Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *primavera de 2017*

Incerteza em Engenharia

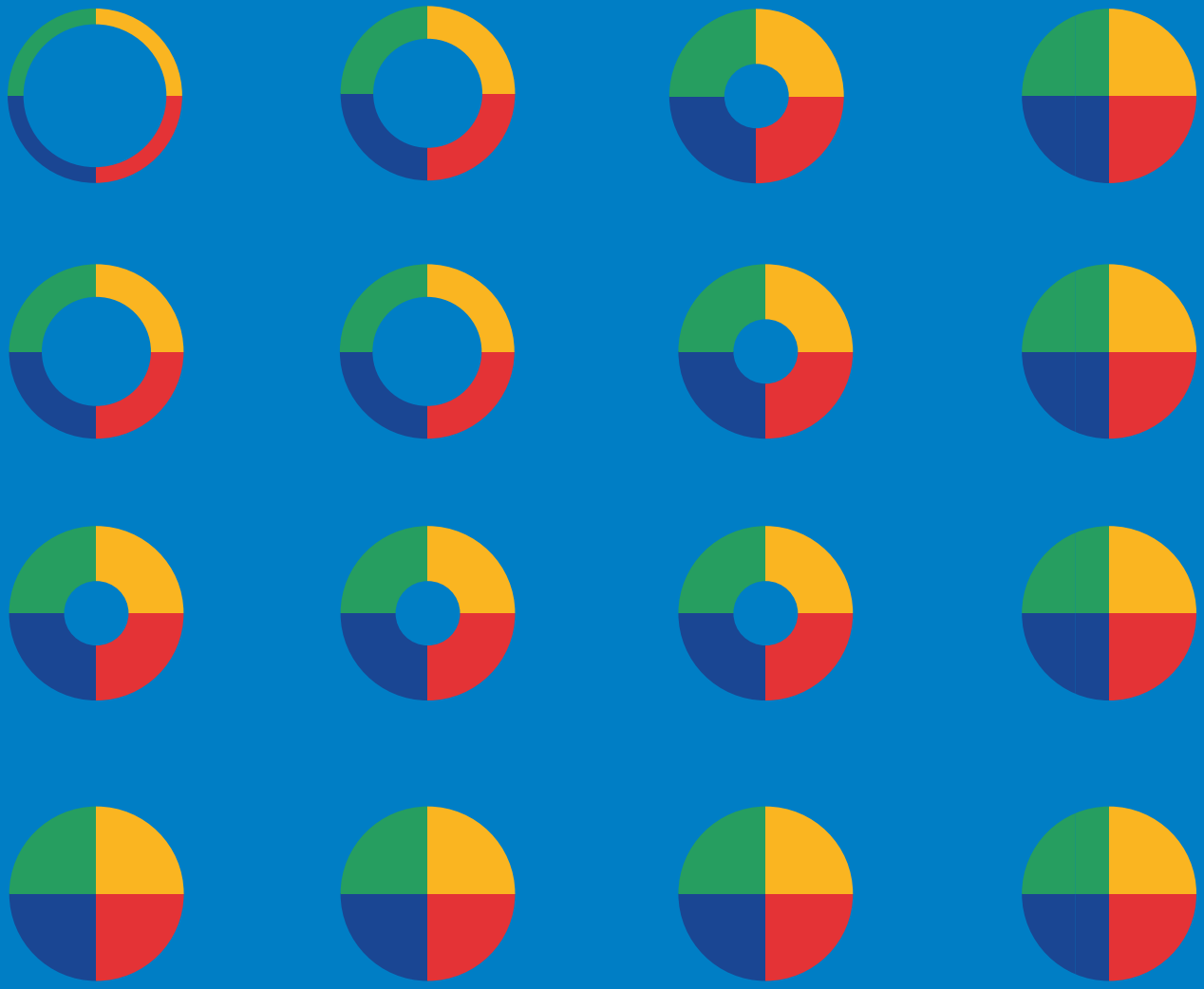
Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@fc.ul.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 500 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

PRÉMIO ESTATÍSTICO JÚNIOR 2017

Candidaturas até
26 DE MAIO DE 2017

SPE
SOCIEDADE PORTUGUESA DE ESTATÍSTICA



CONTACTOS

Sociedade Portuguesa de Estatística
Bloco C6, Piso 4 – Campo Grande
1749-016 Lisboa

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@fc.ul.pt

Editorial

… com (mais) um novo olhar para a Incerteza…

1. Esta edição do boletim corresponde ao "número um da segunda década" de uma nova série iniciada com o Boletim SPE outono 2006. Em jeito de resumo, um historial do caminho percorrido pode ser lido, praticamente em todos os números editados e na página com o título Retrospetiva.

2. Várias têm sido as áreas focadas em cada edição e ao longo do tempo. Começámos pelo Ensino e Aprendizagem da Estatística. Na diversidade editorial, já caminhámos pela Genética e pelo Desporto, nos Boletins primavera e outono de 2015 mas também pelos Métodos Estatísticos em Medicina no outono de 2012.
Desta vez versamos a *Incerteza e a Engenharia.*
A Incerteza é um termo "tão importante" que já potenciou uma "definição oficial" na União Europeia. Segundo a Diretiva Comunitária 2007/589/CE, a Incerteza define-se como "o parâmetro associado ao resultado da determinação de uma quantidade, que caracteriza a dispersão dos valores que poderiam razoavelmente ser atribuídos a essa determinada quantidade (incluindo os efeitos de fatores sistemáticos e aleatórios), expresso em percentagem e que descreve um intervalo de confiança próximo do valor médio, compreendendo 95% dos valores inferidos".
Como seria de esperar, esta definição oficial gera um sem número de textos de clarificação e de opinião.
Como sempre!
A Incerteza é um termo transversal a muitos domínios científicos, desde a Filosofia à Física, mas com muito maior ênfase e exigência de estudo em áreas como a Economia (onde envolve e agrega o risco) e a Estatística. Nestas O Acaso é um bom ponto de partida para abordar a Incerteza. Bem sabemos como acaso, a sua discussão e a sua influência, é gerador dos mais diversificados estudos estatísticos ao longo do tempo. Os autores, nesta edição, conduzem-nos numa viagem que também é guiada pelas diferentes ambientes académicos onde desenvolvem o seu trabalho e a sua investigação.
Assim, O Boletim SPE (também) é fruto do acaso. Ou será da incerteza? Com toda a certeza, é a generosa colaboração dos muitos autores que faz controlar e dominar toda a incerteza da criação e é a força que a vence e faz gerar cada edição do Boletim como exemplar único.

3. Com esta edição, estamos a melhorar e optimizar a distribuição do Boletim SPE. Além de se terem reduzido os custos de expedição, numa primeira etapa, agora atualizamos e melhoramos a rede de distribuição pelas Bibliotecas. Cabe aqui uma palavra de agradecimento a todos os colegas que ajudaram na recolha de informação atualizada e que completou o trabalho que já tinha sido levado a cabo pela secretária da SPE, Ana Maria Ponsard. Esta acção gerou uma melhor abordagem no sentido de o Boletim SPE chegar mais longe, a mais espaços e mais perto de todos os seus potenciais leitores. Assim, o Boletim SPE além de esperar que os leitores se aproximem também, a partir de agora, nas cerca de 100 Bibliotecas, toma a iniciativa de se aproximar deles para mostrar os seus conteúdos e fazer cumprir a sua missão estatutária.
A todos os colegas e leitores se solicita que - em qualquer detalhe ou em iniciativa que achem oportuno - comuniquem no sentido de melhorar a distribuição do Boletim SPE.

4. No passado dia 7 de Fevereiro, inesperada, surgiu a Notícia do falecimento do Prof. Amílcar Sernadas – atualmente distinto Professor do Instituto Superior Técnico (IST) – e que na respetiva secção adiante apresentamos.
No entanto, em testemunho pessoal, desejo salientar e relevar que este desaparecimento fez-me desfilar em memória inúmeros acontecimentos que partilhámos ao longo de muitos anos – com ele e com a sua esposa a Prof. Cristina, também do IST. Refiro alguns que, de facto, pelo pioneirismo, são parte integrante na génese da modernização do ambiente científico, académico e profissional da moderna Academia Portuguesa.

Com o Amílcar participei numa primeira tentativa de realização de um primeiro Mestrado em Probabilidades e Estatística por iniciativa política do Prof. Tiago de Oliveira, secretário de Estado da Investigação Científica. Durante vários meses tivemos aulas e fizemos trabalhos como se de um verdadeiro mestrado se tratasse. A política não permitiu que se efetivasse. Nem a inscrição foi iniciada pelo que docentes e discentes, em espaço emprestado, funcionaram em regime de voluntariado, a favor da Ciência.

Naquela atividade académica que em tudo era nova em Portugal – Mestrado era um nome que ainda não fazia parte do dicionário nos idos anos setenta – o Amílcar foi um elemento fundamental e de uma liderança científica que se foi manifestando ao longo da sua carreira profissional.

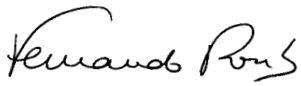
Com o Amílcar e com a sua esposa, tive a oportunidade de assistir ao crescimento do Departamento de Estatística, Investigação Operacional e Computação – o conhecido DEIOC, a origem de intensa atividade científica na Faculdade de Ciências da Universidade de Lisboa nos idos anos oitenta do século passado e, por conseguinte, também de todo o país.

Por último e não menor no que toca à Sociedade Portuguesa de Estatística, o Amílcar, recém doutorado, foi um dos onze Outorgantes na Escritura de Formação da Sociedade Portuguesa de Estatística e Investigação Operacional, em 28 de Novembro de 1980.

Com muita saudade desses tempos de grande inovação e com muita pena, vemos desaparecer, tão precocemente, um daqueles que, em Portugal, foram pioneiros na moderna ciência e do qual se esperava que, ainda muito, poderia gerar Ciência.

A Mãe Natureza, que domina a Incerteza, assim determinou!

O Tema Central do próximo Boletim SPE será *O Tema Central da Estatística – um novo olhar*.



Mensagem da Presidente

Caros Sócios da SPE,

Passado mais um ano do mandato da actual Direcção é tempo de reflectir sobre as actividades passadas e planear as actividades futuras. Este exercício foi feito na preparação da Assembleia Geral Ordinária de 2017 que terá lugar dia 31 de Março.
Num resumo breve, as actividades nortearam-se segundo os dois grandes objectivos traçados por esta Direcção:

Objectivo I - aumentar a sensibilização pública para a Estatística e aumentar a visibilidade da SPE na sociedade;

Objectivo II - aumentar a coesão interna da sociedade e apoiar o desenvolvimento da Estatística em Portugal.

Assim, no âmbito do **Objectivo I** para além das actividades já usuais *Prémio Estatístico Júnior, AEVAE e Explorística, destaca-se a presença da SPE em actividades organizadas por e para estudantes, tais como ENEMATH,* III Iberian Modeling Week e em sessões dedicadas ao ensino e à divulgação no Encontro da SPM.
De entre as actividades que contribuíram para o **Objectivo II** destaco o Prémio SPE 2016 atribuído a Filipa Alexandra Cardoso da Silva com o trabalho intitulado *Processos INGARCH Poisson compostos na modelação de séries temporais de contagem sobredispersas*; e os Prémios Iniciação à Investigação: *Modelação de Admissões e Internamentos na Urgência do Hospital de Braga* de Adriana Vieira, *Classical and Robust Symbolic Principal Component Analysis for Interval Data* de Margarida Vilela e *Barrier Option Pricing under the 2-Hypergeometric Stochastic Volatility Model* de Ruben Sousa.
Outros pontos altos do ano 2016 foram a realização do II Encontro Luso-Galaico de Biometria com aplicações à saúde, ecologia e ambiente em Santiago de Compostela e a comemoração do 36º aniversário da SPE. A SPE tem apoiado diversos eventos científicos que se realizam Portugal e são de interesse para os sócios e tem-se assegurado a representação da SPE em diversas organizações nacionais. Os sócios podem consultar os detalhes das actividades no Relatório de Actividades de 2016 que se encontra disponível, mais uma vez, na página da SPE.
A Direcção quer deixar aqui um agradecimento aos sócios que, certamente à custa de sacrifício pessoal, possibilitaram a realização das actividades no ano de 2016.
Gostaria de chamar a atenção para duas iniciativas que estão em curso no âmbito da Federação das Sociedades Nacionais, FenStatS. Uma diz respeito ao estabelecimento de um programa de acreditação de estatísticos a nível Europeu. Espero poder dar mais informação sobre o assunto em boletins futuros.
A outra diz respeito a um pedido formulado pelo Presidente da FenStatS ao Presidente da ERC, Professor Jean-Pierre Bourguignon, e assinado por 20 Sociedades Nacionais, entre as quais a SPE, para que seja instituído um painel autónomo para a área Estatística. Os sócios podem encontrar mais detalhes sobre este assunto nas páginas deste boletim.
Para finalizar lembro que decorrerá em 2017 o XXIII Congresso da SPE durante o qual se celebrará o Dia Europeu da Estatística. Assim teremos o prazer e a honra de receber em Lisboa o ESAC, representantes do EUROSTAT e do Banco Central Europeu. A organização destas comemorações será feita em colaboração com o INE e a representante Portuguesa na ESAC.
A Sociedade Portuguesa de Estatística é dos sócios e para os sócios e é, essencialmente, o que os sócios fizerem dela.
Assim, a Direcção está aberta a apoiar todas as iniciativas em prol da Estatística em Portugal.

Porto, 25 de Fevereiro de 2017

Cordiais saudações

Maria Eduarda Silva

Notícias

• XXIII Congresso da SPE

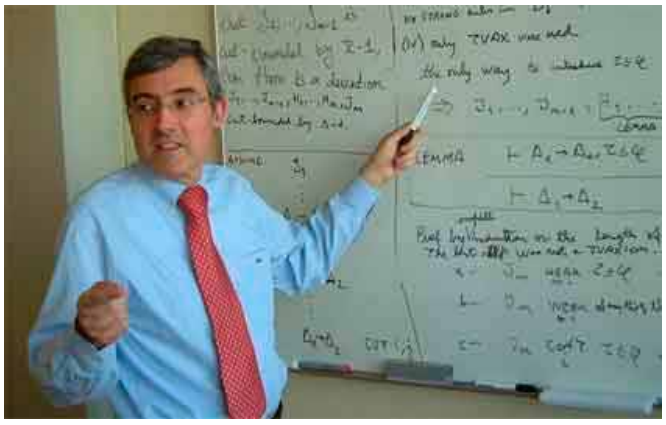
O XXIII Congresso SPE terá lugar de **18 a 21 de Outubro de 2017**, nas instalações do Instituto Universitário de Lisboa (ISCTE-IUL).
A Comissão Organizadora do Congresso é constituída pelos colegas Maria de Fátima Salgueiro, Paula Vicente, Teresa Calapez, Catarina Marques e Elizabeth Reis, da referida instituição. Para além da Presidente da SPE, Maria Eduarda Silva, são membros da Comissão Científica os colegas Conceição Amado (IST), José Manuel Gonçalves Dias (ISCTE-IUL), Maria de Fátima Salgueiro (ISCTE-IUL), Nazaré Mendes-Lopes (Universidade de Coimbra) e Paulo M.M. Rodrigues (Nova School of Business and Economics).
No dia 18 de Outubro haverá um Minicurso sobre Meta-Análise, ministrado pela colega **Fátima Brilhante**, da Universidade dos Açores.
O programa científico do Congresso inclui sessões plenárias apresentadas por oradores convidados; sessões temáticas e sessões livres.
No dia 20 de Outubro Lisboa será a "capital europeia da(s) Estatística(s)", uma vez que se comemorará o *European Statistics Day*, com a presença de representantes do *EUROSTAT*, da *FenStatS* e do *European Central Bank*.
O prazo limite para submissão de resumos para apresentação como comunicação livre (oral ou poster) é o dia 4 de Junho de 2017.
Para mais informações consultar o site do Congresso ***spe2017.ISCTE-IUL.pt*** ou usar o endereço ***spe2017@iscte.pt***.

Até Outubro, em Lisboa!

A Comissão Organizadora.

• Faleceu o Professor Amílcar Sernadas

No dia 7 de fevereiro de 2017 faleceu o Professor Amílcar Sernadas.
Amílcar Sernadas nasceu em 1952, em Angola. Licenciou-se, em 1975, em Engenharia Eletrotécnica no Instituto Superior Técnico. Doutorou-se, em 1980, na Universidade de Londres em *Computer Science*. Obteve o título de Agregado, em 1982, na Faculdade de Ciências da Universidade de Lisboa, na área científica de Análise Numérica e Computação. Era, desde 1990, Professor Catedrático no Departamento de Matemática do Instituto Superior Técnico .



As suas contribuições científicas mais importantes e que o tornaram um reconhecido cientista internacional foram na área da ciência da computação e da engenharia informática.
Com uma carreira científica excelente, o Professor Amílcar Sernadas, em 2016 foi distinguido com o título de Professor Distinto por ocasião do105º Aniversário do Instituto Superior Técnico.
Membro muito interventor em uma intensa atividade científica o seu nome fica registado na folha da génese da moderna Academia portuguesa.

O Prof. Amílcar Sernadas fica também ligado ao início da Sociedade Portuguesa de Estatística de que foi um dos sócios outorgantes na escritura de formação, em 28 de novembro de 1980.

FR

• Prémios Estatístico Júnior 2016

Trabalho classificado em 1º lugar (3º ciclo do Ensino Básico)

Título: ***Estilos de Vida***
Autores: Ana Rita Pôço Henriques
Professora orientadora: Não teve
Estabelecimento de Ensino: Escola Secundária 3º Ciclo Entroncamento

Trabalho classificado em 2º lugar (3º ciclo do Ensino Básico)

Título: ***Situações de Emergência***
Autores: Margarida Pereira Duarte e Maria João Meireles
Professor orientador: Andreia Mónica Moreira Silva.
Estabelecimento de Ensino: E B 2,3 Gonçalo Mendes Maia, Maia

Trabalho classificado em 3º lugar *ex-aequo* (3º ciclo do Ensino Básico)

Título: ***Hábitos de Higiene Oral***
Autores: Alexandre Pinto Silva, Carolina Fernandes Clemente e Maria Leitão Carvalho
Professor orientador: Carlos André Pimentel Lameirinhas
Estabelecimento de Ensino: Colégio Senhor dos Milagres, Leiria

Trabalho classificado em 3º lugar *ex-aequo* (3º ciclo do Ensino Básico)

Título: ***Rotina Diária***
Autores: Luana Isabel Rocha Pereira e Ana Rita Alves Magalhães
Professor orientador: José António Fernandes Freitas
Estabelecimento de Ensino: Escola Básica de Caldas de Vizela, Vizela

Trabalho classificado em 2º lugar *ex-aequo*
(Ensino Secundário)

Título: ***Escola Artur Gonçalves: Pontos Fortes e Pontos Fracos***
Autores: Inês Viriato Santos e Daniela Filipa Rodrigues Godinho
Professora orientadora: Maria Alice da Silva Martins
Estabelecimento de Ensino: Escola Artur Gonçalves, Torres Novas

Trabalho classificado em 2º lugar *ex-aequo*
(Ensino Secundário)

Título: ***O papel da escola e do género na adopção de comportamentos sustentáveis***
Autora: Leonor Veríssimo Carvalho Lince Duarte
Professora orientadora: Cristina Maria Cohen Gonzaga Borges Caseiro Oliveira Carvalho
Estabelecimento de Ensino: Escola Sede Secundária Maria Lamas, Torres Novas

Trabalho classificado em 3º lugar *ex-aequo*
(Ensino Secundário—Cursos Profissionais)

Título: ***Frequência dos Cursos Profissionais***
Autor: Carolina Fernandes Mendes, Mariana Araújo Moura e Pedro Mendes Jacob
Professores orientadores: Sandra Isabel Figueiredo Pimenta e César Torres
Estabelecimento de Ensino: Escola Secundária Cacilhas-Tejo, Almada

Trabalho classificado em 3º lugar *ex-aequo*
(Ensino Secundário—Cursos Profissionais)

Título: ***Características Pessoais***
Autores: Rafael Alexandre Carreira Semedo, Nuno Miguel Ventura Correia Nota Moreira e Thales Barros França Silva.
Professores orientadores: Sandra Isabel Figueiredo Pimenta e César Torres
Estabelecimento de Ensino: Escola Secundária Cacilhas-Tejo, Almada

Nota: Não foi atribuído o 1º lugar (Ensino Secundário)

Sobre a sessão de entrega dos Prémios "Estatístico Júnior 2016"

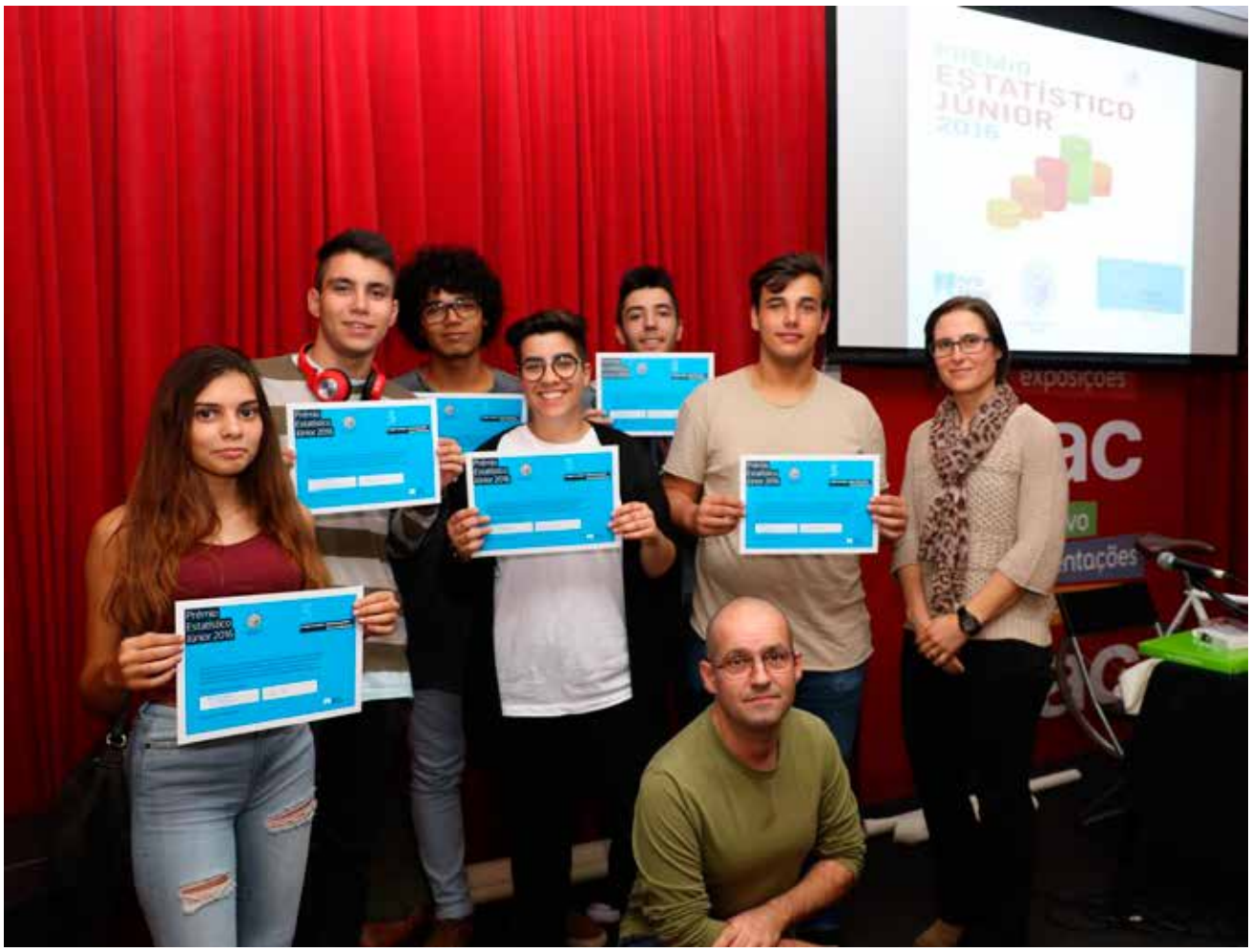
A Sociedade Portuguesa de Estatística promove anualmente o Prémio Estatístico Júnior, com o patrocínio da Porto Editora.
Com esta iniciativa pretende-se incentivar o interesse pelas áreas de Probabilidades e Estatística dos estudantes dos Ensinos Básico e Secundário, e dos Cursos de Educação e Formação (CEF) e de Educação e Formação de Adultos (CEFA).
O Prémio Estatístico Júnior (PEJ) distingue anualmente sete trabalhos e é atribuído aos estudantes que os realizaram e a alguns dos professores orientadores, sendo a sua entrega formal realizada numa sessão que lhe é expressamente consagrada.
A sessão de entrega dos PEJ 2016 decorreu no passado dia 15 de Outubro nas instalações da FNAC (Fórum) em Coimbra. A sessão iniciou-se pelas 16 horas com breves palavras da Presidente da SPE sobre os trabalhos distinguidos, sobre as escolas envolvidas (Almada, Entroncamento, Leiria, Maia, Torres Novas e Vizela) e sobre a proximidade do Dia Europeu da Estatística.
Em seguida tivemos o prazer de ouvir o Doutor Rogério Martins, docente da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. O Doutor Rogério Martins apresenta na SIC-Notícias o programa "Isto é Matemática" onde, através de uma linguagem acessível, descreve situações e exemplos práticos que ligam a Matemática ao quotidiano dos estudantes portugueses.
Desta vez, em directo e ao vivo, o Doutor Rogério Martins propiciou-nos mais um excelente momento de "Isto é Matemática", com uma palestra intitulada "A minha bicicleta calcula áreas!".
Para esta palestra, trouxe a sua bicicleta e a sua mala de prestidigitador. A presença da bicicleta não lhe facilitou a entrada no Fórum e só terá conseguido entrar porque, afinal, … é uma bicicleta-calculadora!
A palestra teve como ponto de partida as trajectórias de bicicletas que não permitem identificar o sentido em que foram produzidas e avançou, sempre de um modo muito vivo e em permanente interacção com o público presente, para os objectos que bóiam de forma estável em qualquer direcção concluindo que estas matérias, aparentemente desconexas, têm em comum uma mesma equação matemática. E, como anunciado, assistimos, com os olhos colados à roda da frente da bicicleta, ao cálculo da área de uma figura geométrica … quando afinal era a roda traseira que estava a trabalhar nesse sentido.
As fotos que se seguem ilustram parte dos vivos instantes com que o Doutor Rogério Martins nos presenteou, recorrendo não só à sua bicicleta mas também aos objectos presentes na sua maleta.
A sessão prosseguiu com a entrega dos prémios aos estudantes e professores distinguidos em 2016, seguida de um lanche ligeiro.

A Direcção da SPE agradece a presença dos estudantes premiados, seus professores e familiares nesta sessão, assim como às Doutoras Maria Eugénia Graça Martins e Manuela Neves, membros do júri dos PEJ 2016. Um agradecimento especial é dirigido ao Doutor Rogério Martins pela interessante palestra com que, generosamente,

enriqueceu o programa desta iniciativa.

P`Direcção da SPE
Esmeralda Gonçalves

(texto escrito ao abrigo do antigo acordo ortográfico)


PRÉMIO
ESTATÍSTICO
JÚNIOR
2016
exposições

• 2.º Encontro da Rede de Bioestatísticos Portugueses e da Secção de Biometria

No passado dia 6 de Janeiro de 2017, teve lugar o 2º Encontro da Rede de Bioestatísticos Portugueses e da Secção de Biometria da SPE no Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa.

À semelhança do 1º Encontro realizado em Junho de 2016, esta 2ª edição contou com a presença de 33 participantes e com as apresentações de 3 oradores convidados: Prof. João Branco do Instituto Superior Técnico que apresentou uma perspectiva histórica sobre o papel da estatística na medicina; Andreia Leite da London School of Hygiene and Tropical Medicine que apresentou o trabalho que tem desenvolvido no âmbito do seu projeto de doutoramento sobre vigilância de segurança de vacinas em tempo quase-real utilizando registos de saúde electrónicos; e Luís Borda de Água do Centro de Investigação em Biodiversidade e Recursos Genéticos que apresentou vários exemplos e desafios da aplicação da estatística na área da ecologia.

Teve lugar ainda a Assembleia Geral da Secção de Biometria da SPE e a eleição da sua Comissão Coordenadora para o biénio 2017-2018 composta pelos seguintes membros: Giovani Silva (Presidente), Laetitia Teixeira (Secretário) e Miguel Pereira (Secretário).

Com o intuito de aumentar a difusão e o número de participantes no Encontro, as apresentações dos oradores convidados foram transmitidas em direto no canal de Youtube da Rede de Bioestatísticos Portugueses. Os vídeos das apresentações encontram-se atualmente disponíveis para visualização através do seguinte link: *http://yt.vu/+redebiostat.*

O Encontro constituiu uma nova oportunidade de conhecer o trabalho efetuado por vários membros da comunidade portuguesa de estatísticos que se dedica à biometria e à bioestatística, bem como de criar e reforçar ligações entre os vários membros da Rede de Bioestatísticos Portugueses e da Secção de Biometria da SPE.

Miguel, Laetitia e Giovani

• Sessão Comemorativa do 36.º Aniversário da SPE

No dia 28 de Novembro 2016 às 14:30h decorreram na Sede as comemorações de mais um aniversário da SPE.

Depois de umas breves palavras de boas vindas pela Presidente, tivemos o prazer de ouvir o Professor João Branco recordar 6 anos de juventude da nossa Sociedade, relatando alguns episódios associados aos mandatos em que foi Presidente.

Seguidamente foram apresentados os seguintes prémios: Prémio SPE 2016 e Prémios Iniciação à Investigação 2016. A laureada do Prémio SPE 2016 foi Filipa Alexandra Cardoso da Silva com o trabalho intitulado *Processos INGARCH Poisson compostos na modelação de séries temporais de contagem sobredispersas*.

Os três trabalhos de Iniciação à Investigação premiados decorreram de teses de Mestrado, tendo sido apresentados os seguintes: *Classical and Robust Symbolic Principal Component Analysis for Interval Data* de Margarida Vilela e *Barrier Option Pricing under the 2-Hypergeometric Stochastic Volatility Model* de Ruben Sousa.

Os cerca de 35 sócios presentes foram surpreendidos com a oferta de caderno de bolso com uma inscrição alusiva à importância dos dados e tiveram a oportunidade de confraternizar durante um lanche convívio.

ES

• Atas do XXII Congresso SPE

Conforme se anuncia na secção Ciência Estatística, está publicada a versão digital das Atas do XXII Congresso SPE.
Pode ser obtida em *http://www.spestatistica.pt/images/spe/LivroAtas_XXII.pdf*

FR

• Prémios "Estatístico Júnior 2017"

Está aberto, **até 26 de Maio de 2017**, o concurso para atribuição de prémios **"Estatístico Júnior 2017"**.
O Regulamento pode ser consultado no final desta edição do Boletim SPE primavera de 2017 ou no sítio da SPE em *http://www.spestatistica.pt/.*

FR

• Prémio SPE 2017

Está aberto, até **30 de junho de 2017**, o concurso para atribuição do **Prémio SPE 2017**
O Regulamento pode ser consultado no final desta edição do Boletim SPE primavera de 2017 ou no sítio da SPE em *http://www.spestatistica.pt/.*

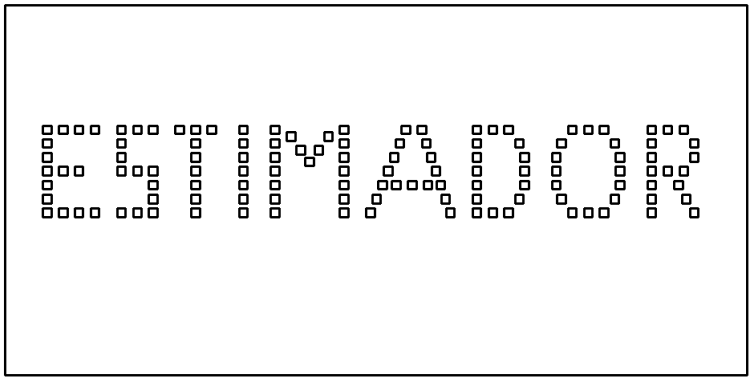
FR

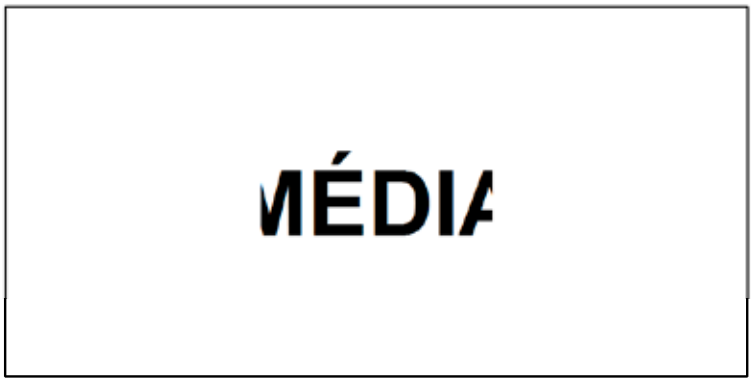
• Bolsas para participação no XXIII Congresso SPE

Pretendendo estimular o estudo e a investigação científica em Probabilidades e Estatística entre os jovens, a SPE atribui um número limitado de bolsas para participação no Congresso da SPE 2017.
O Regulamento pode ser consultado no final desta edição do Boletim SPE primavera 2017 ou no sítio da SPE em *http://www.spestatistica.pt/*

FR

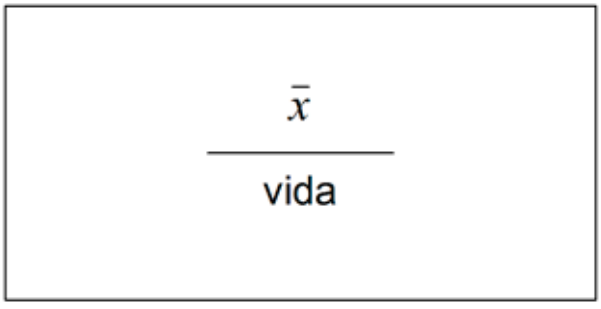
Enigmística de mefqa





No Boletim SPE outono de 2016 (p.10):

redução da variância



sobrevida média

A *SPE* e a *FENStatS*

Maria Eduarda Silva, *mesilva@fep.up.pt*

Presidente da Sociedade Portuguesa de Estatística

A *Federação das Sociedades Estatísticas Nacionais Europeias*, *FENStatS*, constituída em 2011 congrega a grande maioria, 21, das comunidades científicas que na Europa se empenham no desenvolvimento da Ciência Estatística como disciplina autónoma.
Na sua mais recente intervenção, a *FENStatS* acabou de intervir junto do Presidente do *European Research Council*, *ERC*, Professor Jean-Pierre Bourguignon. Com a força da representação de uma enorme maioria, a *FENStatS* defende que a Estatística passe a ter um painel autónomo no sistema *ERC.*
O *ERC* é uma agência financiadora de investigação de ponta, propostas interdisciplinares e ideias pioneiras em áreas novas e emergentes que introduzem abordagens inovadoras e não convencionais.
A iniciativa, à qual se juntou a *SPE*, resultou de uma carta escrita por três estatísticos conhecidos, Adelchi Azzalini, Steve Fienberg, Niels Keiding, Nanny Wermuth aos Presidentes das Sociedades Nacionais.

Estes colegas tendo pertencido a painéis de avaliação de projetos no *ERC*, consideram que a Estatística tem sido gravemente lesada na atribuição quer de bolsas *ERC Starting Grants* quer de bolsas *ERC Advanced Grants*.

Defendem que a situação se deve à ausência de um painel para a Estatística.

Como em muitas outras situações análogas, de facto, também no *ERC* a Estatística existe apenas como descritor no painel da Matemática, PE1-Mathematics.

No documento em referência pode ler-se:
"Our experience working in these committees indicates that a proposal falling in the realm of statistics needs a strong mathematical profile to have any chance of success. Typically, a project in statistical methodology, without the formal theorem-proof component, faces an objection of "lack of theory". Although in principle the Statistics term as a descriptor of the PE1 area refers to the whole of the discipline, the actual working of the panel gives some real chance of success only to proposals in mathematical statistics, except in truly exceptional circumstances."

Por outro lado, a inclusão de ramos da Estatística noutros painéis da árvore de disciplinas do *ERC*, como por exemplo Bioestatística que aparece como LS2 (Genetics, Genomics, Bioinformatics and Systems Biology) não melhora a situação.

De facto, como também é descrito no referido documento:
"The minority condition for Statistics here […Biostatistics which appears as LS2 (Genetics, Genomics, Bioinformatics and Systems Biology)…] is, however, even more marked than in PE1, since the panel members with quantitative expertise have a background in genomics and bioinformatics, which is not surprising given the full title of LS2."

Eu, pessoalmente e como Presidente da *SPE*, identifiquei imediatamente esta descrição com a situação em Portugal.

No entanto, o que mais me admirou foi que quase todas as outras sociedades identificaram uma situação semelhante no respetivo País.

Gerou-se assim um consenso para atuar de alguma forma e o Presidente da *FENStatS*, Maurizio Vichi disponibilizou-se a escrever uma carta em nome das Sociedades Nacionais ao Presidente da *ERC*. A carta foi assinada por 20 Sociedades. A *Royal Statistical Society, RSS* mostrou solidariedade para com a iniciativa apesar de não se identificar com a mesma situação no Reino Unido. Maurizio informou ainda que o Presidente do *ERC* irá apresentar a carta ao Conselho Científico do *ERC*.

Com esta iniciativa a *SPE,* bem com a comunidade científica dos estatísticos portugueses que ela representa, deu um passo no cumprimento de um dos grandes objetivos estatutários, para melhor fazer avançar a Ciência Estatística.

Medições, erros aleatórios e o filtro de Kalman

Marco Costa, *marco@ua.pt*

ESTGA-UA, Escola Superior de Tecnologia e Gestão de Águeda
CIDMA, Centro de Investigação e Desenvolvimento em Matemática e Aplicações
Universidade de Aveiro

1. Introdução

As medições experimentais têm sempre uma incerteza associada. Essa incerteza traduz a impossibilidade de se obter um valor verdadeiro através de um procedimento de medição, em regra, com o auxílio de um ou mais instrumentos de medida que, indubitavelmente induzem erros nas medições. Esses erros podem ser aleatórios no sentido em que são inevitáveis e imprevisíveis, dependendo das condições instrumentais ou ambientais em que as medições são obtidas; ou não aleatórios, por exemplo se forem sistemáticos, sendo suscetíveis de serem corrigidos. São os erros aleatórios que suscitam a necessidade da aplicação de técnicas estatísticas para a sua quantificação ou redução.

A incerteza associada a uma medição única pode ser analisada quanto à sua precisão se considerarmos o erro máximo do instrumento utilizado obtendo-se, desta forma, um majorante para o erro absoluto. Ou, caso seja mais relevante, pode obter-se um majorante para o erro relativo associado a essa medição. Esta é a abordagem que, em regra, se adota no âmbito na Análise Numérica e no tratamento de observações ou de valores resultantes de operações matemáticas sobre medições.

No entanto, em muitas situações em Engenharia dispomos de medições que, além de terem associadas a incerteza inerente ao instrumento ou método de medição, e cujo erro pode ser eventualmente majorado, são resultantes de outros fatores aleatórios que, em regra, não se consegue controlar ou majorar. Por outro lado, em muitos casos práticos dispomos de medições sucessivas que podemos considerar como sendo observações de um processo estocástico, isto é, constituem uma série temporal, e que cuja variabilidade pode ser atribuída a fatores ambientais e/ou instrumentais no âmbito dos procedimentos de medição.

É nestes casos que a análise e a modelação de séries temporais na perspetiva da sua representação em espaço de estados pode ser uma mais valia no tratamento de medições num paradigma estocástico. A modelação de séries temporais, através de modelos de espaço de estados, tem-se evidenciado como uma abordagem bastante útil na análise de séries temporais, cujos principais objetivos sejam a previsão a curto prazo e/ou a obtenção de estimativas "filtradas".

De facto, os modelos de espaço de estados têm uma estrutura *markoviana* que permite obter estimativas de variáveis não observáveis e previsões suscetíveis de serem atualizadas quando nova informação é recolhida por aplicação do conhecido filtro de Kalman (Kalman, 1960). O filtro de Kalman é um conjunto de equações recursivas que são estabelecidas para um modelo que admita uma

representação de espaço de estados que permitem obter, em cada instante, estimativas e previsões e os respetivos erros quadráticos médios.

2. Modelo de espaço de estados e o filtro de Kalman

A variável de interesse X_t, que designamos por *estado*, e que não é observável, mas sobre a qual se pretende obter estimativas e, eventualmente, previsões, evolui ao longo do tempo segundo um modelo autorregressivo da forma

$$X_t = \phi X_{t-1} + \varepsilon_t \qquad (1)$$

isto é, o estado no instante t, X_t, é uma correção do estado no instante anterior "contaminado" com um erro aleatório ε_t. O coeficiente ϕ é designado por coeficiente autorregressivo e o erro aleatório ε_t é assumido como sendo um ruído branco gaussiano, com $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ e $E(\varepsilon_t \varepsilon_s) = 0, \forall t \neq s$.

O estado X_t, não sendo observável, é indiretamente medido após a sua multiplicação por um fator h_t conhecido cujo resultado é "corrompido" por outro ruído branco gaussiano $e_t \sim N(0, \sigma_e^2)$, com $E(e_t e_s) = 0, \forall t \neq s$, cujo resultado observável é a variável Y_t, isto é,

$$Y_t = h_t X_t + e_t. \qquad (2)$$

Em regra, os erros ε_t e e_t são assumidos como não sendo correlacionados, isto é, $E(\varepsilon_t e_s) = 0, \forall t, s$.

A equação (1) é a designada *equação de estado* e descreve a evolução temporal da variável de interesse, enquanto que, a equação (2) é a chamada *equação de observação* relacionando o estado não observável X_t com a variável de *medida* Y_t, esta observável. Os erros ε_t e e_t são comummente designados nas Engenharias apenas como *ruídos*.
Um modelo que admita uma representação sobre a forma das equações (1)-(2) diz-se que admite uma *representação de espaço de estados*.

O modelo (1)-(2) é bastante versátil podendo ser interpretado de diversas formas e em vários contextos. Em particular, podemos entender o estado X_t como o verdadeiro valor da medição no instante t, sendo que a variável Y_t representa a medição efetivamente observada. Este modelo incorpora duas fontes de variabilidade, os dois ruídos brancos, onde ε_t pode representar a variabilidade inerente ao processo que se pretende analisar e o erro e_t pode representar a variabilidade induzida, por exemplo, pelo instrumento de medição e/ou condições ambientais. Noutros contextos, o modelo (1)-(2) pode ser interpretado como um modelo de regressão linear cujo declive não é um parâmetro (determinístico) sendo uma variável aleatória, permitindo uma evolução dinâmica ao longo do tempo.

Os modelos ARMA (Modelos Autorregressivos e de Médias Móveis) são uma classe de modelos que admitem uma representação se espaço de estados, permitindo, por exemplo a obtenção das estimativas de máxima verosimilhança dos parâmetros através desta abordagem. Existem várias extensões do modelo (1)-(2) que têm sido propostas na literatura. Por exemplo, o estado e a variável de medida podem ser vetores aleatórios com dimensões adequadas ou as equações de observação e de estado podem ter outras configurações tais como incluir outros coeficientes ou matrizes de design/planeamento.

O filtro de Kalman é um conjunto de equações recursivas que exprimem recursivamente, em cada instante t, os estimadores de erro quadrático mínimo sob a hipótese de normalidade dos ruídos ε_t e e_t, ou os melhores estimadores lineares centrados quando os ruídos não são normais, uma que são as projeções ortogonais do estado sobre as variáveis observadas até ao instante t; bem como os respetivos erros quadráticos médios das estimativas.

Assim, seja $X_{t|t-1}$ o estimador do estado X_t com a informação disponível até ao instante $t-1$, isto é, baseado em $\mathcal{Y}_{t-1} = (Y_1, Y_2, \ldots, Y_{t-1})$ e $P_{t|t-1}$ o respetivo erro quadrático médio. O preditor de Y_t, baseado em $\mathcal{Y}_{t-1}$, que denotamos por $Y_{t|t-1}$, resulta da linearidade da equação de medição e é dado por

$$Y_{t|t-1} = h_t X_{t|t-1}. \quad (3)$$

Mas quando no instante t temos disponível a medição Y_t podemos atualizar a última estimativa do estado através do erro de predição ou *inovação*, $\eta_t = Y_t - Y_{t|t-1}$, pela equação

$$X_{t|t} = X_{t|t-1} + k_t\, \eta_t \quad (4)$$

onde

$$k_t = P_{t|t-1} h_t (h_t^2 P_{t|t-1} + \sigma_e^2)^{-1} \quad (5)$$

é o designado *ganho de Kalman*. O estimador $X_{t|t}$ permite a obtenção da estimativa filtrada do estado com a informação disponível até ao instante t cujo erro quadrático médio é dado por

$$P_{t|t} = P_{t|t-1}(1 - k_t h_t). \quad (6)$$

No entanto, através da equação de estado, no instante t podemos obter a previsão do estado para o instante seguinte, $t+1$, pela equação

$$X_{t+1|t} = \phi X_{t|t}$$

com erro quadrático médio dado por

$$P_{t+1|t} = \phi^2 P_{t|t} + \sigma_\varepsilon^2. \quad (7)$$

As equações recursivas são inicializadas com os valores $X_{1|0}$ e $P_{1|0}$ que podem decorrer do caso prático ou podem ser estimadas no procedimento da estimação dos restantes parâmetros que referiremos mais à frente. Os estimadores do filtro de Kalman são ótimos no sentido do menor erro quadrático médio quando os ruídos são normais, isto é,

$$X_{t|t} = E(X_t | Y_1, Y_2, \ldots, Y_t) \text{ e } X_{t|t-1} = E(X_t | Y_1, Y_2, \ldots, Y_{t-1}). \quad (8)$$

Sob a normalidade dos ruídos ε_t e e_t, os parâmetros desconhecidos $\Theta = (\phi, \sigma_\varepsilon^2, \sigma_e^2)$, entre outros se existirem, podem ser estimados pelo método da máxima verosimilhança. Nestas condições, a log-verosimilhança de uma amostra $\mathcal{Y}_t = (Y_1, Y_2, \ldots Y_n)$ pode ser escrita através das distribuições condicionais, obtendo-se

$$\ln L(\Theta | \mathcal{Y}_t) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n} \ln(\omega_t) - \frac{1}{2}\sum_{t=1}^{n} \ln(\omega_t^{-1}\eta_t^2)$$

onde $\omega_t = h_t^2 P_{t|t-1} + \sigma_e^2$.

As estimativas de máxima verosimilhança, $\widehat{\Theta}_{MV}$, são obtidas maximizando a log-verosimilhança através de métodos numéricos como o algoritmo EM (Dempster et al., 1977) ou o método de Newton-Raphson (Harvey, 1996), isto é,

$$\widehat{\Theta}_{MV} = \text{argmax}_\Theta \ln L(\Theta | \mathcal{Y}_t)\,. \quad (9)$$

É possível obter-se estimativas para os coeficientes desconhecidos através de *packages* e de rotinas computacionais, como o *package* dlm do ambiente R (Petris, 2010) que incorpora várias funções que permitem modelar e estimar os parâmetros dos modelos de espaço de estados usuais, bem como obter as estimativas e as predições obtidas através do filtro de Kalman.

Sob algumas condições gerais, os estimadores da máxima verosimilhança são assintoticamente normais, tendo-se,

$$\sqrt{n}(\widehat{\Theta}_{MV} - \Theta) \xrightarrow{D} N\,(0, I^{-1}(\Theta)), \qquad (10)$$

onde $I(\Theta)$ é a matriz de informação, cujo elemento (i,j) pode ser aproximado por

$$I_{ij}(\Theta) \approx \frac{1}{2}\sum_{t=1}^{n} \frac{1}{\omega_t^2}\frac{\partial \omega_t}{\partial \theta_i}\frac{\partial \omega_t}{\partial \theta_j} + \sum_{t=1}^{n} \frac{1}{\omega_t}\frac{\partial \eta_t}{\partial \theta_i}\frac{\partial \eta_t}{\partial \theta_j},$$

e as derivadas parciais podem ser obtidas recursivamente aquando da implementação do filtro de Kalman ou através da diferenciação numérica.

3. Alguns comentários

As variâncias dos ruídos

As grandezas das variâncias das equações de observação e de estado, respetivamente σ_e^2 e σ_ε^2, indicam a "confiabilidade" do sensor (isto num contexto mais tecnológico) relativamente à capacidade de estimar o estado.

Por exemplo, se $\sigma_e^2 \gg \sigma_\varepsilon^2$ significa que o erro de observação, em regra associado ao sensor, isto é, associado ao método ou procedimento de medição, é bastante superior ao error inerente ao fenómeno que se pretende estimar, o estado. Caso contrário, se $\sigma_e^2 \ll \sigma_\varepsilon^2$ o sensor diz-se bastante confiável no sentido em que o método de medição acrescenta pouca variabilidade ao estado, logo este será mais fácil de estimar (é observado com pouco ruído).

Por exemplo, na Figura 1 estão representadas duas realizações do modelo de um passeio aleatório mais um erro definido pelas equações

$$Y_t = X_t + e_t$$

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

considerando $\sigma_e^2 = 10$ e $\sigma_\varepsilon^2 = 1$ e $\sigma_e^2 = 1$ e $\sigma_\varepsilon^2 = 5$, o estado e as estimativas do estado obtidas pelo filtro de Kalman.

Podemos verificar que quando σ_e^2 tem uma grandeza bastante maior que σ_ε^2 as estimativas do estado atualizadas pelo filtro de Kalman em cada instante, $X_{t|t}$, tendem a não acompanhar as observações de Y_t, com $t = 1, 2, \ldots, n$, isto é, o filtro aplica a equação de estado, pouco atualizando a previsão $X_{t|t-1}$ quando Y_t é observada.

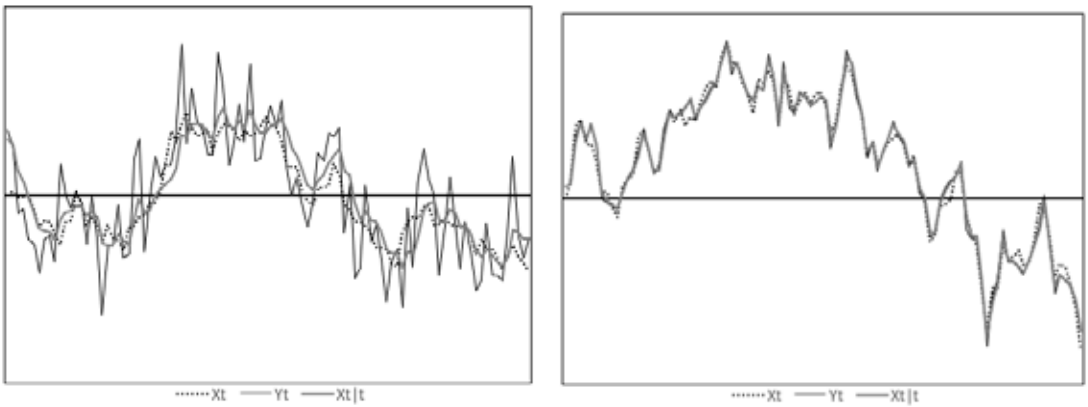


Figura 1: (à esquerda) uma realização de um processo do tipo $\sigma_e^2 \gg \sigma_\varepsilon^2$; (à direita) uma realização de um processo do tipo $\sigma_e^2 \ll \sigma_\varepsilon^2$.

Repare-se que, neste caso, quando $\sigma_e^2 \to +\infty$, temos que o ganho de Kalman tendo para 0, $k_t \to 0$, pelo que a estimativa atualizada será semelhante à previsão, $X_{t|t} \approx X_{t|t-1}$.

Erros não gaussianos

O exemplo anterior ilustra a importância da estimação dos parâmetros de um modelo de espaço de estados, em particular as variâncias dos erros. Quando os erros são normais a estimação dos parâmetros desconhecidos é, em regra, realizada através do método da máxima verosimilhança (Eq. 9), como já discutido anteriormente.

No entanto, em muitas aplicações (ambientais, económicas, industriais, etc.) a normalidade dos erros nem sempre é garantida ou é rejeitada no procedimento de validação do modelo, nomeadamente através da análise da normalidade das inovações que, sob a normalidade dos erros, seguem a distribuição $\eta_t = Y_t - Y_{t|t-1} \sim N(0, \omega_t)$. Por outro lado, em modelo mais complexos (por exemplo, modelos de espaço de estados multivariados que aqui não são abordados por uma questão de simplicidade), nem sempre é garantida a unicidade de um máximo global da função de log-verosimilhança, $\ln L(\Theta|\mathcal{Y}_t)$ (Hamilton, 1994; pág. 387).

Neste sentido, têm sido propostas abordagens alternativas como estimadores não-paramétricos (Alpuim, 1999; Costa & Alpuim, 2010; Gonçalves & Costa, 2013); ou a abordagem Bayesiana com recurso a técnicas de Monte Carlo via cadeias de Markov (Carlin et al., 1992; Shephard & Pitt, 1997; Tanizaki, 2001).

O filtro de Kalman com parâmetros estimados

Quando os erros e o estado inicial X_1 são gaussianos, os estimadores do filtro de Kalman são os melhores estimadores não enviesados no sentido do erro quadrático médio mínimo. No entanto, as propriedades ótimas só podem ser garantidas quando todos os parâmetros do modelo de espaço de estado são conhecidos (Harvey, 1996).

A abordagem mais comum para lidar com modelos de espaço de estados com parâmetros estimados foca-se na análise do erro quadrático médio dos estimadores do filtro de Kalman. Quando o vetor de parâmetros desconhecidos Θ é substituído por uma sua estimativa (por exemplo $\widehat{\Theta}_{MV}$), o erro quadrático médio dos estimadores é subestimado, uma vez que, por exemplo o erro de previsão vem dado por

$$P_{t|t-1}(\widehat{\Theta}) = P_{t|t-1} + E\left[\left(X_{t|t-1} - X_{t|t-1}(\widehat{\Theta})\right)^2\right]. \qquad (11)$$

Nesta perspetiva, a opção recai sobre a estimação da segunda parcela da Eq.10 através de metodologias baseadas no Bootstrap (Pfeffermann & Tiller, 2005; Bandyopadhyay & Lahiri, 2010, Rodríguez & Ruiz, 2012).

Recentemente, Costa & Monteiro (2016) apresentaram alguns resultados sobre a propagação do viés dos estimadores do filtro de Kalman, em particular, deduziram expressões não-recursivas para esse viés. Estes resultados permitem corrigir as estimativas iniciais dos parâmetros através de um procedimento iterativo de correção e, por conseguinte, as estimativas do filtro de Kalman, enquanto que na abordagem anterior corrige-se o erro quadrático médio dos estimadores do filtro de Kalman mantendo as suas estimativas pontuais.

4. Exemplo de aplicação

Considere-se um tanque cujos caudais de fluxo de entrada/saída se desconhecem com rigor, como se representa na Figura 2.

Admita-se que o nível do líquido do tanque é monitorizado através de um instrumento que pode ser mecânico ou eletrónico, e que dispomos de 200 medições obtidas consecutivamente, em períodos de 1 minuto.

Figura 2: tanque com caudais de fluxo de entrada/saída desconhecidos.

Seja Y_t a variável aleatória que representa a mediação do nível do tanque no minuto t obtido pelo instrumento utilizado e X_t o verdadeiro nível do tanque no mesmo instante. Assume-se que esta a medição e o verdadeiro nível do líquido do tanque se relacionam pelas equações

$$Y_t = X_t + e_t$$

$$X_t = \mu + \phi(X_{t-1} - \mu) + \varepsilon_t$$

onde $e_t \sim N(0, \sigma_e^2)$ e $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ são ruídos brancos não correlacionados. Este modelo contempla a hipótese de que o estado X_t seja um processo autorregressivo de ordem 1 estacionário, AR(1), caso $|\phi| < 1$.

Na Tabela 1 apresentam-se as estimativas de máxima verosimilhança, os respetivos erros padrão e os intervalos de confiança a 95% para os parâmetros do modelo.

Tabela 1: Estimativas de máxima verosimilhança e intervalos de confiança a 95% para os parâmetros do modelo.

parâmetro	estimativa	erro padrão	$I_{95\%}(\theta)$
μ	999,9145	0,0074	[999,75 ; 1000,08]
ϕ	0,5443	0,0151	[0,30 ; 0,79]
σ_e^2	0,5611	0,0167	[0,31 ; 0,81]
σ_ε^2	0,3412	0,0190	[0,07 ; 0,61]
$\ln L$	-281,0499		

As estimativas obtidas indicam que o nível do líquido do tanque é um processo estacionário uma vez que $|\hat{\phi}| = |0{,}5443| < 1$, com variância estimada igual a $\hat{\sigma}_{X_t}^2 = \hat{\sigma}_\varepsilon^2/(1-\hat{\phi}^2) = 0{,}4848$. Assim, o ajustamento do modelo indica que a variância estimada do erro que podemos associar ao instrumento, $\hat{\sigma}_e^2 = 0{,}5611$, é superior à variância estimada inerente ao processo que representa o verdadeiro nível do líquido do tanque.

O ajustamento do modelo permitiu estimar o nível médio do líquido do tanque em 999,9145 litros, com um intervalo de confiança a 95% associado igual a [999,75;1000,08].

Com a aplicação do filtro de Kalman é possível obterem-se as estimativas filtradas do filtro de Kalman do nível do líquido do tanque, $X_{t|t}$, em cada instante e, se necessário, os respetivos intervalos de confiança. Na Figura 3 estão representadas as medições observadas do nível do líquido do tanque e as estimativas atualizadas (filtradas) obtidas pelo filtro de Kalman.
Desta forma, é possível fazer-se uma monitorização mais precisa dos caudais de fluxo de entrada/saída do tanque tendo em consideração várias fontes de variabilidade e obtendo estimativas pontuais ou intervalares, se oportunas.

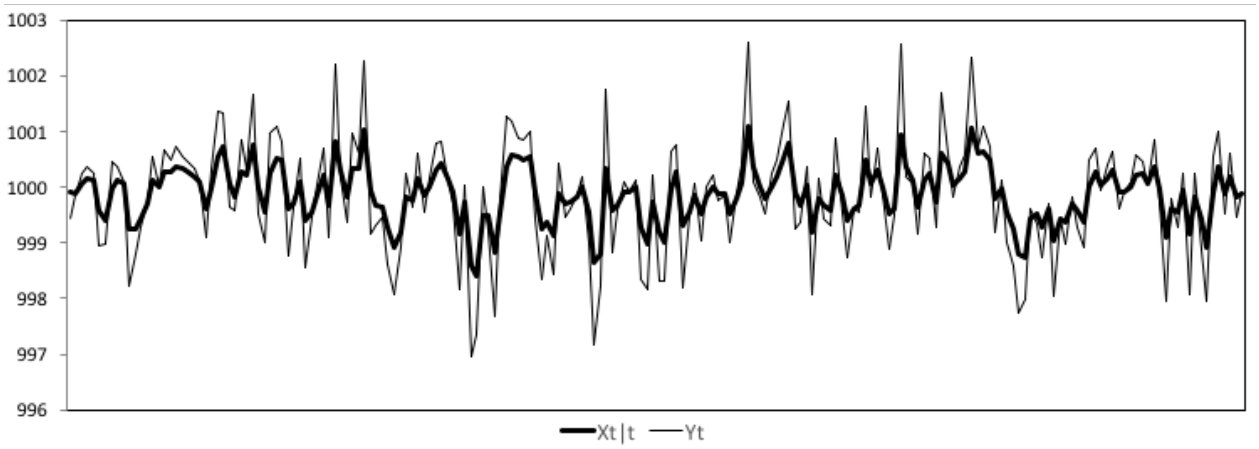


Figura 3: representação das observações no nível do líquido do tanque, Y_t, e as estimativas atualizadas (filtradas) obtidas pelo filtro de Kalman.

Referências

Alpuim, T. (1999). Noise variance estimators in state space models based on the method of moments, *Annales de l'ISUP*, **43** (2-3), 3-23.

Bandyopadhyay, S., Lahiri, S.N. (2010). Resampling-based bias-corrected time series prediction. *Journal of Statistical Planning and Inference,* **140**, 3775–3788.

Carlin, B. P., Polson, N. G., Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association*, **87**:493–500.

Costa, M, Alpuim, T. (2010). Parameter estimation of state space models for univariate observations. *Journal of Statistical Planning and Inference*, **140** (7), 1889-1902.

Costa, M., Monteiro, M. (2016). Bias-correction of Kalman filter estimators associated to a linear state space model with estimated parameters, *Journal of Statistical Planning and Inference*, **176**: 22 - 32.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38.

Gonçalves, AM, Costa, M. (2013). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering, *Stochastic Environmental Research and Risk Assessment*, **27**(5), 1021-1038.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, New Jersey.

Harvey, A.C. (1996). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems, Transactions of the ASME--Journal of Basic Engineering, 82 Series D, 35-45.

Petris, G. (2010). An R Package for Dynamic Linear Models. *Journal of Statistical Software*, **36**(12), 1-16.

Pfeffermann, D., Tiller, R. (2005). Bootstrap approximation to prediction MSE for state space-models with estimated parameters. *Journal of Time Series Analysis*, **21**, 219–236.

Rodríguez, A., Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis*, **56**, 62–74.

Shephard, N., Pitt, M. K. (1997). Likelihood Analysis of Non-Gaussian Measurement Time Series, *Biometrika*, **84(**3), 653-667.

Tanizaki, H. (2001). Estimation of unknown parameters in non- linear and non-Gaussian state-space models. *Journal of Statistical Planning and Inference*, **96**, 301–323.

Testes uniformemente mais potentes não enviesados e o controlo de artigos defeituosos em Engenharia Industrial

Manuel Cabral Morais, *maj@math.ist.utl.pt*

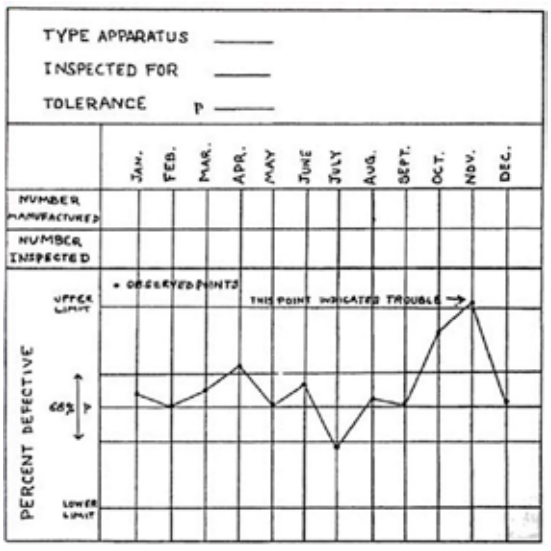
Departamento de Matemática & CEMAT (Centro de Matemática Computacional e Estocástica), Instituto Superior Técnico, Universidade de Lisboa

1. Nota introdutória

Adequação ao uso (*fitness for use*) e conformidade com os requisitos (*conformance to requirements*) são muito provavelmente as definições mais curtas e consensuais de qualidade e podem encontrar-se em duas referências-chave, Juran e Godfrey (1999, pag. 27) e Crosby (1979, pag. 17).

É sabido que a qualidade de um produto constitui para as/os consumidoras/es um factor decisivo na sua aquisição a par do custo do mesmo. No entanto, um facto curioso escapa à maioria das/os consumidoras/es: as preocupações com a qualidade já se faziam sentir no Reino da Babilónia (Gitlow *et al.*, 1989, pag. 8-9). Com efeito, o Código de Hamurábi (c. 1772AC) prova que a não conformidade com os requisitos poderia ter consequências dramáticas. A título de exemplo note-se que um dos artigos deste código é: *If a builder builds a house for a man and does not make its construction firm, and the house which he has built collapse and cause the death of the owner of the house, that builder shall be put to death* (Wikisource, 2017).

Esta forma de assegurar a qualidade era baseada na *lei de talião* que *consiste na rigorosa reciprocidade do crime e da pena*, frequentemente expressa pela máxima *olho por olho, dente por dente*. Cumpre notar que o modo como os inspectores fenícios garantiam a qualidade dos produtos fabricados era igualmente sanguinário, como reportam Gitlow *et al.* (1989, pag. 9): *Phoenician inspectors eliminated any repeated violations of quality standards by chopping off the hand of the maker of the defective product*.



É necessário dar um salto para o início do sec. XX para conhecermos o pai do controlo de qualidade moderno, Walter A. Shewhart. Este físico, engenheiro e estatístico reconhece que os processos industriais produzem dados e entende que estes podem ser analisados usando técnicas estatísticas de modo a averiguar se um processo industrial está ou não sob controlo. Para o efeito, Shewhart propôs a carta de controlo de qualidade num memorando de uma página entregue no dia 16 Maio de 1924 ao seu superior hierárquico na *Bell Telephone* George Edwards. Shewhart alterou, com o diagrama ao lado (disponível em Managers-Net, 2017) e o curto texto que o acompanhava, o curso da história da Engenharia Industrial.

2. Cartas para o número de artigos defeituosos

Uma especificação não satisfeita por uma unidade de um produto corresponde à ocorrência de um defeito, seja ele uma solda irregular, um rebite quebrado ou um erro tipográfico. Mais, é usual classificar-se um artigo inspecionado de defeituoso caso possua pelo menos um defeito.
A carta-*np* com limites 3-sigma é sem sombra de dúvida a carta de controlo de qualidade mais usada na detecção de alterações do número esperado de artigos defeituosos numa amostra de dimensão fixa.
A utilização desta carta pressupõe a recolha regular de amostras de dimensão *n* e o registo sequencial do número observado de artigos defeituosos num gráfico com os limites de controlo

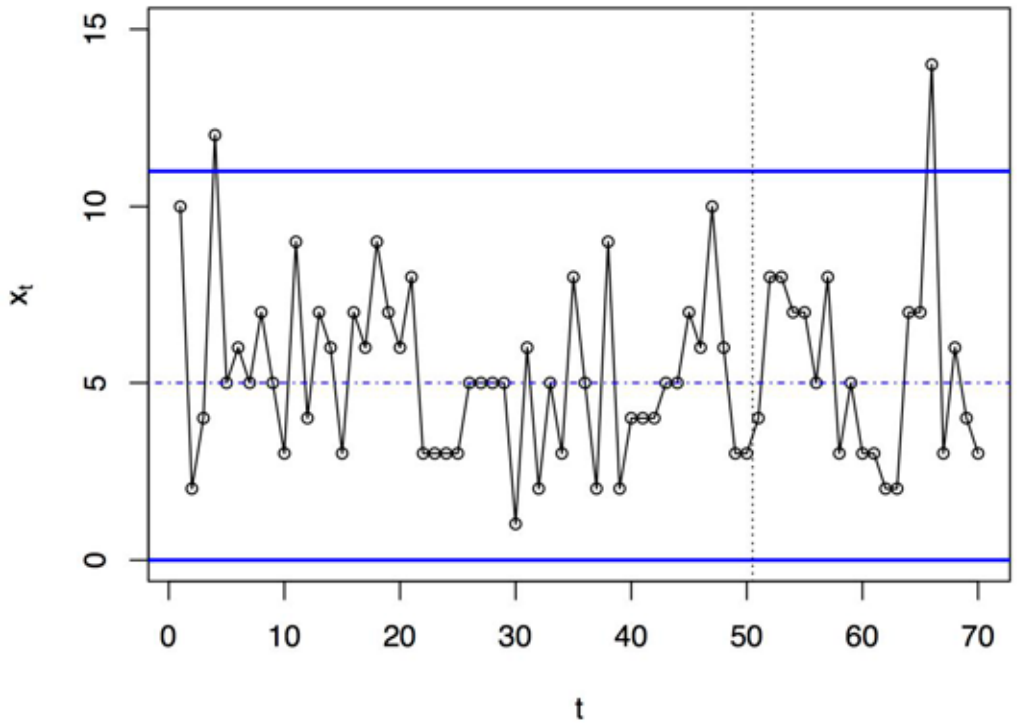
$$np_0 - 3\sqrt{np_0(1-p_0)} \text{ e } np_0 + 3\sqrt{np_0(1-p_0)}.$$

Tendo em conta o carácter inteiro não negativo do número observado de artigos defeituosos, o limite inferior de controlo (*lower control limit*, LCL) e o limite superior de controlo (*upper control limit*, UCL) acima podem reescrever-se à custa do tecto e da parte inteira seguintes:

$$LCL = \lceil max\{0, np\_0 - 3\sqrt{(np\_0\ (1-p\_0))}\}\rceil \text{ e } UCL = \lfloor np_0 + 3\sqrt{np_0(1-p_0)}\rfloor.$$

Atenda-se que p_0 representa o valor alvo da fracção de artigos defeituosos (p) e que a probabilidade de o número de artigos defeituosos na *t*-ésima amostra (X_t) se encontrar no intervalo $[LCL, UCL]$ é bastante elevada ao admitir-se que $X_t \sim binomial(n, p = p_0)$. Posto isto, não surpreende que a *t*-ésima amostra seja responsável pela emissão de um sinal caso o número observado de artigos defeituosos x_t não pertença a $[LCL, UCL]$.

Exemplo 1 — Considere-se a recolha de amostras de dimensão $n = 100$ e que $p_0 = 0.05$. Os limites de controlo são dados neste caso por $LCL = 0$ e $UCL = 11$. De seguida recorreu-se ao R (R Core Team, 2013) para simular um total de 70 amostras independentes, das quais as primeiras 50 sob controlo e as restantes 20 fora de controlo devido a um aumento da fracção de artigos defeituosos para $p = p_0 + 0.006$.



O registo sequencial do número observado de artigos defeituosos leva a concluir que a 4ª amostra é responsável pela emissão de um sinal, na verdade um falso alarme ou não tivessem as 50 primeiras observações geradas admitindo que $X_t \sim_{iid} binomial(n, p = p_0), t = 1, \ldots, 50$.
Para além disso, a 66ª amostra é responsável por um sinal válido já que as 20 últimas observações foram geradas considerando $X_t \sim_{iid} binomial(n, p = p_0 + 0.006), t = 51, \ldots, 70$.

A analogia entre uma carta de controlo de qualidade e um teste de hipóteses é inevitável. De facto, ao recorrer-se à carta-*np* estão a confrontar-se repetidamente as hipóteses nula e alternativa $H_0: p = p_0$ e $H_1: p \neq p_0$, usando a estatística

$$T = \frac{X_t - np_0}{\sqrt{np_0(1-p_0)}} \dot{\sim}_{H_0} normal(0,1)$$

e a região de rejeição de H_0 (escrita para valores de *T*) dada por $W = (-\infty, -3) \cup (3, +\infty)$.
A função potência exacta deste teste é

$$\xi(p) = 1 - \sum_{x=LCL}^{UCL} \binom{n}{x} p^x (1-p)^{n-x}$$

e não atinge o seu valor mínimo quando $p = p_0$. Acrescente-se ainda que $\xi(p_0)$ não é igual a $0.0027 \cong 1 - [\Phi(3) - \Phi(-3)]$ já que a aproximação normal à distribuição binomial não é brilhante mesmo que np_0 e $n(1-p_0)$ sejam superiores a 5.
Para além disso, lida-se neste caso com $LCL = 0$ uma vez que $p_0 \leq 9/(9+n)$ e, consequentemente, $\xi(p) < \xi(p_0)$, para $p < p_0$, ou seja, esta carta-*np* com limites 3-sigma emite falsos alarmes mais frequentemente que sinais válidos na presença de qualquer diminuição em *p*.
De referir que a selecção de uma dimensão amostral mínima n_{min} que garanta que $LCL > 0$, i.e., que satisfaça a condição $n > 9\,(1-p_0)/p_0$, pode conduzir a valores de n_{min} impraticáveis para pequenos valores de p_0. Por exemplo, para $p_0 = 0.01$ tem-se $n_{min} = 892$.
Como se isso não bastasse, a distribuição binomial é assimétrica positiva quando $p \in (0.0.5)$, pelo que a probabilidade $P(X_t < LCL)$ pode ser bastante pequena (ou mesmo nula caso $LCL = 0$), impossibilitando a detecção de diminuições em *p* (i.e., melhorias da qualidade) em tempo razoável.
Ryan e Schwertman (1997) descrevem diversas variantes da carta-*np* com limites 3-sigma que se propõem a mitigar as deficiências de desempenho desta carta de controlo de qualidade. Estas variantes podem ser divididas em duas categorias de acordo com a abordagem adoptada: a transformação dos dados ou a modificação dos limites de controlo.
Refira-se, por exemplo, que a descrição de três delas associadas à transformação dos dados encontra-se em Ryan (1989, pag. 182-186). Elas dizem respeito a transformações baseadas no arco seno e originalmente propostas por:

- Freeman e Tukey (1950), $y = 0.5\left[arcsen\sqrt{x/(n+1)} + arcsen\sqrt{(x+1)/(n+1)}\right]$;
- Hald (1952, pag. 685), $y = arcsen\sqrt{x/n}$;
- Johnson and Kotz (1969, pag. 65), $y = arcsen\sqrt{(x+3/8)/(n+3/4)}$.

Ryan (1989, pag. 186) refere que estas três transformações resultam em cartas com desempenhos similares e que o valor de n_{min} é grosso modo um quarto do requerido por uma carta-*np* com limites 3-sigma, como ilustra Ryan (1989, *Table* 8.3, pag. 187).
Ryan e Schwertman (1997) propõem uma alternativa às variantes acima: a estatística da carta não sofre qualquer transformação e continua a ser o número de artigos defeituosos na amostra, no entanto a dimensão da amostra e os limites de controlo são obtidos por regressão em np_0 e $\sqrt{np_0}$ de forma a que as probabilidades $P(X_t < LCL)$ e $P(X_t > UCL)$ estejam o mais próximas de $0.0027/2$ sob controlo para $p_0 \in (0,0.03]$. Estes autores adiantam que $LCL = 2.9529 + 1.01956\,np_0 - 3.2729\sqrt{np_0}$ e $UCL = 0.6195 + 1.00523\,np_0 + 2.983\sqrt{np_0}$ e facultam uma tabela com diversos valores de *n* e os correspondentes limites de controlo *LCL* e *UCL*, para $p_0 = 0.01,\ 0.03$.
Por forma a efectuar um comentário crucial sobre as variantes da carta-*np* com limites 3-sigma, é necessário de adiantar que, ao avaliar o desempenho de qualquer carta de controlo de qualidade, é costume recorrer ao valor esperado do número de amostras recolhidas até à emissão de um sinal (*vulgo* ARL, *average run length*). Ora, ao admitir-se que $X_t \sim_{iid} binomial(n,p),\ t = 1,2,\ldots$, tal variável aleatória possui distribuição geométrica e $ARL(p) = 1/\xi(p)$.
Todas estas variantes padecem de um problema grave: à semelhança da carta-*np* com limites 3-sigma, levam mais tempo, em média, a detectar determinadas alterações em *p* que a emitir um falso alarme. Com efeito, todas estas cartas possuem função $ARL(p)$ que não atinge um máximo quando $p = p_0$, logo com um *viés*. Por este motivo são designadas de cartas do tipo *ARL-biased*, designação esta que se deve a Pignatiello *et al*. (1995).

3. Uma carta-*np* sem viés

Tanto quanto se pôde apurar, a primeira tentativa de correcção do viés da função $ARL(p)$ da carta-*np* deve-se a Acosta-Mejía (1999). Ao resolver a equação $[d\xi(p)/dp]_{p=p_0} = 0$, este autor obteve a igualdade

$$\frac{p_0^{LCL-1}\,(1-p_0)^{n-LCL}}{\Gamma(n-LCL+1)\,\Gamma(LCL)} = \frac{p_0^{UCL}\,(1-p_0)^{n-UCL-1}}{\Gamma(n-UCL)\,\Gamma(UCL)},$$

onde *LCL* e *UCL* representam os limites de controlo desejados e Γ a função gama. Tal como seria de prever esta equação não conduz, de um modo geral, a limites de controlo inteiros, pelo que Acosta-Mejía (1999) sugere que se escolha o par de inteiros mais próximo do ponto de intercepção entre a curva definida por tal equação e a curva que define todos os pares (LCL, UCL) associados a um valor pré-especificado para o *ARL* sob controlo. A carta-*np* resultante é, inevitavelmente, do tipo *ARL-biased*.

A eliminação do viés da função *ARL* da carta-*np* passa, naturalmente, por associar esta carta a um teste uniformemente mais potente não enviesado (ou *uniformly most powerful unbiased*, UMPU).

O primeiro trabalho que recorre a um teste UMPU para definir uma carta-*np* do tipo *ARL-unbiased* parece dever-se a Morais (2016) e vem na linha da carta do tipo *ARL-unbiased* para o número esperado de defeitos numa amostra, carta esta proposta por Paulino *et al*. (2016a).

De acordo com aquele autor a carta-*np* do tipo *ARL-unbiased*, a *t*-ésima amostra será responsável pela emissão de um sinal com:

- probabilidade um, caso $x_t < LCL$ ou $x_t > UCL$;
- probabilidade γ_{LCL} (resp. γ_{UCL}), caso $x_t = LCL$ (resp. $x_t = UCL$).

Pese em embora a aleatorização da emissão de sinal não seja usual ao recorrer-se a cartas de controlo de qualidade, ela não traz problemas práticos de maior. De acordo com Morais (2016), a aleatorização da emissão do sinal, quando $x_t = LCL$ (resp. $x_t = UCL$), pode fazer-se na prática incorporando a geração de um número pseudo-aleatório da distribuição de Bernoulli com parâmetro γ_{LCL} (resp. γ_{UCL}) no *software* usado para o tratamento dos dados provenientes da linha de produção.

Morais (2016) nota que a descrição do teste UMPU para a probabilidade de sucesso *p* da distribuição binomial se encontra em Lehmann (1959, pag. 128-129, *Example* 1). Acrescenta também que as duas equações que o definem são insuficientes para definir o par de limites de controlo bem como o par de probabilidades de aleatorização.

Importa referir que, ao pretender-se uma carta-*np* do tipo *ARL-unbiased* com probabilidade de emissão de falso alarme igual a α, ou seja, $ARL(p_0) = \alpha^{-1}$, as probabilidades de aleatorização são dadas por

$$\gamma_{LCL} = \frac{d\,e - b\,f}{a\,d - b\,c} \quad \text{e} \quad \gamma_{UCL} = \frac{a\,f - c\,e}{a\,d - b\,c},$$

onde as constantes dependem da função de probabilidade da distribuição $binomial(n, p_0)$, $P_{n,p_0}(x), x = 0,1,\ldots,n$.

A saber:

$$a = P_{n,p_0}(LCL),\ b = P_{n,p_0}(UCL),\ c = LCL \times P_{n,p_0}(LCL),\ d = UCL \times P_{n,p_0}(UCL),$$
$$e = \alpha - 1 + \sum_{x=LCL}^{UCL} P_{n,p_0}(x) \text{ e } f = \alpha \times np_0 - np_0 + \sum_{x=LCL}^{UCL} x \times P_{n,p_0}(x).$$

Posto isto, Morais (2016) define uma grelha de valores de *LCL* e *UCL* e sugere que ela seja percorrida até que o par correspondente $(\gamma_{LCL}, \gamma_{UCL})$ pertença a $(0,1)^2$. Para mais detalhes acerca da dedução destas probabilidades e da definição da grelha de pesquisa, remete-se o/a leitor/a para Morais (2016).

Convém realçar que a obtenção da grelha de pesquisa em Morais (2016) é um pouco menos sofisticada que a descrita em Paulino *et al*. (2016a).

É conveniente destacar que, por mero desconhecimento do autor deste texto aquando da preparação e revisão de Morais (2016), não se tirou partido do pacote *ump* (Geyer e Meeden, 2004, 2005) do R que

adianta valores da função crítica do teste UMPU para a probabilidade de sucesso de uma distribuição binomial, i.e., valores da função

$$\phi(x) = P(\text{Rejeitar } H_0 \mid X_t = x) = \begin{cases} 1, & \text{se } x < LCL \text{ ou } x > UCL \\ \gamma_{LCL}\ (resp.\ \gamma_{UCL}), & \text{se } x = LCL\ (\text{resp. } x = UCL) \\ 0, & \text{se } LCL < x < UCL \end{cases}$$

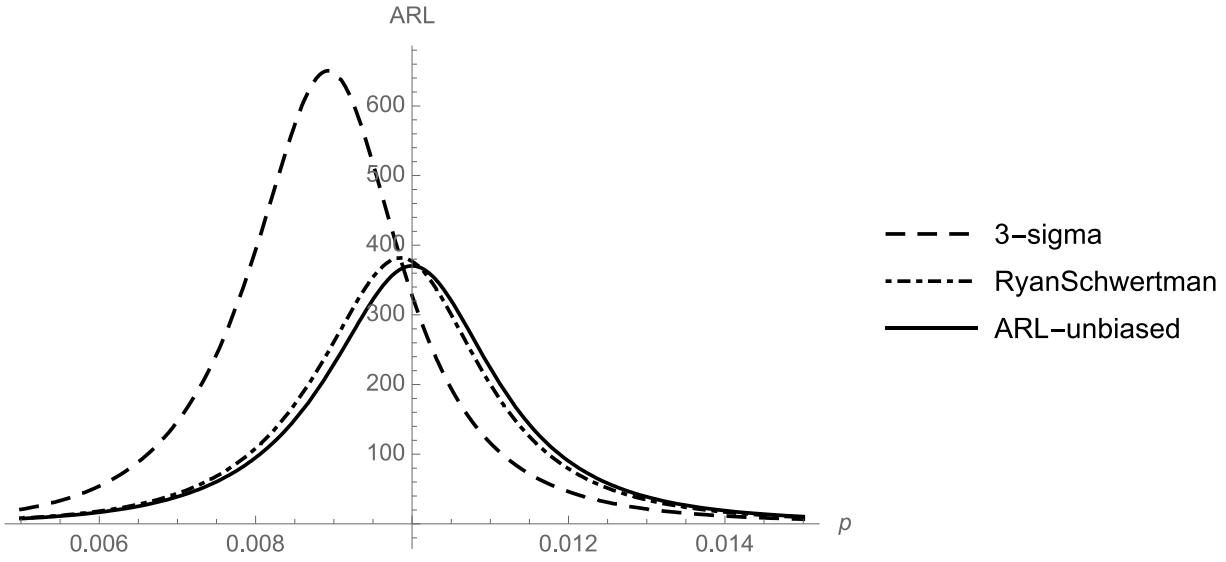
associada à carta-*np* do tipo *ARL-unbiased*. Note-se que é possível obter os valores de *LCL*, *UCL*, γ_{LCL} e γ_{UCL} de forma indirecta e não muito prática, recorrendo ao pacote *ump*, bastando para o efeito solicitar os valores da função crítica para $x = 0,1, \ldots, n$.
A consulta de Geyer e Meeden (2004, 2005) e do código do pacote *ump* do R leva a crer que a grelha de pesquisa de *LCL* e *UCL* difere das descritas em Morais (2016) e Paulino *et al*. (2016a). Como não poderia deixar de ser, os resultados coligidos em Morais (2016) coincidem com os obtidos usando tal pacote.
Seguem-se duas ilustrações de cartas-*np* do tipo *ARL-unbiased*. Na primeira delas tira-se partido do facto de a probabilidade de esta carta emitir um sinal com probabilidade dada por

$$\xi_{unbiased}(p) = \left[1 - \sum_{x=LCL}^{UCL} P_{n,p}(x)\right] + \gamma_{LCL} \times P_{n,p}(LCL) + \gamma_{UCL} \times P_{n,p}(UCL)$$

e de a função *ARL* correspondente ser igual a $1/\xi_{unbiased}(p)$.

Exemplo 2 — Considere-se $n = 1267$, $p_0 = 0.01$ e $\alpha = 0.0027$ e confrontem-se as curvas *ARL* das cartas-*np* seguintes:

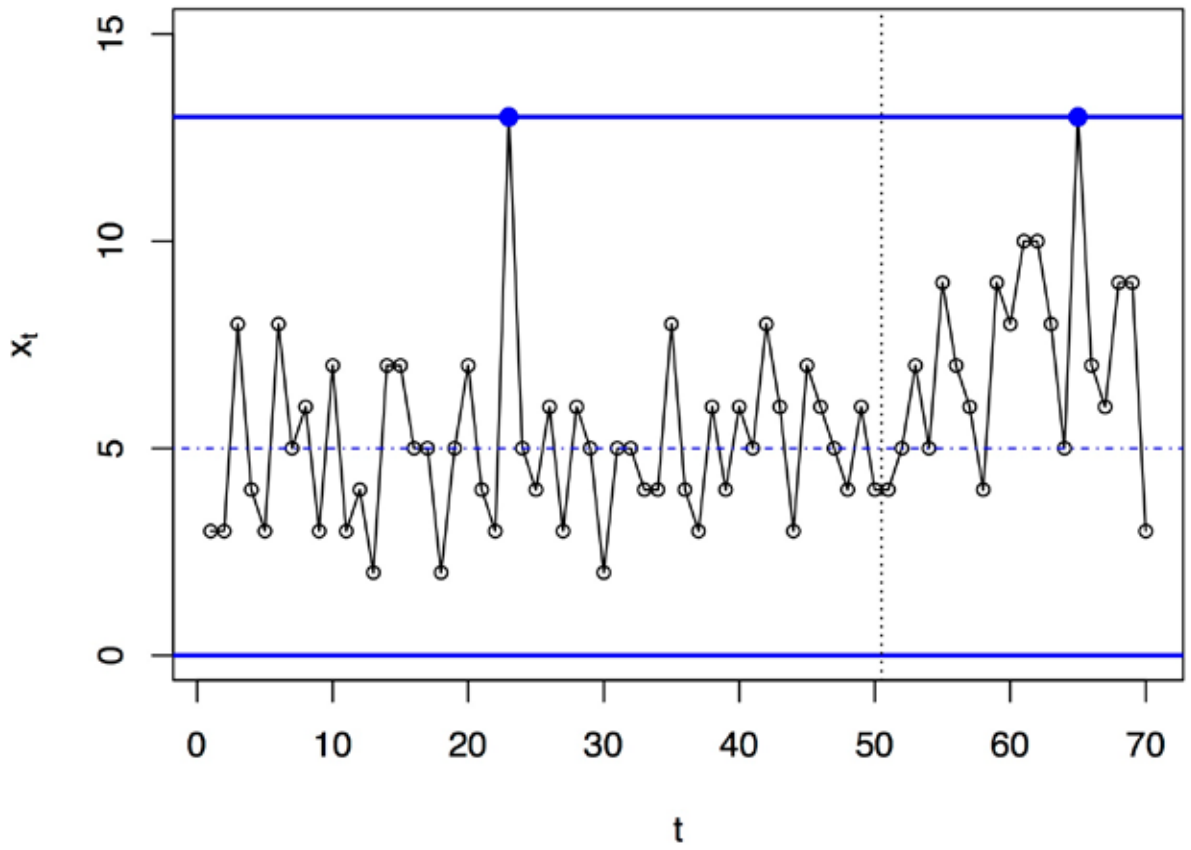
- a carta-*np* com limites 3-sigma $[LCL, UCL] = [3,23]$;
- a carta-*np* proposta por Ryan e Schwertman (1997) que possui limites de controlo $[LCL, UCL] = [4,24]$;
- a carta-*np* do tipo *ARL-unbiased* com $[LCL, UCL] = [4,25]$ e probabilidades de aleatorização $\gamma_{LCL} = 0.076400$ e $\gamma_{UCL} = 0.713818$.



De acordo com Morais (2016) o valor máximo da função *ARL* da carta-*np* com limites 3-sigma (resp. proposta por Ryan e Schwertman, 1997) é aproximadamente igual a 650.419 (resp. 381.718), sendo que tal máximo ocorre para um valor de $p < p_0$.
O valor de ARL sob controlo da primeira das cartas é aproximadamente igual a 327.976 logo distingue-se substancialmente do valor desejado, $\alpha^{-1} \approx 370.4$. O da segunda carta-*np* é aproximadamente igual a 376.811 e pouco se difere de 370.4.

Adiante-se também que o perfil *ARL* da carta-*np* do tipo *ARL-unbiased* possui máximo exactamente igual a α^{-1} quando $p = p_0$, como bem ilustra a figura acima.
Refira-se, por fim, que o valor da dimensão da amostra sugerido por Ryan e Schwertman (1997) é impraticável numa linha de produção.

Exemplo 3 — Retome-se o Exemplo 1 e considere-se novamente $n = 100$, $p_0 = 0.05$ e $\alpha = 0.0027$. Ao executar-se o comando *umpu.binom(0:100, 100, 0.05, 0.0027)* e ao inspeccionar-se o vector de valores da função crítica do teste, conclui-se que os limites de controlo da carta-*np* do tipo *ARL-unbiased* são $[LCL, UCL] = [0,13]$ e as probabilidades de aleatorização $\gamma_{LCL} = 0.289066$ e $\gamma_{UCL} = 0.524741$.



Ao simular-se de novo um total de 70 observações, das quais as primeiras 50 sob controlo e as restantes 20 fora de controlo devido a um aumento de p para $p_0 + 0.006$, registou-se um falso alarme da responsabilidade da 23ª amostra, alarme este emitido devido à aleatorização do sinal associado a uma observação igual a *UCL*. Assinalou-se também a azul um sinal válido da responsabilidade da 65ª amostra com as mesmas características que o falso alarme.

À laia de conclusão

São algumas as vantagens do recurso à carta-*np* do tipo *ARL-unbiased*. Ao contrário da tradicional carta-*np* com limites 3-sigma e qualquer das suas variantes:

- é possível pré-especificar o ARL sob controlo da carta-*np* do tipo *ARL-unbiased*;
- a curva $\mathrm{ARL}(p)$ da carta-*np* do tipo *ARL-unbiased* possui efectivamente um máximo quando $p = p_0$, pelo que, em média, emite sinais válidos mais rapidamente que falsos alarmes;
- a carta-*np* do tipo *ARL-unbiased* é capaz de lidar com a *maldição* do $LCL = 0$ porque depende de duas probabilidades de aleatorização e assim é sempre capaz de detectar melhorias da qualidade em tempo razoável.

Seria interessante obter uma versão *ARL-unbiased* da carta CUSUM (*cumulative sum*) por forma a detectar de forma mais expedita alterações em *p* de pequena ou média magnitude.
Recomenda-se a consulta de Paulino *et al*. (2016b), para uma lista de trabalhos versando cartas do tipo *ARL-unbiased* para parâmetros de características de qualidade com distribuição contínua; a lista de trabalhos reportando-se a distribuições discretas é, infelizmente, muito mais curta. O caso de parâmetros de distribuições mistas é tratado por Morais e Knoth (2016) num ambiente particular, o controlo de alterações da intensidade de tráfego de uma fila de espera.
Termina-se relembrando ao/à leitor/a duas máximas em Engenharia Industrial (Wikipedia, 2017).

- *Produtividade sem qualidade é uma má utilização de recursos.*
- *Qualidade sem produtividade é uma má utilização do tempo*.

(Por decisão pessoal, o autor deste texto não escreve segundo o Acordo Ortográfico de 1990.)

Agradecimentos

O autor muito agradece ao Prof. Fernando Rosado a oportunidade de divulgação deste trabalho e a proposta desse excelente desafio que é escrever um artigo em Português e *Word*.
Este trabalho foi parcialmente financiado pela FCT (Fundação para a Ciência e Tecnologia) através dos projectos UID/Multi/04621/2013, PEst-OE/MAT/UI0822/2014 and PEst-OE/MAT/UI4080/2014.

Referências

Acosta-Mejía, C.A. (1999). Improved p-charts to monitor process quality. *IIE Transactions* **31**, 509-516.

Crosby, P.B. (1979). *Quality is Free: the Art of Making Quality Certain*. New York: MacGraw-Hill.

Freeman, M.F. e Tukey, J.W. (1950). Transformations related to the angular and the squared root. *Annals of Mathematical Statistics* **21**, 607-611.

Geyer, C.J. e Meeden, G.D. (2004). ump: An R package for UMP and UMPU tests. Consultado a 26/01/2017 em www.stat.umn.edu/geyer/fuzz/

Geyer, C.J. e Meeden, G.D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science* **20**, 358-366.

Gitlow, H., Gitlow, S., Oppenheim, A. e Oppenheim, R. (1989). *Tools and Methods for the Improvement of Quality*. Boston: Richard Irwin, Inc.

Hald, A. (1952). *Statistical Theory with Engineering Applications*. New York: John Wiley & Sons.

Johnson, N.L. e Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Boston: Houghton Mifflin Company.

Juran, J.M. e Godfrey, A.B. (1999). *Juran's Quality Handbook* (5ª Edição). New York: MacGraw-Hill.

Managers-Net (2017). Walter Shewhart (1891-1967). Consultado a 23/01/2017 em http://www.managers-net.com/Biography/Shewhart

Morais, M.C. (2016). An ARL-unbiased np-chart. *Economic Quality Control* **31**, 11-21.

Morais, M.C. e Knoth, S. (2016). On ARL-unbiased charts to monitor the traffic intensity of a single server queue. Em: *Proceedings of the XIIth. International Workshop on Intelligent Statistical Quality Control*, 217-242.
URL: http://www.hsu-hh.de/compstat/index_8sVJz3C3s0oQzk3M.html

Paulino, S., Morais, M.C. e Knoth, S. (2016a). An ARL-unbiased c-chart. *Quality and Reliability Engineering International* **32**, 2847-2858.

Paulino, S., Morais, M.C. e Knoth, S. (2016b). On ARL-unbiased c-charts for INAR(1) Poisson counts. *Statistical Papers*. URL: http://rdcu.be/nGs8

Pignatiello, J.J.J., Acosta-Mejía, C.A. e Rao, B.V. (1995). The performance of control charts for monitoring process dispersion. Em: *4th Industral Engineering Research Conference Proceedings*, 320-328.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Ryan, T.P. (1989). *Statistical Methods for Quality Improvement*. New York: John Wiley & Sons.

Ryan, T.P. e Schwertman, N.C. (1997). Optimal limits for attribute control charts. *Journal of Quality Technology* **29**, 86-98.

Wikisource (2017). The Code of Hammurabi (Harper translation). Consultado a 23/01/2017 em https://en.wikisource.org/wiki/The_Code_of_Hammurabi_(Harper_translation)

Wikipedia (2017). Engenharia Industrial | Wikipédia, a enciclopédia livre. Consultado a 25/01/2017 em https://pt.wikipedia.org/wiki/Engenharia_industrial

Simulação de Monte Carlo na avaliação de incertezas de medição

Sandra Ramos, *sfr@isep.ipp.pt*

Instituto Superior de Engenharia do Porto | Centro de Estatística e Aplicações

INTRODUÇÃO

Os métodos de Monte Carlo são uma poderosa ferramenta de simulação estocástica com ampla aplicação em engenharia, já que permitem lidar com problemas com comportamento altamente não linear e que envolvem um grande número de variáveis aleatórias com diferentes distribuições de probabilidade. Neste texto apresenta-se, brevemente, e uma aplicação da simulação de Monte Carlo em engenharia, nomeadamente na avaliação de incertezas de medição.

A medição é algo comum no mundo da engenharia. Em geral, o resultado da medição é uma aproximação (estimativa) do valor da grandeza medida afastando-se do valor verdadeiro dessa grandeza por uma quantidade denominada erro de medição. Os erros de medição podem ser classificados como sistemáticos ou aleatórios. Os primeiros são geralmente causados pelo aparelho de medida, observador e/ou condições ambientais. Por sua vez, nestes últimos, os valores medidos estão dispersos em torno do valor médio.

O facto de que, em situações reais, a diferenciação entre erros sistemáticos e erros aleatórias raramente é possível juntamente com o facto dos erros de medição, devido à sua natureza, não poderem ser conhecidos com exatidão levou à definição de incerteza de medição. De acordo com o Vocabulário Internacional de Metrologia (VIM, 2012), a incerteza de medição é "parâmetro não-negativo que caracteriza a dispersão dos valores da grandeza que são atribuídos à mensuranda (grandeza que se pretende medir) a partir das informações usadas".

O resultado de medição apenas está completo se estiver acompanhado do valor da incerteza de medição.

É frequentemente assumido que a incerteza de medição é constante ou é uma função linear do valor medido. No entanto, isto geralmente não é verdade. A velocimetria por imagem de partículas[1] é um exemplo de uma técnica de medição que possui incertezas locais altamente variáveis e não lineares, não sendo viável, nestes casos, a aplicação do método clássico de avaliação da incerteza de medição descrito no Guia para a Expressão da Incerteza de Medição (GUM, 2003) e aqui designado de método GUM. O método GUM também requer que a variável aleatória que representa os valores possíveis da mensuranda seja normalmente distribuída o que limita bastante a sua aplicabilidade em cenários reais.

A simulação de Monte Carlo (SMC) revelou-se uma alternativa adequada ao método convencional GUM em situações onde os pressupostos de aplicabilidade deste método não são verificados. Os métodos de SMC são também muito úteis como ferramenta de validação do método tradicional cuja aplicação nem sempre é tão sensata como se esperaria.

De seguida faz-se uma muito breve descrição do método GUM seguindo-se uma apresentação resumida do método SMC na avaliação da incerteza de medição.

[1] Método ótico que permite a visualização e a análise do movimento de partículas em fluidos.

A INCERTEZA DE MEDIÇÃO SEGUNDO O MÉTODO GUM

A avaliação convencional da incerteza é baseia-se numa relação funcional entre uma grandeza aleatória (mensuranda - Y) e várias grandezas de entrada, $X_1, X_2, \dots, X_k$: $Y = f(X_1, X_2, \dots, X_k)$. A função f tanto pode ser uma função simples como pode representar uma relação complexa ou ser apenas determinada experimentalmente. Cada grandeza de entrada é considerada uma variável aleatória (v.a.) com uma função densidade de probabilidade (fdp) conhecida.
A incerteza de medição associada com as estimativas das grandezas de entrada é classificada em um de dois tipos, dependendo do método de avaliação: tipo A ou tipo B. No tipo A, a avaliação é feita através da análise estatística de um conjunto de observações experimentais. Neste caso a incerteza padrão é o desvio padrão experimental da média. A avaliação da incerteza padrão do tipo B é feita por outros métodos que não a análise estatística. A incerteza padrão combinada, $u_c(y)$, é estimada através da lei de propagação de incertezas na sua forma geral (obtida a partir de um desenvolvimento em série de Taylor de primeira ordem):

$$u_c(y) = \sqrt{\sum_{i=1}^{k} c_i^2 . u^2(x_i) + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} c_i . c_j . u(x_i) . u(x_j) . r(x_i, x_j)},$$

onde $c_i = \frac{\partial f}{\partial x_i}$ representa o coeficiente de sensibilidade da grandeza x_i, $u(x_i)$ é a incerteza da grandeza de influência x_i e $r(x_i, x_j)$ é o coeficiente de correlação entre x_i e x_j.

É habitual, na prática, apresentar uma medida de incerteza que defina um intervalo com uma maior probabilidade de abrangência (a probabilidade atual é de aproximadamente 68%). A incerteza expandida U é obtida através da expressão $z . u_c(y)$ onde z é o fator de expansão. Métodos de determinação do valor de z podem ver-se em (GUM, 2003).
As condições de aplicabilidade ideais do método GUM são aquelas onde existe um modelo aditivo que relaciona as variáveis de influência X_i com a variável Y e onde X_i sejam v.a. independentes com distribuição normal. Em outras situações o método convencional fornece uma solução aproximada, sendo a qualidade dessa aproximação dependente do desvio em relação à distribuição normal e do grau de não linearidade do modelo.

A SIMULAÇÃO DE MONTE CARLO NA AVALIAÇÃO DA INCERTEZA DE MEDIÇÃO

A aplicação de métodos de SMC na avaliação de incertezas de medição tem sido praticamente limitada aos maiores laboratórios de metrologia devido à exigência computacional e/ou à necessidade de uso de distribuições de probabilidade diferentes das mais comuns. Estes métodos têm particular interesse quando as v.a. de entrada, X_i, não seguem distribuições normais, quando o modelo que relaciona as v.a. X_i com a variável Y é complexo e não linear e, por fim, quando as incertezas sobre as v.a. X_i não são elevadas. Os resultados obtidos por este método tendem para a solução exata e fornecem uma visão abrangente de como a incerteza especificada nas entradas se propaga através do modelo.
O método de SMC propaga as fdp das grandezas de entrada ao invés propagar apenas as incertezas das mesmas conseguindo-se assim obter uma estimativa da fdp de Y em vez de um simples parâmetro estatístico (ver Figura 1).

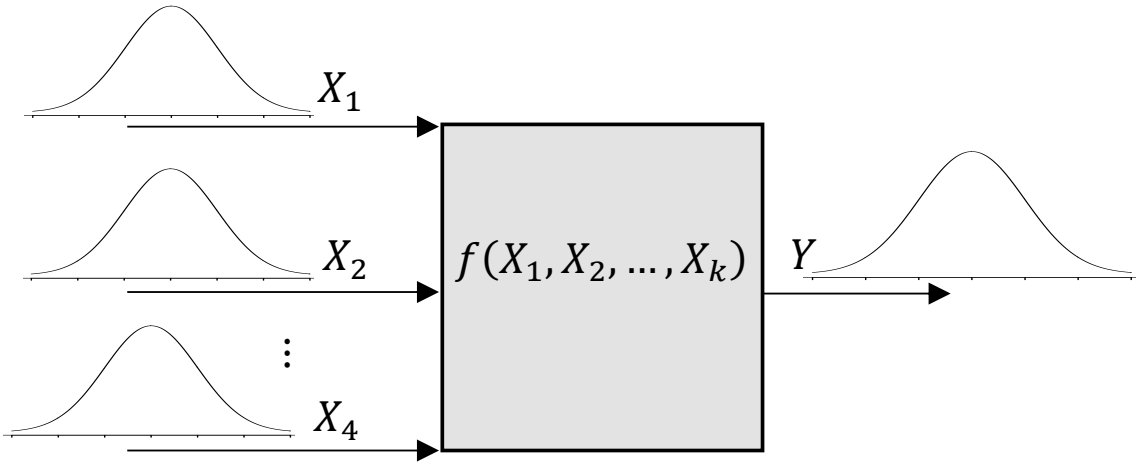


Figura 1: Propagação de fdp.

Tendo em consideração a estimativa da fdp de Y é possível obter um conjunto de quantidades estatísticas, nomeadamente, estimativas do resultado da medição, a incerteza associada e o intervalo de confiança.
A avaliação de incertezas de medição usando o método de SMC envolve os seguintes passos:

1. Definir o modelo de medição em termos de todas as v.a.de influência.
2. Quantificar as características probabilísticas de todas as v.a. (fdp e parâmetros).
3. Gerar M valores a partir de cada uma das fdp especificadas.
4. Obter uma estimativa da fdp de Y a partir do modelo matemático especificado e dos valores gerados no ponto 3.
5. Extrair a informação de interesse da estimativa obtida no ponto anterior.
6. Avaliar a precisão e a eficiência do processo de simulação.

A qualidade dos resultados obtidos depende da representatividade do modelo matemático definido, da qualidade da especificação da distribuição das variáveis de influência, das características do gerador de números pseudo-aleatórios utilizado, do número M de simulações e do procedimento de definição do intervalo de abrangência.
O número M de valores gerados tem forte influência no erro esperado para as estimativas obtidas. Embora a ampliação do valor de M traga um acréscimo do esforço computacional deve-se considerar um valor elevado para este parâmetro de maneira a serem obtidas estimativas mais precisas para o valor da incerteza. Nos exemplos apresentados na bibliografia da área (ver, por exemplo, Cox, et al, 2001) foram consideradas amostras de dimensão superior ou igual a 10^5.
No método de SMC é possível estimar a incerteza expandida mesmo em situações onde a distribuição dos valores possíveis de Y não é a normal. Quanto esta distribuição é simétrica é possível obter um intervalo de abrangência simétrico sendo os seus limites dados, para uma confiança $p = (1-\alpha) \times 100\%$, $\alpha \in]0,1[$ pelos quantis empíricos de ordem $\alpha/2$ (limite inferior) e $1-\alpha/2$ (limite superior). Quando a distribuição dos valores possíveis de Y não é simétrica o método apresentado para o caso simétrico é desadequado, devendo procurar-se o intervalo de abrangência mínimo. De acordo com GUM Suppl (2004), na quantificação dos limites do intervalo de abrangência deve proceder-se da seguinte forma. Seja $0 \leq \beta \leq 1-p$, onde p representa a probabilidade de abrangência desejada. O limite inferior e o limite superior do intervalo de abrangência de probabilidade p são dados, respetivamente, por $\hat{G}^{-1}(\beta)$ e por $G^{-1}(p+\beta)$, onde $\hat{G}^{-1}(.)$ representa o inverso da estimativa da função de distribuição dos valores possíveis de Y. Para que o intervalo anterior seja mínimo, a determinação do valor de β a considerar deve minimizar a diferença $G^{-1}(p+\beta) - G^{-1}(\beta)$ para todo o $\beta \in [0,1-p]$.

ALGUMAS CONSIDERAÇÕES FINAIS

Neste texto foi feita uma breve descrição da SMC para a avaliação de incertezas de medição. Verificou-se que constitui uma alternativa viável ao método GUM que requer condições de aplicabilidade que são, frequentemente violadas na prática. Embora o método de SMC tenha vindo a ser aplicado na estimação de incertezas de medição apresenta algumas deficiências, tais como o elevado tempo de execução para modelos complexos, alguns problemas de convergência e produção de resultados por vezes instáveis. Estes problemas podem ser parcialmente ultrapassados usando técnicas de amostragem mais eficientes.

REFERÊNCIAS BIBLIOGRÁFICAS

COX, M.G. et al. (2001). Use of Monte Carlo Simulation for Uncertainty Evaluation in Metrology. In: *Advanced Mathematical & Computational Tools in Metrology V*. Singapore: World Scientific Publishing.

GUM (2033).*Guia para a Expressão da Incerteza de Medição*. 3ªed. bras. Do *Guide to the Expression of Uncertainty in Measurement*. Rio de Janeiro: INMETRO, ABNT.

GUM Suppl 1. *Guide to the Expression of Uncertainty in Measurement (GUM) – Supplement 1: numerical methods for the propagation of distributions*. In accordance with the ISO/IEC Directives,

Vocabulário Internacional de Metrologia. VIM (2012).

Modelação do atraso dos veículos em cruzamentos semaforizados

Maria Lurdes Simões, *lurdes.simoes@fe.up.pt*
CONSTRUCT e Faculdade de Engenharia da Universidade do Porto
Paula Milheiro Oliveira, *poliv@fe.up.pt*
CMUP e Faculdade de Engenharia da Universidade do Porto

1 Introdução

O controlo de tráfego usa semáforos para, alternadamente, atribuir passagem aos movimentos incompatíveis. As suas principais vantagens residem em permitir estabelecer movimentos ordenados de tráfego, aumentar a capacidade do cruzamento, reduzir a frequência de certos tipos de colisões.

Quando se trata de tráfego rodoviário, nomeadamente de tráfego em cruzamentos, semaforizados ou não, estamos sem sombra de dúvida perante um fenómeno revestido de incerteza. A incerteza nos movimentos dos veículos, a incerteza nos espaçamentos de tempo entre as suas chegadas a um dado ponto da corrente de tráfego, a incerteza nos atrasos que os veículos vão sofrer no seu movimento até de facto abandonarem as correntes de tráfego ou os cruzamentos, são alguns dos aspectos que fazem com que um problema que envolva o tráfego rodoviário deva ser formalizado e tratado no contexto dos processos estocásticos, nomeadamente à luz da teoria de filas de espera.

Historicamente, os semáforos operavam inicialmente apenas com ciclo pré-definido, no qual era dado sinal verde para os movimentos de acordo com um plano de tempos fixado pelo engenheiro de tráfego ou por um técnico. Este plano era desenvolvido com base num historial de volumes de tráfego e tendo em atenção que a duração do ciclo deveria ser tal que, por um lado permitisse escoar a totalidade da procura e que, por outro lado, produzisse o atraso médio por veículo mais baixo. De modo a atribuir mais eficientemente o tempo de verde disponível num cruzamento semaforizado, os semáforos actuados vieram, mais tarde, substituir os semáforos regulados com tempos fixos.

As medidas de desempenho fundamentais em concepção e controlo de cruzamentos semaforizados são o comprimento das filas de espera e os tempos de espera ou atrasos [1]. Estas medidas não avaliam apenas o nível de serviço que é oferecido aos condutores, mas permitem também estimar o consumo de combustível, o ruído produzido e a poluição do ar. Uma importante contribuição é devida a [2], visto como um artigo fundamental sobre as questões da regulação de semáforos, onde se apresenta a simulação do fluxo de tráfego rodoviário duma via na aproximação ao cruzamento semaforizado. Webster [2] ajusta uma curva aos resultados da simulação, a qual descreve o atraso médio por veículo devido ao semáforo. A expressão que descreve esta curva tem-se mostrado fundamental para a regulação de semáforos com ciclo fixo. Havendo qualquer tipo de actuação do semáforo, o estudo das medidas de desempenho e dos processos de chegadas de veículos conduzem a métodos e fórmulas matemáticas mais complexas do que no controlo de ciclo fixo, podendo mesmo acontecer que o problema de modelação do atraso não admita uma solução analítica.

2 Tipos de regulação semafórica

Existem várias formas de controlo semaforizado, que se podem classificar essencialmente em controlo fixo e controlo actuado. Em semáforos com controlo fixo, a sequência de sinais apresentada é pré-

determinada e tem um ciclo de duração fixa, assim como o tempo atribuído a cada sinal. Em controlo actuado, a aproximação dos veículos ao cruzamento é detectada por sensores existentes na via, normalmente colocados próximo da linha de paragem, e a duração do período de sinal verde é ajustada à procura de tráfego. Mais especificamente, podemos considerar que existem de facto três formas de controlo semaforizado, como se descreve a seguir:

Tempos fixos: Em controlo pré-definido ou com ciclo fixo, todos os parâmetros de sinal operacionais estão definidos no controlador, o qual repetidamente executa os tempos fixados. Pode existir um único plano, que será executado continuamente durante todo o dia, ou vários planos que serão executados em diferentes períodos do dia, dependendo do fluxo de tráfego. Estes planos são estabelecidos com base em dados históricos e na experiência do engenheiro de tráfego.

Semi-actuado: Em controlo semi-actuado, a via secundária (com menor volume de tráfego) recebe sinal verde apenas quando é detectada a presença de tráfego. Um sistema de detecção determina quando é que existem veículos na via secundária que se aproximam do cruzamento, o que permite a abertura do sinal verde. Sempre que o intervalo de tempo entre a activação de veículos for superior ao valor da extensão do tempo de verde ou o tempo máximo de serviço expire, o sinal verde passa a ser executado na via principal (com maior volume de tráfego).

Totalmente actuado: Em controlo totalmente actuado, todas as fases são controladas através do uso de detecção. O conjunto de parâmetros de controlo deve incluir um tempo de verde mínimo e um tempo de verde máximo para cada fase. Esta forma de controlo é mais eficiente do que as outras formas, nos casos em que a procura de tráfego varia com o tempo. Para além disso, uma fase pode ser saltada se, num determinado ciclo, não houver a procura respectiva.

Um bom plano de controlo actuado, que responda apropriadamente à procura de tráfego, pode reduzir significativamente o atraso e o consumo de combustível. Isso motivou o estudo da influência das características das chegadas e das partidas de tráfego aos cruzamentos semaforizados com controlo actuado e a pesquisa dirigida para a optimização dos parâmetros da regulação, da localização do detector e da relação entre eles. Ora, os valores a atribuir a estes parâmetros derivam de considerações sobre o comportamento do condutor, as características do veículo e a segurança. De entre os vários parâmetros de controlo fundamentais que podem alterar a operação de um semáforo actuado destacam-se o tempo de verde mínimo, o tempo de verde máximo e a extensão do período de verde.

Os principais factores de controlo que governam a eficiência duma operação semi-actuada dizem respeito ao plano de regulação dos semáforos e à configuração do sensor (detector) e a dificuldade de aplicação do controlo semi-actuado ou totalmente actuado está em seleccionar uma combinação óptima destas operações. A estimação do atraso ao longo das redes semaforizadas é importante como base para um critério a usar na selecção de uma estratégia óptima de controlo de tráfego, de modo a responder às variações dos volumes e condições de tráfego.

3 Conceito e modelação de atraso

Uma das mais importantes consequências resultantes da instalação de semáforos é o atraso e as paragens sofridas pelos veículos, o que justifica a frequente utilização do atraso na definição de critérios de optimização.

3.1 Conceito de atraso

O atraso total de um veículo, d (seg), corresponde à diferença entre o tempo efectivo que ele demora a atravessar o cruzamento e o tempo que demoraria se mantivesse sempre a mesma velocidade (velo-

cidade de cruzeiro), ou seja, em condições ideais, na ausência tanto de incidentes como de controlo de tráfego e sem interacção com veículos a circular na via. Portanto, o atraso sofrido pelo condutor deve-se, principalmente, a factores relacionados com o controlo, mas também à presença de outros veículos e aos eventuais acidentes.

As variáveis independentes que intervêm na definição do atraso podem agrupar-se em 3 conjuntos, englobando parâmetros relativos:

a) às chegadas: débito de chegadas, q (veíc/seg);

b) às partidas: débito de saturação, s (veíc/seg);

c) ao sinal luminoso: duração do ciclo, C (seg); tempo útil de verde, g (seg); tempo de amarelo, A (seg); tempo de limpeza, r (seg).

O grau de saturação, $x_{sat} = \frac{q.C}{s.g}$, é um parâmetro que representa o quociente entre o número médio de veículos que chegam durante um ciclo e o número máximo de veículos que podem passar durante esse período. Este conceito corresponde, na terminologia da teoria das filas de espera, ao conceito de intensidade de tráfego (ou índice de congestionamento) que relaciona, para um certo intervalo de tempo, o número de chegadas de clientes com a capacidade de atendimento do servidor. Para que todos os clientes que chegam ao sistema possam ser atendidos é necessário que a intensidade de tráfego (grau de saturação) seja inferior à unidade. Note-se que, devido ao carácter aleatório das chegadas, mesmo que o grau de saturação (x_{sat}) seja inferior à unidade, existem ciclos saturados (ciclos em que o tempo de verde não é suficiente para escoar todos os veículos), em que a fila de espera não chega a anular-se no fim do tempo de passagem dos veículos. Valores sustentáveis de x_{sat} vão de zero (onde a taxa de fluxo é nula) até 1 (onde a taxa de fluxo é igual à designada capacidade), exclusivé.

A capacidade, $Q = \frac{s.g}{C}$ (veíc/seg), é o número máximo de veículos que podem passar, por unidade de tempo, na linha de paragem, atendendo às condições existentes relativas ao tráfego, à geometria da via e à regulação do semáforo.

O valor do atraso sofrido pelos veículos num cruzamento semaforizado é uma função de vários parâmetros, incluindo a distribuição de chegadas de veículos ao cruzamento. O efeito do tipo de distribuição para as chegadas é mínimo para graus de saturação baixos, mas torna-se significativo quando o fluxo de chegadas se aproxima da capacidade. Neste caso, o atraso real é muito maior do que o atraso previsto para uma situação de chegadas igualmente espaçadas, por causa dos efeitos aleatórios de grandes volumes de tráfego.

3.2 Modelos analíticos do atraso dos condutores

De acordo com estudos comparativos, para controlo de tempos fixos, um modelo de atraso baseado simplesmente em intervalos regulares entre chegadas fornece resultados aceitáveis sempre que o débito de chegadas seja inferior a cerca de 50% da capacidade do sistema. Por outro lado, admitindo que as chegadas dos veículos têm carácter aleatório, a análise do modo como se processa o escoamento dos veículos poderia ser também feita a partir de modelos baseados na teoria de filas de espera. No entanto, na fila do tipo $M/D/1$, considera-se irrealisticamente que, durante o ciclo, as partidas se efectuam a intervalos constantes e iguais a $\frac{C}{s.g}$ (considerado o tempo médio de atendimento), o que conduz a admitir partidas durante o período de vermelho, resultando assim em atrasos inferiores aos verdadeiros. Note-se também que, se não houver chegadas não há partidas. Perante isto, surgiu a necessidade de estudos nesta área que levariam a outros modelos para o atraso dos veículos.

A expressão mais corrente para o cálculo do atraso é sugerida em [2]. Trata-se da seguinte expressão semi-empírica, obtida por simulação em computador:

$$d = \frac{C(1-\frac{g}{C})^2}{2(1-x_{sat}\frac{g}{C})} + \frac{x_{sat}^2}{2q(1-x_{sat})} - 0,65\sqrt[3]{\frac{C}{q^2}}x_{sat}^{(2+5\frac{g}{C})} \quad (1)$$

e é habitualmente denominada Fórmula de *Webster*. A sua obtenção pressupõe que os veículos chegam segundo uma lei de *Poisson*.

Como é do conhecimento geral, o controlo de semáforos actuados tem muito mais parâmetros para os engenheiros fixarem do que o controlo de tempos fixos. Lin [3] refere que a crescente sofisticação na lógica do controlo oferece mais flexibilidade na regulação de semáforos, mas também torna a avaliação do seu desempenho mais difícil. O principal problema é que os atrasos não podem ser facilmente relacionados com os parâmetros de controlo (tempo mínimo e máximo de verde, extensão do tempo de verde ou localização do sensor) em semáforos actuados.

O *Highway Capacity Manual* (HCM) [4] propõe uma metodologia denominada Modelo Generalizado de Atraso (GDM[1]), que considera o facto de, em controlo actuado, a duração do ciclo ser aleatória, contrariamente ao controlo com tempos fixos. Neste modelo intervêm a duração média do ciclo, a duração média do tempo de verde e dois coeficientes de sensibilidade, reflectindo as condições de tráfego e o tipo de controlo semafórico. É notório que, em semáforos actuados, a eficiência do modelo de estimação do atraso dependerá da qualidade das estimativas dos tempos do semáforo. Em 2010, o HCM [4] propunha um método para estimar o comprimento médio do ciclo e o intervalo médio de verde para cada tipo de regulação de semáforos, que apresentava sérios defeitos. Uma fraqueza crítica desse método, segundo Lin [5], consiste em não estabelecer o impacto da regulação dos semáforos e da configuração do detector, pois assume que o comprimento médio do ciclo é apenas função do volume de tráfego associado, tanto com a fase actuada como com a fase não actuada.

Resumindo, os técnicos de tráfego usam habitualmente a Fórmula de *Webster* para estimar o atraso médio sofrido pelos veículos em cruzamentos semaforizados regulados com tempos fixos, o que se revela adequado para graus de saturação inferiores a 70%. Para o caso de controlo actuado, o modelo GDM, apresenta a grande desvantagem de necessitar de estimativas de vários parâmetros que só poderão ser obtidas através de medições em campo ou por simulação.

4 Modelação da fila de espera em cruzamentos semaforizados

4.1 Algumas considerações

Admitindo que as chegadas dos veículos têm carácter aleatório, a análise do modo como se processa o escoamento dos veículos pode ser também feita a partir de modelos baseados na teoria de filas de espera. No entanto, na fila do tipo $M/D/1$, como já tivemos ocasião de referir, considera-se irrealisticamente que, durante o ciclo, as partidas se efectuam a intervalos constantes e iguais a $\frac{C}{s \cdot g}$ (considerado o tempo médio de atendimento), o que conduz a admitir partidas durante o período de vermelho, resultando assim em atrasos inferiores aos verdadeiros. Neste trabalho estamos interessados em mostrar como é que os modelos das filas de espera com pausas do servidor podem ser usados para tratar problemas de tráfego urbano, nomeadamente no que se refere à espera em cruzamentos semaforizados.

Heidemann [1] propõe um modelo analítico em que considera pausas do servidor, mas ainda em condições restritivas em termos da regulação do semáforo. As suas condições de partida são que o processo de chegadas seja um processo de *Poisson*, o cruzamento seja regulado com ciclo fixo, o intervalo entre as partidas dos veículos seja constante. As funções geradoras de probabilidade para o comprimento

[1] em inglês *Generalized Delay Model*

da fila e o atraso do veículo foram deduzidas por investigação das cadeias de *Markov* associadas. Esta técnica permitiu confirmar que a correlação existente no processo de chegada dos veículos não deve ser ignorada, pois conduz a estimações deficientes das medidas de desempenho, especialmente em grandes intensidades de tráfego. A caracterização via processos de chegadas de *Markov* é combinada com método analítico-matricial descrito em [6], conduzindo ao estudo do comprimento da fila de espera em semáforos com ciclo fixo, mas não ainda nos semáforos com ciclo variável, como é o caso dos semáforos semi-actuados, devido à sua complexidade.

Começaremos portanto por apresentar as cadeias de *Markov* do tipo $M/G/1$ e a sua utilização no contexto do tráfego semaforizado regulado por ciclos de comprimento fixo, passando posteriormente ao tratamento das situações de regulação semi-actuada.

Em termos de teoria de espera, cruzamentos regulados por tempos fixos, podem ser vistos como sistemas de espera em que o servidor (semáforo) é desactivado (sinal vermelho) por períodos fixos de tempo. Assim, este sistema pode ser analisado com uma fila do tipo M/G/1 em que ocorre desactivação cíclica do servidor.

Mostramos aqui como é que a teoria de filas de espera com pausas do servidor pode ser explorada de forma a estabelecer um algoritmo de cálculo do atraso médio por veículo no contexto do controlo de tempos fixos. E comparamos os resultados com as fórmulas existente na literatura anteriormente referidas.

4.2 Formulação da cadeia de Markov para controlo fixo

Um cruzamento semaforizado regulado com ciclo fixo, em que tanto o período de verde como o período de vermelho são fixados, é um sistema de espera em que cada veículo, quando chega ao cruzamento, tem de esperar, se existirem veículos à sua frente e chegar durante o período de verde, ou se chegar ao cruzamento durante o período de vermelho. Suponhamos que o período de tempo de verde e o período de tempo de vermelho têm duração $M.T_a$ e $N.T_a$, respectivamente. A constante T_a representa a quantidade de tempo necessária para que um veículo atravesse o cruzamento. Na terminologia dos sistemas de espera, estamos perante um sistema que consiste num servidor (semáforo) que está activo em períodos de duração $M.T_a$ e inactivo por períodos de duração $N.T_a$, sendo o tempo de atendimento igual a T_a segundos, e que apresenta as seguintes especificações:

- Os veículos chegam ao cruzamento de acordo com um processo de *Poisson* de taxa λ;
- O servidor está apto a servir um grupo de quando muito r (número de vias) veículos em cada período de serviço (período de verde);
- O tempo de atendimento, para cada grupo, é constante e igual a T_a segundos;
- O serviço só começa e termina nos instantes de tempo $0, T_a, 2T_a, \ldots, nT_a, \ldots$;
- O ciclo do semáforo consiste num intervalo de verde, de duração $M.T_a$ segundos, seguido de um intervalo de vermelho de duração $N.T_a$ segundos.

Considerar que o tempo de atendimento é constante é pouco realista na aplicação a sistemas de espera de tráfego, em cruzamentos semaforizados, uma vez que apenas os veículos que estão parados em fila de espera, quando o semáforo abre, despendem algum tempo no arranque, o mesmo não se passando com os veículos que não param (porque o sinal está verde e a fila já foi esvaziada), pois passam a linha de paragem à velocidade desejada.

Tomando como referência uma via de acesso ao cruzamento, designemos por $\{L(t), \xi(t)\}$, com $t \geq 0$, um sistema de estados ao longo do instante t, onde $L(t)$ é o comprimento da fila no instante t e $\xi(t)$ o

número de intervalos de duração T_a decorridos até ao instante t, desde a última abertura do sinal verde. O espaço de estados de $L(t)$ é $\mathbb{N}$ e o de $\xi(t)$ é $\{0,1,2,\ldots,M+N\}$. O processo $\{\xi(t)\}$ é também referido como o estado do sinal do sistema. Define-se também, no instante t, o sistema $\{X,T_a\}=\{X_n,t_n;n\in\mathbb{N}\}$, onde $X_n=\{L(t_n),\xi(t_n)\}$ e $t_n=nT_a$. O processo $\{L(t),\xi(t)\}$, com $t\geq 0$, é um processo semi-regenerativo e o processo markoviano de renovamento associado é $\{X,T_a\}$ [9].

A matriz de probabilidades de transição de X_n é uma matriz por blocos, cuja estrutura algo particular fica definida por:

$$Q=\begin{bmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & \cdots \\ A_0 & A_1 & A_2 & A_3 & A_4 & \cdots \\ 0 & A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & 0 & A_0 & A_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{2}$$

onde A_k e B_k, com $k\geq 0$, são matrizes $(M+N)\times(M+N)$.

Neste caso, B_k e A_k podem ser definidas por:

$$(B_k)_{ij}=\begin{cases} a_k, & i=1,2,\ldots M+N-1, j=i+1 \\ a_k, & i=M+N, j=1 \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

e

$$(A_k)_{ij}=\begin{cases} a_k, & i=1,2,\ldots,M-1, j=i+1 \\ a_{k-1}, & i=M,M+1,\ldots,M+N-1, j=i+1 \\ a_k, & i=M+N, j=1 \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

com $a_k=e^{-\lambda T_a}\dfrac{(\lambda T_a)^k}{k!}$, for $k=0,1,2,\ldots$ Define-se a matriz estocástica $A=\sum_{i=0}^{\infty}A_i$, e o vector de probabilidade invariante, π, a ela associado, definido por $\pi A=\pi$ e $\pi\mathbb{1}=1$ com $\mathbb{1}=[1,1,\ldots]^T$. O vector π é único e de componentes positivas, desde que a matriz A seja irredutível. De facto, os elementos de π que resultam da aplicação da teoria das cadeias de *Markov* do tipo $M/G/1$ aos cruzamentos semaforizados regulados por ciclo fixo, são todos iguais entre si e iguais a $1/(M+N)$, no pressuposto das chegadas serem *Poissonianas*.

Definindo o vector $\beta=\sum_{i=1}^{\infty}iA_i\mathbb{1}$ e $\rho=\pi\beta$, a condição necessária e suficiente para a ergodicidade da cadeia X_n é $M-(M+N)\lambda T_a>0$, onde λT_a representa o número médio de chegadas de *Poisson* num intervalo de tempo de duração T_a. E assim, fica garantido que $\rho\leq 1$, isto é, $\dfrac{\lambda T_a(M+N)+N}{M+N}\leq 1$.

Na análise das filas de espera, um passo fundamental no manuseamento de cadeias de *Markov* deste tipo é o cálculo da matriz estocástica G que é solução não negativa minimal[2] da equação matricial não linear (ver detalhes em [7]) $G=\sum_{i=0}^{+\infty}A_iG^i$, onde G é uma matriz $m\times m$, tornando-se essencial para a determinação do vector de probabilidade invariante associado a Q, $u=\left[u^0,u^1,\ldots\right]^T$. Uma vez calculada a matriz G, podemos obter os vectores u^k, com $k\geq 1$, através da Fórmula Recursiva de *Ramaswami* [8]:

$$u^k=\left[u^0\bar{B}_i+\sum_{j=1}^{i-1}\pi^j\bar{A}_{i+1-j}\right]\left(I-\bar{A}_1\right)^{-1},\quad k\geq 1, \tag{5}$$

[2] A equação $G=\sum_{i=0}^{+\infty}A_iG^i$ tem pelo menos uma solução, G, no conjunto das matrizes não negativas tal que $G\mathbb{1}\leq\mathbb{1}$. A matriz G é a solução não negativa tal que $\|G\mathbb{1}-\mathbb{1}\|$ é mínima.

onde

$$\bar{A}_n = \sum_{i=n}^{+\infty} A_i G^{i-n} \quad e \quad \bar{B}_n = \sum_{i=n}^{+\infty} B_i G^{i-n}, \quad n \geq 0 \tag{6}$$

e u^0 é solução do sistema

$$\begin{cases} u^0 &= u^0 K \\ u^0 \kappa &= 1 \end{cases}, \tag{7}$$

$$K = \sum_{h=0}^{+\infty} B_h G^h, \tag{8}$$

$$\kappa = \mathbb{1} + \sum_{i=0}^{+\infty} B_i \sum_{k=0}^{i-1} G^k \mu, \tag{9}$$

$$\mu = \left[I - \sum_{i=1}^{+\infty} A_i \sum_{k=0}^{i-1} G^k \right]^{-1} \mathbb{1}. \tag{10}$$

A matriz G pode ser obtida através do seguinte processo recursivo:

$$G = lim_{n \to +\infty} G_n, \tag{11}$$

$$G_1 = (I - A_1)^{-1} A_0, \tag{12}$$

$$G_{n+1} = \left(I - \sum_{i=1}^{+\infty} A_i G_n^{i-1} \right)^{-1} A_0, \quad n > 1. \tag{13}$$

A sucessão de matrizes assim definida é monótona não decrescente e converge para G. A matriz inversa necessária ao cálculo de G_{n+1}, com $n \geq 0$, existe sempre que $G_n \leq G$.

Uma vez implementados os algoritmos para o cálculo da matriz G e do vector u, torna-se possível estudar as propriedades probabilísticas da cadeia de *Markov* do tipo $M/G/1$, nomeadamente o tempo de permanência no sistema, o tempo de espera e o comprimento da fila. Para que a utilização desta teoria no contexto do tráfego em cruzamentos semaforizados se torne clara, expomos de seguida a forma de se chegar a um modelo analítico do tráfego em cruzamentos regulados por ciclo fixo, em particular, e o procedimento a seguir para o estudo de características como o comprimento da fila de espera e o tempo de espera.

4.3 Algoritmo para cálculo do atraso dos veículos

De [9], o comprimento médio da fila, $E[L]$, é dado por

$$E[L] = \frac{\lambda T_a}{2} + \sum_{i=1}^{M+N} \sum_{k=0}^{+\infty} k u_i^k \tag{14}$$

e o tempo médio no sistema, $E[W]$, é simplesmente obtido por aplicação da fórmula de *Little*,

$$E[W] = E[L]/\lambda. \tag{15}$$

O algoritmo a seguir mostra explicitamente como implementar os cálculos para obter o comprimento médio da fila e o tempo médio de espera num cruzamento regulado por tempos fixos:

1. De acordo com a precisão fixada:

- para $k \geq 0$, calcular as matrizes não nulas B_k (3) até que $B_k = O$ (matriz nula); fixar índice k da última matriz não nula B_k, n_B;
- para $k \geq 0$, calcular as matrizes não nulas A_k (4) até que $A_k = O$ (matriz nula); fixar índice k da última matriz não nula A_k, n_A;
- fixar $n_{max} = \max(n_A, n_B)$;

2. Calcular a matriz G usando as expressões (11) – (13);
3. Calcular os vectores μ, κ e a matriz K usando as expressões (8) – (10);
4. Resolver o sistema (7);
5. Para $n = n_A, n_A - 1, \cdots, 1$ calcular $\bar{A}_n$ e para $n = n_B, n_B - 1, \cdots, 1$ calcular $\bar{B}_n$, usando (6);
6. Para $k = 1, 2, \cdots, n_{max}$ calcular os vectores u^k (de dimensão $M+N$) usando a expressão (5);
7. Calcular $E[L]$ usando a expressão (14) e $E[W]$ usando a expressão (15).

Este algoritmo será aplicado a seguir para calcular os tempos médios de espera no caso de um cruzamento regulados com tempos fixos. Consideramos um cruzamento com duas fases e o seguinte plano de tempos: $g = 30\,s$ and $C = 60\,s$.

O tempo médio de espera (atraso médio) estimado pelo modelo apresentado (referido como modelo de Markov) é apresentado na Figura 4.3 em conjunto com os resultados obtidos por aplicação da fórmula de Webster (1) and the HCM model [4].

O valor $T_a = 2\,s$ é usado como sendo o mais comum em engenharia de tráfego. Os atrasos por veículo apresentados correspondem apenas a uma das vias do cruzamento.

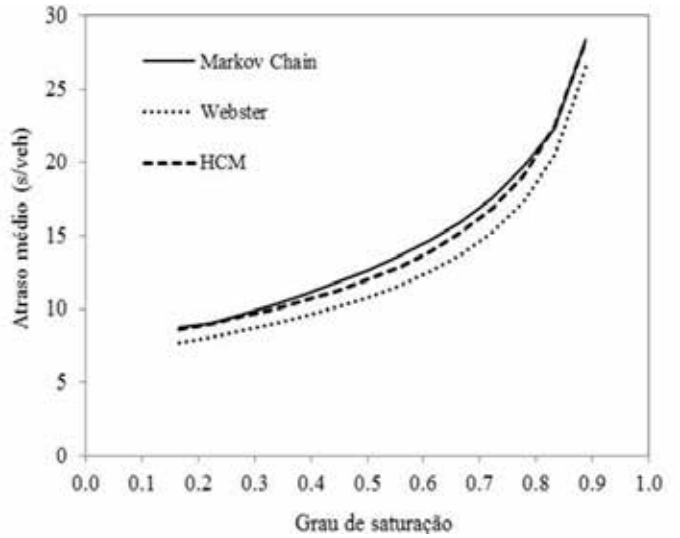


Figura 1: Comparação do atraso estimado pelo modelo de Markov, pela fórmula de Webster e pelo modelo do HCM.

Os resultados sugerem que o modelo de Markov oferece boas estimativas para o atraso médio dos condutores pelas expressões (14) – (15). As estimativas obtidas pela fórmula de Webster são normalmente inferiores às obtidas pelo modelo de HCM, assim como pelo modelo de Markov.

5 Comentários gerais e tópicos de investigação futura

Um algoritmo para calcular o atraso médio dos veículos em cruzamentos regulados por controlo fixo foi desenvolvido aplicando a teoria de filas com pausas do servidor. O modelo estocástico de filas espera com pausas do servidor reflete adequadamente os períodos de sinal vermelho que ocorrem ciclicamente neste tipo de controlo semafórico. A expressão que apresentamos neste trabalho fornece estimativas realistas do atraso médio de um veículo, de acordo com o modelo apresentado no HCM [4].

Este tipo de modelação está a ser estendido para aplicação em cruzamentos regulados com controlo semi-actuado (investigação em curso).

Referências

[1] Heidemann, D. 1994. Queue length and delay distributions at traffic signals. *Transportation Res. B* **28**(5) 377–389.

[2] Webster, F. V. 1958. *Traffic Signal Settings*, Road Research Laboratory, Road Research Laboratory 39, HMSO, London.

[3] Lin, F.-B. and Mazdeyasna, F. 1983. Delay models of traffic-actuated signal controls. *Transportation Research Record* **905** 33–39.

[4] Transportation Research Board (TRB). 2010. *Highway capacity manual* (TRB, National Research Council, Washington D.C.).

[5] Lin, F.-B. 1990. Estimating Average cycle lengths and green intervals of semiactuated signal operations for level-of-service analysis. *Transportation Research Record* **1287** 119–128.

[6] Neuts, M.F. 1989. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker Inc., New York.

[7] Latouche, G. 1994. Newton's iteration for non-linear equations in Markov chains. In *IMA Journal of Numerical Analysis* **14** 583–598.

[8] Ramaswami, V. 1988. A stable recursion for the steady state vector in Markov chains of M/G/1 type. In *Communications Statistics—Stochastic Models* **4(1)** 183–188.

[9] Hu, X.N., Tang, L.C. and Ong, H.L. 1997. A $M/D^X/1$ vacation queue model for signalized intersection. In *Computers Industrial Engineering* **33(3–4)** 801–804.

Regressão Linear com Variáveis Fortemente Correlacionadas

Mário A. T. Figueiredo, *mario.figueiredo@tecnico.ulisboa.pt*
Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Robert D. Nowak, *rdnowak@wisc.edu*
Department of Electrical and Computer Engineering University of Wisconsin, Madison, USA

1 Introdução

A *regressão linear* (RL) é uma das ferramentas básicas da análise estatística de dados, através da qual se infere um modelo que permite prever uma variável (dita *dependente*) como combinação linear de um conjunto de outras variáveis (ditas *independentes*). O critério clássico usado em RL consiste em minimizar a soma dos quadrados dos erros/diferenças entre os valores preditos e os valores observados num conjunto de dados. Outros critérios são também frequentemente usados, tais como a minimização da soma dos valores absolutos dos erros, opção que possui propriedades de robustez perante observações espúrias. Um dos problemas centrais em RL consiste na selecção de um subconjunto de variáveis relevantes para o problema em causa, tarefa de grande importância, por vários motivos: um modelo com um número reduzido de variáveis é computacionalmente mais leve de aplicar; um modelo mais simples apresenta melhores características de generalização; a identificação das variáveis relevantes tem, frequentemente, significado científico (por exemplo, quais os genes envolvidos numa determinada doença). Nas duas últimas décadas, as técnicas de selecção de variáveis baseadas em regularização indutora de esparsidade (ou seja, que dão preferência a modelos com poucas variáveis activas, as quais correspondem a coeficientes de regressão não nulos) tornaram-se o paradigma dominante [Hastie et al., 2015], sendo o LASSO (*least absolute shrinkage and selection operator* [Tibshirani, 1996]) o exemplo mais conhecido.

Em problemas de RL de alta dimensão (elevado número de variáveis independentes, relativamente ao número de amostras), é natural encontrar-se variáveis altamente correlacionadas. Por exemplo, em dados de expressão genética, é comum estar-se na presença de grupos de genes fortemente co-regulados (*co-regulated*). Nestas circunstâncias, a utilização de técnicas de selecção de variáveis simplesmente baseadas em esparsidade (nomeadamente o LASSO) conduzem a resultados não satisfatórios, seleccionando um subconjunto dessas variáveis correlacionadas, ou uma combinação convexa arbitrária das mesmas, e exibindo grande instabilidade. Em muitas aplicações (por exemplo, de natureza científica), é muitas vezes desejável, para efeitos de interpretabilidade, que se identifiquem *todas* as variáveis relevantes para um dado problema de regressão, não apenas um subconjunto; para esse efeito, várias abordagens têm sido propostas recentemente [Bühlmann et al., 2013, Genovese et al., 2012, Shen and Huang, 2010], sendo que a mais conhecida é a *rede elástica* (EN–*ellastic net*) [Zou and Hastie, 2005].

Neste artigo, descreve-se uma abordagem recente para problemas de RL com variáveis fortemente correlacionadas, baseada na chamada norma ℓ_1 ordenada e ponderada (OWL–ordered weighted ℓ_1) [Zeng and Figueiredo, 2014a], [Bogdan et al., 2015], [Figueiredo and Nowak, 2016], a qual generaliza o chamado OSCAR (*octagonal shrinkage and clustering algorithm for regression* [Bondell and Reich, 2007]). Em particular, depois de rever a norma OWL e algumas das suas propriedades fundamentais, descrevem-se resultados teóricos que dão suporte à utilização de regularização OWL em problems de RL com variáveis fortemente correlacionadas. Em particular, apresentam-se condições suficientes para o agrupamento de variáveis fortemente correlacionadas, mostrando que a norma OWL não sofre do problema acima

apontado para o LASSO e que tem a capacidade de identificar explicitamente conjuntos de variáveis correlacionadas.

2 Regressão Linear, Regularização e Esparsidade

Em RL, observam-se n pares $(\boldsymbol{a}_{(1)}, y_1), ..., (\boldsymbol{a}_{(n)}, y_n)$, onde $\boldsymbol{a}_{(i)} \in \mathbb{R}^p$ e $y_i \in \mathbb{R}$ são, respectivamente, o vector de variáveis independentes e a variável dependente (ou resposta) do i-ésimo par. A formulação clássica para RL regularizada combina o erro quadrático (*squared error* ou *residual sum of squares*–RSS) com um regularizador $\Omega : \mathbb{R}^p \to \mathbb{R}_+$, obtendo-se uma função objectivo cujo minimização produz uma estimativa dos coeficientes de regressão,

$$\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \frac{1}{2} RSS(\boldsymbol{x}) + \eta\,\Omega(\boldsymbol{x}), \quad \text{onde} \quad RSS(\boldsymbol{x}) = \sum_{i=1}^{n} \Big(\underbrace{x_0 + \sum_{j=1}^{p} a_{ij} x_j}_{\hat{y}_i} - y_i \Big)^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2, \qquad (1)$$

onde $\|\cdot\|_2$ representa a norma euclideana, a_{ij} é o j-ésimo componente da amostra $\boldsymbol{a}_{(i)}$, a matriz $\boldsymbol{A}$ tem dimensões $n \times p$, com $\boldsymbol{a}_{(i)}$ na linha i. É comum na prática normalizar as variáveis (independentes e dependente) para que tenham média nula, pelo que $\sum_i a_{ij} = 0$ e $\sum_i y_i = 0$, e variância unitária, $\sum_i a_{ij}^2 = 1$, pelo que não há perda de generalidade em considerar a ausência de um termo constante (*offset*) no valor predito $\hat{y}_i$, isto é, tomar $x_0 = 0$ [Hastie et al., 2009].

Podem considerar-se formulações alternativas baseadas na mesma função de erro e regularizador,

$$\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} RSS(\boldsymbol{x}), \text{ sujeito a } \Omega(\boldsymbol{x}) \le \tau \qquad \text{ou} \qquad \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \Omega(\boldsymbol{x}), \text{ sujeito a } RSS(\boldsymbol{x}) \le \delta, \quad (2)$$

por vezes referidas como formulações de Morozov e de Ivanov, respectivamente, enquanto que (1) é chamada formulação de Tikhonov [Lorenz and Worliczek, 2013]. Se Ω for convexo (por exemplo, uma norma ou o quadrado de uma norma), as três formulações são equivalentes no sentido em que, dada uma delas, existe uma escolha do parâmetro de cada uma das outras duas para a qual as soluções coincidem (embora a determinação desse parâmetro possa exigir a resolução do problema).

Se a matriz $\boldsymbol{A}$ possuir característica (*rank*) igual a p, a solução do problema (1) com $\eta = 0$, dita solução de mínimo erro quadrático (*least squares*–LS) é dada por $\hat{\boldsymbol{x}}_{\text{LS}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$. No entanto, se $\boldsymbol{A}$ tiver característica inferior a p (por exemplo, se $n < p$), ou se for muito mal condicionada e como tal a inversão $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ for numericamente instável, $\hat{\boldsymbol{x}}_{\text{LS}}$ é indeterminado ou numericamente impraticável. Esta é uma das motivações para o uso de regularização em RL, em particular para a introdução da regressão *ridge*, na qual $\Omega(\boldsymbol{x}) = \Omega_{\text{ridge}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2$, escolha que implica que a solução de (1) seja dada por $\hat{\boldsymbol{x}}_{\text{ridge}} = (\boldsymbol{A}^T\boldsymbol{A} + \eta\,\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$ [Hoerl and Kennard, 1970].

A regressão *ridge*, embora seja aplicável em cenários de alta dimensão (nomeadamente, para $n \ll p$), produz (quase certamente) estimativas do vector de regressão $\boldsymbol{x}$ no qual todos os componentes são diferentes de zero, pelo que não implementa uma selecção de variáveis explícita. Para esse efeito, o LASSO usa $\Omega(\boldsymbol{x}) = \Omega_{\text{LASSO}}(\boldsymbol{x}) = \|\mathbf{x}\|_1 = \sum_i |x_i|$ (norma ℓ_1), encorajando a esparsidade de $\hat{\boldsymbol{x}}$ [Tibshirani, 1996]. O efeito indutor de esparsidade do regularizador pode ser facilmente entendido no caso em que $\boldsymbol{A}$ é uma matriz ortogonal (ou seja, $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, onde $\boldsymbol{I}$ é uma matriz identidade); neste caso,

$$\hat{\boldsymbol{x}}_{\text{LASSO}} = \text{soft}(\boldsymbol{A}^T\boldsymbol{y}, \eta), \qquad (3)$$

onde a função "soft" é definida componente-a-componente como $\text{soft}(u, \eta) = \text{sign}(u)\max(|u| - \eta, 0)$, pelo que algumas componentes de $\hat{\boldsymbol{x}}_{\text{LASSO}}$ podem ser exactamente zero. Condições de indução de esparsidade do regularizador ℓ_1 para matrizes $\boldsymbol{A}$ mais gerais são mais complexas e não serão aqui abordadas [Hastie et al., 2015]. As propriedades estatísticas da regressão LASSO têm sido objecto de intenso estudo na últimas duas décadas; não sendo esse o foco deste texto, ao leitor interessado é sugerida a consulta do excelente recente livro de Hastie et al. [2015].

3 Regressão com Variáveis Fortemente Correlacionadas

Em problemas caracterizados pela presença de variáveis fortemente correlacionadas, o LASSO produz resultados pouco satisfatórios, em particular no que respeita à interpretabilidade e estabilidade da solução perante pequenas variações nos dados [Bühlmann et al., 2013]. Para obter alguma intuição acerca desta afirmação, considere-se um caso extremo, no qual duas variáveis (digamos j e k) são perfeitamente correlacionadas, ou seja, as colunas j e k de $\boldsymbol{A}$ (designadas $\boldsymbol{a}_j, \boldsymbol{a}_k \in \mathbb{R}^n$, recorde-se que a linha i de $\boldsymbol{A}$ é referida como $\boldsymbol{a}_{(i)}$) são colineares. Dado que a normalização acima assumida significa que $\|\boldsymbol{a}_j\|_2 = \|\boldsymbol{a}_k\|_2 = 1$, esta colinearidade implica que $\boldsymbol{a}_j = \pm\boldsymbol{a}_k$; considere-se, por exemplo, $\boldsymbol{a}_j = \boldsymbol{a}_k$. Nestas circunstâncias, pode escrever-se

$$RSS(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \left\|\boldsymbol{y} - \boldsymbol{a}_j(x_j + x_k) - \sum_{\substack{l=1 \\ l\neq j,k}}^{p} x_l \boldsymbol{a}_l\right\|_2^2 = \left\|\boldsymbol{y} - \boldsymbol{a}_j(x'_j + x'_k) - \sum_{\substack{l=1 \\ l\neq j,k}}^{p} x_l \boldsymbol{a}_l\right\|_2^2 \tag{4}$$

para $x'_j = x_j + d$ e $x'_k = x_k - d$, para qualquer $d \in \mathbb{R}$. Considere-se, sem perda de generalidade, que se sabe que $RSS(\boldsymbol{x})$ tem mínimos no primeiro ortante e adicione-se o regularizador ℓ_1

$$\hat{\boldsymbol{x}}_{\text{LASSO}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^p_+} \left\|\boldsymbol{y} - \boldsymbol{a}_j(x_j + x_k) - \sum_{\substack{l=1 \\ l\neq j,k}}^{p} x_l \boldsymbol{a}_l\right\|_2^2 + \eta(x_j + x_k) + \eta \sum_{\substack{l=1 \\ l\neq j,k}}^{p} x_l, \tag{5}$$

que apresenta o mesmo tipo de invariância (logo ausência de um minimizante único) que $RSS(\boldsymbol{x})$. Por exemplo, se para uma solução $\hat{\boldsymbol{x}}$ de (5), se tiver $\hat{x}_j = b$ e $\hat{x}_k = 0$, qualquer outro vector $\hat{\boldsymbol{x}}'$ com $\hat{x}'_j = \alpha b$ e $\hat{x}'_k = (1-\alpha)b$, para $\alpha \in [0,1]$, é também solução de (5), pois $\hat{x}_j + \hat{x}_k = \hat{x}'_j + \hat{x}'_k = \alpha b + (1-\alpha)b = b$. Em casos mais realistas nos quais a correlação não é perfeita, mas apenas aproximada, a função objectivo não apresenta exactamente esta invariância, mas apenas aproximadamente, ou seja, a localização do(s) mínimo(s) é muito sensível (instável, pouco robusta) a pequenas variações nos dados; esta situação é ilustrada na Figura 1(a), onde é claro que a função objectivo apresenta a quase-invariância acabada de descrever.

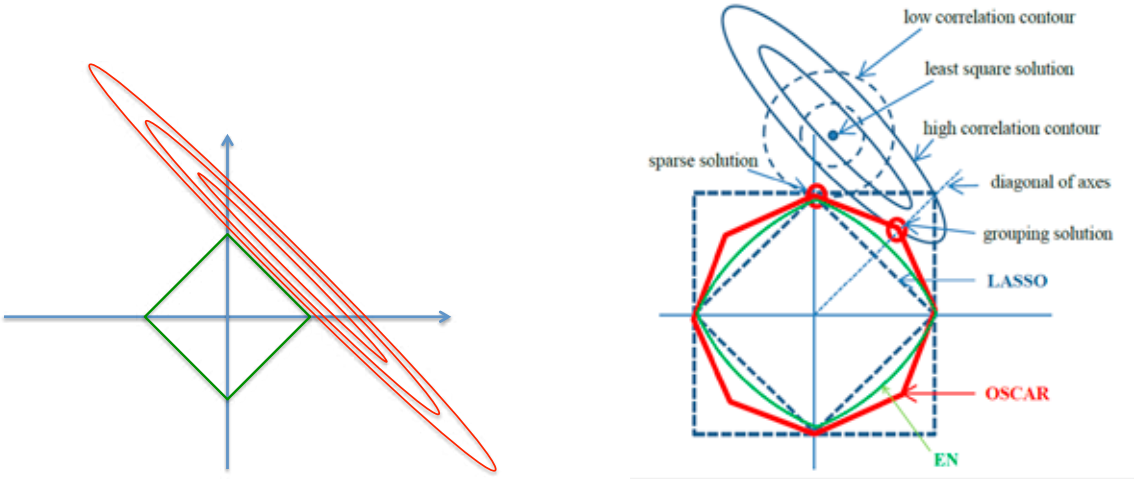


Figura 1: (a) Ilustração do comportamento do LASSO na presença de variáveis fortemente correlacionadas; o quadrado verde é uma curva de nível da norma ℓ_1 em $\mathbb{R}^2$ e as elipses vermelhas são curvas de nível de $RSS(\boldsymbol{x})$. (b) Curvas de nível de Ω_{LASSO}, Ω_{EN} e Ω_{OSCAR}, ilustrando o seu comportamento qualitativamente diferente.

Várias abordagens têm sido propostas para obviar à inadequação do LASSO para lidar com variáveis fortemente correlacionadas, sendo a *rede elástica* (EN–*ellastic net* [Zou and Hastie, 2005]) provavelmente a mais conhecida. Na EN, o regularizador é uma combinação linear da norma ℓ_1 com o quadrado

da norma ℓ_2, ou seja $\Omega_{\text{EN}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 + \beta\|\boldsymbol{x}\|_2^2$, cujo objectivo é quebrar a invariância de Ω_{LASSO} acima descrita. De facto, para qualquer $\alpha \in [0,1]$,

$$\Omega_{\text{LASSO}}\left(\alpha\begin{bmatrix} 0 \\ \pm b \end{bmatrix} + (1-\alpha)\begin{bmatrix} \pm b \\ 0 \end{bmatrix}\right) = |b|,$$

enquanto que

$$\Omega_{\text{EN}}\left(\alpha\begin{bmatrix} 0 \\ \pm b \end{bmatrix} + (1-\alpha)\begin{bmatrix} \pm b \\ 0 \end{bmatrix}\right) = |b| + \beta b^2(1+2\alpha^2-2\alpha),$$

sendo fácil verificar que $(1+2\alpha^2-2\alpha)$ é minimizado para $\alpha = 1/2$, o que mostra que Ω_{EN} dá preferência a soluções com componentes com valores absolutos iguais. Este efeito é ilustrado na Figura 1(b), reflectindo-se na forma das curvas de nível de Ω_{EN} nas quais é evidente a quebra da invariância exibida por Ω_{LASSO}.

Embora a regularização por EN não padeça do problema do LASSO perante um conjunto de variáveis fortemente correlacionadas, as soluções obtidas com este regularizador não indicam explicitamente a presença de um tal conjunto. Para este efeito, foi proposto o OSCAR (*octagonal shrinkage and clustering algorithm for regression* [Bondell and Reich, 2007]), o qual, como o nome sugere, é capaz de agrupar (*cluster*) explicitamente conjuntos de variáveis. Este método consiste em adoptar o seguinte regularizador

$$\Omega_{\text{OSCAR}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 + \gamma\sum_{j=1}^{p}\sum_{k=j+1}^{p}\max\{|x_j|,|x_k|\}, \tag{6}$$

cuja curva de nível em $\mathbb{R}^2$ é octogonal (como se ilustra na Figura 1(b)), sendo essa a justificação para a designação adoptada por Bondell e Reich. Tal como Ω_{EN}, o regularizador Ω_{OSCAR} dá preferência a vectores com componentes iguais (em valor absoluto); por exemplo, no caso bidimensional,

$$\Omega_{\text{OSCAR}}\left(\alpha\begin{bmatrix} 0 \\ \pm b \end{bmatrix} + (1-\alpha)\begin{bmatrix} \pm b \\ 0 \end{bmatrix}\right) = |b| + \gamma|b|\max\{\alpha, 1-\alpha\},$$

cujo mínimo é atingido para $\alpha = 1/2$. No entanto, existe uma diferença qualitativa entre Ω_{OSCAR} e Ω_{EN}: em RL regularizada com Ω_{EN}, existe um limiar para a correlação entre um par de variáveis acima do qual se garante que as estimativas dos coeficientes correspondentes são exactamente iguais, o que não é verdade quando se usa Ω_{EN}. Este resultado vai ser formalmente enunciado na próxima secção, onde se mostra que Ω_{OSCAR} pertence a uma família mais geral de normas e se descrevem algumas resultados teóricos que caracterizam a sua utilização como regularizador em problemas de RL com variáveis fortemente correlacionadas.

4 Norma ℓ_1 Ordenada e Ponderada, Majorização e Convexidade de Schur

Antes de apresentar a norma que dá título a esta secção, introduz-se alguma notação. Dado um vector $\boldsymbol{x} \in \mathbb{R}^p$, usa-se $|\boldsymbol{x}|$ para representar o vector com os valores absolutos dos componentes de $\boldsymbol{x}$ e $x_{[i]}$ para o seu *i*-ésimo maior componente (isto é, $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[p]}$); finalmente, $\boldsymbol{x}_\downarrow$ é o vector obtido por ordenação dos componentes de $\boldsymbol{x}$ em ordem não crescente (ou seja, $x_{[i]}$ é o componente i de $\boldsymbol{x}_\downarrow$, enquanto que $|x|_{[i]}$ é o componente i de $|\boldsymbol{x}|_\downarrow$).

A norma ℓ_1 ordenada e ponderada (*ordered weighted* ℓ_1*–OWL*) é definida como

$$\Omega_{\text{OWL}}^{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i\,|x|_{[i]} = \boldsymbol{w}^T|\boldsymbol{x}|_\downarrow, \tag{7}$$

onde $\boldsymbol{w} \in \mathbb{R}_+^p$ é um vector de pesos, tais que $w_1 > 0$, $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$. É possível mostrar que $\Omega_{\text{OWL}}^{\boldsymbol{w}}$ é de facto uma norma [Zeng and Figueiredo, 2014a] e contém as norma ℓ_1 e ℓ_∞ como casos

particulares, para $w_1 = w_2 = \cdots = w_p = 1$ e para $w_1 = 1, w_2 = w_3 = \cdots = w_p = 0$, respectivamente. É também fácil verificar que tomando $w_i = 1 + \gamma(p-i)$ para $i = 1, ..., p$, tem-se

$$\Omega^{\boldsymbol{w}}_{\text{OWL}}(\boldsymbol{x}) = \Omega_{\text{OSCAR}}(\boldsymbol{x}), \tag{8}$$

pelo que a família de normas OWL também contém Ω_{OSCAR} [Zeng and Figueiredo, 2014a].

Antes de prosseguir, deve referir-se que existem ferramentas computacionais que permitem resolver problemas da forma (1) ou (2), com $\Omega = \Omega^{\boldsymbol{w}}_{\text{OWL}}$, com um custo computacional que escala de acordo com $O(p \log p)$ [Davis, 2015], [Zeng and Figueiredo, 2014b]. A abordagem alternativa que consiste em agrupar (*cluster*) previamente as colunas [Bühlmann et al., 2013], ou identificar todas aquelas que exibem uma correlação acima de um dado limiar escala de acordo com $O(p^2)$, devido à necessidade de calcular correlação (ou outra medida de semelhança) entre todos os pares de colunas de $\boldsymbol{A}$.

Uma ferramenta formal que ajudará a estudar família de normas OWL é teoria da majorização e o conceito de convexidade de Schur [Marshall et al., 2011]. Comecemos por recordar que, sendo $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ dois vectores, diz-se que $\boldsymbol{y}$ *majoriza* $\boldsymbol{x}$ (escreve-se $\boldsymbol{y} \succ \boldsymbol{x}$) se

$$\sum_{i=1}^{p} y_i = \sum_{i=1}^{p} x_i \quad \text{e} \quad \sum_{i=1}^{k} y_{[i]} \geq \sum_{i=1}^{k} x_{[i]}, \text{ para } k = 1, ..., p-1.$$

Intuitivamente, $\boldsymbol{y} \succ \boldsymbol{x}$ se os dois vectores têm a mesma soma e $\boldsymbol{y}$ tem uma distribuição de valores mais heterogénea que $\boldsymbol{x}$. Por exemplo, $(4,0,0,0) \succ (0,3,0,1) \succ (2.9,0.1,0,1) \succ (1,1,1,1)$. Se $\boldsymbol{y}$ difere de $\boldsymbol{x}$ por uma permutação dos seus componentes, então $\boldsymbol{y} \succ \boldsymbol{x}$ e $\boldsymbol{x} \succ \boldsymbol{y}$. A relação de majorização é uma *pré-ordem*, por ser reflexiva ($\forall \boldsymbol{x}$, $\boldsymbol{x} \succ \boldsymbol{x}$) e transitiva ($\boldsymbol{z} \succ \boldsymbol{y}) \wedge (\boldsymbol{y} \succ \boldsymbol{x}) \Rightarrow (\boldsymbol{z} \succ \boldsymbol{x})$) [Marshall et al., 2011].

Sendo $\mathcal{A} \subseteq \mathbb{R}^p$, uma função $\phi : \mathcal{A} \to \mathbb{R}$ é dita convexa no sentido de Schur (*Schur-convexa*) em $\mathcal{A}$, se $\boldsymbol{y} \succ \boldsymbol{x} \Rightarrow \phi(\boldsymbol{y}) \geq \phi(\boldsymbol{x})$. Se se verificar que

$$(\boldsymbol{y} \succ \boldsymbol{x}) \wedge (\boldsymbol{y} \text{ não é uma permutação de } \boldsymbol{x}) \;\Rightarrow\; \phi(\boldsymbol{y}) > \phi(\boldsymbol{x}),$$

então ϕ é dita *estritamente Schur-convexa*. Intuitivamente, uma função Schur-convexa dá preferência (no sentido em que atribui valores menores) a vectores com distribuição de valores mais homogénea. Mostra-se que qualquer função convexa e simétrica (invariante sob permutações dos argumentos) é Schur-convexa [Marshall et al., 2011]; por exemplo, a norma ℓ_1 é Schur-convexa, mas não estritamente.

Considere-se $\boldsymbol{y}$, tal que $y_i > y_j$. Uma *transferência de Pigou-Dalton* (também conhecida como *transferência de Robin dos Bosques*) de valor $\varepsilon \in (0, (y_i - y_j)/2)$ aplicada a $\boldsymbol{y}$ produz um vector $\boldsymbol{x} \prec \boldsymbol{y}$, tal que $x_i = y_i - \varepsilon$, $x_j = y_j + \varepsilon$ e $x_k = y_k$, para $k \neq i, j$ [Dalton, 1920, Marshall et al., 2011, Pigou, 1912]. Uma função $\phi : \mathcal{A} \to \mathbb{R}$ é dita δ*–fortemente Schur-convexa* se $\phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) \geq \delta\varepsilon$, quando $\boldsymbol{x}$ resulta de transferência de Pigou-Dalton de valor ε aplicada a $\boldsymbol{y}$ [Figueiredo and Nowak, 2016].

O resultado fundamental que caracteriza a normal OWL e no qual se baseiam os teoremas que se seguem é o seguinte: $\Omega^{\boldsymbol{w}}_{\text{OWL}}$ é $\Delta_{\boldsymbol{w}}$*–fortemente Schur-convexa*, com $\Delta_{\boldsymbol{w}} = \min\{w_i - w_{i+1},\ i = 1, ..., p-1\}$. Pode mostrar-se que a norma ℓ_1, embora Schur-convexa não o é fortemente (nem estritamente), enquanto que Ω_{EN} é estritamente Schur-convexa, mas não fortemente. Este resultado afirma que a norma OWL dá preferência *forte*, no sentido expresso pela definição de convexidade de Schur forte, a vectores com distribuição de valores homogénea e é o pilar fundamental da demonstração do resultado seguinte [Figueiredo and Nowak, 2016].

Teorema 4.1 *Seja* $\widehat{\boldsymbol{x}}$ *uma solução de* (1)*, com* $\eta = 1$ *e* $\Omega = \Omega^{\boldsymbol{w}}_{OWL}$*, e* $\boldsymbol{a}_i$ *e* $\boldsymbol{a}_j$ *duas colunas* $\boldsymbol{A}$*. Então,*

$$\textbf{(a)} \quad \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \;\Rightarrow\; \widehat{x}_i = \widehat{x}_j \tag{9}$$

$$\textbf{(b)} \quad \|\boldsymbol{a}_i + \boldsymbol{a}_j\|_2 < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \;\Rightarrow\; \widehat{x}_i = -\widehat{x}_j. \tag{10}$$

Claramente, a alínea (b) do teorema é um simples corolário da alínea (a), resultado de uma simples troca de sinais. Note-se que, se duas colunas forem exactamente iguais, $\boldsymbol{a}_i = \boldsymbol{a}_j$, or simétricas, $\boldsymbol{a}_i = -\boldsymbol{a}_j$, qualquer $\Delta_{\boldsymbol{w}} > 0$ é suficiente para garantir que essas duas colunas são agrupadas, ou seja, que os respectivos coeficientes de regressão sejam iguais em valor absoluto.

O corolário seguinte aplica-se ao caso em que as colunas de $\boldsymbol{A}$ estão normalizadas com média nula e norma unitária e resulta de notar que, neste caso, $\|\boldsymbol{a}_i \pm \boldsymbol{b}_j\|_2 = \sqrt{2 \pm 2\rho_{ij}}$, onde $\rho_{ij} = \boldsymbol{a}_i^T \boldsymbol{a}_j$ representa a correlação amostral entre a $i-$ésima e a $j-$ésima variáveis.

Corolário 4.1 *Sob as mesmas hipóteses do Teorema 4.1, se as colunas de* $\boldsymbol{A}$ *verificarem* $\mathbf{1}^T \boldsymbol{a}_k = 0$ *e* $\|\boldsymbol{a}_k\|_2 = 1$*, para* $k = 1, ..., p$*, então,*

$$\textbf{(a)} \quad \sqrt{2 - 2\rho_{ij}} < \Delta_{\boldsymbol{w}} / \|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \tag{11}$$

$$\textbf{(b)} \quad \sqrt{2 + 2\rho_{ij}} < \Delta_{\boldsymbol{w}} / \|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j. \tag{12}$$

O corolário 4.2 recupera o teorema principal de Bondell e Reich (2007), pois no caso em que $\Omega^{\boldsymbol{w}}_{\text{OWL}} = \Omega_{\text{OSCAR}}$, verifica-se que $\Delta_{\boldsymbol{w}} = \gamma$. No entanto, o nosso resultado aplica-se sob condições qualitativamente mais fracas: contrariamente ao resultado demonstrado por Bondell e Reich (2007), aqui não se exige que $\widehat{x}_i$ e $\widehat{x}_j$ sejam ambos diferentes de zero e diferentes de todos as outras estimativas de coeficientes de regressão, nem que $\boldsymbol{A}$ seja tal que $\widehat{x}_k \geq 0$, para $k = 1, ..., p$.

É também possível estabelecer resultados paralelos aos expressos no Teorema 4.1 e no Corolário 4.2 para regressão linear onde erro quadrático é substituído pelo erro absoluto, ou seja, onde o problema de RL é formulado como

$$\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^p} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 + \eta\, \Omega(\boldsymbol{x}). \tag{13}$$

Teorema 4.2 *Seja* $\widehat{\boldsymbol{x}}$ *uma solução de* (13)*, com* $\eta = 1$ *e* $\Omega = \Omega^{\boldsymbol{w}}_{OWL}$*, e* $\boldsymbol{a}_i$ *e* $\boldsymbol{a}_j$ *duas colunas* $\boldsymbol{A}$*. Então,*

$$\textbf{(a)} \quad \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_1 < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \tag{14}$$

$$\textbf{(b)} \quad \|\boldsymbol{a}_i + \boldsymbol{a}_j\|_1 < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j \tag{15}$$

Corolário 4.2 *Sob as mesmas hipóteses do Teorema 4.2, se as colunas de* $\boldsymbol{A}$ *verificarem* $\mathbf{1}^T \boldsymbol{a}_k = 0$ *e* $\|\boldsymbol{a}_k\|_2 = 1$*, para* $k = 1, ..., p$*, então,*

$$\textbf{(a)} \quad \sqrt{n(2 - 2\rho_{ij})} < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \tag{16}$$

$$\textbf{(b)} \quad \sqrt{n(2 + 2\rho_{ij})} < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j. \tag{17}$$

5 Conclusões

Este texto abordou a questão da regressão linear em cenários nos quais se pretende identificar todas as variáveis relevantes, mesmo quando estas são fortemente correlacionadas. Depois de rever o motivo pelo qual a simples regularização indutora de esparsidade não resolve de forma satisfatória esta questão, foram descritas algumas abordagens recentes, com especial ênfase para o OSCAR (*octagonal shrinkage and clustering algorithm for regression* [Bondell and Reich, 2007]). Finalmente, mostrou-se que o OSCAR pertence a uma classe mais geral de regularizadores, a qual foi estudada à luz da teoria da majorização e da convexidade de Schur, permitindo estabelecer resultados teóricos sob a forma de condições suficientes para a identificação de variáveis correlacionadas [Figueiredo and Nowak, 2016].

Referências

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candès. SLOPE – adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9:1103–1140, 2015.

H. Bondell and B. Reich. Regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2007.

P. Bühlmann, P. Rüttiman, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1858, 2013.

H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30:348–361, 1920.

D. Davis. An $o(n\log(n))$ algorithm for projecting onto the ordered weighted ℓ_1 norm ball. Technical report, `arxiv.org/abs/1505.00870`, 2015.

M. Figueiredo and R. Nowak. Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 930–938, 2016.

C. Genovese, J. Jin, L. Wasserman, and Z. Yao. A comparison of the LASSO and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42:80–86, 1970.

D. Lorenz and N. Worliczek. Necessary conditions for variational regularization schemes. *Inverse Problems*, 29(7):075016, 2013.

A. Marshall, I. Olkin, and B. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, 2011.

A. Pigou. *Wealth and Welfare*. Macmillan, London, 1912.

X. Shen and H.-C. Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105:727–739, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58(1):267–288, 1996.

X. Zeng and M. Figueiredo. Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21:1240–1244, 2014a.

X. Zeng and M. Figueiredo. The ordered weighted ℓ_1 norm: atomic formulation, dual norm, and projections. Technical report, `arxiv.org/abs/1409.4271`, 2014b.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320, 2005.

O papel das metodologias probabilísticas e estatísticas no melhoramento da concepção de materiais obtidos por misturas

Paula Milheiro Oliveira, *poliv@fe.up.pt*
Universidade do Porto, Faculdade de Engenharia e CMUP

1 Introduçao

No dia a dia da Engenharia, os profissionais lidam com situações que envolvem medições que não é possível obter de forma exacta ou recorrem a modelos que representam a realidade de uma forma aproximada, no sentido em que a diferença entre o que os valores que o modelo nos fornece e a realidade, numa situação concreta, não é conhecida com extatidão ou apresenta variabilidade que se pode traduzir de forma aleatória. Da medição mais simples à mais complexa, do modelo mais simples ao mais complexo, a incerteza está patente e é necessário reconhecer que não há medições nem modelos perfeitos, livres de erro e que componentes aleatóras são intrínsecas a muitos fenómenos da natureza e a muitos fenómenos hesógenos. Os próprios equipamentos que usamos na medição de grandezas físicas, sejam eles a balança que temos em casa ou aquele que existe no laboratório, dos mais simples aos mais sofisticados, todos eles são de facto inexactos, e são calibrados para se obter uma determinada precisão finita. As condições em que os equipamentos são usados também são muitas vezes impossíveis de controlar completamente e podem influenciar as próprias medições, traduzindo-se por um erro aleatório que podemos modelar de maneira probabilística. São factos bem conhecidos dos probabilistas, dos estatísticos e dos engenheiros. Deste reconhecimento da incerteza inerente resulta a necessidade de saber decidir sobre incerteza.

Neste artigo pretende-se discutir e fazer realçar metodologias de tipo probabilístico e estatístico passíveis de serem usadas na avaliação de adequação de materiais em desenvolvimento ou de propriedades menos percebidas, contribuindo para criar materiais que vão mais ao encontro do que é desejado e permitindo melhorar aspetos da sua concepção/fabricação, conduzindo-nos nomeadamente a custos mais baixos de produção (ver por exemplo Massol-Chaudeur *et al*. (2003), Destandau *et al*. (2006), Prato-Garcia e Buitrón (2013)). Estamos a referir-nos a materiais que são constituídos por uma mistura de substâncias componentes, em que estas substâncias (ditas materiais constituintes) são sujeitas a um procedimento de medição prévio, de forma a obter as percentagens adequadas de cada uma delas na mistura. Se por um lado se pretende que as misturas atinjam certos alvos inicialmente descritos, também é verdade que, geralmente, a robustez da mistura, ou seja "a capacidade de resistir a fontes externas de incerteza ou de variabilidade" é, só por si, uma função objetivo a maximizar.

Propomos uma metodologia que consiste em: (i)definição de uma medida probabilística para a robustez da mistura; (ii)modelação das propriedades de resposta do material; (iii)estimação da probabilidade definida em (i); (iii)comparação das diferentes misturas, incluindo aspectos económicos, de eficiência ou outros aspectos relevantes. A fase que acarreta a modelação das propriedades da mistura (fase (ii)) é realizada em ambiente laboratorial. Previamente à modelação, recomenda-se, sempre que possível, um planeamento experimental, de modo a obter um bom modelo com o menor número de experiências possível.

Descrevemos resumidamente o exemplo da concepção de um tipo de betão e de como algum cuidado científico nesse procedimento pode conduzir a ganhos de produtividade nas empresas que se dediquem ao fabrico desses materiais (para mais detalhes, ver Nunes *et al*. 2006).

2 Uma medida probabilística para a robustez da mistura

Quando pretendemos obter um dado material através de uma mistura de substâncias componentes (por exemplo um betão, uma tinta, um plástico, etc), naturalmente que as quantidades ou as percentagens desses materiais na composição da mistura são determinantes para o desempenho da mesma. Geralmente usa-se uma "receita" para a sua fabricação, tal como na culinária. E, tal como na culinária, há várias receitas possíveis para chegar ao mesmo objectivo que será um material (um bolo) com determinadas características (fofo, com forte sabor a chocolate, que se conserve sem endurecer rapidamente, por exemplo). Os materias constituintes ou substâncias componentes são medidos antes de ser feita a mistura, por pesagem ou por avaliação do volume, geralmente. Começam aqui as fontes de incerteza, porque as medições não são exactas, porque os próprios materiais constituintes variam um pouco de remessa para remessa em termos das suas propriedades (ver Figura 1(a)), porque há outros factores, de que são exemplo os fatores ambientais (por exemplo a humidade do ar) que afectam as reacções químicas na produção do material, porque há equipamentos de fabrico a manusear nem sempre pelo mesmo operador, em certos casos porque o material é transportado em condições variáveis enquanto não estão finalizadas todas as reacções químicas que conduzem ao material no seu estado final (ver Figura 1(b)), enfim um sem número de razões que nos conduzem a admitir a variabilidade na fabricação dos materiais em causa e a incerteza sobre o seu comportamento final.

Na linguagem corrente da engenharia, tratando-se de um material fabricado através de uma mistura, diz-se que a mistura é robusta quando é minimamente afetada por fontes externas de incerteza ou de variabilidade. Uma mistura robusta deverá tolerar desvios correntes inerentes ao processo de fabricação, mantendo as suas propriedades dentro dos limites especificados para o material. Podemos interpretar este conceito matematicamente.

Sejam $Y_1, Y_2, \cdots, Y_n$ as variáveis de resposta consideradas para a mistura e r^i_{inf}, r^i_{sup} os limites de aceitação para a resposta $Y_i\,, i = 1, \cdots, n$, de acordo com as especificações pré-estabelecidas para o material. Consideremos também o critério de aceitação global envolvendo as n variáveis de resposta: $\bigcap_{i=1}^{n} \left\{ r^i_{inf} < Y_i < r^i_{sup} \right\}$. Assumiremos como medida de robustez do material a probabilidade

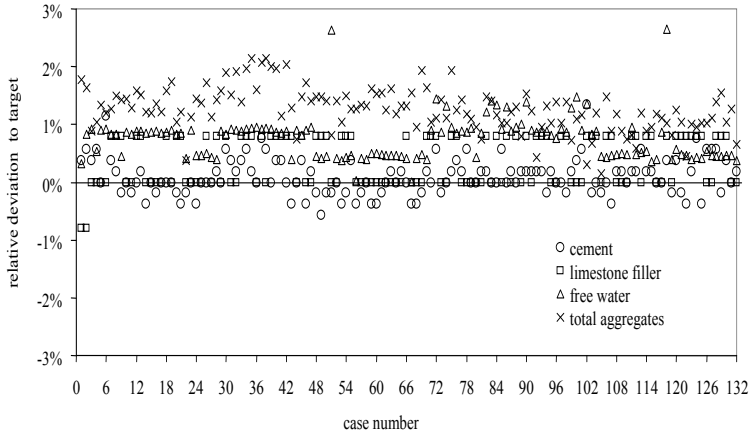
$$pr = P\left(\bigcap_{i=1}^{n} \left\{ r^i_{inf} < Y_i < r^i_{sup} \right\} \right). \tag{1}$$

Admitindo a independência das variáveis de resposta, as contribuições individuais de cada uma das variáveis para a robustez da mistura podem ser obtiddas como sendo as probabilidades de ocorrência de valores dentro dos limites de aceitação para cada uma das variáveis isoladamente:

$$pr_i = P\left(R^i_{inf} < Y_i < R^i_{sup} \right), i = 1, \cdots, n. \tag{2}$$

A possibilidade de comparar estas contribuições é importante, em termos práticos, uma vez que as quantidades pr_i nos dizem quais são as variáveis de resposta, ou, se preferirmos, as propriedades do material, que são mais ou menos sensíveis no processo de produção.

As variáveis de resposta $Y_i\,, i = 1, \cdots, n$ são consideradas variáveis dependentes, sendo as variáveis independentes ou explicativas aquelas que traduzem as quantidades ou as percentagens dos constituintes na mistura, $X_i\,, i = 1, \cdots, p$. A relação entre variáveis de resposta e constituintes da mistura não é geralmente conhecida e tem de ser obtida por via estatística, por exemplo recorrendo a modelos de regressão.



(a)

(b)

Figura 1: Alguns exemplos de fontes de incerteza: (a)variabilidade dos próprios materiais constituintes; (b)condições em que o material é transportado.

3 Modelaçao da resposta do material

Numa situação normal de produção do material podemos assumir que as quantidades de materiais constituintes que compõem uma mistura são variáveis aleatórias que podemos modelar através de distribuições apropriadas, muito frequentemente mas não necessariamente gaussianas. Mediante experiências realizadas em laboratório, em que os materiais são ensaiados em termos da sua resposta, podemos recolher dados sobre as respostas do material, usando observações das variáveis aleatórias constituintes, em remessas controladas. É com base nos ensaios realizados em laboratório, para as diferentes especificações de mistura, que construimos um modelo de regressão para a resposta. Este modelo pode apresentar não linearidades e é relativamente comum recorrer-se a transformações sobre as variáveis explicativas, nomeadamente as potências $1/2$ e -1. Como já tivemos ocasião de referir, haverá vantagens em recorer a um plano experimental, nesta fase, de modo a diminuir o esforço de ensaios a realizar no laboratório e as quantidades de materiais que vão ser utilizados nesses ensaios, que são quase sempre não re-utilizáveis.

Quando a modelação é feita com base nas quantidades absolutas de substâncias constituintes, os modelos de regressão clássicos ou os robustos são habitualmente usados. No entanto convém lembrar que, se, no lugar destas, a modelação for feita com base nas proporções de constituintes na mistura,

e recorrermos a todas as proporções para caracterisar a mistura (ou seja a soma das proporções é 100%), teremos de recorrer a técnicas de modelação de dados composicionais (ver Pawlowsky-Glahn *et al*. (2015)). A decisão sobre modelar quantidades absolutas ou proporções na mistura (e quais das proporções, porque pode ser uma grande lista) é tomada tendo em conta o problema em particular que se está a resolver, pois tem a ver com o tipo de medição que é feita (por peso, por volume, etc), com as fontes de maior variabilidade ou com a variabilidade mais preocupante na concepção do material e ainda com o tipo de procedimentos laboratoriais. Os métodos estatísticos serão naturalmente usados como argumento para eliminar algumas variáveis que não contribuam de forma significativa para a resposta.

Em qualquer dos casos precisaremos de estabelcer um modelo estatístico para as variáveis de resposta que intervêm em (1), que estabeleça a sua relação com os constituintes da mistura.

4 Estimaçao da robustez da mistura e optimizaçao da concepçao

As probabilidades definidas em (1) e (2), poderiam obviamente ser calculadas de forma exata se fosse conhecida a distribuição conjunta dos Y_i, $i = 1, \cdots, n$. Uma vez que se trata de variáveis de resposta (em situação de fabricação corrente não são observadas), o mesmo seria dizer: se fossem conhecidas as distribuições marginais das variáveis independentes X_i, $i = 1, \cdots, p$ e o modelo que relaciona a resposta com as variáveis independentes fosse exatamente conhecido. Como vimos, não é o caso nos problemas em que se pretende melhorar a concepção de materiais. Às variáveis explicativas são ajustadas distribuições, estatisticamente, e a ligação entre estas e as variáveis de resposta é representada por um modelo estatístico também ele ajustado a partir das observações feitas em contexto de laboratório. Resta-nos portanto estimar essas probabilidades ou, por outras palavras, estimar a robustez de cada mistura.

Podemos obter as estimativas das probabilidades definidas em (1) e (2) com base em simulações de Monte Carlo, assumindo as distribuições ajustadas para as variáveis X_i, $i = 1, \cdots, p$ e os modelos de ligação destas às variáveis de resposta, incluindo a parcela de erro aleatório, ou podemos recorrer à reamostragem a partir das observações feitas sobre os constituintes ao longo do processo de produção, e não no laboratório, e analisar a frequência de ocorrência dos acontecimentos nas reamostras para obter as estimativas da probabilidade. Naturalmente que este segundo procedimento será estatisticamente mais robusto.

Uma vez estimadas as probabilidades (1) associadas a diferentes misturas estamos em condições de escolher a mais robusta. Como tivemos ocasião de referir na Secção 2, as estimativas das probabilidades (2) serão úteis em termos práticos, já que nos permitem perceber em que medida é que cada resposta contribui para a robustez da mistura, por exemplo evidenciando as variáveis que mais nos devem preocupar quando tentamos aumentar a robustez de uma dada mistura. Torna-se particularmente interessante, do ponto de vista do gestor de produção, avaliar a situação de possível aumento de robustez sem aumento considerável de custos, o que é bastante fácil de executar, na prática, uma vez que podemos assoociar um custo a cada "opção"de mistura.

5 O exemplo do betao

Esta secção servir-nos-á para exemplificar como é que a metodologia descrita pode ser aplicada na concepção de misturas de betão SCC (betão autocompactável). Os detalhes desta aplicação podem ser consultados em Nunes *et al* (2006) e são aqui apresentados muito resumidamente. Neste caso as variáveis explicativas são:

- X_1 = razão volumétrica água/finos (Vw/Vp);
- X_2 = razão fíler/cimento, em massa, (wf/wc);
- X_3 = razão superplastificante/total de finos, em massa (Sp/p);
- X_4 = relação entre o volume de areia e o volume de argamassa (Vs/Vm).

e as variáveis de resposta consideradas são:

- Y_1 = diâmetro de espalhamento (Desp)
- Y_2 = tempo necessário para escoar a argamassa no funil V (Tfunil)
- Y_3 = resistência à compressão aos 28 dias (fc,28dias)

Procedeu-se à comparação de 3 misturas, denominadas A, B e C. A Tabela 1 mostra as quantidades dos constituintes das diferentes misturas, as respetivas respostas estimadas pelo modelo de regressão obtido em Nunes *et al* (2006), os custos (euro/m^3) associados a cada mistura e a robustez de cada mistura como definida em (1). Note-se que a mistura C é a que apresenta o custo mais elevado e é também a menos robusta, logo deve ser preterida pelo gestor de produção. A mistura B é um pouco mais dispendiosa do que a mistura A (o custo sobe 4%) mas é também um pouco mais robusta. Para decidir entre as misturas A e B o gestor de produção entrará em conta com vários factores, entre os quais a política de satisfação dos clientes que a empresa pretende seguir e os custos reais que a devolução ou a perda de clientes por se encontrarem insatisfeitos poderá representar.

Composição (Kg/m^3)	Mistura A	Mistura B	Mistura C
cimento	333.7	378.7	435.2
filler	344.2	313.5	257.7
água	170.4	173.5	176.4
superplastificante	9.24	9.60	10.47
areia 1	335.0	359.9	354.2
areia 2	406.0	353.7	370.7
agregados	735.3	744.1	729. 3
Custo ($/m^3$)	53.0	55.1	57.9
robustez	0.85	0.86	0.77

Tabela 1: Composição das misturas, custo e robustez.

6 Comentários finais

Descrevemos uma metodologia bastante geral que pode ser seguida para resolver problemas envolvendo a concepção de materiais que são obtidos através de uma mistura e em que é possível variar as quantidades ou as proporções dos constituintes dessa mistura. Pressupõe-se que existem pré-especificações para o material a fabricar, que têm de ser respeitadas, e que se traduzem por gamas de valores para a resposta do dito material, havendo uma preocupação com a sua robustez no sentido que lhe é dado em engenharia de "capacidade de resistir a flutuações diversas que residem nos próprios constituintes ou nas condições de manipulação durante a fabricação e que não se torna viável controlar". Para isso usámos uma medida de robustez que é de tipo probabilístico. A optimização na concepção pode ser feita usando essa medida de robustez.

Agradecimentos
O autor foi parcialmente financiado pelo CMUP (UID/MAT/00144/2013), que é financiado pela FCT usando fundos estruturais nacionais (MEC) e europeus (FEDER), mediante contrato ao abrigo do PT2010.

BIBLIOGRAFIA

Destandau E., Vial J., Jardy A., Hennion M., Bonnet D., Lancelin P. (2006): Robustness study of a reversed-phase liquid chromatographic method for the analysis of carboxylic acids in industrial reaction mixtures *Analytica Chimica Acta*, 572(1), pp. 102-112.

Massol-Chaudeur S., Berthiaux H., Dodds J. (2003): The development and use of a static segregation test to evaluate the robustness of various types of powder mixtures *Food and Bioproducts Processing: Transactions of the Institution of of Chemical Engineers, Part C*, 81(29, pp. 106–118.

Nunes S., Figueiras H., Milheiro–Oliveira P., Sousa–Coutinho J., Figueiras J. (2006): A methodology to assess robustness of SCC mixtures, *Cement and Concrete Research*, 36(12), pp. 2115–2122.

Pawlowsky-Glahn V., Egozcue J. J., Tolosana-Delgado R. (2015): *Modeling and Analysis of Compositional Data*, Wiley.

Prato-Garcia D., Buitrón G. (2013): Improvement of the robustness of solar photo-Fenton processes using chemometric techniques for the decolorization of azo dye mixtures *Journal of Environmental Management*, 131, pp. 66–73.

Métodos Bayesianos para Engenharia

Giovani Loiola da Silva, *giovani.silva@tecnico.ulisboa.pt*

Departamento de Matemática – IST, Universidade de Lisboa
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

1. Introdução

A estatística bayesiana tem sido dominada por outras abordagens inferenciais por muitos anos, sobretudo a inferência frequencista. No entanto, nas últimas décadas, o surgimento de novos métodos bayesianos tem sido impulsionado pela disponibilidade de técnicas computacionais e.g os métodos de Monte Carlo via cadeia de Markov (MCMC) (vide Gilks et al., 1995). Uma das grandes diferenças entre as inferências frequencista e bayesiana é que, contrariamente à abordagem frequencista, os parâmetros desconhecidos não são tratados como sendo valores fixos (situação usual) mas sim como variáveis aleatórias na abordagem bayesiana. Isso permite que distribuições de probabilidades sejam associadas aos parâmetros desconhecidos podendo essas distribuições ser interpretadas como a crença do investigador sobre os verdadeiros valores dos parâmetros.

A semente para da abordagem bayesiana a problemas de inferência foi lançada por Richard Price quando em 1763 publicou a obra póstuma do reverendo Thomas Bayes, intitulada "An Essay towards solving a problem in the Doctrine of Chances". Para detalhes do paradigma quer clássico quer bayesiano, veja-se Paulino, Amaral-Turkman e Murteira (2003), incluindo uma descrição cabal da metodologia bayesiana e aplicações a problemas de interesse em várias áreas do conhecimento tais como Engenharia (vide e.g. o método entropia máxima na procura de objectividade na formulação de distribuições a priori que caracterizem um estado de ignorância dos parâmetros desconhecidos).

Este artigo descreve métodos bayesianos que poderiam ser aplicados a problemas de Engenharia e.g. amostragem de aceitação (Hunter, 1986) e modelo de análise de variância misto (Box e Tiao, 1973). Assim, discute-se a incerteza em Engenharia sob o ponto de vista bayesiano com base em algumas aplicações simples de estatística aplicada. Para aplicações mais específicas e complexas de métodos bayesianos, há várias publicações a registar nomeadamente em delineamento de experiências (Box et al., 2005), fiabilidade (Hamada et al., 2008), aprendizagem automática (Barber, 2012) e em tópicos variados (Box e Tiao, 1973; Gilks et al., 1995; Swiler, 2006; Gelman et al., 2013; Kruschke, 2015).

Entre os muitos *softwares* disponíveis para executar vários aspectos da estatística bayesiana computacional, destaca-se aqui FirstBayes, BUGS, JAGS, BayesX, Stan e INLA (vide Paulino et al., 2003; Amaral-Turkman e Paulino, 2015).
O *software* BUGS (Bayesian inference using Gibbs sampling) tem sido instrumental na conscientização da modelagem bayesiana entre as comunidades académica e comercial internacionalmente, e tem tido um sucesso considerável ao longo de sua vida útil de 20 anos (Lunn et al., 2009).
Usando aproximações de Laplace encaixadas e integradas (INLA), podemos calcular diretamente aproximações muito exatas às distribuições *a posteriori* marginais (Rue et al., 2009).

2. Aplicação a Problemas de Engenharia I

Uma aplicação prática na indústria diz respeito ao problema da amostragem de aceitação. Hunter (1986) ilustrou esse problema considerando um grande lote de itens que aparece no cais de receção e um engenheiro que é convidado a determinar se o lote é aceitável.

O processo de fabricação dos lotes de itens é suposto ter dois estados: A e B. Diz-se que o processo A está bem se o processo produz apenas 1% de itens defeituosos (D), enquanto é indesejável que processo B tenha 3% de itens defeituosos. Suponha que a probabilidade inicial de itens produzidos pelo processo A seja P(A)=0.5, ou seja, não haja evidência prévia antes de se o engenheiro está a lidar com lotes dos processos A ou B.

Suponha que o engenheiro seleciona um item do lote e verifica que esse é não defeituoso ($\overline{D}$). Usando o teorema de Bayes, quanta evidência foi gerada em nome da hipótese de que o processo desconhecido é realmente o processo A?

$$P(A|\overline{D}) = \frac{P(A) \times P(\overline{D}|A)}{P(A) \times P(\overline{D}|A) + P(\bar{A}) \times P(\overline{D}|\bar{A})} = \frac{0.5 \times 0.99}{0.5 \times 0.99 + 0.5 \times 0.97} \cong 0.5051 \quad (1)$$

Outras métricas menos comuns para medir a incerteza são, por exemplo, as chances/vantagens (*odds*), definidas como a razão $\frac{\text{probabilidade}}{1-\text{probabilidade}}$, bem como o logaritmo das chances, por vezes conhecido por "evidência". Note-se que $0 \leq \text{probabilidade} \leq 1$, $0 \leq \text{chances} < \infty$ e $-\infty < \text{log-chances} < \infty$ e que a probabilidade de 0.5 equivale chances unitárias e log-chances nulas (ausência de "evidência"). Assim, as chances de ter sido o processo A gerador do item selecionado não defeituoso é

$$\text{chances}(A|\overline{D}) = \frac{P(A|\overline{D})}{1 - P(A|\overline{D})} \cong 1.0206 \quad (2)$$

enquanto o logaritmo (base 10) das chances é $\log(chances(A|\overline{D})) \cong 0.0089$. Se um novo item é selecionado, sendo esse também um item não defeituoso, a nova probabilidade de A condicional a D em (1), usando P(A)=0.5051 (procedimento de aprendizagem), é

$$\text{nova} \quad P(A|\overline{D}) = \frac{0.5051 \times 0.99}{0.5051 \times 0.99 + 0.4949 \times 0.97} \cong 0.5102 \quad (3)$$

com novas $\text{chances}(A|\overline{D}) \cong 1.0417$ e $\log(chances(A|\overline{D})) \cong 0.0178$. É importante observar que o segundo item não defeituoso adicionou cerca de 0.0089 de "evidência" de processo A. De fato, cada item não defeituoso selecionado sucessivamente adiciona 0.0089 a favor do processo A para lidar com a incerteza.

Analogamente, a probabilidade de ter sido o processo A gerador do primeiro item selecionado, agora defeituoso, é $P(A|D) = 0.25$ e por conseguinte as $\text{chances}(A|D) = 1/3$ e $\log\big(chances(A|D)\big) \cong -0.4771$. Assim, cada item defeituoso selecionado sucessivamente subtrai 0.4771 a favor do processo A. Com cada item sucessivo o engenheiro aprende mais sobre o processo, a "evidência" a favor de A cresce ou diminui de uma forma aditiva simples.

O engenheiro agora adota as seguintes regras de aceitação dos processos:

1) Se a probabilidade do processo A for superior a 0.95 (i.e. "evidência" de pelo menos 1.2876), ele vai decidir que a favor do processo A. Note-se que isso vai exigir pelo menos (1.2876/0.0089) = 145 itens não defeituosos em sucessão.
2) Se a probabilidade de que o processo A for inferior a 0.20 (i.e. "evidência" de no máximo -1.0791), ele decidirá que esse não é um lote fabricado pelo processo A, exigindo somente (1.0791/0.4771) = 2 itens defeituosos em sucessão para a rejeição do processo A.

Por fim, o engenheiro pode agora construir o esquema de amostragem de aceitação sequencial gráfica para lidar com a incerteza, onde cada item identificado como não defeituoso adiciona 0.0089 a sua evidência total e subtrai 0.4771 para cada item defeituoso.

3. Aplicação a Problemas de Engenharia II

Box e Tiao (1973) analisam um conjunto de dados referentes à variação de lote a lote nos rendimentos de corante. Os dados surgem de uma experiência equilibrada em que o rendimento total do produto foi determinado para 5 amostras de cada um de 6 lotes de matéria-prima escolhidos aleatoriamente (exemplo do manual do software WinBUGS/OpenBUGS, Lunn et al., 2009).

O objetivo do estudo foi determinar a importância relativa entre variação de lote versus variação devido a erros de amostragem e analíticos. No pressuposto de que os lotes e as amostras variam independentemente e contribuem aditivamente para a variância do erro total, podemos assumir o seguinte modelo para o rendimento do corante:

$$y_{ij} \sim Normal(\mu_i, \tau_{with}) \quad (4)$$
$$\mu_i \sim Normal(\theta, \tau_{btw}) \quad (5)$$

onde y_{ij} é o rendimento para a amostra j do lote i, μ_i é o rendimento verdadeiro para o lote i, τ_{with} é o inverso da variância dentro do lote σ^2_{with} (ou seja, a variação devida a amostragem e erro analítico), θ é o rendimento médio real para todos os lotes e τ_{btw} é o inverso da variância entre lotes σ^2_{btw}, $i = 1, \dots, 6$, $j = 1, \dots, 5$. A variação total no rendimento do produto é definida por $\sigma^2_{tot} = \sigma^2_{with} + \sigma^2_{btw}$ e as contribuições relativas de cada componente para a variância total são $f_{with} = \sigma^2_{with}/\sigma^2_{tot}$ e $f_{btw} = \sigma^2_{btw}/\sigma^2_{tot}$. Consideram-se aqui distribuições *a priori* não-informativas padrão para θ, τ_w e τ_b, ou seja, Normal (0, 10^{-10}), Gama Inversa(0.001,0.001) e Gama Inversa(0.001,0.001), respetivamente.

Uma representação gráfica do modelo (4)-(5), conhecido por modelo de componentes de variância ou modelo de análise de variância misto encontra-se na Figura 1 em termos do grafo dirigido acíclico (*Directed Acyclic Graph* - DAG). A sua notação é definida como se segue. Os nós retangulares denotam constantes conhecidas. Os nós elípticos representam relações determinísticas (i.e., funções) ou quantidades estocásticas, isto é, quantidades que requerem uma suposição distributiva. Dependência estocástica e dependência funcional são denotadas por setas de um único ângulo e setas de dois gumes, respetivamente. Estruturas repetitivas, como o "loop" de i = 1 a N, são representadas por "placas", que podem ser aninhadas se o modelo for hierárquico.

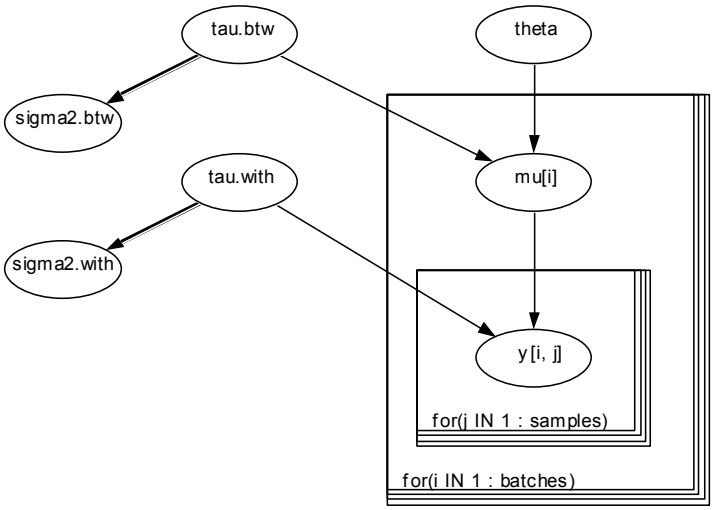


Figure 1: Grafo dirigido acíclico do modelo (4)-(5).

Usando uma notação similar ao pacote estatístico R, pode-se construir um conjunto de comandos para ajustar o modelo (4)-(5) no *software* WinBUGS/OpenBUGS (Lunn et al., 2009), ou seja,

```
model
   {    for(i in 1 : batches) {
                mu[i] ~ dnorm(theta, tau.btw)
                for(j in 1 : samples) {
                y[i,j] ~ dnorm(mu[i], tau.with)
                cumulative.y[i , j] <- cumulative(y[i,j], y[i,j])   }  }
        sigma2.with <- 1 / tau.with
        sigma2.btw <- 1 / tau.btw
        tau.with ~ dgamma(0.001, 0.001)
        tau.btw ~ dgamma(0.001, 0.001)
        theta ~ dnorm(0.0, 1.0E-10)   }
```

Com base em 2 listas de valores iniciais para os parâmetros do modelo, i.e., *list(theta=1500, tau.with=1, tau.btw=1)* e *list(theta=3000, tau.with=0.1, tau.btw=0.1)*, 100000 amostras foram simuladas após um período de aquecimento com 25000 amostras. O resultado desta amostra de tamanho 100000 encontra-se na Tabela 1, nomeadamente a média a posteriori (mean), o desvio padrão (sd) e os quantis de 2.5% (val2.5pc) e de 97.5% (val97.5pc) dos 3 parâmetros de interesse seguida relativamente ao modelo (4)-(5).

Tabela 1: Medidas sumárias dos parâmetros de interesse do modelo (4)-(5)

mean	sd	MC_error	val2.5pc	median	val97.5pc
σ^2_{btw}	2267.0	4057.0	29.28	0.0090	1349.0
σ^2_{with}	3011.0	1099.0	14.55	1550.0	2781.0
θ	1527.0	21.850	0.145	1484.0	1527.0

É importante ter em conta que foi necessária simular uma amostra relativamente longa devido à alta autocorrelação entre valores sucessivamente amostrados de alguns parâmetros, sobretudo as componentes de variâncias. Tais correlações reduzem o tamanho "efetivo" da amostra das distribuições *a posteriori* marginais e, portanto, é necessário um período mais longo para garantir uma precisão suficiente das estimativas posteriores.

Observe-se ainda que a distribuição *a posteriori* para σ^2_{btw} tem uma cauda superior muito longa e consequentemente, a sua média *a posteriori* é consideravelmente maior que a mediana. Box e Tiao (1973) apresentaram a estimativa de σ^2_{with} e σ^2_{btw} via análise de variância sob a abordagem frequencista, dadas respetivamente por 2451 e 1764. Nesse caso, estima-se σ^2_{btw} pela diferença dos quadrados médios entre e dentro do lote divididos pelo número de lotes menos 1. Em alguns casos, isso leva à situação insatisfatória de uma estimativa da variância negativa. Calcular um intervalo de confiança para σ^2_{btw} é também difícil usando a abordagem clássica devido à sua distribuição de amostragem ser complicada.

4. Comentários Finais

Para além do algoritmo de simulação nos métodos MCMC, podem-se usar nomeadamente os métodos com aproximações de Laplace, de quadratura iterativos, e de simulação com reamostragem por importância na resolução bayesiana de muitos problemas práticos i.e. na simulação de distribuições *a posteriori* ou na obtenção de estimativas para quantidades de interesse dos respetivos modelos (vide Paulino et al., 2003; Amaral-Turkman e Paulino, 2015).

Apesar de os métodos MCMC serem uma ótima ferramenta para resolução de muitos problemas práticos na análise bayesiana, algumas questões relativamente à convergência nestes métodos são ainda merecedoras de mais investigação, e.g., quantas iterações deve ter o processo de simulação para

garantir que a cadeia convergiu para o estado de equilíbrio? Obviamente que a resposta definitiva a esta questão poderá nunca ser dada, visto que a distribuição estacionária será na prática desconhecida, porém pode-se sempre avaliar a convergência das cadeias detetando problemas de convergência fora do período de aquecimento da mesma. Cowles e Carlin (1996) compararam os métodos mais populares neste quadro concluindo que, apesar de muitos desses métodos detetar comummente os problemas na convergência que se propuseram identificar, essas técnicas podem também falhar no seu propósito, não sendo possível afirmar qual delas é a mais eficiente (Silva, 2001).

Agradecimentos

Este trabalho foi parcialmente apoiado pela Fundação para a Ciência e a Tecnologia (FCT), projeto UID/MAT/00006/2013.

Referências

Amaral-Turkman, M.A., Paulino, C.D. (2015). *Estatística Bayesiana Computacional - uma introdução*. Edições SPE, Lisboa.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Box, G.E.P., Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

Box, G.E.P., Hunter, W.G., Hunter, J.S. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.). Wiley, New York.

Cowles, M.K., Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91, 883-904.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis* (3rd edn.). CRC Press, London.

Gilks, W.R., Richardson, S., Spiegelhalter, D. (Editors) (1995). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC.

Hamada, M.S., Wilson, A., Reese, C.S., Martz, H. (2008). *Bayesian Reliability*. Springer.

Hunter, J.S. (1986). Bayesian approaches to teaching Engineering Statistics. Paper presented at the International Conference on Teaching Statistics, Victoria, Canada.

Kruschke, J. (2015). *Doing Bayesian Data Analysis* (2nd ed.): *A tutorial with R, JAGS, and Stan*. Academic Press, New York.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28, 3049-3067.

Paulino, C.D., Amaral-Turkman, M.A, Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, 71, 319-392.

Silva, G.L. (2001). *Análise Bayesiana de Modelos de Sobrevivência com Fragilidade*. Tese de Doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Swiler, L.P. (2006). *Bayesian Methods in Engineering Design Problems*. Sandia Report SAND2005-3294. Sandia National Laboratories, Albuquerque, New Mexico.

Incerteza existe!

Dinis Duarte Pestana, *dinis.pestana@fc.ul.pt*

Centro de Estatística e Aplicações, Universidade de Lisboa,
Instituto de Investigação Científica Bento da Rocha Cabral

Fernanda Otília Figueiredo, *otilia@fep.up.pt*

Faculdade de Economia da Universidade do Porto,
Centro de Estatística e Aplicações, Universidade de Lisboa

1 Introdução

Contou-nos a Professora Ivette Gomes que há alguns anos uma aluna que frequentava as aulas de Processos Estocásticos um dia lhe disse, no intervalo, que gostava muito das aulas, mas que achava esquisito que a Professora não tivesse a certeza dos resultados que demonstrava – muitas vezes apresentava um resultado começando por anunciar que quase certamente ...

É consensual que o termo incerteza é muito vago, difícil de definir, podendo ter significados muito diferentes consoante o contexto e área de estudo em que aparece. Mas haverá prova mais evidente de que a incerteza existe?

A convicção de que a incerteza impera neste mundo porque uma entidade irracional tudo controla (*Fortuna Imperatrix Mundi*, princípio e fim de *Carmina Burana* de Carl Orff) é quase universal. Em muitas mitologias os próprios deuses estão submetidos aos caprichosos acasos do destino. A própria Física, que no século XIX parecia um baluarte de certezas, teve que admitir o princípio da incerteza, e Max Born (que afirmou que Física teórica é Filosofia) fez afirmações lapidares como "*I believe that ideas such as absolute certitude, absolute exactness, final truth, etc. are figments of the imagination which should not be admissible in any field of science.*" ou "*The conception of chance enters in the very first steps of scientific activity* [...] *I think chance is a more fundamental conception than causality.*"

2 Como Lidar com a Incerteza?

O acaso pode parecer assustador, a menos que se consiga transformar esse inimigo num aliado; os "jogos de azar" desde há muito são rendosa fonte de lucros para quem usa o acaso como aliado, e ruína para os outros. Se forem gulosos por mão de vaca com grão porventura já repararam que há nela um osso que tem quatro faces quase planas, distintas, podendo, ao ser jogado sobre uma mesa, ficar assente sobre qualquer uma dessas quatro faces – mas não com a mesma probabilidade.

Na Roma antiga faziam-se apostas sobre o resultado do lançamento de quatro astrágalos, dependendo o valor do prémio do "lance"; por exemplo o "lance de Vénus" (as quatro faces viradas para cima serem diferentes umas das outras) saía raramente, pelo que merecia uma maior recompensa.

Figura 1: Astrágalos.

Assim, já na Roma antiga os conceitos de probabilidade e de valor esperado, empiricamente, influenciavam as regras dos jogos de azar.

Esses dois conceitos fundamentais, probabilidade e valor esperado, são claramente assumidos em 1654, quando Pascal e Fermat, em resposta a um desafio feito por um jogador, iniciaram a domesticação do acaso. Mas nos passos iniciais da Probabilidade (Pascal, Fermat, Huyghens, Montmort) assumia-se uma equiprobabilidade dos acontecimentos elementares que restringia a sua aplicação a situações bastante particulares como as que idealmente se esperam em jogos honestos.

J. Bernoulli tinha claramente a perceção de que esse pressuposto de equiprobabilidade é facilitador, e por isso adotou um "princípio da razão insuficiente": se não houver razões suficientes para considerar que os acontecimentos elementares têm probabilidades distintas, é pragmático considerá-los equiprováveis. Mas por outro lado encontrou uma forma de domesticar o acaso quando não faz sentido admitir essa razão insuficiente – observou que em circunstâncias muito gerais a média empírica estabiliza, um resultado que considerava o seu "teorema de ouro", e que foi rebatizado lei dos grandes números por Poisson, com sucesso (não teve sucesso nenhum ao batizar "lei dos pequenos números" a lei que deduziu para o limite de sucessões dicotómicas com probabilidade de sucesso evanescente assumindo estabilidade do valor esperado). A aplicação desse teorema a "provas de Bernoulli" (sucessão de experiências aleatórias independentes com resultado dicotómico, em que a probabilidade dos resultados não varia de experiência para experiência) abriu a possibilidade de usar uma "probabilidade frequencista", cujo espetro de aplicabilidade a situações reais é vasto, e permitiu em particular o desenvolvimento inicial da Estatística, usando o paradigma do coletivo de que von Mises seria o paladino.

O teorema áureo de Bernoulli circulava na correspondência científica desde 1695, mas o seu maravilhoso tratado *Ars Conjectandi* só foi publicado postumamente em 1713 pelo seu sobrinho Nikolaus Bernoulli; entretanto já em 1710 John Arbuthnot tinha enviado uma curiosa carta à Royal Society, em que mostrava que a hipótese de equiprobabilidade dos sexos, nos humanos levaria a que a probabilidade de observação constante de mais nascimentos de rapazes do que raparigas em 72 anos de registos (como acontecia naqueles a que tinha acesso) seria apenas $0.5^{72} \approx 0.000000000000000000000212$ (uma probabilidade muito menor do que o que Poincaré viria a chamar "moralmente impossível"). Daí concluía (com fascinantes comentários sobre a existência da Divina Providência) que aquela hipótese de equiprobabilidade deve ser rejeitada, uma apresentação pioneira de testes estatísticos, muito anterior aos ensaios de significância de Fisher, à ideia de valor-p, e aos testes de hipóteses de Neyman e de E. Pearson.

A probabilidade frequencista permitiu também os arrojados progressos feitos por Abraham de Moivre, que na edição de 1718 do seu *The Doctrine of Chances* escreveu estas palavras lapidares:

"*Further, the same Arguments which explode the notion of Luck may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design: We may imagine Chance and Design to be as if it were in Competition with each other, for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other*". Assim, para Abraham de Moivre já é perfeitamente claro que acaso e necessidade se fundem nas mais diversas proporções, dando origem a padrões de probabilidade muito diversos.

Ainda na herança de J. Bernoulli, Laplace reassumiu o princípio da razão insuficiente, e o facto de escrever em francês possivelmente contribuíu para que se chamasse probabilidade Laplaciana a que se obtém como razão de casos favoráveis por casos possíveis, assumindo equiprobabilidade destes. Mas quem não ler apenas a meia dúzia inicial de páginas do seu *Essai Philosophique sur la Probabilité* percebe que a sua compreensão do acaso é muito mais vasta, devendo a probabilidade ser reavaliada tendo em conta a acumulação de informação disponível (as urnas de Laplace, são um elaborado esquema amostral em duas etapas, que exibem claramente que a probabilidade objetiva pode não ser adequada). Laplace foi também um pioneiro da inferência estatística, e no seu *Mémoire sur la probabilité des causes par les évènements* redescobre os resultados de Bayes sobre a "inversão da probabilidade"e as consequências que tem para a inferência. Esta nova concepção da probabilidade, levada ao extremo por de Finetti na afirmação de que a probabilidade é uma fézada, e por Savage no seu *Foundations of Statistics*, alargou consideravelmente a forma como a Estatística lida com a incerteza.

3 Equiprobabilidade e Para Além

A equiprobabilidade dos acontecimentos elementares corresponde a uma incerteza máxima (entropia máxima) sobre o resultado de uma experiência aleatória, e nesse sentido o correspondente modelo de "uniformidade discreta" não é interessante para fazer escolhas e tomar decisões. No entanto, do ponto de vista metodológico, didático, e de desenvolvimento da teoria da probabilidade, esse tratamento democrático manteve uma enorme importância.

Quando no início do século XX Émile Borel consegue finalmente fazer uma construção rigorosa de probabilidade contínua, é mais uma vez à custa da equiprobabilidade: define uma variável aleatória "uniforme" padrão $U = \sum_{k=0}^{\infty} \frac{X_k}{2^k}$ em que $X_k = \begin{cases} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{cases}$, independentes, enunciando o que é atualmente referido como "princípio de Borel" para definir aquela variável (detalhes em Pestana e Velosa, 2010, pp. 305-306).

Claro que esta interessante definição de uma variável contínua como acumulação infinita de variáveis dicotómicas, que na essência é a representação dos reais de [0,1] na base 2, mostra claramente que é a de entropia máxima no suporte [0,1]. Mas as novas aquisições da Teoria da Probabilidade – nomeadamente os conceitos de variáveis e vetores aleatórios, e suas funções de distribuição – permitiram de forma rigorosa ir muito para além da equiprobabilidade, usando condicionamento e transformações de variáveis e de vetores aleatórios para aceder a padrões de probabilidade adequados para modelar a fusão de contingência e necessidade a que aludiu Abraham de Moivre.

Merece particular destaque a transformação uniformizante: denote-se U a variável aleatória uniforme padrão, seja F uma função contínua crescente de 0 para 1. Então $X = F^{-1}(U)$ é uma variável aleatória com função de distribuição F (e este resultado pode ser generalizado usando a inversa generalizada $F^{\leftarrow}$ quando a função de distribuição F não é invertível).

Quer isto dizer que se pode partir da uniforme padrão para qualquer outro padrão de aleatoriedade, e que a uniforme padrão é uma ponte entre quaisquer dois padrões de aleatoriedade univariada

(desde que se conheça forma explícita da função de distribuição). Um resultado que foi fulcral para o desenvolvimento da Estatística Computacional e para o simulacro da realidade que é a simulação, que imita a multiplicação dos pães e dos peixes do Evangelho, pois permite fazer muito (?) de muito pouco, quando se quer usar grandes números e se tem pequenas amostras.

4 Incerteza, Metrologia, Amostragem e Controle da Qualidade

Probabilidade e valor esperado foram duas formas incipientes de lidar com o acaso, informação e entropia tornaram-se auxiliares importantes, e amostragem e planeamento de experiências tornaram-se incontornáveis, por serem instrumentos adequados para tornar o acaso nosso aliado. Só podemos abordar um número limitado de temas, e naturalmente devido aos nossos interesses e experiência elegemos Metrologia, Amostragem e Controle da Qualidade.

Começamos por completar a 2^a citação que acima fizemos de palavras de Max Born: "*The conception of chance enters in the very first steps of scientific activity in virtue of the fact that no observation is absolutely correct*". Guardamos para agora a citação mais completa, porque assim chama melhor a atenção para as medições serem essencialmente diversas das contagens, no sentido em que a sua precisão é limitada pelos instrumentos de medição e frequentemente também por erro humano, ou incúria. Qualquer medição é na sua essência imperfeita, incerta; por isso, a ciência da medição (não só como medir, como também avaliar a imprecisão das medições feitas) tornou-se parte da formação científica de cientistas e de utilizadores da Ciência e das Técnicas. Adiante referiremos alguns progressos que a Metrologia trouxe, nomeadamente ao nível da normalização. E como a incerteza existe, devemos sempre usar como regra de ouro

- controlar tudo o que for posssível controlar,
- aleatorizar o que não for possível controlar, com os objetivos de equilibrar, reduzir a variabilidade, limitar esses enviesamentos,
- e o que não for possível controlar nem aleatorizar, bloquear.

São estes princípios norteadores que presidiram ao desenvolvimento da Amostragem e do Planeamento de Experiências. Em Controle da Qualidade é sobretudo Amostragem que está em jogo (por exemplo, amostragem de aceitação), sendo em geral o objetivo tomar uma decisão, sendo necessário estimar algum parâmetro com um grau de precisão adequado aos nossos objetivos. Naturalmente o problema central é a escolha de uma estratégia de amostragem adequada (em que questões muito pragmáticas como facilidade de obtenção de dados, custo, etc, estão na ordem do dia – a ponto de haver mesmo disciplinas amostrais em que se mede uma coisa que é fácil e/ou barato medir para avaliar outra que é difícil e dispendioso medir, no fundo o recurso à maravilhosa ideia da regressão, de Galton), e a consequente determinação da dimensão da amostra que se deve recolher para se conseguir, com uma probabilidade elevada pré-fixada, o grau de precisão desejado.

Como se referiu anteriormente, em qualquer processo de medição existe sempre alguma imprecisão no resultado, a qual pode surgir devido à falta de precisão do equipamento de medição ou à falta de calibração do mesmo, pode estar associada ao processo de medição e ambiente em que ele ocorre, às próprias caraterísticas do que está a ser medido ou a outras fontes desconhecidas incluidas no processo de medição. Como avaliar e contornar esta incerteza no resultado? O primeiro passo a dar é identificar as possíveis fontes de incerteza, e só depois é possível quantificá-la, e eventualmente desenvolver métodos que incorporem a variabilidade a ela associada. Neste contexto, podemos fazer referência aos seguintes trabalhos: Stoto (1988), num estudo estatístico

efetuado sobre o envelhecimento da população, identifica possíveis fontes de incerteza e apresenta técnicas interessantes para a estimar e reportar. Em NASA (2010) podemos encontrar um conjunto de princípios e métodos para análise e gestão da incerteza em processos de medição. Lopes *et al.* (2016) descreve uma metodologia detalhada para identificação de fontes de incerteza e sua caraterização num estudo sobre medição e caraterização de incerteza em medidas de desempenho de processos.

Devido à necesssidade de existir um procedimento para avaliação da incerteza em metrologia, a Sociedade Internacional de Standardização, ISO (*International Organization for Standardization*), publicou o guia GUM (*Guide to the Expression of Uncertainty in Measurement*), o qual passou a ser utilizado a partir de 1995. Em 1999, de acordo com requisitos especificados pela ECA (*European Co-operation for Accreditation*), foi desenvolvido o programa GUM *Workbench*, o qual disponibilizou ferramentas úteis para calcular a incerteza total num processo de medição resultante da combinação das incertezas individuais das várias componentes do processo. Este procedimento envolve oito passos, descritos detalhadamente em Losinger (2004). Se tivermos em atenção a Diretiva Comunitária 2007/589/CE, a incerteza é definida como o parâmetro que carateriza a dispersão dos valores que poderiam razoavelmente ser atribuídos ao resultado da medição de uma determinada quantidade, tendo em atenção os efeitos de fatores sistemáticos e aleatórios. Assim, a incerteza de uma medição pode exprimir-se por uma faixa de valores provavelmente próximos do valor real, representados por um intervalo do tipo "valor $\pm$ variabilidade", ou através de um gráfico. Por exemplo, Jackson (2008) apresenta representações gráficas com sombreamento para representar a incerteza.

A incerteza impera também na área de Controle da Qualidade. O Controle Estatístico da Qualidade consiste na utilização de um conjunto de métodos com o objetivo de melhorar continuamente um produto, processo ou serviço. Este objectivo, em termos estatísticos, pode ser expresso em termos de redução de variabilidade. Muitos dos procedimentos de Controle da Qualidade estão sujeitos a diversas fontes de incerteza/variabilidade, existindo também várias ferramentas frequentemente utilizadas com o objetivo de a avaliar. Refira-se, por exemplo, os fluxugramas ou organigramas, os diagramas de causa-e-efeito e os diagramas de Pareto, que são ferramentas típicas em Controlo da Qualidade. Nestes diagramas as várias etapas do processo são enumeradas detalhadamente e as possíveis fontes de problemas em cada etapa são referidas, de modo a facilitar a identificação e correção de problemas, caso ocorram, contribuindo assim para melhorar o processo. Representações gráficas como o diagrama de dispersão, a caixa de bigodes, o diagrama de pontos e o diagrama de caule e folhas, além das medidas de dispersão, permitem também detetar possíveis erros nos dados, analisar a sua variabilidade e compará-los com valores nominais pretendidos.

Por volta de 1920-1930 Dodge e Romig, dos *Bell Laboratories*, desenvolveram a amostragem de aceitação como alternativa à inspeção a 100%, estando cientes de que esta abordagem podia induzir em erro/incerteza na tomada de decisão. A escolha do processo de seleção e dimensionamento da amostra de forma a garantir uma dada precisão para o resultado, expressam preocupações no sentido de contornar esta incerteza. A curva caraterística operacional associada aos planos de amostragem de aceitação permite também minimizar a incerteza na tomada de decisão, pois exprime a probabilidade de aceitação de um lote face a vários cenários possíveis no que respeita à verdadeira incidência de defeituosos nesse lote. Outros exemplos de fontes de incerteza nos resultados de amostragem e análise dos mesmos são, por exemplo, a subjetividade do analista em análises sensoriais, a falta de precisão e/ou calibração nos instrumentos de medição em análises de cromatografia, e também valores de referência mal calculados em estudos comparativos.

As cartas de controlo, a ferramenta mais popular de Controle Estatístico de Processos, são representações gráficas que permitem ver como as observações do processo ou os valores de uma estatística adequada marcados sequencialmente na carta, se afastam do valor pretendido, fixando margens de variabilidade aceitável que ajudam à tomada de decisão. A primeira carta de controlo, a carta de médias $\overline{X}$, surgiu com Shewhart em 1924, nos *Bell Laboratories*, para monitorizar

um processo normal, sendo os limites de controlo da carta do tipo "valor nominal do processo ± variabilidade". Em muitas das situações práticas as hipóteses subjacentes à implementação desta carta – dados do processo normais e valores nominais para o valor médio e desvio padrão do processo conhecidos – não são realistas. A distribuição subjacente aos dados pode exibir características bem diferentes das observadas numa normal, sendo desconhecida e difícil de estimar, o mesmo acontecendo com os valores nominais do processo. Este processo de modelação dos dados e estimação dos valores nominais do processo é obviamente uma fonte de incerteza que afeta o verdadeiro desempenho da carta de controlo. Para minimizar o efeito da incerteza induzida pela modelação dos dados tem sido proposto na literatura utilizar-se famílias de distribuições em vez de distribições específicas, assim como a utilização de distribuições por troços de modo a permitir uma melhor modelação dos dados quer na parte central quer nas caudas, especialmente quando temos valores muito extremos nos dados em análise (veja-se, por exemplo, Azzalini (1985, 2005), Jamalizadeb *et al.* (2011) e Figueiredo e Gomes (2013)). Outra forma de contornar esta incerteza é a utilização de métodos de estimação/inferência semiparamétrica, e a implementação de cartas de controlo robustas, ou de cartas não paramétricas (veja-se, por exemplo, Chakraborti *et al.* (2004) e Chakraborti *et al.* (2011)).

No que respeita à incerteza no desempenho da carta devido à estimação dos valores nominais do processo, pode aconselhar-se a utilização de estimadores robustos, combinações lineares de estimadores, ou classes de estimadores com um parâmetro de afinamento de modo a selecionar em cada caso o melhor estimador da classe face aos dados em estudo. A utilização de procedimentos dinâmicos de estimação de forma a permitir a atualização dos valores nominais do processo é também uma opção. Este método tem sido sugerido na determinação do valor de referência associado à implementação de uma carta CUSUM. Efetuar uma análise de sensibilidade é também bastante útil, pois os valores nominais do processo que foram fixados podem não ser os corretos. Neste contexto de implementação de cartas de controlo, o próprio processo de amostragem utilizado na recolha dos dados na fase de estimação e das amostras subsequentes consideradas na fase de monitorização propriamente dita, assim como o processo de análise dos mesmos, é também uma fonte incerteza. Alguns detalhes sobre estas problemáticas podem ser encontrados em Huber (1964), Hampel (1971, 1974), Figueiredo e Gomes (2004), Saisana *et al.* (2005), Jensen *et al.* (2006), Chakraborti *et al.* (2009), Wu et al. (2009), Li and Wang (2010) e Psarakis *et al.* (2014), entre outros trabalhos.

Existe também incerteza na avaliação do desempenho e comparação de diferentes cartas, decorrente do cálculo dos erros tipo I e tipo II, ou das medidas de *performance* associadas à tomada de decisão, quando a incerteza do sistema de medição não está contemplada nestes erros ou medidas de desempenho, e não pode ser negligenciada. As regras de sequências (*runs*) usadas conjuntamente com as regras usuais de tomada de decisão numa carta de controlo permitem contornar alguma possível incerteza na tomada de decisão, ao analisarem a aleatoriedade ou não dos dados representados na carta e a existência ou não de padrões.

Na perspetiva de gestão total da qualidade, TQM (*Total Quality Management*), a função prejuízo, a função utilidade e a razão sinal/ruído são procedimentos utilizados também com o objetivo de medir a incerteza.

Mais detalhes sobre os principais procedimentos e ferramentas usuais em Controle Estatístico da Qualidade podem ser encontrados, por exemplo, em Montgomery (2009) e Gomes *et al.* (2010).

Para concluir é de referir que a teoria dos conjuntos difusos como alternativa, ou pelo menos como complementar, à utilização de métodos estatísticos e probabilísticos, assim como a utilização de metodologias Bayesianas, têm sido abordagens frequentemente consideradas por vários autores para resolver problemas que envolvem incerteza. Entre outros trabalhos, veja-se por exemplo, Laviolette *et al.* (1995), Lira e Woger (2006) e Willink (2007).

Agradecimentos: Investigação parcialmente financiada por fundos nacionais através da **FCT**—Fundação para a Ciência e a Tecnologia, no âmbito do projecto UID/MAT/00006/2013 (CEA/UL).

Referências

[1] Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions*, **27**, 186–90; reeditado em Kendall and Plackett (1977).

[2] Azzalini, A. (1985). A Class of distributions which includes the normal ones. *Scandinavian J. of Statistics*, **12**, 171–178.

[3] Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian J. of Statistics*, **32**, 159–188.

[4] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Royal Soc. London*, **53**, 370–418. Há reedição recente por Pearson, E. S., and Kendall, M. G. (1970). *Studies in the History of Statistics and Probability*, Griffin, London.

[5] Bernoulli, J. (1713). *Ars Conjectandi* (reedição moderna: Editions Culture et Civilisation, 1968, e tradução inglesa preparada por Bing Sung, Harvard University, Cambridge, MA, 1966).

[6] de Moivre, A. (2015). *The Doctrine of Chances*, reedição da 3^a edição de 1756, Andesite Press.

[7] Chakraborti, S., van der Laan, P. e van de Wiel, M.A. (2004). A class of distribution-free control charts. *Journal of the Royal Statistical Society C – Applied Statistics*, **53**, 3, 443–462.

[8] Chakraborti, S., Human, S.W. e Graham, M.A. (2009). Phase I Statistical Process Control Charts: An Overview and Some Results. *Quality Engineering*, **21**, 1, 52–62.

[9] Chakraborti, S., Human, S.W. e Graham, M.A. (2011). Nonparametric (distribution-free) quality control charts. In N. Balakrishnan, Ed., *Methods and Applications of Statistics: Engineering, Quality Control and Physical Sciences*, 298–329.

[10] Figueiredo, F. e Gomes, M.I. (2004). The total median in Statistical Quality Control. *Applied Stochastic Models in Business and Industry*, **20**, 339–353.

[11] Figueiredo, F. e Gomes, M.I. (2013). The skew-normal distribution in SPC. *Revstat*, **11**, 83–104.

[12] Finetti, B. de (1974). *Theory of Probability*, Wiley, New York.

[13] Gomes, M.I., Figueiredo, F. e Barão, M.I. (2010). *Controlo Estatístico da Qualidade*, Edições SPE, 2^a edição.

[14] Hampel, F.R. (1971). A general qualitative definition of robustness. *Annals of Mathematics and Statistics*, **42**, 1887–1896.

[15] Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Society*, **69**, 383–393.

[16] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematics and Statistics*, **35**, 73–100.

[17] ISO (1995). *Guide to the Expression of Uncertainty in Measurement – GUM.*

[18] Jackson, C.H. (2008). Displaying Uncertainty with Shading. *The American Statistician*, **62**, 4, 340–347.

[19] Jamalizadeb, A., Arabpour, A.R. e Balakrishnan, N. (2011). A generalized skew two-piece skew normal distribution. *Statistical Papers*, **52**, 431–446.

[20] Jaynes, E.T. (2003). *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge.

[21] Jensen, W.A., Jones-Farmer, L.A., Champ, C.W. e Woodall, W.H. (2006). Effects of parameter estimation on control chart properties: A literature review. *Journal of Quality Technology*, **38**, 349–364.

[22] Laplace, P.H.S. de (1774). Mémoire sur la probabilité des causes par les évènements. *Mém. Acad. R. Sci. Paris (Savants Etrangers)*, **6**, 621–656; **8**, 27–65. (Tradução inglesa (S. M. Stigler) em *Statist. Sci.*, **1**, 359–378.)

[23] Laplace, P.H.S. de (1814). *Théorie Analytique des Probabilités*, 2ème ed., Mme Veuve Courcier, Paris.

[24] Laviolette, M., Seaman, J.W.Jr., Barrett, J.D. e Woodall, H.W. (1995). A Probabilistic and Statistical View of Fuzzy Methods. *Technometrics*, **37**, 3, 249–261.

[25] Lira, I. e Woger, W. (2006). Comparison between the conventional and Bayesian approaches to evaluate measurement data. *Metrologia*, **43**, S249–S259.

[26] Li, Z. e Wang, Z. (2010). Adaptive CUSUM of the Q chart. *International Journal of Production Research*, **48**, 1287–1301.

[27] Lopes, I.S., Sousa, S.D. e Nunes, E. (2015). Methodology for uncertainty characterization of performance measures. *International Journal of Quality and Reliability Management*, **33**, 9, 1346–1363.

[28] Losinger, W.C. (2004). A Review of the GUM Workbench. Dealing with Uncertainty: Statistics for an Aging Population. *The American Statistician*, **58**, 2, 165–167.

[29] Montgomery, D.C. (2009). *Introduction to Statistical Quality Control*, 6th edition. John Wiley & Sons, Inc..

[30] NASA (2010). *NASA Measurement Quality Assurance Handbook ANNEX 3 – Measurement Uncertainty Analysis, Principles and Methods.*

[31] Pascal, B. e Fermat, P. – A troca de correspondência sobre a questão do Cavaleiro de Meré encontra-se reproduzida em David, F.N. (1962). *Games, Gods and Gambling*, Griffin, London.

[32] Pestana, D.D. e Velosa, S. (2010). *Introdução à Probabilidade e à Estatística*, 4^a edição revista, Fundação Calouste Gulbenkian, Lisboa.

[33] Pierre, H. M. (1713). *Essai d'Analyse des Jeux de Hazard* (existe edição moderna acessível, Chelsea, New York, 1980, ou Editions Jacques Gabay, Paris).

[34] Psarakis, S., Vyniou, A.K. e Castagliola, P. (2014). Some recent developments on the effects of parameter estimation on control charts. *Quality and Reliability Engineering International*, **30**, 1113–1129.

[35] Saisana,M., Saltelli, A. e Tarantola, S. (2005). Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators. *Journal of the Royal Statistical Society, Series A*, **168**, 2, 307–323.

[36] Stoto, M.A. (1988). Dealing with Uncertainty: Statistics for an Aging Population. *The American Statistician*, **42**, 2, 103–110.

[37] von Mises, R. (1981). *Probability, Statistics and Truth*, Dover, New York.

[38] Willink, R. (2006). On the uncertainty of the mean of digitized measurements. *Metrologia*, **44**, 73–81.

[39] Wu, Z., Jiao, J., Yang, M., Liu, Y. e Wang, Z. (2009). An enhanced adaptive CUSUM control chart. *IIE Transactions*, **41**, 642–653.

Pós-Doc

Tese de doutoramento: New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions

(Boletim SPE outono de 2012, p.62)

Paulo Canas Rodrigues, *paulocanas@gmail.com*

CAST—Center for Applied Statistics and Data Analytics, University of Tampere, Finland
e
Departamento de Estatística, Universidade Federal da Bahia, Salvador, Brasil

Caros Colegas,

Realizei as minhas provas públicas de doutoramento na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa no dia 28 de fevereiro de 2012. Foi uma data extremamente importante que permitiu concluir um longo projeto de vida (pessoal e profissional). Foi uma fase da minha vida muito enriquecedora e gratificante já que tive a possibilidade e privilégio ter feito parte de quatro excelentes grupos de investigação ao passar cerca de oito meses na Polónia, seis meses nos Estados Unidos e dois anos e meio na Holanda, além do tempo passado em Portugal. Mas nem tudo foram viagens, foi um processo longo e difícil com algumas adaptações e mudanças de planos que no final apenas desejava ver terminado. Várias coisas mudaram ao longo destes quatro anos (e mais uns meses) mas também houve algo constante: os grandes mestres e extraordinárias pessoas que me acompanharam, o Professor Stanislaw Mejza e o Professor João Tiago Mexia!
Mas, como sempre costumo dizer: terminar o doutoramento é de alguma forma similar a obter a carta de condução. A partir dessa data já podemos conduzir sozinhos!
Durante os quase cinco anos (estou a escrever este texto a dez dias de completar esses cinco anos) que passaram desde as provas de doutoramento, muito mudou na minha vida. Após dar aulas em Portugal cerca de um ano, em abril de 2013 comprei viagem só de ida para Salvador da Bahia para tomar posse como "professor adjunto" (equivalente a professor auxiliar) na Universidade Federal da Bahia. O processo para conseguir esta posição foi bastante penoso, quer as provas do concurso, quer o tempo de espera para poder assinar contrato, mas isso daria uma longa história! Apenas penso ser aqui relevante referir que em novembro/dezembro de 2011 tive que fazer uma escolha: terminar o doutoramento imediatamente ou fazer uma pausa para estudar para um concurso em Salvador e viajar para fazer as provas. Decidi dar prioridade ao concurso e foi, sem qualquer dúvida, uma excelente escolha!
Entretanto algo voltou a mudar. No dia 8 de setembro de 2014 recebi um email que tive que ler várias vezes. O Diretor de uma das Escolas da University of Tampere, na Finlândia, onde, ate então, não conhecia ninguém, escreveu-me um email a dizer que a universidade estaria a no processo de criação de um centro de investigação em estatística aplicada e análise de dados. Até aqui tudo normal, mas o segundo parágrafo do email escrevia o seguinte:
"*I am now in search of a leader to take charge of the applied statistics center which will be university level function hosted by School of Information Sciences. It has been brought to my attention that you might have the qualifications and interest to this type of task. Would you be interested in this and if so would you have a chance to have a deeper discussion about the topic over the phone in next few weeks?"*.
Naturalmente nem hesitei em dizer que estaria sim interessado nesta posição. Após vários emails, conversas e entrevistas via Skype, fui convidado para dar um seminário e fazer uma entrevista nas

instalações da University of Tampere, em janeiro de 2015. Umas semanas mais tarde recebi uma oferta formal para ser o Research Director deste centro de investigação, agora denominado de CAST (Center for Applied Statistics and Data Analytics). Não poderia recusar tal desafio e não recusei! Primeiro trabalhei cerca de um ano em part-time remotamente e com algumas visitas à Finlândia. Depois pedi "Licença para tratar de interesses particulares" (uma espécie de licença sem vencimento) na Universidade Federal da Bahia e estou agora em Tampere, na gelada Finlândia!

Uma mudança entre dois extremos. Penso que em qualquer escala que se consiga imaginar, Salvador e Tampere estão em extremos opostos!

Entretanto também tenho aumentado as minhas atividades e contribuição para a comunidade científica sendo atualmente, entre outros: (i) Chair da recém criada Latin American Regional Section (LARS) da International Association for Statistical Computing (IASC); (ii) Vice-Presidente da Região Brasileira da Sociedade Internacional de Biometria; (iii) Membro do Comité Excecutivo da IASC; (iv) Council Member do International Statistical Institute (ISI); (v) Council Member da International Society for Business and Industrial Statistics; (vi) Chair do ISI Young Statisticians Committee; e (vii) Membro do Scientific Board do European PhD in Socio-Economic and Statistical Studies.

Termino com duas das minhas frases favoritas. A primeira, uma citação do Walt Disney e que é uma reflecção extraordinária:

If you can dream it, you can do it

A segunda, escutei-a pela primeira vez nas palavras do meu mentor quando estava na Holanda durante o meu doutoramento. Parece ser uma cotação do teólogo político inglês Algernon Sidney que ficou famosa após Benjamin Franklin a ter usado em 1736 no seu almanaque anual "Poor Richard's Almanack":

God helps those who help themselves

Saudações académicas e pessoais!

Paulo Canas Rodrigues

Tese de Doutoramento: Uma Aplicação da Metodologia ROC na Análise de Dados de *Microarrays*

(Boletim SPE primavera de 2013, p.79)

Carina Silva, *carina.silva@estesl.ipl.pt*

Escola Superior de Tecnologia da Saúde de Lisboa (ESTeSL)– Instituto Politécnico de Lisboa
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

Já se passaram quase cinco anos??!! De facto, o tempo tem um comportamento *significativamente* diferente quando se está a realizar um doutoramento e após este! Foi no dia 29 de outubro de 2012 que realizei as minhas provas na Reitoria da Universidade de Lisboa. As minhas orientadoras Professora Antónia Turkman e Lisete Sousa, partilharam comigo esta travessia. Muito lhes devo, não só a minha evolução científica, mas também o rigor e a ética. São minhas professoras, mentoras e amigas. Lembro-me quando escrevi o primeiro artigo sem ser sob a alçada das minhas orientadoras. E agora?? Só me fez lembrar quando queremos muito ser independentes dos pais e à primeira "dor de barriga" só queremos voltar para casa. Mas pude continuar a contar com a sua colaboração, apoio e incentivo.

O tema do meu doutoramento obrigou-me a sair da minha área de conforto em várias vertentes. Primeiro porque tive de aprender conceitos de biologia, perceber a parte tecnológica que produz os dados (afinal o que é um *microarray*) e usar o R mais na lógica da programação, do que usar apenas as suas bibliotecas. Este foi com certeza o maior impacto. Sinto que houve um crescimento científico o que me levou a um estado esquizofrénico que afinal não sei nada. E por isso, muito humildemente, vou continuando a tentar aprender.

Todos certamente se revêem no percurso árduo que é realizar este projeto. Sendo docente na ESTeSL não tive direito a dispensa letiva, mas atenção, na altura a Presidência permitia a redução de duas horas na carga horária! Mas a colaboração com as mil quinhentas e cinquenta e sete comissões mantiveram-se. Na realidade o que sentia, era que fazia o doutoramento nas horas vagas, o que na prática correspondia na sua grande maioria aos fins de semana e durante a semana à noite. Este processo durou tanto tempo, que quando terminei, tinha uma sensação estranha quando durante os fins de semana tinha mais tempo para usufruir da família e amigos. Durante a realização do doutoramento também fui mãe. Embora não tenha sido fácil, mas ajudou-me a focar, pois tinha de terminar o mais rapidamente possível para poder usufruir em pleno (pensando ingenuamente que iria ter mais tempo!!). Antes da conclusão do doutoramento já era Professora Adjunta, categoria que ainda mantenho. Uns meses após a conclusão fui Coordenadora da Área Científica de Matemática e fui convidada a integrar a Comissão Executiva do Gabinete de Gestão da Qualidade da ESTeSL, a qual ainda integro. Como consequência, tive uma redução de 50% na carga horária letiva e um aumento de 500% de trabalho. Neste momento sou Diretora do Departamento das Ciências Naturais e Exatas. Na realidade a parte burocrática e de gestão foi tomando uma dimensão cada vez maior, o que teve o seu impacto na parte da investigação. Uma vez mais continuo a fazer investigação nas horas vagas. Enquanto reclamava que não tinha tempo para nada e que era um desgaste muito grande, o que ouvia da parte de quem tinha concluído era que o volume de trabalho iria aumentar. O que é certo é que só se acredita quando passamos pelas situações. No entanto, existe uma diferença. A sensação de "espada na cabeça", como

eu dizia muitas vezes, não é tão latente. Mas a dispersão, essa é muito grande. A minha "to do list" parece que padece de "virose da multiplicação".
Uma vez que o tema de doutoramento está relacionado com a genética, permitiu-me integrar o grupo de investigação em Genética e Metabolismo da ESTeSL e vou fazendo investigação em colaboração com outros colegas nas diversas áreas da saúde: Radiologia, Genética, Ciências Forenses, Farmácia, Ortóptica, Cardiopneumologia e muitas outras. Sendo Estatística numa escola de saúde, há sempre uma procura enorme de quem precisa que "trate" dos dados. Quantos de vós já ouviu a célebre expressão "passo por aí no gabinete, são só 5 minutinhos", e não sei quantas horas depois ainda se está a tentar perceber o problema. Mas é uma oportunidade para aprender sem dúvida.

Sendo docente no politécnico e com as acreditações dos cursos pela A3ES, sinto que de facto o meu doutoramento teve impacto. A expressão utilizada era "agora já contas para os ETIS", isto é, contribuía para a percentagem exigida pela A3ES de docentes doutorados. Também me recordo da primeira vez que fui júri num doutoramento. Lembro-me de ter pensado que tinha estado sentada tão há pouco tempo naquela cadeira. A obtenção do grau também me permitiu passar de colaboradora do CEAUL para membro integrado e desde o início deste ano faço parte da Comissão Executiva do CEAUL, o que muito me honra. Esta ligação ao centro para mim é fundamental, pois permite estar junto dos meus pares e poder evoluir como investigadora.

Estes quase 5 anos passaram a correr. Entre docência, gestão e investigação não há mãos a medir. Gostaria de ter mais tempo para a investigação sem dúvida. Mas quero acreditar (possivelmente de uma forma ingénua) que daqui a uns tempos seja possível optar por uma carreira mais virada para a investigação ou para a gestão. Continuar a exigir dos docentes do ensino superior esta dispersão trará consequências num futuro próximo (li um artigo há uns tempos que apontava que mais de 60% dos professores do ensino superior sofre de "burnout").
Em jeito de conclusão posso dizer que o doutoramento para além de ser um projeto científico é uma prova de resistência. E de facto se não formos puxados para além daquilo que pensamos que são os nossos limites, também não evoluímos. E o que não nos mata, torna-nos mais fortes.

Carina Silva

Ciência Estatística

• Artigos Científicos Publicados

Beirlant, J., Fraga Alves, M.I. and Gomes, M.I. (2016). Tail fitting for truncated and non-truncated Pareto-type distributions. *Extremes*, 19:3, 429–462.

Fraga Alves, I., Neves, C. and Rosário, P. (2017). A general estimator for the right endpoint with an application to supercentenarian women's records. *Extremes,* 20:1, 199-237.

Gouveia, S., Scotto, M.G., Weiss, C.H. and Ferreira, P.J.S.G. (2017). Binary autoregressive geometric modelling in a DNA context. *Journal of the Royal Statistical Society*, Series C 66, 253-271.

Mendes, S. e outros (2016): CO-tucker: a new method for the simultaneous analysis of a sequence of paired tables. *Journal of Applied Statistics*, 1-27.

Pereira, L.N. (2016). An introduction to helpful forecasting methods for hotel revenue management. *International Journal of Hospitality Management*, 58, 13-23.

• Capítulos de Livros

Fraga Alves, M.I. and Neves, C. (2016). Extreme Value Theory: An Introductory Overview. In *Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications*, (Longin Ed.). Handbook Series in Financial Engineering and Econometrics (Ruey Tsay Adv.Ed.). pp 53-95. John Wiley & Sons.

• Livros

Título: *ESTATÍSTICA: Progressos e Aplicações.*
Atas do XXII Congresso da Sociedade Portuguesa de Estatística
Editores: Cordeiro, C., Ribeiro, C., Sousa, C., Gonçalves, M.H, Antunes, N. e Silva, M.E.
Ano: 2016. Editora: Edições SPE. ISBN: 978-972-8890-39-4. (ISBN online: 978-972-8890-39-1) Depósito Legal: 417937/16.

Título: *As Sondagens: Princípios, Metodologias e Aplicações*
Autores: Coelho, P. S., Pereira, L. N., Pinheiro, J. A. e Xufre, P.
Ano: 2016. Editora: Escolar Editora. ISBN: 978-972-59-2514-0

• Teses de Mestrado

Título: Modelos para Acontecimentos Múltiplos
Autor: Ivo Miguel Sousa Ferreira, *ivooferreiraa@gmail.com*
Orientadora: Ana Maria Abreu

Título: A Classe de Distribuições de Panjer e a Modelação do Risco Coletivo
Autor: Duarte Nuno da Silva Sousa, *duartesousa1993@hotmail.com*
Orientadora: Sandra Mendonça

Título: Modelação de propriedades de pastas de betão autocompactável com comportamento heterocedástico
Autor: Diogo Filipe de Bastos Sousa Ribeiro, *up199300364@fc.up.pt*
Orientadoras: Paula Milheiro de Oliveira e Sandra Nunes

Título: Estudo de sobrevivência de doentes com cancro de colo de útero
Autora: Paula Rosa Lopes, *rosyfogo09@gmail.com*
Orientador: Luís Machado

Título: Análise do Erro de Estimação em Filtros de Bloom Lineares
Autora: Célia Ferreira, *celiacf64@gmail.com*
Orientadores: Raquel Menezes e Carlos Baquero

Título: Caracterização de Alterações Visuais na fase Inicial da Esclerose Múltipla
Autora: Olga Joana Miguel, *joaniinha@hotmail.com*
Orientadores: Inês Sousa e Paulo Pereira

Título: Modelos Longitudinais e de Sobrevivência para Recidiva do Cancro da Mama
Autora: Liliana Coelho, *lilianarcoelho@hotmail.com*
Orientadora: Inês Sousa

Título: Métodos para a Deteção e Modelação de Tendências - Uma aplicação a variáveis ambientais
Autor: Olexandr Batuin, *olexandr.baturin@gmail.com*
Orientadores: Arminda Manuela Gonçalves e Marco Costa

Título: Métodos econométricos de desagregação temporal - migração de código TSP para ambiente R
Autor: Jorge Vieira, *jorgealexandrevieira@gmail.com*
Orientadoras: Raquel Menezes e Alda Manso Rito

Título: Exploração e extensão do package *prob* existente no R
Autor: Nuno Martins, *nunomartins_11@hotmail.com*
Orientadora: Cecília Azevedo

Título: Modelos de estimação aplicados às estatísticas do Comércio Internacional
Autora: Lídia Sá, *lidiamariasa@gmail.com*
Orientadores: Raquel Menezes, Susana Faria e Cristina Neves

• Teses de Doutoramento

Título: Métodos estatísticos aplicados à modelação de dados oncológicos.
Autor: Ricardo Miguel Vieira de São João, *rsj@net.sapo.pt*.
Orientadores: Ana Luísa Trigoso Papoila da Silva e Bruno Cecílio de Sousa.

A minha tese foi dedicada aos "Métodos Estatísticos Aplicados à Modelação de Dados Oncológicos" e teve como principal fonte de informação o registo de cancro de base populacional da região sul de Portugal (ROR-Sul). A pertinência dos dados é justificada pelo facto do registo permitir aferir na sua população, o risco desta sofrer ou vir a sofrer de uma dada neoplasia. Foram considerados todos os casos diagnosticados com cancro do cólon, recto e estômago no período de 1998-2006. Dentre as minhas motivações, gostaria de realçar: a importância crescente que o cancro assume na nossa sociedade, em particular nestas neoplasias, tocando muitas vezes pessoas próximas e que amamos; o desafio da modelação da incidência do cancro por metodologias estatísticas distintas.
Como objetivo principal destaco a apresentação e aplicação de metodologia estatística, implementada em código R, permitindo uma análise sob diferentes prismas. Os modelos abordados foram:
Modelos Idade-Período-Coorte *(Age-Period-Cohort models- APC)* por permitem modelar a taxa de incidência segundo três variáveis temporais. No cancro, para além do período (ano de diagnóstico), a idade à data do diagnóstico e a coorte de nascimento, são variáveis temporais que poderão prestar um contributo adicional na caracterização das taxas de incidência. Os referidos modelos ultrapassam o problema de relações não lineares e/ou de mudanças súbitas na tendência linear das taxas.
Modelos de projeção para as taxas no período 2007-2010, sendo apresentadas estimativas do impacto económico. Conhecido o comportamento das taxas de incidência, uma questão subsequente prende-se com a sua projeção em períodos futuros. Embora as projeções tenham alguma incerteza associada, auxiliam no planeamento de políticas de saúde.
Modelos de regressão de Sobrevivência Relativa (SR) por ser uma medida objetiva que nos permite dar uma estimativa da probabilidade de sobrevivência caso o cancro em análise, num cenário hipotético, seja a única causa de morte. Nos modelos foram considerados as variáveis: período de *follow-up* de 5 anos, o sexo, a idade e cada uma das regiões que constituem o ROR-Sul.
Modelos Bayesianos Hierárquicos Espaço-Temporais com o objetivo de aferir se as variações nas taxas de incidência observadas entre os concelhos inseridos na área do registo poderiam ser explicadas quer pela variabilidade temporal e geográfica quer por fatores socioeconómicos ou, ainda, pelos desiguais estilos de vida.
Pretendo que os resultados da tese sejam um contributo para um retrato mais nítido destas neoplasias em Portugal e desejo que a sua leitura possa servir de estímulo na constante batalha contra o cancro.

Ricardo São João

Prémio SPE 2017

Está aberto, até **30 de junho de 2017**, o concurso para atribuição do **Prémio SPE 2017**, de acordo com o seguinte regulamento:

1. Pretendendo estimular a atividade de estudo e investigação científica em Probabilidades e Estatística entre os jovens, é instituído o **Prémio SPE 2017**.

2. O **Prémio SPE 2017** é constituído por uma quantia de 1000 euros.

3. Ao **Prémio SPE 2017** podem concorrer trabalhos originais sobre temas de Probabilidades e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição.

4. Os autores dos trabalhos candidatos ao **Prémio SPE 2017** devem ser estudantes ou investigadores em alguma instituição portuguesa, devem ser sócios da SPE e não devem ter completado os 35 anos de idade até 31 de dezembro de 2017. Os autores não devem ter recebido o Prémio SPE nas quatro edições anteriores. O trabalho deve ser escrito em português e não poderá exceder 25 páginas A4.

5. As candidaturas deverão vir acompanhadas do trabalho concorrente e do *curriculum vitae* dos autores e ser dirigidas à Presidente da SPE. Podem ser enviadas por correio electrónico para **spe@fc.ul.pt** ou, em carta registada, para a morada a seguir indicada. O carimbo do correio validará a data de entrega.

   Sociedade Portuguesa de Estatística, Bloco C6,
   Piso 4 - Campo Grande
   1749-016 LISBOA

6. A decisão de admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição será da responsabilidade da Direção da SPE.

7. Os critérios de seleção pautar-se-ão pela exigência e precisão nos vários aspetos que o júri considerar pertinentes, nomeadamente: i) qualidade e clareza do texto; ii) inovação e rigor científico; iii) contribuição para o desenvolvimento da área de Probabilidades e Estatística nos planos teórico, metodológico e/ou aplicado.

8. O júri é soberano nas suas decisões, não havendo lugar a recurso.

9. O trabalho galardoado com o **Prémio SPE 2017** será apresentado em sessão plenária pelo seu autor ou autores durante o XXIII Congresso da SPE.

10. O júri reserva-se o direito de não atribuir o **Prémio SPE 2017**.

PRÉMIOS "ESTATÍSTICO JÚNIOR 2017"
REGULAMENTO

Está aberto, **até 26 de Maio de 2017**, o concurso para atribuição de prémios **"Estatístico Júnior 2017"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2017"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos dos Ensinos Básico e Secundário pelas áreas das Probabilidades e Estatística.

2. Os candidatos aos prémios **"Estatístico Júnior 2017"** devem ser alunos do 3.° Ciclo do Ensino Básico, do Ensino Secundário, dos Cursos de Educação e Formação (CEF) ou dos Cursos de Educação e Formação de Adultos (CEFA), no ano letivo 2016-2017.

3. As candidaturas podem ser individuais ou em **grupo com um máximo de 3 alunos**. Do grupo pode ainda fazer parte um professor, do grau de ensino em que o trabalho se insere, ao qual caberá o papel de orientador.

4. Os candidatos devem apresentar um trabalho cuja temática deve estar relacionada com as áreas de Probabilidades ou Estatística.

5. O **trabalho** deverá ser constituído por um texto escrito em Português com um máximo de 10 páginas A4 dactilografadas e um *poster* formato A2 que resuma os principais aspetos do trabalho.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2017**" a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.° Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário e um primeiro classificado de entre os trabalhos candidatos dos Cursos CEF ou CEFA. Os prémios são constituídos por lotes de livros presentes nas notas de encomenda da Porto Editora (à exceção de manuais escolares e livros auxiliares), no valor de 500€ para os classificados em primeiro lugar e de 200€ para os classificados em segundo e terceiro lugares.

7. Ao professor orientador do trabalho classificado em 1º lugar, em cada grau de ensino, é atribuída uma anuidade grátis como sócio da SPE, ajudas de custo para participação na Sessão de Entrega do Prémio e lotes de livros presentes nas notas de encomenda da Porto Editora (à exceção de manuais escolares e livros auxiliares), no valor de 350€.

8. Aos grupos proponentes dos trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente *poster* que será exposto na Sessão de Entrega do Prémio.

9. A candidatura é composta pelo **Boletim de Candidatura**, devidamente preenchido, e pelo **trabalho (**poster e texto**)**. A candidatura, dirigida ao Presidente da SPE, deverá ser enviada

    a. **impressa em papel para efeitos da avaliação** para:
    Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa
    b. **em formato digital (pdf) por e-mail para spe@fc.ul.pt**

10. O carimbo do correio validará a data de entrega do trabalho, sendo os autores notificados por e-mail sobre a sua receção no prazo de uma semana.

11. A admissibilidade e apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição e nomeação será da responsabilidade da Direção da SPE.

12. O júri é soberano nas decisões, não havendo lugar a impugnação ou recurso.

13. A atribuição dos prémios **"Estatístico Júnior 2017"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada numa sessão expressamente dedicada a essa entrega.

14. Os prémios **"Estatístico Júnior 2017"** poderão não ser atribuídos.

15. O boletim de candidatura e este regulamento podem ser obtidos em

http://www.spestatistica.pt

Apoio Porto Editora

Bolsas para participação no Congresso SPE 2017

Pretendendo estimular o estudo e a investigação científica em Probabilidades e Estatística entre os jovens, a SPE atribui um número limitado de bolsas para participação no Congresso da SPE 2017, de acordo com o seguinte regulamento:

1. Os candidatos devem ser estudantes de mestrado ou de doutoramento inscritos no ano lectivo 2016/2017 em alguma instituição portuguesa.
2. São também admitidos candidatos que tenham completado o respectivo ciclo de estudos durante o ano de 2016.
3. Os candidatos não devem ter completado os 35 anos de idade até 31 de dezembro de 2017.
4. A bolsa é constituída pela inscrição no Congresso e por uma quantia de 100 euros.
5. A candidatura consta de um resumo alargado (documento em pdf com 2 a 4 páginas) para uma comunicação oral e de uma carta de apresentação onde deve constar uma breve biografia. O resumo deve ser escrito em Português. A candidatura deve ser enviada à Direção da SPE para spe@fc.ul.pt até 15 de junho de 2017. Os candidatos devem fazer prova das condições de admissibilidade descritas em 1, 2 e 3.
6. A decisão será comunicada a 15 de julho de 2017.

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

www.spestatistica.pt

O MUNDO DA **ESTATÍSTICA**

ORGANIZAÇÃO PARTICIPANTE

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR