical problems of using prevalence of infection in malaria elimination settings. Other problems do exist

Table 3: Posterior probability of each malaria elimination stage using Jeffreys and Uniform noniformative prior distribution for π.

Study	Stage	Jeffreys	Uniform
Swaziland	Elimination	0.347	0.268
	Pre-elimination	0.653	0.732
	Other	0.000	0.000
Iran	Elimination	0.917	0.777
	Pre-elimination	0.83	0.222
	Other	< 0.01	0.01

in this setting such as how sensitivity and specificity of the diagnostic tests can be estimated when only a few sampled individuals are expected to be found infected and how these statistical parameters affect the estimation of the prevalence of infection. Planning data collection to maximise the chance of sampling infected individuals is another issue to be tackled in practice because malaria transmission might be affected by different seasonal and environmental factors. Since all these problems are difficult to be controlled by malaria epidemiologists, alternative approaches for measuring malaria reduction and potentially elimination have been proposed such as the one using antibody-based measures [9]. The basic notion is that the host immune system is capable of reacting to malaria infection via production of antibodies targeting specific parasite antigens. Since these antimalarial antibodies can persist in time at reasonably stable concentrations, they are extremely useful to inform on the past malaria exposure of individuals living in endemic areas. Two epidemiological measures arise in this context. The so-called seroprevalence is the proportion of malaria exposed individuals defined by a sufficiently high antibody concentration. The seroconversion rate is the frequency by which seronegative individuals become seropositive, thus, being considered as a proxy of the underlying ma-

laria transmission intensity. In this setting, malaria elimination and eradication brings the problem of accurately detecting seropositive individuals in the data. Moreover, since seroconversion rate can be integrated in stochastic models describing different disease transmission dynamics, it is important to understand whether data has enough power to distinguish models that assume either a very low but stable transmission intensity over time or the occurrence of an elimination event somewhere in the past [10]. These questions will be investigated in a near future.

References

[1] World Health Organization (2015). *2015 World Malaria Report*. World Health Organization, Geneva.

[2] Stresman, G., Kobayashi, T., Kamanga, A., Thuma, P. E., Mharakurwa, S., Moss, W. J., Shiff, C. (2012). Malaria research challenges in low prevalence settings. *Malaria Journal* 11, 353.

[3] Wu, L., van den Hoogen, L. L., Slater, H., Walker, P. G., Ghani, A. C., Drakeley, C. J., Okell, L. C. (2015). Comparison of diagnostics for the detection of asymptomatic Plasmodium falciparum infections to inform control and elimination strategies. *Nature* 528:S86-93.

[4] Hsiang, M. S., Hwang, J., Kunene, S., Drakeley, C., Kandula, D., Novotny, J., Parizo, J., Jensen, T., Kemere, J., Dlamini, S., Moonen, B., Angov, E., Dutta, S., Ockenhouse, C., Dorsey, G., Greenhouse, B. (2012). Surveillance for Malaria Elimination in Swaziland: A National Cross-Sectional Study Using Pooled PCR and Serology. *PLoS ONE* 7, e29550.

[5] Zoghi, S., Mehrizi, A. A., Raeisi, A., Haghdoost, A. A., Turki, H., Safari, R., Kahanali, A. A., Zakeri, S. (2012). Survey for

asymptomatic malaria cases in low transmission settings of Iran under elimination programme. *Malaria Journal* 11:126.

[6] Gonçalves, L., Rosário de Oliveira, M., Pascoal C., Pires, A. (2012). Sample size for estimating a binomial proportion: comparison of different methods. *Journal of Applied Statistics* 39, 2453–2473.

[7] Brown, L. D., Cai, T. T., Das Gupta, A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *Annals of Statistics* 30, 160–201.

[8] Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics* 8, 817–840.

[9] Corran, P., Coleman, P., Riley, E., Drakeley, C. (2007). Serology: a robust indicator of malaria transmission intensity? *Trends in Parasitology* 23, 575–82.

[10] Sepúlveda, N., Stresman, G., White, M. T., Drakeley, C. (2015). Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication. *Journal of Immunology Research* 2015, 738030.

Comportamento extremal de um modelo INMA(q) segmentado

Rui Sequeira
Dep. de Matemática, Fac. de Ciências e Tecnologia, Universidade de Coimbra, *ruisequeira_@hotmail.com*

Maria da Graça Temido
CMUC, Dep. de Matemática, Fac. de Ciências e Tecnologia, Universidade de Coimbra, *mgtm@mat.uc.pt*

Palavras–chave: teoria de valores extremos, classe de Anderson, sucessões estacionárias inteiras

Resumo: Estudamos o máximo de um modelo INMA(q) segmentado. Depois de validadas condições de independência assintótica e de dependência local apropriadas, é obtida, como limite em distribuição do máximo, a função de distribuição de Gumbel discreta.

1 Introdução

Muitas das séries temporais que se encontram na prática são, pela sua natureza, constituídas por variáveis aleatórias (v.a.'s) inteiras não negativas. Este tipo de dados surgem naturalmente associados a processos de contagem de interesse estatístico em diversas áreas.
Entre todas as classes de modelos de contagem, encontra-se uma classe baseada num operador aleatório aplicável a inteiros, denominado *operador binomial thinning*. Recordemos que este operador, introduzido por [3], transforma um número real $\beta \in [0,1]$ e uma variável aleatória (v.a.) inteira positiva Z na v.a. inteira positiva $X := \beta \circ Z = \sum_{i=1}^{Z} B_i(\beta)$, onde a *sucessão de contagem* $\{B_i(\beta)\}$ ([10]) é uma sucessão de v.a.'s independentes com distribuição de Ber-

[10]Neste trabalho o índice das sucessões percorre o conjunto $\mathbb{N}$.

noulli de parâmetro β e independente de Z. Ao substituir a multiplicação escalar usual por esta operação aleatória foi possível construir modelos análogos aos modelos ARMA (do inglês *auto-regressive moving average*), bem como a muitas das suas generalizações. Sendo os modelos ARMA já uma referência clássica no domínio das v.a.'s contínuas, a ideia de encontrar modelos análogos para dados de contagem tem atraído a atenção de inúmeros autores, entre os quais citamos [1], [4], [5] e [7].

Este trabalho é dedicado ao estudo do comportamento extremal de uma sucessão média móvel de ordem finita de v.a.'s inteiras positivas, $\{X_n\}$, proposta por [7], que designamos INMA(q) segmentada (do inglês *integer moving average of order q*).([11]) Esta sucessão fortemente estacionária (f.e.) é definida por

$$X_n = \begin{cases} \beta_0 \circ Z_n & \text{c.p. } b_0 \\ \beta_0 \circ Z_n + \beta_1 \circ Z_{n-1} & \text{c.p. } b_1 \\ \quad \cdots & \\ \beta_0 \circ Z_n + \ldots + \beta_{q-1} \circ Z_{n-q+1} & \text{c.p. } b_{q-1} \\ \beta_0 \circ Z_n + \ldots + \beta_{q-1} \circ Z_{n-q+1} + Z_{n-q} & \text{c.p. } b_q \end{cases} \tag{1}$$

com $\beta_i \in]0{,}1[$, $0 \leq i \leq q-1, \beta_q = 1$, $b_0 = \beta_0$, $b_i = (1-\beta_0)\ldots(1-\beta_{i-1})\beta_i$, $1 \leq i \leq q$, e $\{Z_n\}$ é uma sucessão de v.a.'s inteiras positivas e identicamente distribuídas (i.i.d.). Mais, considera-se que todas as operações aleatórias envolvidas pelo operador $\circ$ são independentes. O comportamento extremal desta sucessão é estudado em [4], considerando que $\{Z_n\}$ tem distribuição marginal geométrica.

Ao longo deste trabalho, representamos por P_Z a função geradora de probabilidades (f.g.p.) de qualquer v.a. inteira Z. Relativamente à v.a. operada $X = \beta \circ Z$, destacamos as seguintes propriedades.

Lema 1.1 *([5]) Para $X = \beta \circ Z$, tem-se: $E(X) = \beta E(Z)$; $P_X(1+h) = P_Z(1+\beta h)$; $E((1+h)^Z) = 1 + hE(Z) + o_h(1)$, $h \to 0$; $\beta_1 \circ (\beta_2 \circ Z) =^d \beta_2 \circ (\beta_1 \circ Z) =^d (\beta_1\beta_2) \circ Z$; $\beta \circ (Z+Y) =^d \beta \circ Z + \beta \circ Y$.*

[11] Em [7] esta sucessão é designada Geometric INMA(q) uma vez que as inovações $\{Z_n\}$ têm distribuição geométrica, o que não sucede neste trabalho.

Notamos que $\{X_n\}$ se pode escrever na forma

$$X_n = \sum_{j=0}^{q} \mathbb{I}_{A_j} \left(\sum_{i=0}^{j} \beta_i \circ Z_{n-i} \right), \tag{2}$$

onde A_0,A_1,...,A_q constituem uma partição do espaço Ω e $\mathbb{I}_{A_j}$ representa a indicatriz de A_j. As v.a.'s $\mathbb{I}_{A_j}$, $j = 0,1,\ldots,q$, são dependentes mas, para cada $j = 0,1,\ldots,q$, $\mathbb{I}_{A_j}$ e $\sum_{i=0}^{j} \beta_i \circ Z_{n-i}$ são independentes.

2 Estacionaridade forte da sucessão

Proposição 2.1 *A sucessão INMA(q) definida por (1) é f.e.*

Dem.: Escreva-se X_n na forma (2). Para $j \in \{0,1,\ldots,q\}$, seja $Y_n^{(j)} = \sum_{i=0}^{j} \beta_i \circ Z_{n-i}$. Uma vez que

$$E\left(s_0^{Y_n^{(j)}} s_1^{Y_{n+1}^{(j)}} \ldots s_k^{Y_{n+k}^{(j)}} \right) =$$

$$\begin{aligned}
&= \prod_{i=0}^{k-1} E\left(\prod_{l=0}^{i} s_l^{\beta_{j-i+l} \circ Z_{n-j+i}} \right) \prod_{i=0}^{j-k} E\left(\prod_{l=0}^{k} s_l^{\beta_{j-k-i+l} \circ Z_{n-j+k+i}} \right) \\
&\qquad \times \prod_{i=0}^{k-1} E\left(\prod_{l=0}^{i} s_{k-l}^{\beta_{i-l} \circ Z_{n+k-i}} \right) \\
&= \prod_{i=0}^{k-1} E\left(\prod_{l=0}^{i} s_l^{\beta_{j-i+l} \circ Z_{n-j+i+t}} \right) \prod_{i=0}^{j-k} E\left(\prod_{l=0}^{k} s_l^{\beta_{j-k-i+l} \circ Z_{n-j+k+i+t}} \right) \\
&\qquad \times \prod_{i=0}^{k-1} E\left(\prod_{l=0}^{i} s_{k-l}^{\beta_{i-l} \circ Z_{n+k-i+t}} \right) \\
&= E\left(s_0^{Y_{n+t}^{(j)}} s_1^{Y_{n+1+t}^{(j)}} \ldots s_k^{Y_{n+k+t}^{(j)}} \right)
\end{aligned}$$

decorre que $\{Y_n^{(j)}\}_n$ é f.e.. Facilmente se prova que $\{W_n^{(j)}\}_n$ com $W_n^{(j)} = \mathbb{I}_{A_j} Y_n^{(j)}$ é também f.e. Como $X_n = \sum_{j=0}^{q} W_n^{(j)}$ é uma função mensurável de uma sucessão f.e., conclui-se que $\{X_n\}$ é f.e. ■

3 Cauda das margens da sucessão

Para funções de distribuição (f.d.'s) F inteiras, com extremo superior do suporte infinito ($w_F = \infty$), tais que

$$\lim_{n\to+\infty} \frac{1-F(n-1)}{1-F(n)} = r > 1, \tag{3}$$

como a Binomial Negativa, o máximo de n v.a's i.i.d. com função de distribuição (f.d.) F não possui distribuição limite não degenerada na classe de leis max-estáveis, sob nenhum tipo de normalização. Recordamos que se $\lim_{x\to\omega_F^-} F(x) = 1$, então $\lim_{x\to\omega_F^-} \frac{1-F(x)}{1-F(x^-)} = 1$ é condição necessária e suficiente para que exista uma sucessão real $\{u_n\}$ e $\tau > 0$ tais que $\lim_{n\to+\infty} F^n(u_n) = e^{-\tau}$. Todavia, [2] prova que uma f.d. discreta, com $w_F = \infty$, satisfaz (3) se e só se existe uma sucessão real $\{b_n\}$ tal que, com $u_n = x + b_n$, se tem

$$\exp(-r^{-x-1}) \le \liminf_{n\to+\infty} F^n(u_n) \le \limsup_{n\to+\infty} F^n(u_n) \le \exp(-r^{-x}), \tag{4}$$

para qualquer $x \in \mathbb{R}$. A classe das f.d.'s que verificam (3), introduzida em [2], recebe o nome de classe de Anderson e é aqui denotada por $\mathcal{C}_{\mathcal{A}}(r)$. Por outro lado, [8] prova que para f.d.'s F discretas, com $w_F = \infty$, existe uma sucessão crescente de inteiros $\{k_n\}$ a verificar

$$k_{n+1}/k_n \to r \ge 1,\ n \to +\infty, \tag{5}$$

e existe uma sucessão real $\{b_n\}$ tais que

$$\lim_{n\to+\infty} F^{k_n}(x + b_n) = G(x) := \exp(-r^{-\lfloor x \rfloor}),\ x \in \mathbb{R}, \tag{6}$$

se e só se F verifica (3). À f.d. limite G chamamos Gumbel discreta. De modo a usarmos resultados de [5], no que segue, consideramos a subclasse de $\mathcal{C}_{\mathcal{A}}(r)$, constituída pelas f.d.'s que satisfazem

$$1 - F(z) \sim K \lfloor z \rfloor^{\xi} r^{-\lfloor z \rfloor}, \; z \to +\infty, \tag{7}$$

onde $\xi \in \mathbb{R}$, $K > 0$ e $r > 1$. Esta classe, que inclui a f.d. Binomial Negativa e a f.d. Gumbel discreta, será denotada por $\mathcal{C}^*_{\mathcal{A}}(r)$.
O lema que apresentamos de seguida deve-se a [5] e estabelece essencialmente que a soma de duas v.a independentes na classe $\mathcal{C}^*_{\mathcal{A}}(r)$ pertence ainda à mesma classe.

Lema 3.1 *([5])* 1. *Suponhamos que a v.a. Y_1 satisfaz (7), e que é independente da v.a. inteira Y_2 que verifica $E((r^*)^{Y_2}) < +\infty$, para algum $r^* > r$. Então*

$$P(Y_1 + Y_2 > z) \sim K E(r^{Y_2}) \lfloor z \rfloor^{\xi} r^{-\lfloor z \rfloor}, \; z \to +\infty. \tag{8}$$

2. *Suponhamos que Y_1 e Y_2 são v.a.'s independentes e verificam (7) com $\xi_i \in \mathbb{R}, K_i > 0$ e $r > 1$ para $i \in \{1,2\}$. Se $\xi_1 = \xi_2 = \xi < -1$, a soma $Y_1 + Y_2$ satisfaz (7) com $K = K_1 E(r^{Y_2}) + K_2 E(r^{Y_1})$ e se $\xi_1 > -1, \xi_2 > -1$, a soma $Y_1 + Y_2$ satisfaz (7) com $K = (r-1) K_1 K_2 \frac{\Gamma(\xi_1+1)\Gamma(\xi_2+1)}{\Gamma(\xi_1+\xi_2+2)}$ e $\xi = \xi_1 + \xi_2 + 1$.*

A distribuição de cauda da v.a. operada $X = \beta \circ Z$ também se deve a [5] e é especificada a seguir.

Lema 3.2 *([5]) Se Z é uma v.a. com f.d. F_Z pertencente a $\mathcal{C}^*_{\mathcal{A}}(r)$, então a v.a. $X = \beta \circ Z$ tem f.d. F_X a satisfazer*

$$1 - F_X(z) \sim A \lfloor z \rfloor^{\xi} (r^*)^{-\lfloor z \rfloor}, \; z \to +\infty,$$

onde $r^* = 1 + \frac{r-1}{\beta}$, $A = K\beta \left(\frac{r}{r^*\beta}\right)^{\xi+1}$, isto é, F_X pertence a $\mathcal{C}^*_{\mathcal{A}}(r^*)$.

No Teorema seguinte estabelecemos que se as margens de $\{Z_n\}$ pertencerem a $\mathcal{C}^*_{\mathcal{A}}(r)$, o mesmo sucede com $\{X_n\}$.

Teorema 3.3 *Seja $\{X_n\}$ a sucessão f.e. definida por (1). Se a f.d. de $\{Z_n\}$ satisfizer (7), com $\xi \neq -1$, então existe uma constante A_q tal que*

$$P(X_n > z) \sim A_q \lfloor z \rfloor^{\xi} r^{-\lfloor z \rfloor},\ z \to +\infty.$$

Dem.: Tem-se

$$P(X_n > z) = \sum_{j=0}^{q} b_j P\left(\sum_{i=0}^{j} \beta_i \circ Z_{n-i} > z\right),\ z \in \mathbb{R}^+. \tag{9}$$

De acordo com o Lema 3.2, a f.d. de $\beta_i \circ Z_{n-i}$ pertence a $\mathcal{C}^*_{\mathcal{A}}(r^*)$, com β substituído por β_i. Seja $\beta^{(j)}_{max} = \max\{\beta_0, \ldots, \beta_j\}$. Seguindo de perto [5], escreva-se

$$\sum_{i=0}^{j} \beta_i \circ Z_{n-i} = \sum_{\beta_i = \beta^{(j)}_{max}} \beta_i \circ Z_{n-i} + \sum_{\beta_i \neq \beta^{(j)}_{max}} \beta_i \circ Z_{n-i},$$

para $j \in \{0, 1, \ldots, q-1\}$. Então, aplicando o Lema 3.1 repetidas vezes obtemos

$$P\left(\sum_{i=0}^{j} \beta_i \circ Z_{n-i} > z\right) \sim C_j \lfloor z \rfloor^{\xi} (r'_j)^{-\lfloor z \rfloor},\ z \to +\infty,$$

onde $r'_j = 1 + \frac{r-1}{\beta^{(j)}_{max}} > r$ e C_j é uma constante dependente das constantes iniciais. Por outro lado, uma vez que $\sum_{i=0}^{q} \beta_i \circ Z_{n-i} = Z_{n-q} + \sum_{i=0}^{q-1} \beta_i \circ Z_{n-i}$ e $\sum_{i=0}^{q-1} \beta_i \circ Z_{n-i}$ tem f.g.p. finita, pelo Lema 3.1, obtemos

$$P\left(\sum_{i=0}^{q} \beta_i \circ Z_{n-i} > z\right) \sim C_q \lfloor z \rfloor^{\xi} r^{-\lfloor z \rfloor},\ z \to +\infty,$$

com $C_q = (r-1)k \prod_{i=0}^{q-1} E(1-\beta_i+\beta_i r)^{Z_{-i}}$ se $\xi > -1$ e $C_q = k \prod_{i=0}^{q-1} E(1-$

$\beta_i + \beta_i r)^{Z_{-i}}$ se $\xi < -1$. Devido a (9), concluímos que

$$\begin{aligned} P(X_n > z) &= \sum_{j=0}^{q} P(\sum_{i=0}^{j} \beta_i \circ Z_{n-i} > z) b_j \\ &\sim \sum_{j=0}^{q-1} C_j \lfloor z \rfloor^{\xi} (r'_j)^{-\lfloor z \rfloor} b_j + C_q [z]^{\xi} r^{-\lfloor z \rfloor} b_q \\ &= C_q [z]^{\xi} (r)^{-\lfloor z \rfloor} \beta_{q-1} \prod_{i=0}^{q-1} (1 - \beta_{i-1})(1 + \mathrm{o}_z(1)),\ z \to +\infty, \end{aligned}$$

pois $r'_j > r$ implica $(\frac{r'_j}{r})^{-\lfloor z \rfloor} \to 0,\ z \to +\infty$, para $j \in \{0,1,\ldots,q-1\}$. A prova fica completa com $A_q = C_q \beta_{q-1} \prod_{i=0}^{q-1} (1 - \beta_{i-1})$. ■

4 Resultado principal

Se considerarmos sucessões de v.a.'s estritamente estacionárias, em [6] prova-se que, sob certas restrições, $\{P(M_n \leq u_n)\}$ possui uma f.d. limite não degenerada igual à que teria se as v.a.'s da sucessão fossem i.i.d. Referimo-nos às condições $D(u_n)$, que confere à sucessão independência assintótica, e a $D'(u_n)$, sob a qual as margens $\{X_n\}$ assumem um comportamento oscilatório como ocorre no caso i.i.d. Com o objectivo de estender este resultado ao caso das f.d. F discretas, com w_F infinito, que verificam $\lim_{n \to +\infty} \dfrac{1 - F(x_n)}{1 - F(x_{n-1})} = r > 1$, onde $\{x_n\}$ coincide com o suporte de F, em [9] são adaptadas estas condições de Leadbeter, provando-se que, com $\{k_n\}$ a satisfazer (5), as sucessões $\{P(M_{k_n} \leq u_n)\}$ e $\{F^{k_n}(u_n)\}$ possuem a mesma f.d. limite. Trata-se das condições $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$ apresentadas adiante. Antes de tais definições, recordamos ainda que em [5] se prova que se a sucessão estacionária $\{X_n\}$ tiver f.d. marginal inteira, F, pertencente a $\mathcal{C}_{\mathcal{A}}(r)$, e verificar as condições $D(x + b_n)$ e $D'(x + b_n)$, então (4) ocorre.

Definição 4.1 *([9]) Seja $\{k_n\}$ uma sucessão de inteiros estritamente crescente e $\{u_n\}$ uma sucessão real. A sucessão de v.a.'s $\{X_n\}$ satisfaz a condição $D_{k_n}(u_n)$ se, para quaisquer inteiros $1 \leq i_1 < ... < i_p < j_1 < ... < j_q \leq k_n$, com $j_1 - i_p > \ell_n$, se tem $\left|P\left(\bigcap_{s=1}^{p}\{X_{i_s} \leq u_n\}, \bigcap_{m=1}^{q}\{X_{j_m} \leq u_n\}\right) - P\left(\bigcap_{s=1}^{p}\{X_{i_s} \leq u_n\}\right) \times \right.$ $\left. \times P\left(\bigcap_{m=1}^{q}\{X_{j_m} \leq u_n\}\right)\right| \leq \alpha_{n,\ell_n}$, onde $\lim_{n\to+\infty} \alpha_{n,\ell_n} = 0$ para algum $\ell_n = o_n(k_n)$.*

Definição 4.2 *([9]) Sejam $\{k_n\}$ e $\{s_n\}$ sucessões de inteiros estritamente crescentes que verificam $\lim_{n\to+\infty} \frac{k_n}{s_n} = +\infty$, e $\{u_n\}$ uma sucessão real. A sucessão f.e. $\{X_n\}$ satisfaz a condição $D'_{k_n}(u_n)$ se satisfizer $D_{k_n}(u_n)$ e*

$$\lim_{n\to+\infty} k_n \sum_{j=2}^{[k_n/s_n]} P(X_1 > u_n, X_j > u_n) = 0.$$

Para uma sucessão de inteiros estritamente crescente $\{k_n\}$ a satisfazer (5), [9] prova que se $\{X_n\}$ é uma sucessão f.e. que verifica $D_{k_n}(u_n)$ e $D'_{k_n}(u_n)$, então $\lim_{n\to+\infty} k_n(1 - F(u_n)) = \tau < +\infty$ se e só se $\lim_{n\to+\infty} P(M_{k_n} \leq u_n) = e^{-\tau}$. Como consequência de (6) e desta equivalência obtemos o resultado seguinte.

Teorema 4.3 *Seja $\{X_n\}$ uma sucessão f.e. com f.d. marginal em $\mathcal{C}_{\mathcal{A}}(r)$. Se existir uma sucessão de inteiros positivos a verificar (5) e uma sucessão real $\{b_n\}$ tais que $D_{k_n}(x + b_n)$ e $D'_{k_n}(x + b_n)$ ocorrem, para qualquer x real, então*

$$\lim_{n\to+\infty} P(M_{k_n} \leq x + b_n) = \exp(-r^{-\lfloor x \rfloor}),\ x \in \mathbb{R}.$$

Estamos agora em condições de estabelecer o resultado principal deste trabalho. Há porém que provar previamente a convergência de $P_{X_n+X_{n-j}}(1 + h)$, para h convenientemente escolhido, seguindo os argumentos de [5].

Lema 4.4 *Seja* $\{X_n\}$ *a sucessão definida por (1) com margens em* $\mathcal{C}^*_{\mathcal{A}}(r)$. *A f.g.p. de* $X_n + X_{n-j}$, $j = 1, \ldots, q$, *é finita, para* h *tal que* $(1+\beta^*)h + \beta^* h^2 < r - 1$ *com* $\beta^* = \max\{\beta_i, i = 0, 1, \ldots, q-1\}$.

Dem.: Comecemos por observar que

$$P_{X_n+X_{n-j}}(1+h) = E[(1+h)^{X_n+X_{n-j}}]$$

$$= E[E[(1+h)^{X_n+X_{n-j}}|Z_n, Z_{n-1}, \ldots, Z_{n-q-j}]]$$

$$= E[E[(1+h)^{X_n}|Z_n, \ldots, Z_{n-q-j}]E[(1+h)^{X_{n-j}}|Z_n, \ldots, Z_{n-q-j}]].$$

Relativamente às esperanças condicionais tem-se, com $B_i(h) = 1 + \beta_i h$,

$$\begin{aligned}
& E[(1+h)^{X_n}|Z_n, \ldots, Z_{n-q-j}]E[(1+h)^{X_{n-j}}|Z_n, \ldots, Z_{n-q-j}] = \\
= & \sum_{l=0}^{q}\sum_{i=0}^{q} b_j b_i \prod_{k=0}^{l} B_k^{Z_{n-k}}(h) \prod_{k=0}^{i} B_k^{Z_{n-k-j}}(h) \\
= & \sum_{l=0}^{j-1}\sum_{i=0}^{q} b_j b_i \prod_{k=0}^{l} B_k^{Z_{n-k}}(h) \prod_{k=0}^{i} B_k^{Z_{n-k-j}}(h) \\
+ & \sum_{m=0}^{q-j}\sum_{i=0}^{m} b_{m+j} b_i \prod_{k=0}^{j-1} B_k^{Z_{n-k}}(h) \prod_{k=0}^{i} (B_{k+j}(h)B_k(h))^{Z_{n-k-j}} \\
& \times \prod_{k=0}^{m} B_k^{Z_{n-k-j}}(h) + \sum_{m=0}^{q-j}\sum_{i=m+1}^{q} b_{m+j} b_i \prod_{k=0}^{j-1} B_k^{Z_{n-k}}(h) \\
& \times \prod_{k=0}^{i} (B_{k+j}(h)B_k(h))^{Z_{n-k-j}} \prod_{k=0}^{i} B_k^{Z_{n-k-j}}(h).
\end{aligned}$$

Então, devido à independência das variáveis da sucessão $\{Z_n\}$, ob-

temos

$$
\begin{aligned}
& P_{X_n+X_{n-j}}(1+h) = \\
= & \sum_{l=0}^{j-1}\sum_{i=0}^{q} b_j b_i \prod_{k=0}^{l} E\left(B_k^{Z_{n-k}}(h)\right) \prod_{k=0}^{i} E\left(B_k^{Z_{n-k-j}}(h)\right) + \\
+ & \sum_{m=0}^{q-j}\sum_{i=0}^{m} b_{m+j} b_i \prod_{k=0}^{j-1} E\left(B_k^{Z_{n-k}}(h)\right) \times \\
& \times \prod_{k=0}^{i} E\left((B_{k+j}(h)B_k(h))^{Z_{n-k-j}}\right) \prod_{k=0}^{m} E\left(B_k^{Z_{n-k-j}}(h)\right) + \\
+ & \sum_{m=0}^{q-j}\sum_{i=m+1}^{q} b_{m+j} b_i \prod_{k=0}^{j-1} E\left(B_k^{Z_{n-k}}(h)\right) \times \\
& \times \prod_{k=0}^{i} E\left((B_{k+j}(h)B_k(h))^{Z_{n-k-j}}\right) \prod_{k=0}^{i} E\left(B_k^{Z_{n-k-j}}(h)\right).
\end{aligned}
$$

Uma vez que esta última expressão envolve somas e produtos finitos, resta provar a convergência das funções geradoras de probabilidades (f.g.p.'s) envolvidas, concretamente, de termos do tipo $E\left(B_k^{Z_{n-k}}(h)\right)$ e de $E\left((B_{k+j}(h)B_k(h))^{Z_{n-k-j}}\right)$. Ora, uma vez que

$$
\frac{P(Z=n)}{P(Z=n+1)} = \frac{\frac{1-F_Z(n-1)}{1-F_Z(n)} - 1}{1 - \frac{1-F_Z(n+1)}{1-F_Z(n)}} \to \frac{r-1}{1-1/r} = r,\ n \to +\infty,
$$

isto é, P_Z tem raio de convergência r, pela segunda propriedade do Lema 1.1, concluímos que $P_{\beta\circ Z}(1+h)$ converge se $h < (r-1)/\beta$. Assim $E\left(B_k^{Z_{n-k}}(h)\right) = E((1+\beta_k h)^Z)$ é convergente se $\beta_k h < r-1$ e

$$
E\left((B_{k+j}(h)B_k(h))^{Z_{n-k-j}}\right) = E((1+(\beta_{k+j}+\beta_k)h + \beta_{k+j}\beta_k h^2)^Z)
$$

é convergente se $(\beta_{k+j}+\beta_k)h + \beta_{k+j}\beta_k h^2 < r-1$. Mas $(1+\beta^*)h + \beta^* h^2 < r-1$ implica $(\beta_{k+j}+\beta_k)h + \beta_{k+j}\beta_k h^2 < r-1$ bem como $\beta_k h < r-1$, para quaisquer j e k em $\{0,1,\ldots,q-1\}$. ■

Teorema 4.5 *Se a f.d. marginal de* $\{Z_n\}$ *pertencer à classe* $\mathcal{C}^*_{\mathcal{A}}(r)$, *então, com* $k_n = [n^{-\xi} A_q^{-1} r^n]$, *tem-se*

$$P(M_{k_n} \leq x + n) \longrightarrow \exp(-r^{-\lfloor x \rfloor}),\ n \to +\infty,\ x \in \mathbb{R}.$$

Dem.: Do Teorema 3.3 decorre que também $\{X_n\}$ tem margens na classe $\mathcal{C}^*_{\mathcal{A}}(r^*)$ e assim

$$k_n(1 - F_X(x+n)) \sim n^{-\xi} r^{n - \lfloor x+n \rfloor} (n+x)^{\xi} \sim r^{-\lfloor x \rfloor},\ n \to +\infty.$$

Uma vez que a sucessão é q-dependente a condição $D_{k_n}(u_n)$ é trivialmente verificada. Para estabelecer a condição $D'_{k_n}(u_n)$ usa-se primeiro a q-dependência do processo para obter

$$\begin{aligned} k_n \sum_{j=2}^{[k_n/s_n]} & P(X_1 > u_n, X_j > u_n) \\ &= k_n \sum_{j=2}^{q+1} P(X_1 > u_n, X_j > u_n) + \frac{1}{s_n} (k_n P(X_1 > u_n))^2 \\ &\leq k_n \sum_{j=2}^{q+1} P(X_1 + X_j > 2u_n) + o_n(1), \end{aligned}$$

atendendo a que $\{k_n(1 - F_X(x+n))\}$ é limitada e a que $s_n \to +\infty$. Pela desigualdade de Markov e devido ao lema anterior, obtemos

$$\begin{aligned} P(X_1 + X_j > 2u_n) &= P((1+h)^{X_1 + X_j} > (1+h)^{2u_n}) \\ &\leq \frac{E((1+h)^{X_1+X_j})}{(1+h)^{2u_n}} \leq \frac{C}{(1+h)^{2u_n}}, \end{aligned}$$

para h tal que $(1+\beta^*)h + \beta^* h^2 < r - 1$. Uma vez que existe h tal que $(1+\beta^*)h + \beta^* h^2 < r - 1$ e $(1+h)^2 > r$, seja $\theta > 1$ tal que $(1+h)^2 = r^\theta$. Temos então

$$k_n \sum_{j=2}^{q+1} P(X_1 > u_n, X_j > u_n) \leq C_1 \frac{n^\xi r^n}{r^{\theta n}} = o_n(1), n \to +\infty.$$

■

Agradecimentos

O trabalho da segunda autora foi parcialmente apoiado pelo Centro de Matemática da Universidade de Coimbra - UID/MAT/00324/2013, financiado pelo Governo Português através da FCT/MCTES e co-financiado pelo Fundo Europeu de Desenvolvimento Regional através do Acordo de Parceria PT2020.

Referências

[1] Al-Osh, M. e Alzaid, A. (1988). Integer-valued moving average (INMA) process. *Stat. Papers* 29, 281-300.

[2] Anderson, C.W. (1970). Extreme value theory for a class of discrete distribution with applications to some stochastic processes. *Journal of Applied Probability* 7, 99–113.

[3] Steutel, F.W., van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability* 7, 893–899.

[4] Hall, A. (1996). Maximum term of a particular autoregressive sequence with discrete margins. *Communications in Statistics - Theory and Methods* 25, 721–736.

[5] Hall, A. (2003). Extremes of integer-valued moving average models with exponential type tails. *Extremes* 6, 361-379.

[6] Leadbetter, M. R., Lindgren, G. e Rootzén, H. (1983). Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, Berlin.

[7] McKenzie, E. (1986). Auto regressive-moving-average processes with negative binomial and geometric marginal distribution. *Advances in Applied Probability* 18, 679–705.

[8] Temido, M.G. (2002). Domínios de atracção de funções de distribuição discretas. In Carvalho, L. *et* al. (eds): *Novos Rumos em Estatística*, 415–426, Edições SPE.

[9] Temido, M.G., Canto e Castro, L. (2003). Max-semistable laws in extremes of stationary random sequences. *Theory of Probability and its Applications* 47, 365–374.

Metodologias de classificação baseadas em testes compostos: um estudo comparativo via simulação

Ricardo Sousa
Escola Superior de Tecnologia da Saúde de Lisboa, Instituto Politécnico 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*

Miguel Felgueiras
Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, Centro de Investigação em Gestão para a Sustentabilidade, *mfelg@ipleiria.pt*

Palavras–chave: classificação, custo relativo, especificidade, sensibilidade.

Resumo: Neste trabalho são comparados, via simulação, os desempenhos de distintas metodologias de classificação baseadas em testes compostos, nomeadamente o custo relativo (eficiência) e a probabilidade de erro de classificação (fiabilidade).

1 Introdução

Os testes compostos são efetuados usando um sangue combinado, isto é, uma mistura de sangue de n indivíduos. De facto, quando a taxa de prevalência da infeção é diminuta, estes testes podem ser utilizados na identificação de todos os indivíduos infetados da população com o objetivo de economizar recursos, uma vez que são necessários menos testes. O preço a pagar por este aumento da eficiência traduz-se na diminuição da fiabilidade, uma vez que a probabilidade de ocorrência de erros de classificação nos testes compostos é superior à dos testes individuais, verificando-se, nomeadamente, uma sensibilidade inferior [5]. Granado [2] comparou distintas metodologias de classificação (ensaios individuais, algoritmos hierárquicos e não hierárquicos, com e sem utilização de *masterpool*, cf. [4]), nomeadamente quanto ao seu custo relativo, definido como o número esperado de testes para a classificação de 100 indivíduos, e à sua probabilidade de erro de classificação medida pela especificidade e sensibilidade associadas a cada metodologia. Todavia, o seu estudo é restrito a testes qualitativos, nos quais é suficiente identificar a presença ou a ausência de uma qualquer substância no sangue composto, considerando ausência de efeito de diluição e, por conseguinte, utilizando a sensibilidade e a especificidade dos testes compostos iguais às correspondentes medidas dos testes individuais. Neste trabalho estende-se este estudo a testes quantitativos, nos quais a classificação é realizada em função da quantidade de determinada substância, sendo o indivíduo classificado como infetado se essa quantidade for superior a um determinado ponto de corte t, incluindo o efeito de diluição na modelação e recorrendo aos dois procedimentos de testes quantitativos compostos utilizados em [7, 8, 9]. Tendo em consideração que o peso da cauda direita da distribuição é o fator capital para a determinação do desempenho dos testes compostos (cf. [3, 6, 7]), a análise será restrita a três distribuições com caudas pesadas (Pareto, Weibull e Lévy), sendo a distribuição de um indivíduo infetado igual à de um individuo saudável com alteração de localização, de forma a serem comparados diferentes valores para a medida ϕ, de qualidade

do teste individual, definida em [8].
Deste modo, foram simulados potenciais cenários de caracterização de uma população, de forma a identificar, para cada caso, a melhor metodologia de classificação a aplicar. Para tal, na Secção 2 são apresentadas as diferentes metodologias de classificação e na Secção 3 são explicados os procedimentos aplicados na realização dos testes quantitativos compostos. Por fim, na Secção 4, são apresentados e comentados alguns dos resultados obtidos nas simulações realizadas.

2 Metodologias de classificação

Os testes compostos são utilizados em variadíssimas situações e têm como principal objetivo aceder à informação contida nas amostras individuais a custos reduzidos. Foram introduzidos na estatística durante a segunda gerra mundial por Dorfman [1] com o objetivo de determinar a dimensão n ótima para cada grupo em função da taxa de prevalência. Desta forma, Dorfman pretendeu minimizar o número de testes necessários para a identificação de todos os soldados americanos infetados com sífilis. Na *metodologia de Dorfman*, é realizado um teste composto a cada grupo. Se o resultado do teste composto for negativo, todos os elementos do grupo estão livres da infeção. No caso de um resultado positivo, um ou mais elementos do grupo estão infetados e todos terão de ser testados individualmente. Posteriormente, surgiram algoritmos mais complexos, onde perante um teste composto positivo se procede à divisão sucessiva das amostras compostas em subamostras de menor dimensão até que, em última análise, sejam realizados testes individuais (*algoritmos hierárquicos*). Estas metodologias partilham o princípio fundamental da *metodologia de Dorfman*, a qual consiste em iniciar a deteção de infetados com testes compostos e apenas realizar testes individuais nos indivíduos suspeitos. Deste modo, a *metodologia de Dorfman*, que tem 2 etapas, pode ser alargada a 3 ou mais etapas, onde as amostras com resultado positivo são novamente testadas em grupos de menor dimensão até que, chegando à última etapa, são realizados testes individuais. Neste trabalho, como exemplo de um *algoritmo*

hierárquico, optou-se pela divisão da amostra composta com resultado positivo em duas subamostras de igual dimensão (se n for ímpar as subamostras terão dimensão $\frac{n-1}{2}$ e $\frac{n+1}{2}$), sendo posteriormente aplicados testes individuais aos elementos das subamostras com resultado positivo.
Caso o número de indivíduos seja um quadrado perfeito, os *algoritmos não hierárquicos*, baseados em *arrays*, são a alternativa mais comum. A sua versão mais simples corresponde a uma tabela quadrada, denotada por 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 P_r e P_c o número de linhas e colunas com testes positivos, respetivamente. Se $\max(P_r,P_c) = 0$ os n^2 indivíduos são classificados como saudáveis. Se $\min(P_r,P_c) \geq 1$ realizam-se testes individuais a todos os indivíduos situados nas interseções de linhas e colunas infetadas. Por fim, se $\min(P_r,P_c) = 0$ e $\max(P_r,P_c) \geq 1$ será necessário testar individualmente todos os indivíduos na(s) linha(s) (ou coluna(s)) positiva(s). A notação A2 $(n:1)$ deriva de estarmos a utilizar *arrays* de duas dimensões (poder-se-ia utilizar de 3 ou mais dimensões) e $n:1$ representa a aplicação de testes a grupos com n indivíduos (1.ª fase) e posteriormente testes individuais (2.ª fase). Uma variante desta metodologia, denominada por MA2 $\left(n^2:n:1\right)$, inclui a realização prévia de um teste composto aos 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 A2 $(n:1)$. Assim, começamos com um teste a n^2 indivíduos (1.ª fase), depois (caso o primeiro seja positivo) são realizados $2n$ testes a grupos com n indivíduos (2.ª fase) e, por fim, testes individuais aos indivíduos suspeitos (3.ª fase), daí $n^2:n:1$.

3 Testes compostos quantitativos

Consideremos uma população composta por N indivíduos com probabilidade de infeção comum e igual a p. Deste modo, podemos ca-

racterizar os membros da população através de variáveis aleatórias (v.a.) $X_i, i = 1,2,\ldots,N$, com distribuição de Bernoulli de parâmetro p, onde X_i assume os valores 1 ou 0 consoante o i-ésimo indivíduo esteja ou não infetado. Ao longo do presente trabalho, nomeadamente nas simulações, assumimos que não existe interação entre os indivíduos, pelo que consideramos que as v.a. X_i são independentes. 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 D_0 se o indivíduo estiver saudável, sendo caracterizada por outra distribuição D_1 caso esteja infetado, isto é, $X_i = 0 \Rightarrow Y_i \frown \mathrm{D}_0$ e $X_i = 1 \Rightarrow Y_i \frown \mathrm{D}_1$ para $i = 1,2,\ldots,N$. É esta distinção que permitirá classificar o i-ésimo indivíduo com a informação de y_i (valor observado de Y_i). Usualmente um indivíduo é classificado como infetado se a análise acusar uma elevada quantidade da substância no sangue e, portanto, se for superior a um determinado ponto de corte predefinido t, i.e. $Y_i > t \Rightarrow X_i^+$, onde X_i^+ representa um teste positivo para o i-ésimo indivíduo. Caso contrário, o indivíduo é considerado saudável ($Y_i \leq t \Rightarrow X_i^-$, onde X_i^- representa um teste negativo para o i-ésimo indivíduo).

Para a obtenção de uma amostra composta procede-se à divisão dos N indivíduos em m grupos de dimensão n e retira-se uma determinada quantidade de sangue de cada um dos n elementos do grupo, que posteriormente é misturada de forma homogénea, pelo que vamos observar no sangue composto um valor próximo da média da quantidade de substância (valores Y_i) dos elementos que compõem esse grupo. Se o teste ao grupo der negativo (represente-se por $X^{[-,n]}$ um teste com resultado negativo aplicado a um grupo de n indivíduos) conclui-se que nenhum elemento do grupo está infetado ($\sum_{i=1}^{n} X_i = 0$), ou seja, que $Y_i \leq t$ para $i = 1,2,\ldots,n$, e portanto $\mathrm{M}_n = \max(Y_1,\ldots,Y_n)$ verifica $\mathrm{M}_n \leq t$. Caso contrário (denote-se por $X^{[+,n]}$ um teste com resultado positivo aplicado a um grupo de n indivíduos), existirá pelo menos um elemento contaminado no grupo ($\sum_{i=1}^{n} X_i \geq 1$) e, por conseguinte, o máximo do grupo deverá ultrapassar o ponto de corte t ($\mathrm{M}_n > t$). Constata-se assim que, ao

efetuarmos um teste quantitativo composto, pretendemos determinar se existe algum elemento infetado no grupo e, como tal, estamos interessados em averiguar se o máximo do grupo é superior ao ponto de corte t. Uma vez que apenas temos acesso ao valor da média do grupo, o grupo será classificado como infetado caso a média do grupo seja superior a um determinado ponto de corte t'. Por este motivo, a correlação existente entre o máximo e a média terá um papel preponderante na avaliação da qualidade do teste (cf. [6]).

Todavia, a realização de testes compostos pode originar erros de classificação, que são frequentemente avaliados pela sensibilidade e pela especificidade do teste. Assim, denote-se por $\varphi_{e}^{[n]} \in (0,1]$ a especificidade de um teste composto — probabilidade de obter um teste negativo num grupo saudável, i.e. $\mathbb{P}\left(X^{[-,n]} | \sum_{i=1}^{n} X_i = 0\right)$; e seja $\varphi_{s}^{[n]} \in (0,1]$ a sensibilidade do teste composto — probabilidade de obter um teste positivo num grupo infetado, i.e. $\mathbb{P}\left(X^{[+,n]} | \sum_{i=1}^{n} X_i \geq 1\right)$. Consequentemente $1-\varphi_{s}^{[n]}$ representará a probabilidade de um falso negativo e $1-\varphi_{e}^{[n]}$ a probabilidade de um falso positivo. Notemos igualmente que a sensibilidade de um teste composto depende do número de indivíduos infetados no grupo. Se representarmos por $\varphi_{s}^{[k,n]}$, com $k = 1, \ldots, n$, a sensibilidade quando há k indivíduos infetados no grupo, i.e. $\varphi_{s}^{[k,n]} = \mathbb{P}\left(X^{[+,n]} | \sum_{i=1}^{n} X_i = k\right)$, então $\varphi_{s}^{[1,n]} \leq \varphi_{s}^{[2,n]} \leq \ldots \leq \varphi_{s}^{[n,n]}$ (efeito de diluição e consequente rarefação, sendo $\varphi_{s}^{[n]}$ uma média ponderada dos $\varphi_{s}^{[k,n]}$ (cf. [5]).

Analisemos agora os procedimentos de realização dos testes quantitativos compostos. Conforme referido, num teste composto pretende-se averiguar a existência de pelo menos um indivíduo infetado e, por conseguinte, utilizando o procedimento aplicado nos testes individuais (na qual um indivíduo é classificado como infetado se $y_i > t$), a análise conjunta vai identificar se o máximo do grupo é superior ao ponto de corte t. Dado que só dispomos de informação da média do grupo, os testes compostos efetuados baseiam-se na quantificação da informação sobre o máximo dada pela média do grupo, através da definição de um ponto de corte t' para a média do grupo. Note-se que a eventual existência de pelo menos um indivíduo infetado no grupo

implica que o máximo excede t, o que terá obviamente influência no valor da média.
Por outro lado, analogamente ao que acontece nos testes de hipóteses com os erros de tipo I e de tipo II, não é possível aumentar simultaneamente a sensibilidade e a especificidade do teste e, por conseguinte, a estratégia passa por controlar uma das medidas e observar o comportamento da outra. No procedimento tradicional de realização de um teste composto (Procedimento M_1) são consideradas as hipóteses:

$$\mathrm{H}_0 : \sum_{i=1}^{n} X_i = 0 \quad \text{versus} \quad \mathrm{H}_1 : \sum_{i=1}^{n} X_i \geq 1$$

para as quais o tamanho do teste

$$\alpha = \mathbb{P}\left(X^{[+,n]} | \sum_{i=1}^{n} X_i = 0\right) = 1 - \varphi_{e}^{[n]}$$

corresponde à probabilidade de um falso positivo. Neste procedimento estamos a controlar a especificidade uma vez que $\varphi_{e}^{[n]} = 1-\alpha$. Contudo, em algumas situações (doenças contagiosas) é importante controlar os falsos negativos. Assim, a permuta das hipóteses no procedimento M_1 dá origem ao procedimento M_2 cujas hipóteses são:

$$\mathrm{H}_0 : \sum_{i=1}^{n} X_i \geq 1 \quad \text{versus} \quad \mathrm{H}_1 : \sum_{i=1}^{n} X_i = 0.$$

Neste procedimento o tamanho do teste é dado por

$$\alpha = \mathbb{P}\left(X^{[-,n]} | \sum_{i=1}^{n} X_i \geq 1\right) = 1 - \varphi_{s}^{[n]}$$

que corresponde à probabilidade de um falso negativo, pelo que estamos a controlar a sensibilidade uma vez que $\varphi_{s}^{[n]} = 1 - \alpha$. A possibilidade de termos entre 1 e n infetados em H_0 dificulta a determinação do ponto de corte. Assim, para contornar esta dificuldade, implementamos um procedimento simplificado (Procedimento M_2^*) cujas hipóteses são:

$$H_0 : \sum_{i=1}^{n} X_i = 1 \quad \text{versus} \quad H_1 : \sum_{i=1}^{n} X_i = 0.$$

O tamanho do teste é dado por

$$\alpha = \mathbb{P}\left(X^{[-,n]} | \sum_{i=1}^{n} X_i = 1\right) = 1 - \varphi_{s}^{[1,n]}$$

e, por conseguinte, α irá determinar o valor de $\varphi_{s}^{[1,n]}$. Estamos, mais uma vez, a controlar a sensibilidade do teste. Sublinhemos que os resultados da aplicação do procedimento M_2^* são muito semelhantes aos do procedimento M_2 uma vez que a probabilidade de encontrar um grupo com mais do que um infetado é muito diminuta [5] quando a taxa de prevalência é baixa. Além disso, atendendo a que a presença de um indivíduo infetado corresponde ao pior cenário (sensibilidade mínima), pois $\varphi_{s}^{[1,n]} \leq \varphi_{s}^{[n]}$, estamos também a controlar a sensibilidade total.

4 Simulações: resultados e comentários

As simulações foram realizadas recorrendo ao *software* R, tendo-se utilizado um nível de significância $\alpha = 0{,}05$ e uma taxa de prevalência $p = 0{,}01$ (outros valores foram utilizados, considerando sempre taxas de prevalência baixas, tendo-se obtido conclusões semelhantes às apresentadas). Foram efetuadas 10^5 réplicas em cada simulação e foram considerados os casos em que D_0 engloba os modelos Weibull, Pareto e Lévy. Restringimos, neste trabalho, a análise às situações em que as distribuições D_0 (saudáveis) e D_1 (infetados) são iguais com uma alteração de localização, isto é, $D_1 = \mu_0 + D_0$. O valor μ_0, que define a distância de localização entre os saudáveis e os infetados, é determinado de forma a garantir que $\phi = \varphi_{s}^{[1]} = \varphi_{e}^{[1]}$ com $\phi = \{0{,}95; 0{,}999\}$, onde $\varphi_{s}^{[1]}$ e $\varphi_{e}^{[1]}$ representam respetivamente a sensibilidade e a especificidade do teste individual e ϕ a medida de qualidade dos testes individuais proposta em [8].

Para a determinação do ponto de corte utilizou-se o percentil 95 de 10^5 grupos simulados de n indivíduos saudáveis no procedimento M_1, e no procedimento M_2^* utilizou-se o percentil 5 de 10^5 grupos simulados de n indivíduos dos quais $n-1$ são saudáveis e 1 está infetado.

Para analisar o desempenho em cada situação, foram determinados a sensibilidade φ_{s}, a especificidade φ_{e}, o valor preditivo positivo VPP

Tabela 1: Simulação utilizando metodologias hierárquicas

	Metodologia de Dorfman $Y_i \frown$ **Lévy**									
	Procedimento $\mathbf{M_1}$					Procedimento $\mathbf{M_2}$				
$\phi = ,95$	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
$n = 2$	5,99	**97,37**	2,24	99,04	55,05	**94,86**	94,99	16,02	99,95	61,64
$n = 3$	5,74	**98,19**	3,17	99,02	38,33	**95,21**	95,09	16,68	99,95	50,32
$n = 5$	2,65	**98,80**	2,18	99,02	25,03	**95,06**	94,86	15,89	99,80	47,86
$n = 7$	2,82	**99,07**	2,96	99,03	19,29	**94,86**	94,92	15,12	99,93	51,45
$n = 10$	4,94	**99,28**	6,49	99,04	14,09	**95,54**	94,93	15,99	99,95	60,02
$n = 20$	5,19	**99,50**	9,45	99,05	10,11	**95,98**	94,91	15,97	99,96	80,64
$n = 30$	1,11	**99,61**	2,79	99,01	8,19	**95,32**	95,03	15,93	99,94	89,68
$\phi = ,999$	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
$n = 2$	100	**97,31**	27,00	100	56,98	**95,28**	99,86	86,74	99,95	52,05
$n = 3$	100	**98,12**	34,39	100	41,20	**95,56**	98,97	47,70	99,96	38,73
$n = 5$	100	**98,63**	41,11	100	29,53	**95,99**	98,74	42,13	99,96	28,88
$n = 7$	13,04	**99,06**	12,99	99,09	19,64	**95,58**	98,60	41,66	99,95	27,03
$n = 10$	6,53	**99,23**	7,77	99,08	15,30	**95,31**	98,48	38,23	99,95	26,55
$n = 20$	4,40	**99,53**	8,84	99,02	9,75	**96,53**	97,33	27,15	99,96	41,59
$n = 30$	6,45	**99,54**	12,6	99,04	8,89	**95,99**	97,10	25,47	99,96	48,75
	Metodologia Hierárquica $Y_i \frown$ **Pareto** (5)									
	Procedimento $\mathbf{M_1}$					Procedimento $\mathbf{M_2}$				
$\phi = ,95$	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
$n = 2$	33,85	**97,48**	12,22	99,3	58,36	**91,76**	95,16	16,42	99,91	71,21
$n = 3$	20,16	**98,21**	10,29	99,18	40,09	**89,41**	95,14	15,76	99,89	62,84
$n = 5$	13,66	**98,82**	10,58	99,12	25,10	**89,27**	95,24	16,07	99,89	58,45
$n = 7$	11,37	**99,04**	10,47	99,13	18,80	**87,95**	95,26	15,45	99,88	60,01
$n = 10$	9,52	**99,24**	11,14	99,10	13,98	**88,03**	95,25	15,67	99,87	64,90
$n = 20$	7,32	**99,46**	11,98	99,07	8,54	**88,54**	95,26	15,79	99,88	77,09
$n = 30$	6,49	**99,55**	12,59	99,06	6,74	**88,90**	95,27	15,93	99,88	82,72
$\phi = ,999$	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
$n = 2$	100	**97,49**	28,16	100	60,24	**93,28**	99,90	89,96	99,93	53,14
$n = 3$	100	**98,21**	35,59	100	42,81	**92,47**	99,82	83,48	99,93	37,26
$n = 5$	100	**98,8**	45,64	100	28,97	**92,26**	99,72	76,93	99,92	25,27
$n = 7$	100	**99,00**	49,92	100	23,50	**92,46**	99,60	69,76	99,92	20,88
$n = 10$	99,96	**99,11**	51,15	99,92	19,54	**92,44**	99,39	60,50	99,92	18,79
$n = 20$	91,92	**99,18**	42,53	99,6	14,42	**92,34**	98,61	40,06	99,92	22,45
$n = 30$	60,62	**99,17**	37,26	99,49	13,07	**93,26**	97,94	31,20	99,93	30,16

e o valor preditivo negativo VPN de cada metodologia (conforme definidos em [8, 9]), bem como o custo relativo CR (número médio de testes realizados para a classificação de cada 100 indivíduos). As Tabelas 1 e 2 apresentam alguns do resultados obtidos nas simulações, tendo sido escolhidas distintas situações para ilustrar as conclusões que apresentamos. Contudo, tais conclusões não são baseadas unicamente nestes resultados (foram aplicadas todas as metodologias em todos os casos analisados). Todavia, por restrição de espaço, não é possível expor todos os resultados no presente artigo.
Deste modo, da análise aos resultados obtidos salientamos as seguintes conclusões:

→ somente em casos com testes individuais com excelente desempenho ($\phi \approx 1$) podem ser aplicadas, com fiabilidade, metodologias

Tabela 2: Simulação utilizando metodologias não hierárquicas

	Array sem masterpool $Y_i \frown$ **Weibull** (0,5)									
	Procedimento $\mathbf{M_1}$					Procedimento $\mathbf{M_2}$				
ϕ = ,95	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
n = 2	24,34	**97,14**	8,27	99,18	103,79	**94,27**	94,89	16,35	99,94	109,26
n = 3	16,37	**97,75**	6,36	99,21	70,27	**92,54**	95,00	14,71	99,93	79,03
n = 5	11,8	**98,50**	7,21	99,12	43,39	**87,67**	95,09	15,00	99,87	58,45
n = 7	9,78	**98,80**	7,58	99,09	31,94	**86,94**	95,08	15,15	99,86	56,23
n = 10	7,80	**99,08**	7,82	99,08	23,24	**87,38**	95,10	15,11	99,87	61,26
n = 20	4,21	**99,55**	8,53	99,04	12,23	**87,07**	95,18	15,32	99,86	75,98
n = 30	2,61	**99,73**	8,97	99,02	8,12	**87,95**	95,19	15,62	99,87	82,59
ϕ = ,999	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
n = 2	100	**97,07**	24,33	100	104,57	**98,93**	99,84	85,61	99,99	101,15
n = 3	100	**97,93**	32,21	100	70,90	**99,09**	99,81	83,90	99,99	68,10
n = 5	100	**98,63**	42,60	100	44,13	**97,85**	99,64	73,56	99,98	42,10
n = 7	100	**99,00**	50,19	100	32,57	**98,74**	99,56	68,58	99,96	31,24
n = 10	100	**99,34**	60,25	100	23,83	**93,24**	99,50	65,22	99,93	23,35
n = 20	52,41	**99,53**	53,00	99,52	13,55	**91,66**	99,08	50,31	99,91	17,91
n = 30	36,25	**99,56**	45,79	99,35	10,11	**92,94**	98,35	36,41	99,93	24
	Array com masterpool $Y_i \frown$ **Weibull** (0,25)									
	Procedimento $\mathbf{M_1}$					Procedimento $\mathbf{M_2}$				
ϕ = ,95	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
n = 2	4,32	**98,6**	3,15	98,99	31,77	**91,37**	94,97	16,06	99,90	55,65
n = 3	1,14	**99,38**	1,78	99,03	14,99	**89,74**	95,07	15,19	99,89	52,38
n = 5	0,53	**99,70**	1,75	99,02	6,48	**84,47**	95,09	14,62	99,84	49,51
n = 7	0,34	**99,84**	2,15	99,00	3,70	**85,16**	95,03	14,83	99,84	46,76
n = 10	0,19	**99,91**	2,11	98,99	2,13	**84,11**	95,11	14,97	99,83	49,24
n = 20	0,12	**99,96**	2,98	99,01	0,88	**84,85**	95,11	14,91	99,84	69,05
n = 30	0,08	**99,80**	3,57	99,00	0,54	**85,44**	95,20	15,29	99,85	79,32
ϕ = ,999	φ_s	φ_e	VPP	VPN	CR	φ_s	φ_e	VPP	VPN	CR
n = 2	100	**98,63**	43,33	100	36,35	**95,67**	99,87	88,25	99,95	30,56
n = 3	100	**99,23**	57,27	100	22,25	**93,79**	99,87	88,04	99,94	18,69
n = 5	100	**99,53**	67,96	100	16,53	**93,86**	99,85	85,98	99,94	15,12
n = 7	62,25	**99,61**	61,84	99,62	10,55	**93,90**	99,84	85,76	99,94	16,09
n = 10	62,44	**99,67**	65,47	99,62	9,45	**92,38**	99,81	82,94	99,92	17,03
n = 20	76,51	**99,60**	65,91	99,76	11,23	**92,16**	99,68	74,27	99,92	14,67
n = 30	92,71	**99,44**	62,37	99,93	14,92	**92,54**	99,48	62,47	99,92	15,27

baseadas em testes compostos. De facto, se $\phi = 0{,}95$, ou ainda mais baixo, a probabilidade de erros de má classificação torna-se bastante elevada.

→ O procedimento M_2 é mais estável do que M_1 quando aumentamos a dimensão n do grupo (apesar de M_1, de uma forma geral, ter menores valores de CR, isto é, ser mais eficiente).

→ O procedimento M_1 é, em alguns casos, muito instável dado que um reduzido aumento da dimensão do grupo provoca uma redução significativa no valor da sensibilidade do teste (efeito de diluição e consequente rarefação).

→ Com $\phi = 0{,}999$ a quase totalidade dos testes demonstra uma boa performance. Contudo, com $\phi = 0{,}95$, M_1 tem resultados fracos

e M_2 tem resultados razoáveis. Deste modo, M_2 parece ser mais robusto no que respeita ao valor de ϕ, mas quando o teste individual é fiável (ϕ elevado) M_1 atinge uma melhor performance em algumas medidas.

Por fim, saliente-se que não há nenhuma metodologia de classificação que apresente sempre melhores resultados que as restantes, pois nas simulações realizadas encontramos situações nas quais cada uma das metodologias apresentou melhor performance que as suas concorrentes. Assim, cada situação deve ser analisada de forma a ser identificada a melhor metodologia (e dimensão dos grupos) a aplicar.

5 Conclusões

A seleção da melhor metodologia depende dos objetivos pretendidos, nomeadamente se o principal objetivo é a eficiência ou a fiabilidade. De facto, o recurso a metodologias de classificação baseadas em testes compostos permite, em situações com taxas de prevalência baixa, obter um diminuição significativa no número de testes a realizar, mas, por vezes, à custa de um aumento significativo da probabilidade de má classificação. Todavia, conforme ficou demonstrado nas simulações realizadas, quando a qualidade dos testes individuais é elevada (ϕ próximo da unidade) a fiabilidade mantém-se em valores elevados, sendo mais vantajoso o recurso a este tipo de metodologias. Porém, é sempre necessária uma análise casuística, pois a metodologia mais adequada para cada situação depende das suas características. Além disso, há que evitar o risco, associado a estas metodologias, de cair em situações com baixa fiabilidade, uma vez que em algumas situações a sensibilidade e/ou especificidade são extremamente sensíveis a variações da dimensão dos grupos.

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto UID/MAT/00006/2013.

Referências

[1] Dorfman, R. (1943). The detection of defective members in large populations. *Ann. Math. Statistics* 14, 436–440.

[2] Granado, A. (2014). *Análises clínicas compostas: um estudo crítico via simulação*. Dissertação de Mestrado, Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

[3] 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*, Springer, 55–63.

[4] 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.

[5] 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*, 179–189.

[6] 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* 44(10), 2473–2491.

[7] Santos, R., Martins, J.P. e Felgueiras, M. (2014). Medidas para avaliar a utilização de testes compostos. *Atas do XXI Congresso Anual da Sociedade Portuguesa de Estatística*, 267–278.

[8] Santos, R., Felgueiras, M., Martins, J.P. (2015). Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification. In Kitsos, C. *et al.* (eds.): *Risk Assessment Challenges: Theory and Practice, Springer Proceedings in Mathematics and Statistics* 136, 85–98.

[9] Santos, R., Martins, J.P., Felgueiras, M. (2015). An Overview of Quantitative Continuous Compound Tests. In Bourguignon, J.P.*et al.* (Eds.): *Dynamics, Games and Science, CIM Series in Mathematical Science*s 1, 627–641.

Efeito de uma variável explicativa na modelação de uma trajetória latente: Estudo de simulação

Paula C.R. Vicente
ULHT - Escola de Ciências Económicas e das Organizações;
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit(BRU-IUL), Lisboa, Portugal, *p951@ulusofona.pt*

Maria de Fátima Salgueiro
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit(BRU-IUL), Lisboa, Portugal, *fatima.salgueiro@iscte.pt*

Palavras–chave: Dados Longitudinais, Modelos com Trajetória Latente, *Planned Missing Design*

Resumo: Este trabalho consiste num estudo de simulação, com o objetivo de perceber qual a dimensão mínima necessária para a amostra, de modo a que o efeito de uma variável explicativa seja detetado, ao estimar um modelo com trajetória latente. São geradas amostras com dados completos e com um padrão de omissões resultantes de um *planned missing design*.

1 Introdução

A modelação de trajetórias de mudança ao longo do tempo é muitas vezes o objetivo dos investigadores em diversas áreas do conhecimento, seja no estudo do desenvolvimento da personalidade, seja para uma melhor compreensão de como evoluem os comportamentos sociais, para um determinado conjunto de indivíduos. Todavia, este objetivo requer um extenso conjunto de dados, que consiste em medidas repetidas de variáveis, e na análise desses dados recorrendo a modelos longitudinais, designadamente modelos com trajetória la-

tente. Usualmente conhecidos por *Latent (Growth) Curve Models*, permitem capturar informações sobre as diferenças inter-individuais na mudança intra-individual ao longo do tempo, sendo ainda possível incorporar no modelo variáveis explicativas das trajetórias - ver Bollen e Curran [1].

A existência de não respostas é um fenómeno bastante comum em estudos por inquérito, sendo praticamente impossível que não ocorra quando o estudo é em painel. Num estudo longitudinal, a principal causa da ocorrência de omissões é o abandono dos participantes, podendo ainda ocorrer entradas tardias no estudo ou não resposta intermitente. No entanto, as omissões também podem resultar do desenho do estudo planeado pelo investigador. Num *planned missing design*, os dados em falta ocorrem como uma opção do investigador, com a finalidade de minimizar o esforço de inquirição e consequentemente de aumentar a qualidade dos dados disponíveis, Enders [3]. Um painel rotativo é um exemplo de um estudo em que as principais omissões são planeadas pelo investigador.

Na escolha da técnica estatística para lidar com dados que apresentam não respostas é necessário ter em consideração o mecanismo de omissão dos dados. De acordo com Rubin [6] existem três mecanismos de omissão de dados: i) completamente aleatório (MCAR - *Missing Completely At Random*); ii) aleatório (MAR - *Missing At Random*); e iii) não aleatório (MNAR - *Missing Not At Random*). Quando é assumido um mecanismo de omissão de dados MAR ou MCAR um dos métodos de estimação mais utilizados é o *Full Information Maximum Likelihood* (FIML), Schafer e Graham [7].

É frequente na área das ciências sociais existirem dificuldades na recolha / obtenção de amostras com uma dimensão adequada face ao tipo de modelação desejado / recomendado. Assim, com este trabalho pretende-se determinar qual a dimensão mínima da amostra necessária para que seja identificado o efeito de uma variável explicativa na modelação de uma trajetória latente, tanto no caso de estudos com dados completos como em estudos que apresentem omissões resultantes do desenho amostral definido pelo investigador.

Para tal é realizado um estudo de simulação recorrendo ao pacote estatístico Mplus 7, Muthén e Muthén [5].

2 Metodologia

2.1 Modelos com Trajetória Condicionada

Os modelos com trajetória latente constituem um técnica frequentemente utilizada no estudo da mudança usando dados longitudinais. Os modelos não condicionais permitem descrever uma trajetória individual para cada indivíduo, uma trajetória média para o conjunto dos indivíduos em estudo, bem como a variabilidade em torno dessa trajetória média. Os modelos com trajetória condicionada permitem a incorporação de variáveis explicativas da trajetória. Esta trajetória latente é estimada a partir da estrutura de médias e de variâncias-covariâncias entre as medidas repetidas das variáveis observadas (y) - ver Bollen e Curran [1].

O modelo com trajetória condicionada representado na figura 1,

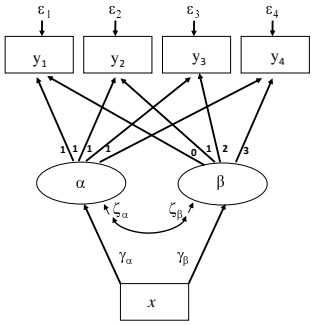


Figura 1: Diagrama de um modelo com trajetória latente condicionada, com quatro momentos temporais

é descrito pela seguinte equação de trajetória, para o indivíduo $i = 1, \cdots, N$, no momento $t = 1, \cdots, T$

$$y_{it} = \alpha_i + \lambda_t \beta_i + \varepsilon_{it}, \tag{1}$$

em que y_{it} é o valor da variável observada y para o indivíduo i, no momento t, α_i e β_i são, respetivamente, o intercepto e o declive aleatórios do indivíduo i, dados por

$$\alpha_i = \mu_\alpha + \gamma_\alpha x_i + \zeta_{\alpha_i} \tag{2}$$

e

$$\beta_i = \mu_\beta + \gamma_\beta x_i + \zeta_{\beta_i}, \tag{3}$$

com μ_α e μ_β a designar as médias do intercepto e do declive, respetivamente, no caso de existir apenas uma variável explicativa x. A equação da trajetória condicionada pode também ser obtida por

$$y_{it} = (\mu_\alpha + \lambda_t \mu_\beta) + (\gamma_\alpha + \lambda_t \gamma_\beta) x_i + (\zeta_{\alpha_i} + \lambda_t \zeta_{\beta_i} + \varepsilon_{it}), \tag{4}$$

em que γ_α e γ_β são os coeficientes da variável explicativa na equação do intercepto e do declive e podem ser interpretados de forma análoga a um modelo de regressão linear. ε_{it} representa o termo residual da trajetória traçada para o indivíduo i, no momento t, e λ_t é especificado como igual a $(t-1)$ quando é considerada uma trajetória linear. É pressuposto do modelo que o termo residual da trajetória ε_t tem distribuição normal com média zero e matriz de variâncias-covariâncias diagonal Θ_ε. Os termos residuais do intercepto e do declive, ζ_{α_i} e ζ_{β_i}, têm distribuição normal com média zero e variâncias dadas por $\psi_{\alpha\alpha}$ e $\psi_{\beta\beta}$, respetivamente, e com covariância $\psi_{\alpha\beta}$. Estes termos residuais não estão correlacionados nem com o termo residual da trajetória, nem com a variável explicativa. Note-se que no modelo com trajetória condicionada, $\psi_{\alpha\alpha}$ e $\psi_{\beta\beta}$ são variâncias condicionais, respetivamente do intercepto e do declive.
A fiabilidade do indicador y no momento t representa a proporção da variância do indicador que é explicada pelas variáveis latentes

que definem a trajetória, e é dada por

$$\frac{(\psi_{\alpha\alpha} + \lambda_t^2\psi_{\beta\beta} + 2\lambda_t\psi_{\alpha\beta})}{(\psi_{\alpha\alpha} + \lambda_t^2\psi_{\beta\beta} + 2\lambda_t\psi_{\alpha\beta} + \theta_{\varepsilon_t})}, \tag{5}$$

sendo θ_{ε_t} a variância do termo residual ε_t.

2.2 Estimação FIML

A estimação de um processo de mudança usando um modelo com trajetória latente, quando existem omissões nos dados pode ser realizada recorrendo ao método FIML. Ao contrário de outros métodos que inputam ou substituem os valores omissos, este método utiliza toda a informação disponível durante a estimação. Quando o mecanismo de omissão dos dados é aleatório e se pode assumir que os dados seguem distribuição normal multivariada, o método FIML produz estimativas dos parâmetros e erros padrão que são consistentes e eficientes.
A função a maximizar, na presença de dados completos é, para a observação i,

$$logL_i = \frac{-k}{2}log(2\pi) - \frac{1}{2}log|\Sigma| - \frac{1}{2}(y_i - \mu)^T\Sigma^{-1}(y_i - \mu) \tag{6}$$

onde k é o número de variáveis, y_i é o vetor para a observação i, μ é o vetor das médias populacionais e Σ é a matriz das variâncias-covariâncias. Com dados omissos, a função para a observação i vem dada por

$$logL_i = \frac{-k_i}{2}log(2\pi) - \frac{1}{2}log|\Sigma_i| - \frac{1}{2}(y_i - \mu)^T\Sigma_i^{-1}(y_i - \mu_i) \tag{7}$$

onde k_i representa o número de variáveis completas (ou com valor) para aquela observação e os μ_i e Σ_i estão associados apenas aos dados disponíveis. O cálculo para a função $logL_i$ para a observação i depende apenas das variáveis e dos parâmetros para os quais esse

elemento tem dados completos, Enders, [3]. A função de verosimilhança final corresponde à soma de N funções de verosimilhança, para os N elementos da amostra, sendo dada por

$$logL(\mu,\Sigma) = \sum_{i=1}^{N} logL_i. \tag{8}$$

3 Estudo de simulação

Para realizar o estudo de simulação foi escolhido o pacote estatístico Mplus. Este *software* permite gerar m amostras de dados a partir da estrutura postulada para o modelo com trajetória latente, com parâmetros populacionais definidos a priori pelo investigador. Para cada uma das m amostras geradas, e de uma forma integrada, é estimado um modelo com trajetória latente, obtendo-se, deste modo, m estimativas para cada um dos parâmetros do modelo. Se nas amostras geradas existem omissões, para efeitos de estimação é utilizada uma abordagem FIML. Para cada um dos parâmetros do modelo o Mplus disponibiliza, entre outras medidas, a média das estimativas, calculada a partir das m amostras independentes que foram geradas, bem como a proporção de amostras nas quais um teste de significância a 5% se mostrou significativo. Nos casos em que o valor do parâmetro se assume não nulo, esta proporção de amostras corresponde a uma estimativa da potência do teste. O enviesamento relativo na estimação de cada parâmetro, $B_R(\widehat{\theta})$, pode ser calculado utilizando o valor considerado como parâmetro populacional, θ, e a média das estimativas dos parâmetros obtidos nas várias amostras geradas, $E(\widehat{\theta})$, da seguinte forma

$$B_R(\widehat{\theta}) = \frac{E(\widehat{\theta}) - \theta}{\theta}. \tag{9}$$

O valor obtido de enviesamento pode ser multiplicado por 100, para obter a percentagem de enviesamento existente na estimação do parâmetro em análise. De acordo com Hoogland e Boomsma [4], apenas

para valores absolutos de enviesamento relativo inferiores a 0.05, isto é 5%, se pode concluir não existir enviesamento na estimação de um parâmetro.

4 Resultados

Neste estudo de simulação foram geradas, a partir da estrutura de modelos com trajetória latente com 3 e 4 momentos temporais, 1000 amostras de dados, com 50, 250, 500 ou 1000 observações cada. De referir que, como consequência dos pressupostos do modelo sobre a distribuição de probabilidade dos termos residuais, os dados gerados a partir da estrutura imposta por um modelo com trajetória latente têm distribuição normal. São utilizados como valores dos parâmetros populacionais que definem o modelo, $\mu_\alpha = 0$, $\mu_\beta = 0$, $\psi_{\alpha\alpha} = 1$, $\psi_{\beta\beta} = 0.2$ e $\psi_{\alpha\beta} = 0$, sendo a variância dos termos residuais fixada a valores que permitem obter uma fiabilidade de 0.5, para cada um dos indicadores, em cada um dos momentos temporais. Para o coeficiente de regressão sobre o intercepto foi estabelecido o valor 0.5, tendo a variável explicativa sido considerada como dicotómica. Para o coeficiente de regressão sobre o declive foram considerados valores de 0.1, 0.25 e 0.35. De acordo com Cohen [2], estes valores correspondem, respetivamente, a um efeito pequeno, médio e grande. São geradas amostras com dados completos e amostras com omissões resultantes do desenho do estudo planeado pelo investigador, tal como especificado na tabela 1. O padrão de omissões utilizado para a geração dos dados pretende reproduzir um padrão comum em dados obtidos através de um painel rotativo com uma dinâmica de rotatividade de 1/4 da amostra, implicando diferentes percentagens de omissões nos vários momentos temporais considerados.

A tabela 2 apresenta os resultados quanto ao enviesamento relativo na estimação dos coeficientes de regressão, bem como a proporção de amostras para a qual é rejeitada a hipótese nula de que esse parâmetro assume o valor zero num teste a 5%, que se obtêm quando são geradas amostras com dados completos de 50, 250, 500 e 1000 ob-

Y_1	Y_2	Y_3
✓	✓	-
✓	✓	✓
✓	✓	✓
-	✓	✓

Y_1	Y_2	Y_3	Y_4
✓	✓	✓	-
✓	✓	✓	✓
-	✓	✓	✓
-	-	✓	✓

Tabela 1: *Planned missing design* para estudos com três (à esquerda) e quatro (à direita) momentos temporais (✓designa valor observado; - designa valor omisso)

servações, considerando diferentes valores para os efeitos da variável explicativa sobre a trajetória e considerando um modelo com quatro momentos temporais. A análise destes resultados permite dizer que o enviesamento relativo na estimação do valor dos coeficientes de regressão sobre o intercepto e sobre o declive é inferior a 5% (valor que se encontra no limite do negligenciável), qualquer que seja a dimensão da amostra e o valor assumido pelo coeficiente de regressão sobre o declive da trajetória estimada. Todavia, para amostras de muito pequena dimensão, $N = 50$, quando o efeito da variável explicativa sobre a trajetória é pequeno, $\gamma_\beta = 0.1$, o valor do enviesamento relativo é de 8.1%, superior ao considerado negligenciável. Quanto à potência do teste da hipótese de que o coeficiente de regressão sobre o declive não é significativo, pode observar-se que este valor aumenta com o aumento da dimensão da amostra e com o valor populacional considerado para o parâmetro em discussão. De referir que, para se obter uma potência de teste de aproximadamente 60% é necessária uma amostra com 1000 observações quando é considerado um efeito pequeno, isto é, $\gamma_\beta = 0.1$. No entanto, quando é considerado um efeito médio, $\gamma_\beta = 0.25$, apenas é necessária uma amostra de 250 observações para obter uma potência do teste de 78.5%.

Na tabela 3 são apresentados os resultados quando é considerado um modelo com trajetória latente condicionada com apenas três momentos temporais. A análise destes resultados permite concluir que os valores de enviesamento relativo obtidos na estimação do efeito da

		Enviesamento				Potência do teste			
γ_β	N	50	250	500	1000	50	250	500	1000
0.1	γ_α	<5%	<5%	<5%	<5%	0.395	0.976	1.000	1.000
	γ_β	8.1%	<5%	<5%	<5%	0.115	0.208	0.345	0.603
0.25	γ_α	<5%	<5%	<5%	<5%	0.395	0.976	1.000	1.000
	γ_β	<5%	<5%	<5%	<5%	0.284	0.785	0.967	1.000
0.35	γ_α	<5%	<5%	<5%	<5%	0.395	0.976	1.000	1.000
	γ_β	<5%	<5%	<5%	<5%	0.459	0.969	1.000	1.000

Tabela 2: Percentagem de enviesamento relativo na estimação dos coficientes de regressão de uma variável explicativa da trajetória e potência do teste (em modelos com quatro momentos temporais e amostras com dados completos)

variável explicativa sobre a trajetória estimada são superiores a 5%, quando a dimensão da amostra é inferior a 1000 observações e o efeito considerado é pequeno, 0.1. A potência do teste de significância a este mesmo parâmetro é inferior a 25% para amostras de dimensão inferior a 500 observações. No entanto, para um efeito de média magnitude (0.25), a potência do teste é de 60.5%, sendo de 86.7% se o efeito considerado é de magnitude grande, para uma amostra com 250 observações.

Em seguida são apresentados os resultados quanto ao enviesamento relativo na estimação dos coeficientes de regressão, bem como quanto à potência do teste, obtidos quando são geradas amostras com omissões devidas ao desenho amostral planeado pelo investigador, com 50, 250, 500 e 1000 observações, considerando diferentes valores para os efeitos da variável explicativa sobre a trajetória e recorrendo a um modelo com trajetória latente condicionada com quatro momentos temporais - ver tabela 4. Da comparação destes valores com os que são apresentados na tabela 2, em que foram geradas amostras com dados completos, é possível concuir que a potência do teste diminui para todos os casos considerados; por exemplo, se o efeito considerado é pequeno, para uma amostra com 1000 observações a potência

		Enviesamento				Potência do teste			
γ_β	N	50	250	500	1000	50	250	500	1000
0.1	γ_α	<5%	<5%	<5%	<5%	0.417	0.957	0.999	1.000
	γ_β	5.1%	7.5%	7%	<5%	0.081	0.135	0.250	0.426
0.25	γ_α	<5%	<5%	<5%	<5%	0.417	0.957	0.999	1.000
	γ_β	<5%	<5%	<5%	<5%	0.177	0.605	0.861	0.991
0.35	γ_α	<5%	<5%	<5%	<5%	0.417	0.957	0.999	1.000
	γ_β	<5%	<5%	<5%	<5%	0.286	0.867	0.994	1.000

Tabela 3: Percentagem de enviesamento relativo na estimação dos coficientes de regressão de uma variável explicativa da trajetória e potência do teste (em modelos com três momentos temporais e amostras com dados completos)

do teste reduz de 60.3% para 46.5%. Quanto ao enviesamento, e tal como obtido para amostras com dados completos, apenas o valor estimado do coeficiente de regressão do declive, em amostras de pequena dimensão e no caso de ser considerado um efeito de pequena magnitude, apresenta um valor de enviesamento relativo que não pode ser negligenciado, isto é, um valor de 7.1%. De referir ainda que este valor é inferior ao obtido para amostras com dados completos, consequência de algumas das amostras terem apresentado problemas aquando da estimação do modelo.

Quando é considerado um modelo com trajetória latente com três momentos temporais para gerar amostras com omissões os resultados são apresentados na tabela 5. É possível concluir que os valores de potência de teste diminuem face aos obtidos para amostras com dados completos, bem como quando comparados com os obtidos na estimação de um modelo com quatro momentos temporais e dados com omissões. De referir ainda que o enviesamento relativo obtido na estimação do coeficiente de regressão sobre o declive é não negligenciável, quando é considerado um efeito de pequena magnitude, qualquer que seja a dimensão da amostra.

		Enviesamento				Potência do teste			
γ_β	N	50	250	500	1000	50	250	500	1000
0.1	γ_α	<5%	<5%	<5%	<5%	0.281	0.849	0.987	1.000
	γ_β	7.1%	<5%	<5%	<5%	0.101	0.157	0.277	0.465
0.25	γ_α	<5%	<5%	<5%	<5%	0.281	0.849	0.987	1.000
	γ_β	<5%	<5%	<5%	<5%	0.218	0.641	0.907	0.997
0.35	γ_α	<5%	<5%	<5%	<5%	0.281	0.849	0.987	1.000
	γ_β	<5%	<5%	<5%	<5%	0.371	0.909	0.995	1.000

Tabela 4: Percentagem de enviesamento relativo na estimação dos coficientes de regressão de uma variável explicativa da trajetória e potência do teste (em modelos com quatro momentos temporais e amostras com omissões)

		Enviesamento				Potência do teste			
γ_β	N	50	250	500	1000	50	250	500	1000
0.1	γ_α	<5%	<5%	<5%	<5%	0.345	0.900	0.998	1.000
	γ_β	7.4%	7.4%	7.1%	6%	0.089	0.123	0.186	0.338
0.25	γ_α	<5%	<5%	<5%	<5%	0.345	0.900	0.998	1.000
	γ_β	<5%	<5%	<5%	<5%	0.156	0.488	0.768	0.969
0.35	γ_α	<5%	<5%	<5%	<5%	0.345	0.900	0.998	1.000
	γ_β	<5%	<5%	<5%	<5%	0.229	0.762	0.964	1.000

Tabela 5: Percentagem de enviesamento relativo na estimação dos coficientes de regressão de uma variável explicativa da trajetória e potência do teste (em modelos com três momentos temporais e amostras com omissões)

5 Discussão

Este trabalho baseia-se num estudo de simulação realizado em Mplus 7, com o objetivo de perceber qual a dimensão míníma da amostra que permite detetar o efeito de uma variável explicativa dicotómica aquando da estimação de um modelo com trajetória latente condicionada, com dados completos e com dados que apresentam não respostas que configuram um *planned missing design*. Os resultados obtidos permitem concluir que, no caso de existirem omissões, são necessárias amostras maiores para detetar efeitos de igual magnitude, face a amostras com dados completos. Conclusão análoga para modelos com menor número de momentos temporais. Quando se consideram efeitos de menor magnitude da variável explicativa sobre a trajectória, são necessárias amostras de maior dimensão para que estes efeitos se revelem significativos. Por outro lado, se as amostras são de pequena dimensão e o efeito é de pequena magnitude, existe um enviesamento relativo na estimação do parâmetro que não pode ser negligenciado, agravando-se a situação se os dados apresentam omissões, particularmente em modelos com um menor número de momentos temporais. Como em qualquer estudo de simulação, também neste existem limitações, que podem ser consideradas questões de investigação em aberto para trabalho futuro, nomeadamente, o lidar com o efeito de uma variável explicativa na modelação de uma trajetória latente, quando os dados apresentam omissões cujo mecanismo é não aleatório.

Referências

[1] Bollen, K.A., Curran, P.J. (2006). *Latent Curve Models - A Structural Equation Perspective*. John Wiley & Sons, New Jersey.

[2] Cohen, J.(1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates Publishers, New York.

[3] Enders, C.K.(2010). *Applied Missing Data*. The Guilford Press, New York.

[4] Hoogland, J.J., Boomsma, A. (1998). Robustness studies in covariance structure modelling: An overview and meta-analysis.*Sociological Methods and Research* 26, 329–367.

[5] Muthén, L.K., Muthén, B.O. (1998-2012). *Mplus user's guide, 7th edition.* Los Angeles, CA: Muthén & Muthén.

[6] Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63, 581–592.

[7] Schaffer, J.L., Graham, J. (2002). Missing Data: Our View of the state of the art.*Psychological Methods* 7, 147–177.

Autores

XXII CONGRESSO
SPE2015
SOCIEDADE PORTUGUESA
DE ESTATÍSTICA