Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *primavera de 2018*

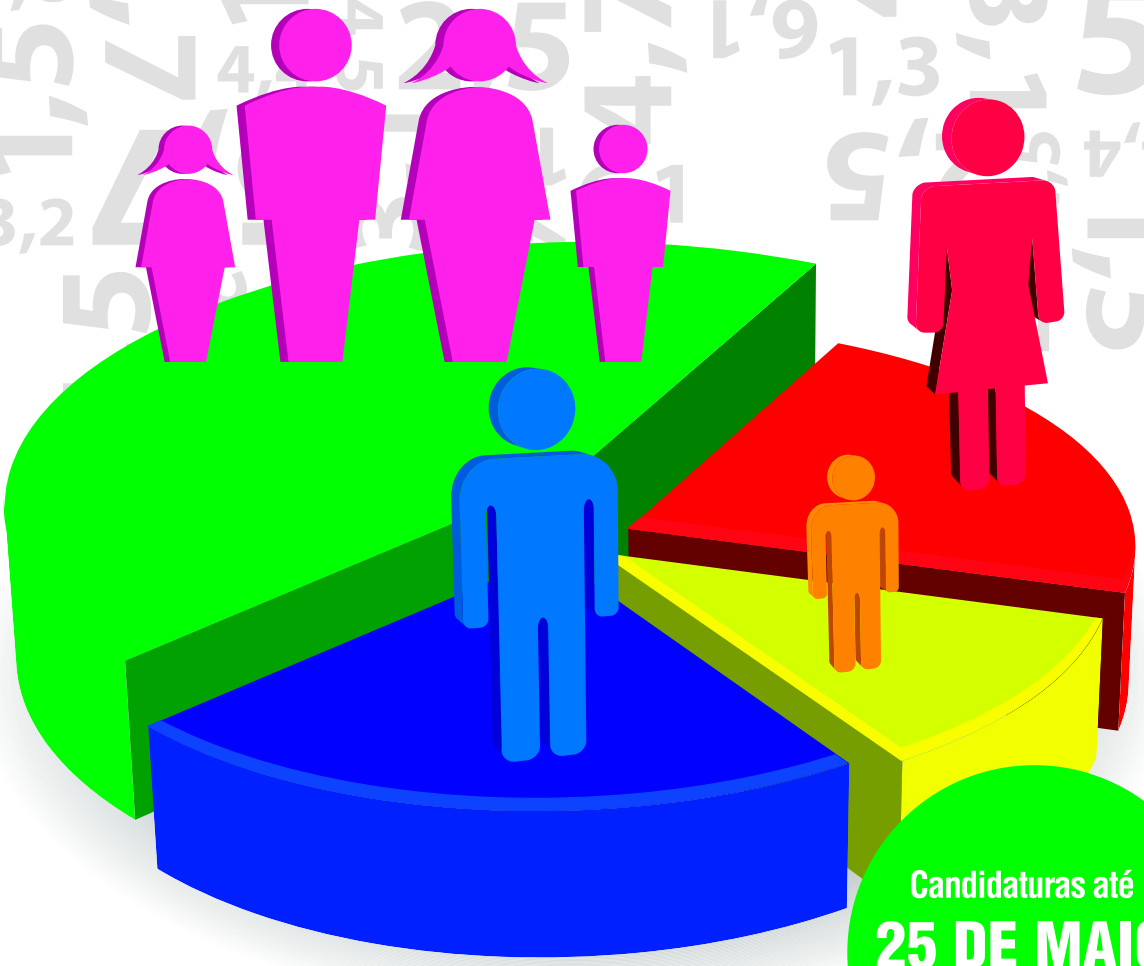
Estatística Multivariada – perspetiva no século XXI

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe*@spestatistica.pt
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 400 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO ESTATÍSTICO JÚNIOR 2018



Candidaturas até
25 DE MAIO DE 2018

CONTACTOS

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

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@spestatistica.pt

Com o apoio:
Porto Editora

Editorial

…para o Bem da Ciência e da Estatística…

1. A atual Direção SPE presidida por Maria Eduarda Silva – empossada em 9 de fevereiro e de que na respetiva secção damos Notícia – teve a gentileza de me convidar para continuar... É uma honra que de imediato aceitei com o intuito de prosseguir e melhorar o trabalho. Nele pode-se introduzir mais alguma reflexão sobre a "problemática editorial". Para tal, a oportunidade de uma nova Direção é incentivadora. E assim, a Direção já reuniu para o efeito e convidou o Editor para participar. Várias "pequenas novas ideias" foram avançadas para concretização e, em breve, podemos dar notícia; com o objetivo principal de reduzir custos, uma dificuldade inerente aos tempos que se vivem nos mais diversos domínios. O Boletim SPE está consolidado na sua maqueta editorial. Ela assenta basicamente em três secções: *Notícias* (científicas e da comunidade), O *Tema Central* e *SPE e a Comunidade*. O *Tema Central* foi iniciado no outono de 2006 e *SPE e a Comunidade* na primavera de 2008. O *Tema Central,* de facto é uma "imagem de marca" do Boletim; fundamentalmente pela força de desejar ser uma "atualização e ponto da situação" para determinado assunto, em termos de grande divulgação pela comunidade científica. A criação deste espaço, como escrevi em editorial, acrescentou matéria científica que podemos situar num objetivo vasto de divulgação da Estatística entre os sócios mas também destes para toda a comunidade.

Foi assim há já 12 anos. O amadurecimento adquirido ao longo de muitos anos bem como a opinião interventiva que tenho recebido dos sócios e leitores do Boletim SPE, permitem concluir sobre o bom modelo editorial assim construído.

Mas, tudo isto, sem prejuízo de um desiderato de melhor racionalização – por exemplo dos custos e da eficácia editoriais. Esta será também uma mais-valia do Boletim SPE em favor da SPE. Decerto, em breve haverá notícias.

2. Faleceu o Prof. Fernando Nicolau; uma triste notícia que o Boletim deve fazer incluir no Memorial dos Estatísticos em Portugal. Passou o seu tempo.

Como para todos nós um tempo formado por uma sucessão infinita de pequenos instantes. Nestes e aos mais diversos níveis o Fernando Nicolau construiu muitos momentos de pioneirismo – desde a liderança administrativa de coordenação académica nos órgãos diretivos de Escolas Universitárias até à criação e projeção de associações científicas congregadas em torno dos dados e da Ciência Estatística. A seu modo liderou projetos inovadores, sem dúvida com o maior interesse para a comunidade científica do seu tempo. Neste Boletim apresentamos um breve relato curricular.

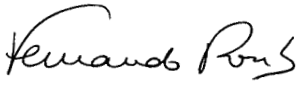
A todos os níveis merece a nossa homenagem!

Com o Fernando e a sua esposa, a Prof. Helena Nicolau, muitos de nós, tivemos a feliz oportunidade de participar nas mais diversas atividades inovadoras em Portugal nos domínios da implementação, divulgação e formação académica em Estatística e Análise de Dados.

Com muita saudade desses tempos de grande inovação e com muita pena, vemos desaparecer, precocemente, um daqueles que, em Portugal, foram pioneiros na moderna Ciência Estatística.

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

O Tema Central do próximo Boletim SPE será *Equações diferenciais estocásticas e algumas aplicações*



Mensagem da Presidente

Caros sócios da SPE,

Os novos órgãos administrativos da SPE, eleitos em Assembleia Geral que decorreu durante o XXIII Congresso da SPE, tomaram posse no dia 09/02/2018 em sessão realizada na sede da SPE. Neste momento de transição quero agradecer a todos os colegas que se empenharam em cargos do mandato findo e a todos os que colaboraram com a Direção no desenvolvimento de atividades em prol da SPE e da estatística em Portugal. Quero agradecer aos colegas Marília Antunes e Tiago Marques, que por razões profissionais deixam o Conselho Fiscal, o trabalho desenvolvido com a Direção anterior. Quero agradecer muito particularmente à Patrícia Bermudez que deixa, por vontade própria, o cargo de tesoureira. A Patrícia deparou-se ao longo do último mandato com situações difíceis que resolveu com empenho, voluntarismo e persistência. MUITO OBRIGADA, Patrícia em nome de todas nós e, muito particularmente em meu nome. Quero, ainda, agradecer aos novos elementos dos Órgãos Sociais terem aceite o desafio para participar nesta aventura.

Estamos, assim, no início de um mandato determinados a continuar a envidar todos os esforços para bem servir a Estatística em Portugal. Os principais problemas e desafios enumerados neste boletim em 2015 mantém-se mas as condições que temos para abordar estes problemas degradaram-se, dada a carga crescente de trabalho a que os docentes do Ensino Superior que constituem a maioria dos sócios da SPE têm vindo a ser sujeitos, dificultando o envolvimento participado dos sócios na vida da sociedade.
Termino certa do empenhamento dos sócios para com a SPE e a Estatística. A Sociedade é dos sócios e para os sócios e é, essencialmente, o que os sócios fizerem dela.

Porto, 25 de Fevereiro de 2018

Cordiais saudações

Maria Eduarda Silva

Notícias

• Novos Órgãos Sociais da Sociedade Portuguesa de Estatística

Em sessão realizada na sede da SPE, tomaram posse no dia 10 de fevereiro de 2018, os elementos constituintes dos seus órgãos administrativos eleitos no passado dia 7 de novembro.
A constituição dos novos órgãos administrativos da SPE para o triénio 2018 – 2020 é a seguinte:

Mesa Assembleia Geral

Presidente: Maria Antónia Turkmann, Universidade de Lisboa
Primeiro Vogal: Carlos Macedo, Instituto Nacional de Estatística
Segundo Vogal: Russell Alpizar-Jara, Universidade de Évora

Direcção

Presidente: Maria Eduarda Silva, Universidade Porto
Vice-Presidente: Isabel Simões Pereira, Universidade de Aveiro
Tesoureiro: Conceição Amado, Universidade de Lisboa
Primeiro Vogal: Cláudia Nunes Philippart, Universidade de Lisboa
Segundo Vogal: Maria Esmeralda Gonçalves, Universidade de Coimbra

Conselho Fiscal

Presidente: Graça Themido, Universidade de Coimbra
Primeiro Vogal: Carla Henriques, Instituto Politécnico de Viseu
Segundo Vogal: Maria João Polidoro, Instituto Politécnico do Porto

Direção cessante

Direção SPE 2018 - 2020

• Comissões Especializadas e Representações na SPE

1. Secção Biometria

Presidente: Giovani Silva, Universidade de Lisboa-IST
Secretários: Laetitia Teixeira, Universidade do Porto- ICBAS
Miguel Pereira, Imperial College of London

2. CEE (Comissão Especializada de Educação)

Maria Eugénia Graça Martins (Coordenadora)
Maria Manuela Neves
Andreia Hall
Claúdia Nunes
Cristina Rocha
Fernanda Otilia Figueiredo

3. CENE (Comissão Especializada de Nomenclatura Estatística)

Carlos Daniel Paulino (Coordenador)
Dinis Pestana
João Branco

4. Exploristica

Pedro Campos (Coordenador)
Conceição Rocha
Paulo Infante

5. AEVAE (A Estatística vai à escola)
Coordenadores: Tiago Marques
Carla Henriques
Carla Santos
Cristina Dias
Fátima Brilhante
Sandra Mendonça

6. Representação no *IAVE*
Maria Eugénia Graça Martins (Avaliação de propostas de exames)
Cristina Rocha Martins (CC)
Fernanda Otília Figueiredo (Auditoria de Exames)

7. Representação na *CNM (Comissão Nacional de Matemática)*
Isabel Pereira

8. Representação no *CIM (Centro Internacional de Matemática)*
Esmeralda Gonçalves

9. Representação na *Rede Portuguesa de Matemática para a Indústria*
Cláudia Nunes

10. Representação na *FENSTATS*
Maria Eduarda Silva

11. Representação no *ISI – International Statistical Institute*
Maria Eduarda Silva

12. Representação no *IASE*
Pedro Campos

13. Representação no *Espaço Matemático em Língua Portuguesa (EMeLP)*
Andreia Hall

14. Representação na *Bernoulli Society*
Paulo Eduardo Oliveira

15. Representação no *Committee of European Statistics Accreditation*
Feridum Turkman

16. Co-editor Springer Book Series: Studies in Theoretical and Applied Statistics
Maria Eduarda Silva

17. Committee of internal cooperation
Maria Eduarda Silva

18. European Statistical Advisory Committee (ESAC)
Maria Eduarda Silva

• Faleceu o Professor Fernando Nicolau

No dia 12 de dezembro de 2017 faleceu o Professor Fernando Nicolau.
Fernando Augusto Antunes da Costa Nicolau, nasceu em Lisboa em 7 de Agosto de 1942. Era casado com a Prof. Helena Bacelar Nicolau.
Licenciado em Ciências Matemáticas pela Faculdade de Ciências da Universidade de Lisboa obteve, em 1971, o Diplôme d'Études Approfondies (DEA) em Estatística Matemática, no Institut de Statistique des Universités de Paris (ISUP) da Universidade de Paris VI; e em 1972, Docteur 3ème Cycle em Estatística Matemática, opção Análise de Dados, ISUP, Universidade de Paris VI (Pierre et Marie Curie).
Em 1981, obteve o grau de Doutor em Ciências, especialidade Probabilidades e Estatística, na Universidade de Lisboa (Faculdade de Ciências).
Em 1997, qualificou-se com o título de Agregado em Matemática, na Universidade Nova de Lisboa (Faculdade de Ciências e Tecnologia).

Fernando Nicolau iniciou a sua carreira profissional, em 1965, na Faculdade de Ciências da Universidade de Lisboa.
Foi Professor Associado do Departamento de Matemática da Universidade de Aveiro e Professor Associado com Agregação, de nomeação definitiva, do Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, desde 1995.
Nestas Universidades liderou os mais diversos projetos científicos.
Foi Presidente da direcção da APCE - Associação Portuguesa de Ciências Estatísticas.
Foi Vice-Reitor da Universidade Aberta.

As principais áreas de interesse científico de investigação e ensino do Prof. Fernando Nicolau foram **Estatística e Análise de Dados Multivariados** e **Análise Classificatória**.

Principalmente nestes domínios desenvolveu as suas mais importantes contribuições para a Ciência Estatística e publicou uma longa lista de trabalhos científicos nas mais variadas revistas nacionais e internacionais.
O nome do Prof. Fernando Nicolau fica também ligado ao início da Sociedade Portuguesa de Estatística de que, durante muitos anos, foi um membro muito ativo. Foi membro de diversas sociedades científicas internacionais.
Foi Coordenador Científico de alguns Laboratórios de Estatística e Análise de Dados.
Além disso, foi sócio fundador da Associação Portuguesa de Classificação e Análise de Dados - CLAD da qual era Presidente da Assembleia Geral.
Membro muito interventor e com uma intensa atividade científica o seu nome fica registado na folha da génese da moderna Academia portuguesa.

FR

• III Encontro Luso-Galaico de Biometria

A Sociedade Portuguesa de Estatística (SPE) e a Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO) estão a organizar, em colaboração com o Departamento de Matemática da Universidade de Aveiro, o III Encontro Luso-Galaico de Biometria (EBio2018) que decorrerá, entre 28 e 30 de junho de 2018.
Pretende-se com este encontro, dirigido a profissionais e utilizadores da Estatística, académicos, investigadores e estudantes, difundir os mais recentes avanços no desenvolvimento e aplicação de métodos estatísticos e matemáticos em Biologia, Medicina, Ecologia, Psicologia, Farmacologia, Agricultura, Meio Ambiente e outras Ciências da Vida.
O programa científico do Encontro inclui um minicurso, uma mesa-redonda, sessões plenárias, sessões convidadas e comunicações (orais e em painel) selecionadas.
Assim, apelamos à vossa participação através da submissão de trabalhos que podem ser apresentados nos idiomas português, galego ou inglês.

DATAS IMPORTANTES:

Submissão de resumos: 8 de abril de 2018
Notificação de aceitação: 11 de maio de 2018
Inscrição a preço reduzido e inclusão no livro de atas: 25 de maio de 2018
Para mais informações consultar o Website **http://ebio2018-pt.weebly.com/**

• Prémios "Estatístico Júnior 2018"

A Sociedade Portuguesa de Estatística promove estes prémios como incentivo à atividade de estudo em Probabilidades e Estatística entre os jovens.
A Sociedade Portuguesa de Estatística, uma vez mais, com o apoio da Porto Editora promove estes prémios. Assim, está aberto, **até 25 de Maio de 2018**, o concurso para atribuição de prémios **"Estatístico Júnior 2018"**.
O Regulamento pode ser consultado nesta edição do Boletim SPE primavera de 2018 ou no sítio da SPE em *http://www.spestatistica.pt/.*

FR

• Prémio SPE 2018

A Sociedade Portuguesa de Estatística, uma vez mais, promove este prémio como incentivo à atividade de estudo e investigação científica em Probabilidades e Estatística entre os jovens.
Assim, está aberto, até **31 de agosto de 2018**, o concurso para atribuição do **Prémio SPE 2018.**
O Regulamento pode ser consultado no final deste Boletim SPE primavera de 2018 ou no sítio da SPE em *http://www.spestatistica.pt/.*

FR

• Prémio "Iniciação à Investigação"

A Sociedade Portuguesa de Estatística instituiu o prémio **Iniciação à Investigação**, que premeia trabalho desenvolvido em Probabilidades e Estatística no âmbito de teses de mestrado.
Assim, está aberto, **até 31 de agosto de 2018**, o concurso para atribuição do prémio **"Iniciação à Investigação"**.
O Regulamento pode ser consultado no final desta edição do Boletim SPE primavera de 2018 ou no sítio da SPE em *http://www.spestatistica.pt/*.

FR

• Retrospetiva do Boletim SPE

O *Boletim SPE* através dos seus "Tema Central"

- Primavera de 2017 - Destaque: Incerteza em Engenharia
- Outono de 2016 - Destaque: O Tema Central da Estatística
- Primavera de 2016 - Destaque: Séries Temporais e suas aplicações
- Outono de 2015 - Destaque: Estatística em Genética
- Primavera de 2015 – Destaque: Estatística no Desporto
- Outono de 2014 – Destaque: Estatística no Ensino Básico e Secundário
- Primavera de 2014 – Destaque: (Um) Ano Internacional da Estatística
- Outono de 2013 – Destaque: A "Escola Bayesiana" em Portugal
- Primavera de 2013 – Destaque: Estatística não-paramétrica
- Outono de 2012 – Destaque: Métodos Estatísticos em Medicina
- Primavera de 2012 – Destaque: Estatística no Ensino Superior Politécnico
- Outono de 2011 – Destaque: Análise de Sobrevivência
- Primavera de 2011 – Destaque: Sondagens e Censos
- Outono de 2010 – Destaque: Estatística Espacial
- Primavera de 2010 – Destaque: Data Mining - Prospecção (Estatística) de Dados
- Outono de 2009 – Destaque: Modelos Econométricos
- Primavera de 2009 – Destaque: Investigação (em) Estatística
- Outono de 2008 – Destaque: Processos Estocásticos
- Primavera de 2008 – Destaque: ALEA - Um sítio do nosso mundo
- Outono de 2007 – Destaque: Bioestatística
- Primavera de 2007 – Destaque: A "Escola de Extremos" em Portugal
- Outono de 2006 – Destaque: Ensino e Aprendizagem da Estatística

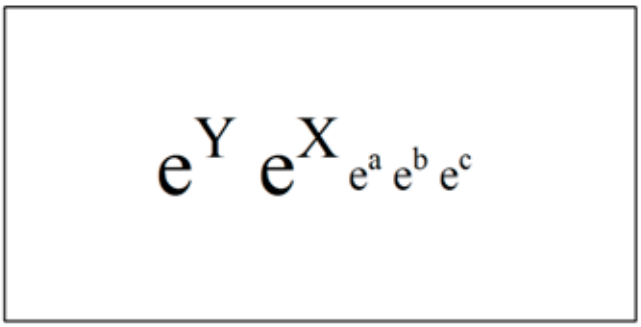
também disponíveis em *http://www.spestatistica.pt/index.php/publicacoes-57/boletins*

Enigmística de mefqa

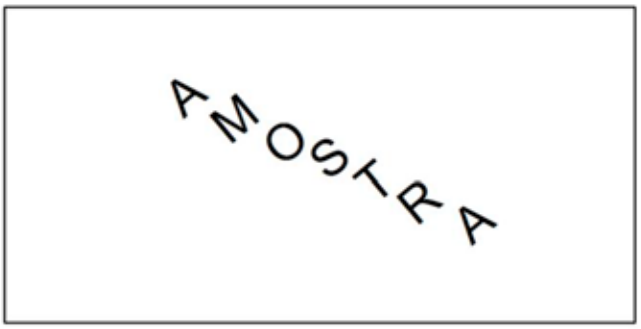
E M F O F D E E C L T S S

27 / 03 / 1857
S C I E N C E

No Boletim SPE outono de 2017 (p. 25):



Família Exponencial



Amostra Enviesada

Uma revisão sobre dados parcialmente sintéticos: Modelo de Regressão Linear Multivariada

Ricardo Moura, *pinto.moura@marinha.pt* e *rp.moura@fct.unl.pt*

CINAV, Centro de Investigação Naval, Marinha
CMA, Centro de Matemática e Aplicações, Universidade Nova de Lisboa

Nos nossos dias, uma simples utilização de um smartphone pode gerar uma multiplicidade de dados. Estes dados são guardados de forma quase automática e cada vez mais várias entidades, empresas e instituições "exigem" acesso a esta informação para a estudar e analisar. Contudo, a divulgação desses dados de uma forma desmedida e descontrolada poderá pôr em causa a confidencialidade de cada um dos indivíduos/unidades à qual a informação pertence. Posto isto, para se respeitar o princípio do segredo estatístico (Lei nº 22/2008, de 13 de Maio, Lei do Sistema Estatístico Nacional) para além da proteção física dos dados, isto é, dados que são guardados e apenas acessíveis a quem tenha a devida autorização, várias instituições nacionais ou internacionais usam habitualmente técnicas de controlo de divulgação estatística (CDE) com a finalidade de proteger a informação dos dados existentes que seja considerada confidencial, reduzindo o risco de se identificar um indivíduo (REGULATION (EC) No 223/2009, 2009) podendo dessa forma tornar públicos esses dados. Adição de ruído, arredondamentos, supressão local e geração de dados sintéticos são alguns exemplos de técnicas de CDE usados no EUROSTAT e no US CENSUS BUREAU antes de se disponibilizarem publicamente os dados. No contexto deste texto, irá ser aprofundada a técnica de geração de dados sintéticos, onde, de um modo sucinto, se substituem os dados originais por versões sintéticas destes. Para além de ser uma técnica relativamente recente, uma das suas maiores vantagens é a possibilidade de preservar as propriedades estatísticas do modelo, ao contrário de outras técnicas de CDE (Drechsler, 2011), e, portanto, instituições governamentais mundiais incentivam a sua investigação.

Poder-se-á dizer que Little (1993) e Rubin (1993) foram os pioneiros na exploração desta técnica por terem sido os primeiros a sugerir o uso de dados sintéticos gerados através de imputação múltipla (Rubin, 1987) como técnica de CDE, isto é, substituindo os dados originais por um conjunto de múltiplas versões sintéticas dos dados originais que podem ser divulgadas publicamente pois não possuem informação suficiente para comprometer a confidencialidade do indivíduo respondente. A viabilização de procedimentos que permitam a análise destes dados sintéticos gerados por imputação múltipla foi disponibilizada por Reiter (2003) e Raghunathan *et al*. (2003), motivados por uma perspetiva bayesiana assente em distribuições aproximadas que permitem o estudo de qualquer parâmetro ou vetor de parâmetros. No entanto, em certos casos (Kinney S. , *et al.*, 2011; Kinney S. , *et al*., 2011; Kinney, Reiter, & Miranda, 2014), devido ao elevado risco de divulgação da identidade do respondente não é possível divulgar múltiplas versões dos dados originais, exigindo-se a divulgação de apenas uma versão sintética destes, isto é, recorrendo apenas a dados gerados por imputação única. Motivados pela inexistência de procedimentos de análise inferencial destes dados, Klein e Sinha (2015; 2015; 2016) desenvolveram procedimentos exatos para a análise inferencial de dados sintetizados por imputação única, para vários modelos estatísticos incluindo o modelo de regressão linear múltipla. Em 2017, Moura *et al*. (2017a; 2017b; 2018) alargaram este estudo ao panorama multivariado de dados parcialmente sintetizados ao desenvolverem procedimentos exatos de inferência a dados sintéticos gerados pelos métodos *Posterior Predictive Sampling* (PPS), *Fixed-Posterior*

Predictive Sampling (FPPS) e *Plug-in Sampling* (Plug-in) sob o modelo de Regressão Linear Multivariada (RLM).

Geração de dados parcialmente sintéticos

Quando se refere que os dados são parcialmente sintetizados, trata-se de apenas gerar versões sintéticas dos valores registados por indivíduo que se consideram sensíveis, passíveis de comprometer a confidencialidade dos indivíduos, deixando os outros valores inalterados, protegendo sem comprometer a qualidade final dos dados divulgados. Assumindo, então, que um conjunto de dados estatísticos segue um modelo RLM, considera-se, no contexto da proteção da identidade, que as variáveis resposta serão as variáveis que põe em risco a confidencialidade e as variáveis explicativas serão as variáveis cujos valores registados poderão permanecer intactos por não violar esse pressuposto.

Para que se possa compreender melhor como se processa a técnica de geração de dados parcialmente sintéticos, será demonstrado o procedimento a tomar perante um conjunto de dados que seguem um modelo RLM. Consideremos que foram registados os valores relativamente a $m+p$ variáveis de n "indivíduos", dispostos de tal forma numa matriz

$$\begin{bmatrix} y_{1,1} & \cdots & y_{1,m} & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,m} & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}.$$

Considere-se agora o vetor $\boldsymbol{y} = (y_1, \dots, y_m)'$, contendo as m variáveis consideradas sensíveis e o vetor $\boldsymbol{x} = (x_1, \dots, x_m)'$ as p variáveis não-sensíveis. No modelo RLM, assume-se que $\boldsymbol{y}|_{\boldsymbol{x}} \sim N_m(\boldsymbol{B}'\boldsymbol{x}, \boldsymbol{\Sigma})$, onde $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$ são parâmetros desconhecidos, denominados por matriz dos coeficientes de regressão e matriz de covariância, respetivamente. Dessa forma, é possível resumir o modelo RLM a

$$\boldsymbol{Y}_{m\times n} = \boldsymbol{B}'_{m\times p}\boldsymbol{X}_{p\times n} + \boldsymbol{E}_{m\times n} \qquad (1)$$

onde $n \geq m + p$,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{bmatrix}, \qquad \boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}$$

e $\boldsymbol{E}_{m\times n} \sim N_{mn}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$.

De modo a não divulgar os valores originais das variáveis resposta, o que se pretende é substituir a matriz $\boldsymbol{Y}$, por uma ou mais versões sintetizadas por imputação única ou múltipla tendo por base o modelo (1). No caso de se proceder à geração de dados pelo método PPS, essas versões são obtidas recorrendo à distribuição à posteriori de $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$, imputando no modelo as estimativas destes parâmetros geradas aleatoriamente através dessas distribuições. Quando se gera pelo método Plug-in recorre-se diretamente às estimativas usuais que são imputadas no modelo diretamente para se gerar as versões sintéticas de $\boldsymbol{Y}$. Em Moura *et al*. (2017a; 2018) pode-se observar com maior detalhe como se processa essa geração, ao qual se apresenta de seguida um resumo.

Vamos denominar as versões criadas por FPPS, $\boldsymbol{W}_1, \dots, \boldsymbol{W}_M$ e as por Plug-in, $\boldsymbol{V}_1, \dots, \boldsymbol{V}_M$. Focando em primeiro lugar o caso da imputação única, é gerada apenas uma versão sintética a ser disseminada, tendo dessa forma apenas um $\boldsymbol{W} = \boldsymbol{W}_1$ (neste caso, o método PPS e FPPS coincidem) e um $\boldsymbol{V} = \boldsymbol{V}_1$.

Geramos $\boldsymbol{W}$, tendo em conta que $\boldsymbol{w}_i = (w_{1i}, \dots, w_{mi})'$ serão distribuídos independentemente como

$$\boldsymbol{w}_i|_{\widetilde{\boldsymbol{B}},\widetilde{\boldsymbol{\Sigma}}} \sim N_m\left(\widetilde{\boldsymbol{B}}'x_i, \widetilde{\boldsymbol{\Sigma}}\right), i = 1, \dots, n$$

onde $\widetilde{\boldsymbol{B}}$ e $\widetilde{\boldsymbol{\Sigma}}$, são gerados aleatoriamente através das distribuições à posteriori de $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$, e geramos $\boldsymbol{V}$, tendo em conta que $\boldsymbol{v}_i = (w_{1i}, \dots, w_{mi})'$ serão distribuídos independentemente como

$$\boldsymbol{v}_i|_{\widehat{\boldsymbol{B}},\boldsymbol{S}} \sim N_m\left(\widehat{\boldsymbol{B}}'x_i, \boldsymbol{S}\right), i = 1, \dots, n$$

onde $\widehat{\boldsymbol{B}}$ e $\boldsymbol{S}$ são os estimadores usuais de $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$. Como forma de ilustrar, os dados que se tornarão públicos serão

$$\begin{bmatrix} w_{1,1} & \cdots & w_{1,m} & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,m} & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \text{ ou } \begin{bmatrix} v_{1,1} & \cdots & v_{1,m} & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & \cdots & v_{n,m} & x_{n,1} & \cdots & x_{n,p} \end{bmatrix},$$

sejam estes dados gerados pelo método FPPS ou pelo método Plug-in, respetivamente.

No caso da imputação múltipla, ou seja, se se pretender a divulgação de M versões da matriz $\boldsymbol{Y}$, repete-se o processor M vezes, dando origem a $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M$, por FPPS, considerando os valores de $\widetilde{\boldsymbol{B}}$ e $\widetilde{\boldsymbol{\Sigma}}$ fixos, e $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_M$, por Plug-in, ou seja, divulgando, para o caso FPPS

$$[\boldsymbol{W}_1, \boldsymbol{X}], \ldots, [\boldsymbol{W}_M, \boldsymbol{X}]$$

e, para o caso Plug-in,

$$[\boldsymbol{V}_1, \boldsymbol{X}], \ldots, [\boldsymbol{V}_M, \boldsymbol{X}].$$

Para gerar múltiplos conjuntos de dados parcialmente sintéticos, a diferença entre os métodos FPPS e o PPS reside na imputação das estimativas de $\widetilde{\boldsymbol{B}}$ e $\widetilde{\boldsymbol{\Sigma}}$ imputadas no modelo, fixa-se os mesmos valores ao longo do método FPPS e geram-se M valores diferentes para cada um dos M conjunto de dados gerados por PPS.

Distribuições exatas dos dados parcialmente sintéticos por imputação única

Em Moura *et al*. (2017a; 2017b; 2018) é possível aceder à distribuição exata das versões sintetizadas por PPS e Plug-in, bem como a distribuição exata dos estimadores dos parâmetros desconhecidos $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$, para cada um dos métodos. No que diz respeito a estes parâmetros, segundo o ponto de vista de um analista, os estimadores são de certa forma similares aos estimadores usuais $\widehat{\boldsymbol{B}} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{Y}'$ e $\boldsymbol{S} = \frac{1}{n-p}(\boldsymbol{Y} - \widehat{\boldsymbol{B}}'\boldsymbol{X})(\boldsymbol{Y} - \widehat{\boldsymbol{B}}'\boldsymbol{X})'$ dos dados originais, tornando a obtenção das respetivas estimativas num processo bastante simples e familiar.
No caso FPPS, os estimadores dos parâmetros serão

$$\boldsymbol{B}^{\#} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{W}' \text{ e } \boldsymbol{S}^{\#} = \frac{1}{n-p}(\boldsymbol{W} - \boldsymbol{B}^{\#\prime}\boldsymbol{X})(\boldsymbol{W} - \boldsymbol{B}^{\#\prime}\boldsymbol{X})'$$

e, no caso Plug-in, os estimadores serão

$$\boldsymbol{B}^{*} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{V}' \text{ e } \boldsymbol{S}^{*} = \frac{1}{n-p}(\boldsymbol{V} - \boldsymbol{B}^{*\prime}\boldsymbol{X})(\boldsymbol{V} - \boldsymbol{B}^{*\prime}\boldsymbol{X})'.$$

Tanto $\boldsymbol{B}^{\#}$ como $\boldsymbol{B}^{*}$ são estimadores de máxima verosimilhança centrados de $\boldsymbol{B}$, e $\boldsymbol{S}^{\#}$ e $\boldsymbol{S}^{*}$ são estimadores centrados de $\boldsymbol{\Sigma}$, ou seja, os valores esperados destes estimadores são os mesmos valores esperados dos estimadores dos dados originais:

$$E(\boldsymbol{B}^{\#}) = E(\boldsymbol{B}^{*}) = \boldsymbol{B};$$
$$E(\boldsymbol{S}^{\#}) = E(\boldsymbol{S}^{*}) = \boldsymbol{\Sigma}.$$

Considerando as variáveis aleatórias

$$\boldsymbol{T}^{\#} = \frac{|(\boldsymbol{B}^{\#} - \boldsymbol{B})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{B}^{\#} - \boldsymbol{B})|}{|(n-p)\boldsymbol{S}^{\#}|}, \tag{2}$$

para o caso FPPS, e

$$\boldsymbol{T}^{*} = \frac{|(\boldsymbol{B}^{*} - \boldsymbol{B})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{B}^{*} - \boldsymbol{B})|}{|(n-p)\boldsymbol{S}^{*}|}, \tag{3}$$

para o caso Plug-in, é possível efetuar análises inferenciais à matriz $\boldsymbol{B}$, tendo em conta que a distribuição de (2) é estocasticamente equivalente a

$$\left\{\prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i\right\} |2\boldsymbol{I}_m + \boldsymbol{\Omega}| \tag{4}$$

e que a distribuição de (3) é estocasticamente equivalente a

$$\left\{\prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i\right\} |(n-p)\boldsymbol{\Psi}^{-1} + \mathbf{I}_m| \tag{5}$$

onde $F_i \sim F_{p-i+1,n-p-i+1}$ são variáveis independentes entre si e independentes de $\boldsymbol{\Omega}$ e $\boldsymbol{\Psi}$, cuja distribuição de $\boldsymbol{\Omega}$ é equivalente à de $\boldsymbol{A}_1^{\frac{1}{2}}\boldsymbol{A}_2^{-1}\boldsymbol{A}_1^{\frac{1}{2}}$ com $\boldsymbol{A}_1 \sim W_m(\boldsymbol{I}_m, n+\alpha-p-m-1)$ e $\boldsymbol{A}_2 \sim W_m(\boldsymbol{I}_m, n-p)$ variáveis independentes (distribuições Wishart) e a de $\boldsymbol{\Psi}$ é $\mathrm{W}_m(\mathbf{I}_m, \mathrm{n}-\mathrm{p})$.
A partir do disposto acima, um analista pode construir distribuições empíricas de (2) e de (3) através de simulações de Monte Carlo e usá-las para efetuar, como por exemplo, o estudo da significância do

modelo, bem como testar uma combinação linear da matriz dos coeficientes de regressão (Moura, Klein, Coelho, & Sinha, 2017a; Moura, Sinha, & Coelho, 2017b; Moura, Klein, Zylstra, Coelho, & Sinha, 2018).

No caso de se estar perante um caso em que são disponibilizadas publicamente mais do que uma versão sintética, um dos procedimentos mais simples é recolher estimativas para os parâmetros tomando cada versão separadamente e depois utilizar a média destas para calcular uma estatística similar a (2) e (3) cujas distribuições apenas diferem nos graus de liberdade da distribuição F. O outro procedimento consiste em agrupar as múltiplas versões sintéticas numa só matriz

$$\begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{W}_2 & \boldsymbol{X} \\ \vdots & \vdots \\ \boldsymbol{W}_M & \boldsymbol{X} \end{bmatrix} \text{ ou } \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{X} \\ \boldsymbol{V}_2 & \boldsymbol{X} \\ \vdots & \vdots \\ \boldsymbol{V}_M & \boldsymbol{X} \end{bmatrix}$$

e proceder de forma similar ao procedimento exemplificado para a imputação única (Moura, Klein, Coelho, & Sinha, 2017a; Moura, Sinha, & Coelho, 2017b; Moura, Klein, Zylstra, Coelho, & Sinha, 2018).

Para o caso de imputação múltipla por PPS, os procedimentos são algo mais complexos e poderão ser consultados em (Moura, Sinha, & Coelho, 2017b).

Discussão das simulações

As simulações realizadas em Moura *et al.* (2017a; 2017b; 2018), demonstram que, em qualquer um dos casos, FPPS, PPS ou Plug-in, e em qualquer uma das situações, imputação única ou múltipla, os procedimentos disponibilizados exibiram precisões muito próximas de 0.95 quando estabelecido um nível de confiança de 0.95 ($\gamma = 0.05$), mesmo que as amostras apresentem dimensões reduzidas, como era previsível, visto terem por base distribuições exatas. Estes foram comparados com a precisão que se obteria, quando aplicável (apenas para casos de imputação múltipla), através dos procedimentos assintóticos de Reiter (2003) adaptados ao estudo de matrizes de parâmetros, verificando-se que esta adaptação só atingia a precisão pretendida para valores de n grandes.

É habitual, comparar os diferentes procedimentos medindo o "tamanho" das regiões de confiança recorrendo ao volume destas, no entanto, a região de confiança para a matriz dos coeficientes de regressão é na verdade sempre infinito, por consequência, considerou-se necessário propor uma outra medida, denominado *raio* (2017a; 2017b; 2018). Para o caso de imputação única quando os dados são gerados por FPPS, o *raio* será dado por

$$\Upsilon^{\#} = d^{\#}_{m,n,p,\alpha,\gamma} \times |(n-p)\boldsymbol{S}^{\#}|$$

e, quando gerados por Plug-in,

$$\Upsilon^{*} = d^{*}_{m,n,p,\gamma} \times |(n-p)\boldsymbol{S}^{*}|,$$

onde $d^{\#}_{m,n,p,\alpha,\gamma}$ e $d^{*}_{m,n,p,\gamma}$ serão os quantis obtidos a partir de (4) e (5) associados ao nível γ de confiança, para os casos FPPS e Plug-in, respetivamente. Para o caso de imputação múltipla, os *raios* são similares.

As simulações realizadas demonstraram os procedimentos exatos criados para analisar dados sintéticos gerados por FPPS apresentam *raios* maiores, sendo estes aproximadamente duas vezes e meia superior aos *raios* provenientes dos procedimentos para o método Plug-in. Esta avaliação dos procedimentos poderia levar o leitor a concluir que se deveria optar apenas em divulgar dados gerados por Plug-in por se obter regiões de confiança menores que aquelas provenientes do método FPPS havendo dessa forma um conjunto de dados com maior qualidade de informação, no entanto, importa não esquecer que para além da qualidade está também em jogo a proteção da privacidade que pode ser reduzida ao aumentarmos essa qualidade.

Posto isto, para aferir esse nível de confidencialidade e recorrendo a microdados de uso público respeitantes ao suplemento de março de 2000 do *Current Population Survey (CPS),* habitualmente usados neste contexto, foram geradas, repetidamente, múltiplas versões sintéticas da secção que se pressupõe ser sensível, através dos métodos FPPS, PPS e Plug-in, e foram calculados os valores respeitantes a três medidas que permitem estudar o nível de confidencialidade. Resumidamente, as três medidas usadas (Moura, Klein, Coelho, & Sinha, 2017a; Moura, Sinha, & Coelho, 2017b; Moura, Klein, Zylstra, Coelho, & Sinha, 2018) permitem observar, em primeiro lugar, qual a proximidade

global entre os dados sintéticos e os dados originais, em segundo, a proximidade entre os valores sintéticos e os originais por indivíduo e, por fim, a proximidade elementar entre cada um dos valores originais e o seu respetivo valor sintético.
Dos resultados obtidos a partir das três medidas é possível observar que o método Plug-in apresentou uma maior proximidade entre dados sintéticos e originais, ou seja, representando uma maior probabilidade de se pôr em risco a confidencialidade do indivíduo, quando comparado com o método PPS e o FPPS, sendo este último aquele que apresenta um maior nível de confidencialidade. Isto contrasta com a qualidade da informação disponível por cada um dos métodos, como foi visto anteriormente. Existe sempre uma relação inversa entre a qualidade da informação disponibilizada e o nível de proteção oferecida, sendo uma tarefa árdua decidir qual das duas se quer privilegiar.
No que se refere ao número de versões sintéticas a publicar, notou-se que à medida que se aumenta o número de elementos do conjunto de versões sintéticas que se tornarão públicas o risco de estar a revelar o que deveria ser protegido quase duplica. Este facto demonstra a importância de, em certas situações, ser exigido pelas instituições disponibilizar apenas uma versão sintética dos dados originais em vez de múltiplas versões.

Qualidade dos procedimentos em condições não ideais

Em termos práticos, existe sempre a possibilidade de o conjunto de dados original não satisfazer todas as condições do modelo RLM. Com esse intuito, em Moura *et al.* (Moura, Klein, Zylstra, Coelho, & Sinha, 2018), foram aplicados os mesmos métodos de geração sintética e procedimentos para análise dos dados, para o caso Plug-in sob o modelo RLM, a dados originais onde a matriz $\boldsymbol{Y}$ não era normalmente distribuída, sendo provenientes, na verdade, de uma população com distribuição do tipo t-Student multivariada ou do tipo skew normal. A precisão calculada através de simulações idênticas às anteriores apresentou-se bastante próxima do valor 0.95 estipulado, registando-se um aumento dessa proximidade à medida que se aumenta a dimensão da amostra. Desta forma, os resultados levam-nos a concluir que os procedimentos apresentados são robustos, demonstrando a qualidade dos procedimentos.
Quando um analista pretende fazer um estudo aos dados disponíveis, ao nível da regressão, este não se limita a estudar a regressão das variáveis que a instituição considerou sensíveis nas variáveis consideradas não-sensíveis, analisando a correlação entre qualquer combinação de variáveis. Por esse motivo, também se considerou oito casos diferentes de regressão com diferentes escolhas de variáveis como variáveis resposta e explicativas. Analisando os resultados obtidos nos oito diferentes casos, observou-se que a estimativa obtida do conjunto de dados parcialmente sintético está sempre muito próxima da estimativa proveniente dos dados originais e que a precisão, especialmente no caso de imputação única, mantém-se muito próxima do valor 0.95 estipulado usando os procedimentos exatos desenvolvidos.
Prevê-se que se poderá obter resultados similares se se fosse aplicados os métodos FPPS e PPS para a geração de dados sintéticos e concluir-se-ia da mesma forma a qualidade dos procedimentos para a sua análise em condições não-ideiais.

Conclusão

Prevendo o aumento exponencial de informação reservada nas instituições mundiais e o aumento da requisição de acesso a esta, a disponibilização de processos de análise dos dados, advindo quer de dados sintéticos gerados por imputação múltipla ou advindo pela geração por imputação única, é de extrema importância.
Os procedimentos agora disponíveis permitem a análise estatística de dados gerados por imputação única sob o modelo RLM e, por se basearem em distribuições exatas, a sua precisão é também exata mesmo perante amostras de dimensão pequena.
O raio de ação destes procedimentos não se limita ao estudo dos dados sob o modelo RLM, no panorama da geração sintética de dados por Plug-in, estes procedimentos podem também ser usados em conjuntos de dados cuja população sigam uma outra distribuição, prevendo-se o mesmo para os casos FPPS e PPS. A sua aplicação não é estática a uma escolha fixa de variáveis resposta e variáveis

explicativas podendo ser aplicada a qualquer uma combinação de modelos de regressão sem grande perda de precisão.

Perspetiva-se, facilitar o trabalho do analista disponibilizando no futuro distribuições assintóticas das distribuições exatas da variável aleatória usada para testar a matriz dos coeficientes de regressão, para que não se esteja a recorrer a distribuições empíricas destas, bem como a procedimentos para analisar a matriz de covariância do modelo.

Referências

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation (Vol. 201)*. Springer Science & Business Media.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical.

Kinney, S., Reiter, J., & Miranda, J. (2014). Improving the Synthetic Longitudinal Business Database. US Census Bureau. *Center for Economic Studies*, 12-14.

Kinney, S., Reiter, J., Reznek, A. P., Miranda, J., Jarmin, S., R., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review 79.3*, 362-384.

Klein, M., & Sinha, B. (2015). Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models. *Sankhya B 77.2*, 293-311.

Klein, M., & Sinha, B. (2015). Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model. *Thailand Statistician 13.1*, 33-47.

Klein, M., & Sinha, B. (2015). Likelihood-based inference for singly and multiply imputed synthetic data under a normal model. *Statistics & Probability Letters 105*, 168-175.

Klein, M., & Sinha, B. (2016). Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models. *Journal of Privacy and Confidentiality 7.1*, 43-98.

Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*, 407-426.

Moura, R., Klein, M., Coelho, C. A., & Sinha, B. (2017). Inference for Multivariate Regression Model based on synthetic data generated under Fixed-Posterior Predictive Sampling: comparison with Plug-in Sampling. *Revstat*, 155-186.

Moura, R., Klein, M., Zylstra, J., Coelho, C. A., & Sinha, B. (2018). *Inference for multivariate regression model based on synthetic data generated using plug-in sampling*. Washington USA: US Census Bureau.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19*, 1-16.

REGULATION (EC) No 223/2009. (2009). *Official Journal of the European Union, 87*, 164-173.

Reiter, J. (2003). Inference for Partially Synthetic Public Use Microdata Sets. *Survey Methodology 29*, 181-188.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics 9*, 461-468.

Testes sobre a estrutura de matrizes de covariância

Filipe J. Marques, *fjm@fct.unl.pt*
Carlos A. Coelho, *cmac@fct.unl.pt*

Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT NOVA)
Centro de Matemática e Aplicações (CMA)

1. Introdução

A estrutura da matriz de covariância pode revelar caraterísticas importantes de uma determinada distribuição ou, no caso amostral, da estrutura dos dados. Vários modelos nas mais diversas áreas de investigação assumem como pressupostos estruturas para a matriz de covariância dos erros que podem ser simples ou ter alguma complexidade. Por este motivo, é importante ter ferramentas que nos permitam realizar, com a precisão adequada, testes sobre estruturas de matrizes de covariância. Se considerarmos uma população $N_p(\underline{\mu}, \Sigma)$, temos como alguns exemplos de estruturas mais simples:

1. Independência: $\Sigma = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$
2. Esférica: $\Sigma = \sigma^2 I_p$
3. Igualdade de variâncias e de covariâncias: $\Sigma = \sigma^2 \left((1-\rho)I_p + \rho E_{pp}\right)$ (onde $-\frac{1}{p-1} < \rho < 1$ e E_{pp} é uma matriz de ordem p com todas as entradas iguais a 1)
4. Circular: $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$ (para $p = 6$)
5. Indepêndencia de grupos de variáveis: $\Sigma = bdiag(\Sigma_{11}, \dots, \Sigma_{kk}, \dots, \Sigma_{mm})$, onde Σ_{kk} é uma matriz de ordem p_k, com $p_1 + \cdots + p_k + \cdots + p_m = p$.

Claro que a estrutura de Σ pode-se tornar mais complexa por composição das estruturas acima. O interesse no estudo destas estruturas ditas mais complexas é hoje em dia potenciado pela também complexidade de novos modelos, nomeadamente modelos mistos. Veremos mais à frente como podem ser feitos testes a este tipo de estruturas.

Para realizar testes, quer a estruturas mais simples quer a estruturas complexas das matrizes de covariância, é possível deduzir as estatísticas de razão de verosimilhanças, de forma mais ou menos trabalhosa, contudo a questão coloca-se nas distribuições exatas destas estatísticas, as quais são normalmente de estrutura demasiado elaborada, o que torna difícil a sua implementação computacional e por isso pouco úteis na prática. Em geral, as estatísticas de razão de verosimilhanças, usadas em testes sobre a estrutura de matrizes de covariância, têm uma distribuição igual à do produto de variáveis aleatórias independentes com distribuição Beta. Existe uma vasta literatura sobre este tópico

onde constam diferentes representações para esta distribuição como são os casos das representações em série (Tang e Gupta, 1984; Moschopoulos, 1986), das representações através de funções G de Meijer (Meijer, 1946; Nagar et al.,1985) ou funções H de Fox (Fox, 1961; Springer, 1979; Carter e Springer, 1977), entre outras. Contudo, hoje em dia, com toda capacidade computacional existente, ainda pode ser um problema obter quantis ou *p-values* precisos para estas distribuições. Em Coelho e Alberto (2012) os autores apresentam uma revisão de literatura muito detalhada sobre produto de variáveis aleatórias independentes com distribuição Beta. Neste artigo os autores desenvolvem distribuições quase-exatas precisas e computacionalmente implementáveis para produto de variáveis aleatórias independentes com distribuição Beta. No que diz respeito a testes sobre a estrutura de matrizes de covariância é bem conhecido que a distribuição do logaritmo da estatística de razão de verosimilhanças pode ser aproximada por um qui-quadrado, eventualmente multiplicado por um fator de correção. Estas aproximações podem ser melhoradas se considerarmos as aproximações obtidas por Box (1949) que são usualmente apresentadas como misturas de duas distribuições Gama. Contudo, o desempenho destas aproximações é limitado, principalmente se considerarmos cenários extremos como aqueles em que temos amostras de dimensão reduzida e/ou um número elevado de variáveis. Uma alternativa diferente são as aproximações ponto-de-sela (Daniels, 1954; Booth *et. al*, 1995). Contudo estas têm a desvantagem de não produzirem uma expressão nem para a função densidade nem para a função distribuição, mas apenas aproximações para pontos específicos, e a literatura mostra que estas podem ser francamente melhoradas. Mais recentemente, surgiram as aproximações quase-exatas (Coelho, 2004) que têm sido bastante utilizadas para aproximar a distribuição de estatísticas de razão de verosimilhanças utilizadas para realizar testes sobre a estrutura de matrizes de covariância e também em problemas relacionados com a distribuição de produtos, somas e combinações lineares de variáveis aleatórias. As aproximações quase-exatas podem ser utilizadas em estruturas simples como as já apresentadas ou em estruturas mais complexas. O procedimento para o desenvolvimento destas aproximações será apresentado em detalhe na secção seguinte.

2. Testes sobre matrizes de covariância com estruturas complexas

Muitas estruturas complexas podem ser interpretadas como composições de testes mais simples. Por exemplo, o teste de esfericidade apresentado anteriormente pode ser visto como a composição de dois testes; o teste à independência de várias variáveis e o teste de igualdade de variâncias, aliás em Anderson (2003) o autor utiliza esta mesma estratégia para obter a estatística de razão de verosimilhanças do teste. Em Coelho e Marques (2009) os autores mostram com é possível desenvolver distribuições quase-exatas para estruturas ditas complexas. A ideia geral é a seguinte: suponhamos que pretendemos testar uma determinada estrutura complexa e especificada na hipótese nula H_0 versus a correspondente hipótese alternativa H_1, a ideia fundamental é tentar decompor, de forma adequada, a hipótese nula inicial numa sequência de hipóteses nulas parciais. Suponhamos então que é possível fazer a decomposição de H_0 em m hipóteses nulas parciais, que podem ter que obedecer a uma determinada ordem, e cuja decomposição pode ser apresentada através da seguinte notação

$$H_0 \equiv H_{0m|1,\cdots,m-1} \circ \ldots \circ H_{02|1} \circ H_{01}$$

como referido em Coelho e Marques (2009) esta notação representa que testar H_0 é equivalente a testar sequencialmente as m hipóteses $H_{0j|1,\cdots,j-1}$ $(j = 1,\ldots,m)$, testando primeiro H_{01}, em seguida $H_{02|1}$, depois $H_{03|1,2}$, e assim sucessivamente, onde testar $H_{0j|1,\ldots,j-1}$ representa testar H_{0j} assumindo que as hipótese H_{01} até $H_{0,j-1}$ não são rejeitadas. Note-se que, de uma forma geral, fazendo uma decomposição adequada de H_0 tem-se, sob esta hipótese nula, que as estatísticas de razão de verosimilhanças $\Lambda_{j|1,\cdots,\mathrm{j}-1}$ usadas para testar as hipóteses parciais $H_{0j|1,\ldots,j-1}$ $(j = 1,\ldots,m)$ são independentes. Tendo por base esta decomposição a estatística de razão de verosimilhanças, Λ, usada para testar a hipótese nula global H_0 é dada por

$$\Lambda = \prod_{j=1}^{m} \Lambda_{j|1,\cdots,\mathrm{j}-1}.$$

Tendo em conta a independência das estatísticas $\Lambda_{j|1,\cdots,\mathrm{j}-1}$ sob H_0 podemos determinar a expressão do h-ésimo momento de Λ como o produto dos h-ésimos momentos das estatísticas $\Lambda_{j|1,\cdots,\mathrm{j}-1}$ ou seja

$$E[\Lambda^h] = \prod_{j=1}^{m} E[\Lambda_{j|1,\cdots,\mathrm{j}-1}^h].$$

A partir desta última expressão é possível obter a função caraterística da variável aleatória $W = -\log \Lambda$ da seguinte forma

$$\Phi_W(t) = E\left[\mathrm{e}^{\mathrm{i}tW}\right] = E\left[\Lambda^{-\mathrm{i}t}\right] = \prod_{j=1}^{m} E\left[\Lambda_{j|1,\cdots,\mathrm{j}-1}^{-\mathrm{i}t}\right] = \prod_{j=1}^{m} E\left[\mathrm{e}^{\mathrm{i}tW_{j|1,\cdots,\mathrm{j}-1}}\right] = \prod_{j=1}^{m} \Phi_{W_{j|1,\cdots,\mathrm{j}-1}}(t)\,, t \in \mathbb{R}$$

onde $\Phi_{W_{j|1,\cdots,\mathrm{j}-1}}(t)$ representa a função caraterística de $\mathrm{W}_{\mathrm{j}|1,\ldots,\mathrm{j}-1} = -\log \Lambda_{\mathrm{j}|1,\ldots,\mathrm{j}-1}, j = 1, \ldots, m$. A fatorização obtida deste modo para a função caraterística de W é o procedimento base para o desenvolvimento das aproximações quase-exatas para W e para Λ. O passo seguinte para a construção destas aproximações é obter uma nova fatorização da função caraterística de W de forma a que se aproximarmos um dos fatores por outra função característica possamos obter uma nova função caraterística à qual corresponda uma distribuição conhecida e fácil de utilizar na prática. Apresentamos na secção seguinte um exemplo deste procedimento.

3. Exemplo

Para ilustrar o procedimento descrito na secção anterior vamos apresentar sumariamente o teste estudado em (Marques e Coelho, 2015). Por uma questão de simplicidade vamos omitir algumas expressões podendo estas ser consultadas com detalhe na referência acima. Suponhamos então que, dada uma amostra extraída de uma população $N_p(\underline{\mu}, \Sigma)$ estamos interessados em testar a seguinte hipótese nula

$$H_0: \Sigma = \begin{pmatrix} \Sigma_{11} & 0 & 0 \\ 0 & \Sigma_{22} & 0 \\ 0 & 0 & \Sigma_{33} \end{pmatrix} \qquad (1)$$

isto é, pretende-se testar se a matriz de covariância Σ tem uma estrutura diagonal por blocos em que Σ_{11} é uma matriz de ordem $p_1$sem uma estrutura especifica, Σ_{22} , de ordem p_2, tem uma estrutura esférica, ou seja, $\Sigma_{22} = \sigma^2 I_{p_2}$ (Anderson, 2003; Marques e Coelho, 2008) e Σ_{33}, de ordem p_3 tem uma estrutura circular representada por Σ_C (Olkin e Press, 1969; Marques e Coelho, 2013) e onde $p = p_1 + p_2 + p_3$.

É importante referir que o pressuposto de normalidade, em alguns casos, poder ser estendido a outras distribuições, por exemplo em Anderson et al. (1986) os autores, para uma classe de distribuições elípticas, obtém as estatísticas de razão de verosimilhanças para alguns testes sobre estruturas de matrizes de covariância e referem que a distribuição é a mesma que a do caso Normal.

Considerando o procedimento apresentado na secção anterior, vamos decompor a hipótese nula em (1) em três hipóteses nulas parciais, a primeira utilizada para testar a independência dos três grupos de variáveis

$$H_{01}: \Sigma_{ij} = 0\,, \qquad i \neq j, \qquad i, j = 1, \ldots, 3 \qquad (2)$$

a segunda para testar a estrutura esférica do segundo bloco diagonal da matriz de covariância de ordem p_2

$$H_{02|1}: \Sigma_{22} = \sigma^2 I_{p_2} \text{ (assumindo que } H_{01} \text{ não é rejeitada)} \qquad (3)$$

e a terceira para testar a estrutura circular do terceiro bloco diagonal de ordem p_3

$$H_{03|1}: \Sigma_{33} = \Sigma_C \text{ (assumindo que } H_{01} \text{ não é rejeitada)}. \quad (4)$$

Assim, com base no Lema 10.3.1 apresentado em Anderson (2003), a estatística de razão de verosimilhanças, Λ, usada para testar H_0 em (1) é dada pelo produto das estatísticas de razão de verosimilhanças utilizadas para testar as hipóteses nulas parciais apresentadas em (2), (3) e (4). Pelo que, usando as expressões das estatísticas de teste utilizadas para testar H_{01}, $H_{02|1}$ e $H_{03|1}$designadas respetivamente por Λ_1, $\Lambda_{2|1}$e $\Lambda_{3|1}$ e dadas em Marques e Coelho (2015), Anderson (2003, sec. 9.2, 10.7) e Olkin e Press (1969, sec. 3.3) obtem-se

$$\Lambda = \Lambda_1 \times \Lambda_{2|1} \times \Lambda_{3|1}.$$

Pode encontrar todos os detalhes sobre a expressão de Λ na expressão (4) em Marques e Coelho (2015). Dada a independência das estatísticas Λ_1, $\Lambda_{2|1}$e $\Lambda_{3|1}$, sob H_0, a expressão do h-ésimo momento pode ser obtida como o produto das expressões dos h-ésimos momentos das estatísticas Λ_1, $\Lambda_{2|1}$e $\Lambda_{3|1}$, disponíveis em Marques e Coelho (2015), Anderson (2003, sec. 9.3, 10.7) e Olkin e Press (1969, sec. 3.3). Assim,

$$\mathrm{E}[\Lambda^h] = E[{\Lambda_1}^h] \times E[{\Lambda_{2|1}}^h] \times E[{\Lambda_{3|1}}^h].$$

Consideremos agora a variável aleatória $W = -\log \Lambda$, cuja função caraterística é dada por

$$\begin{aligned}\Phi_W(t) = E\left[\mathrm{e}^{\mathrm{i}tW}\right] = E\left[\Lambda^{-\mathrm{i}t}\right] = E[{\Lambda_1}^{-\mathrm{i}t}] \times E[{\Lambda_{2|1}}^{-\mathrm{i}t}] \times E[{\Lambda_{3|1}}^{-\mathrm{i}t}] \\ = \Phi_{W_1}(t) \times \Phi_{W_{2|1}}(t) \times \Phi_{W_{3|1}}(t)\end{aligned}$$

onde Φ_{W_1}, onde $\Phi_{W_{2|1}}$ e onde $\Phi_{W_{3|1}}$ são, respetivamente, as funções caraterísticas das variáveis aleatórias $W_1 = -\log \Lambda_1$, $W_{2|1} = -\log \Lambda_{2|1}$ e $W_{3|1} = -\log \Lambda_{3|1}$. Como já referido, o objetivo agora é encontrar uma fatorização de Φ_W de forma que, mantendo a maior parte intacta, e aproximando um dos fatores por outra função característica possamos obter uma nova função caraterística à qual corresponda uma distribuição conhecida e manejável. Em Marques e Coelho (2015) os autores mostram que é possível escrever Φ_W da seguinte forma:

$$\Phi_W(t) = \Phi_{W_1^*}(t) \times \Phi_{W_2^*}(t) \quad (6)$$

onde $\Phi_{W_1^*}$ é a função característica da soma de um dado número de variáveis aleatórias independentes com distribuição Gama com parâmetros de forma inteiros, o que corresponde a uma distribuição designada por Gama Inteira Generalizada (GIG) obtida em Coelho (1998) e $\Phi_{W_2^*}$ é a função caraterística da soma, de um dado número, de variáveis aleatórias com distribuição Logbeta (note-se que se X tem distribuição Beta de parâmetros a e b então dizemos que $-\log X$ tem uma distribuição Logbeta com os mesmos parâmetros). Usando os resultados em Tricomi e Erdélyi (1951) sabemos que uma simples distribuição Logbeta pode ser aproximada por uma mistura infinita de distribuições Gama, pelo que a abordagem seguida passa por aproximar a função característica $\Phi_{W_2^*}$ em (6) por uma mistura de distribuições Gama cuja função caraterística é dada por

$$\Phi_{\widetilde{W}_2}(t) = \sum_{j=0}^{m} \pi_j \, \lambda^{r+j} (\lambda - \mathrm{i}t)^{-(r+j)} \quad (7)$$

de forma a que $\widetilde{W}_2$ tenha os mesmos m primeiros momentos de W_2^*. Obtem-se assim como função caraterística aproximada de Φ_W

$$\Phi_W(t) \approx \Phi_{NE}(t) = \Phi_{W_1^*}(t) \times \Phi_{\widetilde{W}_2}(t).$$

No que se segue designaremos a função caraterística Φ_{NE} como função caraterística quase-exata. Na expressão de $\Phi_{\widetilde{W}_2}$ em (7) o parâmetro λ é a taxa de uma mistura de duas distribuições Gama que acerta os primeiros quatro momentos de W_2^* e r é igual à soma dos segundos parâmetros das distribuições Logbeta que caraterizam a distribuição de $\Phi_{W_2^*}$ em (6) para mais detalhes veja-se Coelho et al. (2010).

Fixados os parâmetros λ e r os pesos π_j são determinados de forma a que $\widetilde{W}_2$ tenha os mesmos m primeiros momentos de W_2^*, ou seja, são as soluções do seguinte sistema de equações

$$\frac{\partial^h}{\partial t^h}\Phi_{W_2^*}(t)\bigg|_{t=0} = \frac{\partial^h}{\partial t^h}\Phi_{\widetilde{W}_2}(t)\bigg|_{t=0}, \qquad h = 1, \dots, m, \qquad \text{com } \pi_m = 1 - \sum_{j=0}^{m-1} \pi_j.$$

Note-se que este sistema de equações é de resolução simples com um software de cálculo matemático.

Finalmente, seguindo esta construção, obtemos como função caraterística quase-exata

$$\Phi_{NE}(t) = \sum_{j=0}^{m} \pi_j \left\{ \Phi_{W_1^*}(t)\, \lambda^{r+j} (\lambda - \mathrm{i}t)^{-(r+j)} \right\}. \tag{8}$$

Para um valor de j fixo a expressão $\Phi_{W_1^*}(t)\, \lambda^{r+j} (\lambda - \mathrm{i}t)^{-(r+j)}$ corresponde à função característica da soma de duas variáveis aleatórias independentes; W_1^* com distribuição GIG e uma variável aleatória com distribuição Gama com taxa λ e parâmetro de forma $r+j$. Se r for um número inteiro a soma destas duas variáveis aleatórias continua a ter uma distribuição GIG, se por outro lado r não for inteiro a distribuição da soma é uma Gama Quase-Inteira Generalizada (GQIG) obtida em Coelho (2004). Pelo que a distribuição correspondente à função caraterística Φ_{NE} em (8) é uma mistura de distribuições GIG ou uma mistura de distribuições GQIG consoante r seja inteiro ou não.

Em geral, as aproximações obtidas através deste processo apresentam elevado grau de precisão e são assimptóticas não só relativamente ao tamanho da amostra mas também a outros parâmetros envolvidos, como por exemplo o número de variáveis. Para avaliar as qualidade destas aproximações, em Marques e Coelho (2015), os autores utilizam uma medida de proximidade dada por

$$\Delta = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \frac{\Phi_W(t) - \Phi_{NE}(t)}{t} \right| \, \mathrm{d}t \,. \tag{8}$$

Esta medida, baseada nas funções características exata e aproximada, fornece um valor numérico para o limite superior da distância entre a função distribuição exata e a aproximada. Podem observar-se, a partir da Tabela 1 em Marques e Coelho (2015), os valores da medida em diferente cenários. Estes valores ilustram a qualidade das aproximações e também as suas propriedades assimptóticas.

Referências

Anderson, T. W. (2003) - *An Introduction to Multivariate Statistical Analysis*. 3rd ed., J. Wiley & Sons, New York.

Anderson, T., Fang, K., Hsu, H. (1986) - Maximum-Likelihood Estimates and Likelihood-Ratio. Criteria for Multivariate Elliptically Contoured Distributions. *The Canadian Journal of Statistics*,14, 55-59.

Booth, J. G., Butler, R. W., Huzurbazar, S., Wood, A. T. A. (1995) - Saddlepoint approximations for p-values of some tests of covariance matrices. *Journal of Statistical Computation and Simulation*, 53, 165-180.

Box, G. E. P. (1949) - A general distribution theory for a class of likelihood criteria. *Biometrika*, 36, 317–346.

Carter, B. D., Springer, M. D. (1977) - The distribution of products, quotients and powers of independent H-function variates. *SIAM J. Appl. Math.* 33, 542-558.

Coelho, C. A. (1998) - The Generalized Integer Gamma Distribution - A Basis for Distributions in Multivariate Statistics. *Journal of Multivariate Analysis*, 64, 86–102.

Coelho, C. A. (2004) - The Generalized Near-Integer Gamma Distribution: A Basis for 'Near-Exact' Approximations to the Distribution of Statistics which are the Product of an Odd Number of Independent Beta Random Variables. *Journal of Multivariate Analysis*, 89, 191-218.

Coelho, C. A., Arnold, B. C., Marques, F. J. (2010) - Near-exact distributions for certain likelihood ratio test statistics. *Journal of Statistical Theory and Practice* 4, 711-725.

Coelho, C. A., Alberto, R. P. (2012) - On the Distribution of the Product of Independent Beta Random Variables - Applications. *Technical Report, CMA*, 12.

Daniels, H. E. (1954) - Saddlepoint Approximations in Statistics. *Ann. Math. Statist.*, 25, 631-650.

Fox, C. (1961) - The G and H functions as symmetrical kernels. *Trans. Amer. Math. Soc.*, 98, 395-429

Marques, F. J., Coelho, C. A. (2008) - Near-exact distributions for the sphericity likelihood ratio test statistic. *Journal of Statistical Planning and Inference*, 138, 726-741.

Marques, F. J., Coelho, C. A. (2015) - Testing elaborate block-structures in covariance matrices by splitting the null hypothesis - an overview. *Proceedings of the 60th ISI World Statistics Congress, 26-31 July 2015, Rio de Janeiro, Brazil*, 1-6.

Meijer, C. S. (1946) - On the G-function I–VIII. *Proc. Koninklijk Nederlandse Akademie van Weteenschappen* 49, 227-237, 344-356, 457-469, 632-641, 765-772, 936-943, 1063-1072, 1165-1175.

Moschopoulos, P. G. (1986) - New Representations for the Distribution Function of a Class of Likelihood Ratio Criteria. *Journal of Statistical Research*, 20, 13-20.

Nagar, D. K., Jain S. K., Gupta A. K. (1985) - Distribution of LRC for testing sphericity of a complex multivariate Gaussian model. *Internat. J. Math. & Mathematical Sci.*, 8, 555–562.

Springer, M. D. (1979) - *The Algebra of Random Variables*. New York: J. Wiley & Sons.

Tang, J., Gupta, A. K. (1984) - On the distribution of the product of independent beta random variables. *Statistics & Probability Letters*, 2, 165-168.

Tricomi, F. G., Erdélyi, A. (1951) - The asymptotic expansion of a ratio of Gamma functions. *Pacific Journal of Mathematics* 1, 133-142.

Big Outlier(s)

Fernando Rosado, *fernando.rosado@fc.ul.pt*

DEIO, Faculdade de Ciências
Universidade de Lisboa

Introdução

Como nota prévia e a jeito de justificação, registo que nesta série do *Boletim* SPE, iniciada em 2006, é esta a primeira vez que participo como autor, sem prejuízo de alguma pequena abordagem teórica que tenha explanado em Editoriais. O *Boletim SPE* em cada edição elege um *Tema* Central. A sequência com todos os temas centrais pode ver-se, por exemplo na mais recente edição, no *Boletim SPE* outono de 2017, p. 64. Para cada um dos temas selecionados, como Editor, tenho contactado estatísticos seniores que, como "co-Editores", ajudam a estabelecer e construir uma lista dos autores convidados para incluir na referida secção. Assim, em todas as edições do Boletim *SPE,* ficamos com a devida atualização científica da respetiva área temática – um ponto de situação, divulgação à comunidade e perspetivas. Foi o que aconteceu, relativamente ao presente Tema Central[1]

O estudo da Estatística Multivariada desperta dois grandes subtemas diretamente relacionados com a dimensão e a dimensionalidade dos dados – este mais teórico do que aquele, embora ambos igualmente importantes na construção dos resultados. No entanto, o estudo da dimensão que invoca diretamente o volume da informação e dos dados estatísticos é, atualmente, mais importante do que aqueloutro estudo da dimensionalidade que investiga a verdadeira dimensão do espaço onde os dados foram gerados e o menor número de variáveis que podem garantir um estudo prático decisivo a partir desses dados estatísticos. É sobre aquele que nos vamos debruçar.

Este texto insere-se, também, em *Uma Perspetiva no século XXI* e então é, acima de tudo, um olhar para o futuro.

A temática agora abordada vem na sequência de duas edições do *Boletim SPE* que se debruçaram sobre *O Tema Central da Estatística*. Permito-me sugerir uma leitura revisitada e cuidada desses textos onde os diversos autores, juniores e seniores, apaixonadamente, registaram excelentes, indeléveis reflexões científicas e profissionais e que podemos situar mesmo para além da Estatística. São de uma riqueza única que "apetece resumir". De um modo simples e, por consequência, (seguramente) enviesado arrisco (apenas) sequenciar títulos (mais) significativos também pela "estranheza" das, muito oportunas, palavras utilizadas[2]:

---

[1] Mas, desta vez, com um detalhe acrescido: um dos "co-editores" com enorme gentileza, na mensagem de resposta ao convite para ajudar a construir "a lista dos autores" referiu que eu próprio deveria ser incluído. Respondi que o meu estatuto de "aposentado", enfim, já me afasta "do centro da investigação" e isso limita a iniciativa e o eventual interesse de umas modestas linhas temáticas sobre Estatística Multivariada. De facto, foi resposta de "pouca dura"; porque, apesar de ser há quase 30 anos, foi nesse domínio e numa época pioneira em Portugal que tive a oportunidade científica de alguma intervenção na área agora abordada – a criação de uma nova disciplina de licenciatura, *Análise de Dados Multivariados*, a que se seguiu uma também pioneira iniciação ao *Estudo Estatístico de Outliers,* também Multivariados. A junção destes dois temas e o contexto atual, como se verá, alteraram a resposta inicial. Assim, "a motivação pela investigação científica em Portugal" e a minha condição de "Professor Aposentado com Acordo de Cooperação" com a Universidade de Lisboa fez-me repensar e aceitar o "convite". Esta função ativa "fez despertar" uma nova resposta que conduziu ao presente texto; com a modesta intenção de testemunho científico, com alguma transmissão de saber de experiência feito além de, um óbvio, incentivo à investigação da temática.

[2] Até parece combinado mas, como editor, posso assegurar que não foi. O acaso, diz-se – a única coisa que não acontece por acaso – assim quis manifestar mais uma das suas apelativas intervenções.

Data Science um desafio para os estatísticos?
Reflexões estatísticas
O Futuro da Estatística
Data Science, Big Data e um novo olhar sobre a Estatística
Estatística – "Espelho meu, espelho meu, que futuro terei eu?
Novo olhar sobre a Estatística, imaginar o mundo
A Revolução dos Dados
A tirania dos jargões
Desafios da Estatística para o século XXI
A minha utopia sobre o Tema Central da Estatística.

Os referidos, são textos memoriais do ponto de vista de reflexões na, e da, Ciência Estatística.

Os títulos anteriores também constituem uma acrescida motivação para que, modestamente, me inclua nesta edição como autor. Na realidade trata-se do futuro da Estatística e, mais ainda, do Estatístico. Perante isto, no enquadramento, estas linhas pouco ou nada acrescentam. No entanto perante o novo desafio que envolve "o multivariado" algumas notas breves com (também) alguma história desejo acrescentar. Uma justificação!

Uma evolução no domínio científico – do ***Data Analysis*** ao ***Big Data***

O *Boletim SPE*, ao longo das suas edições mas em especial nas mais recentes, tem versado sobre os grandes temas de investigação nos diversos domínios da Estatística. Pela generalidade que pressupõe e também pela atualidade dos grandes assuntos que nela se incluem, a Estatística Multivariada é, seguramente, uma área muito apelativa e onde os maiores desafios são colocados, como veremos.

Nos Editoriais das edições outono de 2016 e de 2017 referi um pouco daquele que pode ser um olhar sobre esses desafios. Nestas linhas, noutra vertente, tenciono aprofundar um pouco.

Para um melhor enquadramento e também para se poder concluir do enorme avanço que se tem verificado na *Análise de Dados Estatísticos*, iniciamos com (um pouco) a sua história.

Data Analysis e o seu futuro promissor foi assunto criado há mais de 50 anos por Tukey (1962) a que se seguiram uma infinidade de livros e artigos científicos. Simples e apelativo, de modo rápido, tudo começou a avançar. O grande motor científico, na realidade, era a velocidade e a capacidade de cálculo apoiada nas máquinas recentemente criadas – os computadores, que evoluíam rapidamente. Uma (r)evolução perante os métodos científicos tradicionais.

Mas, passados todos esses anos é importante contrapor: uma evolução, um Avanço ou uma Continuidade (científica)? Avanço em que direção? "Tudo" passou a girar à volta dos dados. Evolução no domínio científico, não necessariamente na Ciência Estatística de onde, às vezes, parece que algumas áreas estão a ficar de fora: Já se desligaram? Assim, mais uma vez e como sempre, surge a dicotomia entre a Investigação Fundamental e a Investigação(?) Aplicada. Qual o benefício desta em proveito daquela? Certamente que muito fraco! Nos primeiros 20 anos, nos fóruns internacionais a questão corrente era "pró ou contra" e às vezes mais radical: O que fazem os Analistas de Dados? Na década de oitenta assistiu-se a uma "aceitação biunívoca" com alguma reserva pelos "mais teóricos"[3].

A *Análise Multivariada*, como sabemos, estuda dados estatísticos contendo observações em duas ou mais variáveis medidas[4] num conjunto de objetos.

A *Estatística Multivariada*, por sua vez, iniciou-se nesse mesmo ponto de partida científico e avançou no domínio das suas diversas especificidades – umas mais teóricas e outras de índole mais prática que, genericamente, podemos agrupar na *Análise de Dados Multivariados*. Do ponto de vista teórico, mesmo passados quase quarenta anos, Mardia *et al* (1979) mantém-se atual[5] o que pode

[3] No início da década de 1980, dois encontros que testemunhei, em Hong-Kong e em Barcelona, foram palco de aceitação mútua, de participação e de início de discussão científica por parte "dos grandes nomes" que até aí se recusavam.

[4] O avanço científico, registe-se, também se tem concretizado no número, cada vez maior, de variáveis em estudo e resultante (apenas) das capacidades tecnológicas de cálculo quer ao nível de hardware quer de software, Estas, possivelmente, podem ser o *mobile* do *boom* que gerou e conduziu ao *Big Data*.

[5] Ao longo do tempo "apenas" têm sido reproduzidas reimpressões do original o que avaliza a excelência da obra clássica fundamental. Este é um exemplo, entre outros, de livros teóricos basilares para a investigação fundamental que, ela sim, apoia e é o suporte da investigação aplicada. Outras obras similares podemos acrescentar invocando pioneiros como M. S. Bartlett, M. G. Kendall, R. A. Fisher, P. C. Mahalanobis ou C. R. Rao.

significar que, desde logo no início se atingiu um "conhecimento total". Do ponto de vista prático, aí sim, muito se tem avançado e por diversos caminhos desde a pioneira *Análise de Dados*[6]. No entanto, como noutras áreas, a bibliografia fundamental teórica da Estatística Multivariada mantém atualidade, mesmo passados dezenas de anos sobre a sua edição; um garante da excelência, por um lado, mas também revelador de um valor estatístico que o tempo e o avanço científico torna difícil de superar – os patamares atingidos, também na Ciência Estatística ficam mais altos, o que os torna mais difícil de superar; mas não impossível! Estas reflexões foram já abordadas, por diversos autores, em Rosado (2005).

Mas, o caminho iniciado pela Análise de Dados foi, ao mesmo tempo, percorrido pelas mais diversas áreas até hoje onde, em enorme competição científica, chegámos à *Data Science* – A Ciência de Dados[7]. E aqui, um pouco nebulosos ainda, surgem os mais diversos "conceitos" para os quais basta ser inovadores para se afirmarem; mesmo que careçam de suporte científico, na maior parte das vezes. A era digital afirma-se! E, como sempre, "gera crise". Mas a realidade científica já evolui em *Machine Learning, Data Science,* ou *Big Data*[8]. Novos desafios, mas que nada trazem de novo.

Big Outlier(s)

Em 1978, Barnett e Lewis publicaram a primeira edição de *Outliers in Statistical Data* – livro de base para o estudo de *outliers* em dados estatísticos tanto do ponto de vista teórico como prático. Nesta obra fundamental foi, pela primeira vez, agregada e sistematicamente organizada toda a vasta literatura.

Em 1994 foi publicada a terceira e última edição e nela foram incluídas novas abordagens para dados univariados e multivariados, apresentando ainda tópicos especiais nos métodos bayesianos e em sucessões cronológicas com os aditivos e os inovadores.

As "observações difíceis" de uma amostra sempre desafiaram os estatísticos. O conceito de *outlier* tem fascinado (em especial) os cientistas que numa primeira abordagem querem interpretar os dados.

Na época pioneira, o registo da informação, ainda com mais ênfase permitia admitir como erros todas as observações que ao experimentador parecessem mal vindas. E as reacções foram desde os seguidores da "incondicional inclusão" – como admitem Barnett e Lewis na primeira edição da obra acima referenciada – porque "nunca devemos violar a santidade dos dados" atrevendo-nos a julgar as suas propriedades até aqueles que sempre usam "na dúvida deita-se fora" como regra prática.

Em 1976, Barnett publicou "*The Ordering of Multivariate Data*"[9]

Numa perspectiva actual os pontos de vista são mais sofisticados. A teoria estatística dos *outliers* já possui diversas metodologias de tratamento de observações discordantes ou contaminantes; têm sido propostos modelos de discordância que permitem explicar a geração dos dados; os procedimentos robustos têm tido bastante avanço (cf. Barnett and Lewis (1994)). Em Rosado (2006), numa perspetiva de século XXI desenvolve-se uma base teórica e prática para o estudo de observações discordantes e muito em especial sobre os métodos e modelos de discordância; também para as questões de redução de dimensionalidade.

---

[6] Em Rosado (1991), apresentei o Programa, Conteúdos e Métodos de Ensino Teórico e Prática da disciplina *Análise de Dados Multivariados (ADM)*, em provas de agregação na Universidade de Lisboa. Em ADM para além do uso dos avanços computacionais à época também se insistia bastante na componente teórica quer na Estatística Descritiva Multivariada quer nas Técnicas de Redução de Dimensionalidade.

[7] Pela generalidade, a Ciência dos Dados já não é simplesmente uma área exclusiva dos Estatísticos mas "uma grande competição" onde eles, pela excelência, se têm de afirmar. Desde *Data Science* até *Big Data (ou Big Outlier) t*odos estes novos termos merecem ser analisados (e introduzidos?) no Glossário Estatístico da SPE!

[8] A facilidade de divulgação é inversa do rigor que nela se deve exigir. Muito se diz sobre estas novas terminologias e às vezes pouco se acrescenta na clarificação do conceito. Alguma contenção é atitude avisada!

[9] Barnett (1976) é um estudo fundamental cujo lema é "order properties… exist only in one dimension" e com discussão pelos melhores especialistas. É um artigo de referência que desperta para a importância da ordenação na detecção de observações discordantes. Conjugado com a dimensão dos dados estatísticos esse artigo "atravessa" muitos domínios, novos à época, como o estudo de dados multivariados e a sua relação com as subordens. O "termo *outlier*" surge "no contexto" onde vai adquirindo cada vez mais importância à medida que se avança no estudo desse texto. Este pode ser um sinal, a palavra-chave, para o despertar de um novo campo de investigação (nessa década ainda) sem história em Portugal (e muito novo no mundo científico de então!). E assim pode acontecer (mais) um acaso científico! Este, (verificado em 1982) levaria à elaboração de Rosado (1984) – para obtenção de doutoramento na área dos *outliers*, o primeiro em Portugal.

Mais recente, todas as reflexões e propostas de Rosado (2014) podem ser usadas para *Big Outliers(s)*, em particular: "a necessidade de *outliers*", "um caminho de investigação" ou " A Força desses Menores".

Conclusão

Na Ciência em geral e na Estatística Multivariada em particular, é possível comparar os desafios de ontem e de hoje? Ontem existiam mais incentivos à investigação, desde logo as bolsas; hoje existe (muito) mais informação e o seu acesso que (também) é estimulante. Ontem não havia (tanto) software! Hoje, há software a mais?! A teoria de ontem continua a ser resposta teórica de hoje. Nesse ponto de vista pouco se avançou.

Big Outlier na investigação fundamental é (apenas e não mais do que!) um *Outlier* e como tal deve ser estudado. No futuro, que a velocidade e premência rapidamente transformam em presente, *outliers* continuarão a ocupar um lugar no centro da Ciência Estatística e nos seus Métodos Estatísticos, quaisquer que eles sejam, porque uma observação discordante sempre será um desafio para o analista e pode largamente influenciar todos os seus relatórios para as mais importantes decisões.

Falamos de excelência na investigação!

Mas, muito está a mudar, os desafios orientadores das mais diversas funções profissionais estão seguramente alterados perante a visão tradicional[10]. E para os Estatísticos também!

Em todos os níveis científicos e profissionais os momentos que se vivem são de mudança constante e veloz e isto também muito em consequência da "rapidez digital" que carateriza a sociedade atual; desde logo pelas terminologias inovadoras que utilizamos e que, em termos gerais, revisitámos neste texto.

No entanto, a Ciência em geral e a Estatística em particular é uma nobre atividade, necessária ao corpo e ao espírito, indispensável ao bem-estar e à felicidade. Mas, a ciência é cara. Só os ricos a podem praticar … e os pobres se a praticam ficam mais pobres. Embora exigindo grande esforço e dedicação a solução deve estar em (apesar de tudo) fazer ciência para caminhar na saída daquele dilema. E o mesmo se passa na *Teoria dos Outliers*.

Contudo, quando tudo está dito e feito, mesmo e talvez ainda mais para o(s) *Big Outlier(s)*, o principal tema no estudo de observações (supostamente) suspeitas continua a ser aquele que desafiou os pioneiros investigadores – *O que é um (Big) Outlier e como tratar essa observação?* E a resposta será sempre: Investigação Fundamental como suporte para excelência nas Aplicações.

Referências e Bibliografia

Barnett, V. (1976) – The Ordering of Multivariate Data (with discussion). *Journal of Royal Statistical Society* A, p. 318-354.

Barnett, V. and Lewis, T. (1994) – *Outliers in Statistical Data*. 3rd edition. Wiley.

Mardia, K. V., Kent, J. T. e Bibby, J. M. (1979) – *Multivariate* Analysis. Academic Press.

Rosado, F. (1984) – *Existência e Detecção de Outliers – Uma Abordagem Metodológica*. Tese de Doutoramento. Universidade de Lisboa.

Rosado, F. (1991) – *Análise de Dados Multivariados*. Programa de Disciplina; conteúdos e métodos. Universidade de Lisboa.

Rosado, F. (2005) – *Memorial da Sociedade Portuguesa de Estatística.* Edições SPE.

Rosado, F. (2006) – *Outliers em Dados Estatísticos.* Edições SPE.

Rosado, F. (2014) – Outliers: The Strength of Minors. *News Advances in Statistical Modelling and Applications,* Pacheco, A. *et al* (Editores), p. 17-29.

Tukey, J. (1962) – The Future of Data Analysis. The *Annals of Mathematical Statistics*, Vol. 33, No. 1 p. 1-67.

[10] Na verdade, para algumas organizações, os dados de hoje tornaram-se uma parte tão explosiva do negócio que já criaram um (novo) cargo de "Diretor de Dados" (CDO - Chief Data Officer). Estes novos profissionais vão ter (porque ainda não têm) um perfil consolidado em áreas que, assim se deseja, tenham uma forte formação Estatística. Só assim "a decisão" será bem fundamentada. Tudo isto requer que o Estatístico se afirme pela excelência na sua formação; o que exige novos planos curriculares, também na Análise de Dados e (talvez) bem diferentes de Rosado (1991), acima referido. São desafios para os quais, no essencial, a Sociedade Portuguesa de Estatística tem a responsabilidade de, como líder, enfrentar e ajudar resolver para a excelência dos estatísticos portugueses.

Uma curta reflexão sobre o futuro da Estatística Multivariada

Jorge Cadima, *jcadima@isa.ulisboa.pt*

Instituto Superior de Agronomia, Universidade de Lisboa
CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa

Este curto texto visa, de forma despretensiosa e não exaustiva, contribuir para a discussão em curso sobre a estatística e o seu futuro, dando particular atenção à estatística multivariada.

Sobre as raízes

A estatística multivariada em sentido lato, ou seja, englobando a análise de dados multivariados e as técnicas inferenciais e de base probabilística, tem cerca de um século de História. O conhecimento deste percurso ajuda a compreender os novos desafios dos nossos dias e a antever as tendências da sua evolução.

Como em qualquer outra área do conhecimento, o desenvolvimento da estatística multivariada foi marcado pelos problemas concretos que motivavam o estudo de várias variáveis. Mas também pela disponibilidade de dados sobre os quais assentar esse estudo; pelos conhecimentos teóricos que permitiam (ou não) dar-lhes resposta; e pelas limitações das capacidades computacionais disponíveis.

Até ao final da primeira metade do Século XX, o principal fator limitante residia na capacidade de cálculo. Embora a própria possibilidade de recolha de grandes volumes de dados fosse mais limitada do que na actualidade, eram as dificuldades computacionais que representavam o maior obstáculo ao desenvolvimento da estatística multivariada. As limitações computacionais contribuíram certamente para incentivar os notáveis avanços no plano teórico que marcaram esse período, assentes em áreas matemáticas como a teoria das probabilidades e a álgebra linear e teoria de matrizes.

Não foi automática a ideia de que uma colecção multivariada de dados pode ser tratada como uma matriz do tipo indivíduos x variáveis. Essa conceptualização gradual abriu portas à possibilidade de recorrer ao corpo crescente de resultados matriciais (e de simultaneamente contribuir para o seu ulterior desenvolvimento). Métodos como a Análise em Componentes Principais ou a Análise (linear) Discriminante de Fisher assentam nestes desenvolvimentos teóricos, embora de forma inicialmente titubeante[i]. Ao mesmo tempo, as matrizes de covariâncias e de correlações ganharam papel central na estatística multivariada, em parte associadas à sua presença na função densidade da distribuição Multinormal. Neste período, o desenvolvimento da estatística em geral, e da estatística multivariada em particular, teve uma forte marca matemática, nomeadamente de teoria das probabilidades. Essa faceta probabilística encobriu por vezes os aspectos geométricos e de álgebra linear sobre os quais assentam muitos dos tradicionais métodos multivariados de análise de dados, de tal forma que ainda hoje é para muitos utilizadores nebulosa a distinção entre exigências de pressupostos probabilísticos e o fundo geométrico que pode existir independentemente desses pressupostos.

Com os avanços na capacidade de cálculo da segunda metade do Século XX, ganham importância métodos de forte componente computacional. Por vezes, trata-se de métodos essencialmente empíricos, e mais permeáveis a múltiplas opções de percurso (que afetam os resultados), de que são exemplo bem conhecido as Análises Classificatórias (*Clustering*). Noutros casos, geram inesperados desenvolvimentos conceptuais, como é o caso das técnicas de reamostragem, tipo *bootstrap*.

Já neste período se verificaram controvérsias sobre a natureza da estatística e das suas ferramentas, de que é exemplo o texto *The Future of Data Analysis* de John W. Tukey, em 1962[ii]. É um tema que, na viragem do milénio, ganha ainda maior importância e predominância.

A explosão computacional e de informação

Os avanços quantitativos, quer na capacidade e velocidade de cálculo, quer no volume de dados disponível em muitas áreas de aplicação (está já consagrada a expressão *big data*) estão a gerar uma transformação qualitativa em muitas áreas da estatística multivariada. Como é usual neste tipo de situações, os processos de transformação e adaptação são por vezes conturbados, e merecem algumas considerações.

A existência de grandes volumes de informação não é uma novidade em si mesma, sobretudo em certas áreas de aplicação. Casos paradigmáticos são os censos populacionais ou os registos meteorológicos. Curiosamente, discute-se hoje se é possível substituir os censos populacionais (recorrendo a fontes indirectas de recolha de informação, mas também à amostragem), dado o seu elevado custo e dificuldades organizativas. No entanto, é uma realidade que a multiplicação de fontes de recolha de informação (por exemplo, os dados meteorológicos recolhidos por satélite) significa que o volume de dados disponível está em acelerado crescimento, mesmo nestas áreas. Mas o facto recente saliente é a tendência para estes grandes volumes de informação passarem a estar disponíveis num número cada vez maior de áreas de aplicação. Este processo objetivo resulta do papel crescente da informática, da Internet e outras redes de comunicação, do número cada vez mais vasto de bases de dados em múltiplas áreas, do crescente acesso a dados de satélite com aplicações nas mais diversas áreas, e de sensores de baixo custo facilmente instaláveis em meios de recolha móveis (*drones*) e/ou em computadores de custo quase nulo e facilmente instaláveis nas mais diversas situações (tipo *Raspberry Pi*).

Paradoxalmente, o aumento vertiginoso de enormes massas de dados ameaça recriar, num patamar mais elevado, uma situação parecida com a que existia há algumas décadas atrás, quando a capacidade de analisar e interpretar os dados disponíveis era insuficiente, embora desta vez as insuficiências estejam mais do lado dos recursos humanos, do que do lado das insuficiências técnicas.

Esta explosão de dados é, em grande medida, uma explosão de dados *multivariados*. Objetivamente, torna a estatística multivariada mais actual do que nunca. Ao mesmo tempo, coloca novas exigências no que respeita às metodologias de análise e tratamento e tem, de forma aparentemente contraditória, dado origem a um novo debate questionando a estatística e o seu futuro. Um exemplo dessa controvérsia pode encontrar-se num número recente da revista *Journal of Computational and Graphical Statistics* (2017, Vol.26, No.4).

Se, por um lado, métodos clássicos da análise exploratória (descritiva) de dados multivariados assumem novo protagonismo, num contexto onde pequenas amostras são cada vez mais substituídas por amostras de muito grande dimensão ou mesmo censos completos de um dado universo, por outro lado, a quantidade por vezes gigantesca de dados coloca novos desafios. Alguns desses desafios prendem-se com necessidades de identificar características especiais de dados no meio dum mar de informação – por vezes duma autêntica *overdose* de informação. A expressão *data mining*, hoje na moda, refere-se a esta realidade. As características que se procura identificar podem ser de tipo tradicional – como a identificação de observações atípicas (*outliers*) em geral multivariadas, ou a classificação (*clustering*) de indivíduos em grupos mais ou menos homogéneos, ao abrigo de algum critério, também frequentemente de índole multivariado. Mesmo assim, podem assumir traços específicos do contexto (como a atipicidade de algum objecto numa imagem de milhões de pixeis, associada a conceitos de contiguidade espacial, mas também de identificação do próprio objecto). Os objectivos podem também corresponder a problemas menos tradicionais, resultantes do enorme volume de dados e das ligações de diverso tipo que se estabelecem entre os registos. Estão neste caso o reconhecimento de padrões (*pattern analysis* ou *machine learning*), por exemplo na análise de comportamentos de clientes ou utilizadores de algum serviço, ou a necessidade de algoritmos especialmente concebidos para grandes volumes de dados com propriedades especiais, como é o caso de dados esparsos (*sparse data*), ou com exigências especiais (tais como a robustez face à presença de

observações atípicas). Em alguns casos, tem sido mesmo sugerida a amostragem das imensas quantidades de dados disponíveis, gerando novas aplicações de problemas de inferência estatística.

Uma Ciência dos Dados?

Inevitavelmente, estes problemas – velhos ou novos – associados a grandes conjuntos de dados, necessitam do contributo de áreas que até um passado recente estavam menos ligadas à estatística multivariada: algoritmos computacionais, teoria de grafos, inteligência artificial, entre outros. Em alguns casos, a análise destes grandes conjuntos de dados passou a ser feita quase exclusivamente por pessoas que não têm uma formação de base em estatística, mas sim formações noutras áreas, nomeadamente informáticas e das aplicações em questão (e.g., economia, gestão, marketing).

O contributo de pessoas com diferentes formações sempre foi enriquecedor, em todos os aspectos de actividade científica. No contexto em apreço, veio colmatar a ausência de trabalho da comunidade estatística 'tradicional' em certos problemas. Mas existe um perigo de divórcio das comunidades que trabalham sobre problemas análogos, de desconhecimento mútuo sobre as respectivas actividades e avanços, conducentes à 'redescoberta da roda', ou à separação gradual das terminologias e conceitos. Donoho (2017)[iii] escreve: «Para os estatísticos, o fenómeno [...] pode parecer intrigante. Os estatísticos vêem administradores a promover, como se fossem novas, as actividades que os estatísticos têm levado a cabo diariamente, ao longo de toda a sua carreira, e que já eram consideradas usuais quando esses mesmos estatísticos frequentavam os seus estudos de graduação. […] A profissão estatística encontra-se num momento de perplexidade: as atividades que os têm preocupado ao longo de séculos estão hoje sob os holofotes, mas essas atividades são apresentadas como sendo novinhas em folha, e são executadas (embora não na realidade inventadas) por recém-chegados e estranhos».

Há quem defenda a necessidade de novos paradigmas, real ou supostamente inovadores, que passam inclusivamente pela substituição da designação 'estatística' por outras, em princípio mais abrangentes, como 'ciência dos dados' (é o caso do próprio Donoho, 2017). Mas por vezes, resvala-se para a teorização de fazer tábua rasa de práticas e conhecimentos consagrados. Veja-se a frase de outro promotor da *Data Science*, citada de forma crítica por Donoho (p.747): «a Ciência de Dados sem a estatística é possível, até mesmo desejável». Uma tal evolução, que ignora décadas de conhecimento e de avanços, não é benéfica para o progresso dos conhecimentos científicos, e muitas vezes apenas esconde ignorância. O nihilismo científico não é uma boa receita.

Este divórcio pode também, em parte, ser resultante da natureza fechada e opaca de duas áreas destacadas onde predominam os grandes volumes de dados: meios comerciais preocupados com proteger o sigilo da sua informação (ou mesmo da natureza dos seus estudos), e meios militares ou de serviços secretos e de recolha de informações. As revelações sobre as actividades de recolha e monitorização massiva de comunicações pela National Security Agency dos EUA, ou sobre as alterações da Google aos seus motores de busca de forma a filtrar os resultados apresentados nas pesquisas, são disso exemplo. Para além dos problemas éticos e políticos associados (violação de direitos básicos, perigo de controlo e manipulação da sociedade e dos cidadãos), esta realidade levanta o perigo de apropriação de áreas do conhecimento, com base em considerações extra-científicas.

Seria irónico que esta apropriação ocorresse precisamente na estatística, que é o palco de um dos maiores casos de êxito dum paradigma científico e social diferente: o *software* R, assente nos princípios colaborativos do *software* livre, e que se transformou, indiscutivelmente, numa espécie de *língua franca* da comunidade estatística. O R, que tem as suas raízes no trabalho pioneiro de John Chambers, é hoje, não apenas uma ferramenta fundamental para fazer estatística, mas um campo de inovação constante e de desenvolvimento e sistematização de conceitos estatísticos. É um exemplo cimeiro da capacidade da comunidade estatística se transformar e acompanhar as evoluções em curso, nomeadamente em campo informático, sem ser em oposição a, ou negando, o seu próprio passado.

A diversificação dos dados

Uma outra característica associada à recolha massiva de dados diz respeito à diversificação de contextos com dados de natureza especial, alargando por vezes o próprio conceito de 'dados'.

Já foi referido o caso de *dados esparsos*, que deu origem, entre outras, a uma Análise em Componentes Principais para dados esparsos[iv]. Em muitas áreas, como a microbiologia e genética, surgem grandes conjuntos de dados com um enorme número de variáveis observadas em relativamente poucos indivíduos (o problema de p >> n), não permitindo a aplicação de alguns conceitos tradicionais de estatística multivariada (como as distâncias de Mahalanobis, Branco & Pires, 2011[v]) e dando origem à necessidade de métodos específicos, com o uso de regularizações. Noutras aplicações, o ponto de partida não são matrizes de dados do tipo indivíduos x variáveis, mas medidas de similaridade entre entidades. Tais tipos de dados, já anteriormente estudados por técnicas como as análises classificatórias, encontram hoje um grande campo de expansão com métodos *kernel*, do tipo *support vector machines*. Pode referir-se também a importância crescente de *dados funcionais*, resultantes de recolha contínua de registos (por exemplo meteorológicos). A cada vez maior utilização de *dados espaciais*, resultantes da facilidade atual de georreferenciar dados por meio de sistemas do tipo GPS, coloca seguramente desafios ainda por explorar, de adaptação de tradicionais métodos multivariados à incorporação dos conceitos da estatística espacial.

Mais radicalmente inovadores são os avanços no campo dos *dados simbólicos*, uma área com desenvolvimentos no nosso país. Por dados simbólicos entende-se dados que em lugar de serem valores individuais, são entidades mais complexas, que podem conter informação sobre a sua própria variabilidade, tais como intervalos ou histogramas. A estatística multivariada está bem presente neste campo de desenvolvimento, que tem raízes sólidas na comunidade estatística 'tradicional' (Brito, 2014)[vi].

Especialização e integração

A discussão actual sobre o futuro da estatística é, em parte, um processo natural ligado ao avanço de conhecimentos e à sua especialização. Em qualquer campo do conhecimento, a diversificação e especialização crescentes dos conhecimentos tem tendência a gerar áreas específicas que, em processos mais ou menos conturbados, procuram a sua autonomia. Como salientam Hofman e VanderPlas (2017)[vii], as próprias origens da estatística, e a sua autonomização face à matemática, são o fruto dum processo análogo.

Mas autonomia não deve significar costas voltadas. O que até seria paradoxal quando a tendência cada vez mais evidente em todas as áreas do conhecimento é a da necessidade da integração de conhecimentos de tipo diversificado. Especialização e integração estão em relação dialética. O desenvolvimento de numerosos algoritmos computacionais assentes em conceitos biológicos ou da esfera produtiva, como as redes neuronais, algoritmos genéticos, ou o arrefecimento controlado (*simulated annealing*) mostram como a integração aduba o desenvolvimento científico, com um impacto directo na estatística e, em particular, na estatística multivariada.

Mas a integração exige que haja conhecimentos específicos e especializados, que possam ser integrados. É impossível ser-se especialista em todos os campos. E poucos terão a capacidade de ser especialistas em várias áreas. A integração pressupõe que cada um saiba um pouco de várias áreas, mas bastante mais de uma área específica. A ideia de que todos podemos ser especialistas em integração é pouco frutuosa, e arrisca-se a ser uma receita para que ninguém perceba nada. É um perigo que ronda por aí.

O ensino

Esta questão traz-nos ao tema do ensino da estatística, tema de acesos e salutares debates. É uma evidência que o ensino da estatística, e muito em particular da estatística multivariada, tem de incorporar uma aprendizagem de conceitos básicos da informática e algoritmia, e provavelmente também de teoria de grafos e outras áreas aparentadas. Esta evolução tem vindo a verificar-se nos últimos anos, embora seja necessária uma maior sistematização e destaque. A linguagem de programação do R fornece uma excelente oportunidade para uma incorporação natural, e desde a primeira hora, da informática no ensino da estatística. Mas tal como faz hoje pouco sentido um ensino da estatística sem uma forte componente informática e algorítmica, faria pouco sentido deitar pela

borda fora décadas de conceitos consagrados na estatística, tanto mais que continuam a ser usados de forma geral, embora por vezes cega.

No entanto, não é fácil encontrar a forma de diversificar e alargar conteúdos, ao mesmo tempo que se preserva profundidade em áreas fundamentais. Sobretudo, e pensando no nosso país, quando a reforma de Bolonha conduziu ao desaparecimento repentino dum nível inteiro de ensino correspondente aos antigos Mestrados, desaparecimento que a actual integração de componentes lectivas nos Doutoramentos encontra dificuldades em suprir.

Roger Peng (2017)[viii], referindo-se às propostas de reorganização dos conteúdos de cursos em 'ciência dos dados', afirma que «muitas das atividades são coisas que se presume que os estudantes poderão 'descobrir por si próprios', sem qualquer instrução formal. Mas, pelo contrário, a teoria assintótica *exige* uma instrução formal». Salienta o perigo de que o ensino se transforme num «desfile de casos especiais» e a impossibilidade de «dar instrução formal sobre todos os casos especiais», tornando o ensino nestes moldes «difícil e potencialmente muito ineficiente». Peng refere um aspecto de potencial interesse, que diz respeito à criação de um, até aqui inexistente ou incipiente, quadro formal para a tipologia de dados (por vezes designada *data cleaning*).

Sem conclusão...

Seja como for, a análise, tratamento e modelação de dados é uma realidade incontornável, que não só não irá desaparecer, como terá uma importância cada vez mais central nas nossas vidas. E a estatística multivariada (qualquer que venha a ser a sua designação) não irá desaparecer, nem diminuir em importância. Pelo contrário. Cabe-nos a nós contribuir para que ela seja feita com base no muito e importante conhecimento fundamental passado, presente e futuro, de quem trabalhou e trabalha com a matéria prima dos dados. E baseada nos melhores princípios científicos e para bem da comunidade, sem deixar que seja apropriado e desviado por interesses extra-científicos que, em última análise, serão seguramente nocivos para o bem comum.

Agradecimentos

O autor agradece os preciosos comentários e sugestões do Pedro Duarte Silva e João Branco, que não os comprometem com as opiniões aqui expressas, mas que muito enriqueceram esta reflexão.

Por opção, o autor usa a antiga ortografia.

---

i Aquele que é considerado por alguns o texto fundador da Análise em Componentes Principal (Hotelling, H., 1933, Analysis of a complex of statistical variables into Principal Components, *Journal of Educational Psychology*, Vol.24, pgs. 417-441 + 498-520) revela claramente como ainda eram incipientes os conceitos de teoria de matrizes nesta definição pioneira de ACP.

ii Tukey, John W. (1962) The future of Data Analysis. *The Annals of Mathematical Statistics*, 33, 1-67.

iii Donoho, David (2017) 50 Years of Daya Science, *Journal of Computational and Graphical Statistics*, Vol. 26, No. 4, 745-766, https://doi.org/10.1080/10618600.2017.1384734.

iv Zou H, Hastie T, Tibshirani R. 2006 Sparse principal components. *Journal of Computational and Graphical Statistics* 15, 262–264. doi:10.1198/jcgs.2006.s7

v Branco, J.A. & Pires, A.M. (2011) Travelling through multivariate data spaces with Mahalanobis distance, JOCLAD 2011, Vila Real).

vi Brito P. 2014 Symbolic data analysis: another look at the interaction of data mining and statistics. WIREs Data Mining Knowl. Discov. 4, 281–295. (doi:10.1002/widm.1133)

vii Hofmann, H. & VanderPlas, S. (2017) All of this has happened before. All of this will happen again: Data Science. *Journal of Computational and Graphical Statistics*, Vol. 26. No.4, 775-778, https://doi.org/10.180/10618600.2017.1385474.

viii Peng, Roger D. (2017) Comment on '50 years of data Science'. *Journal of Computational and Graphical Statistics*, Vol. 26. No.4, 767, https://doi.org/10.180/10618600.2017.1385470.

Estatística Multivariada – uma perspetiva muito pessoal

Carlos A. Coelho, *cmac@fct.unl.pt*

Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT NOVA)
Departamento de Matemática (DM)
Centro de Matemática e Aplicações (CMA)

1. Introdução

Antes de mais quero aqui deixar um agradecimento ao Professor Fernando Rosado pelo trabalho que tem desenvolvido em prol do Boletim da SPE e da divulgação da Estatística, e nomeadamente da investigação realizada na área da Estatística em Portugal, divulgação esta que tem já desde há vários anos vindo a ser realizada neste Boletim, sob perspetivas, é claro, inevitavelmente pessoais, mas que têm enriquecido o nosso património cultural na área da Estatística, pelos olhares que nos propõem e nos deixam sobre as várias facetas desta estimulante área do conhecimento.
É assim com bastante receio de poder não estar à altura de outros autores que aqui têm deixado as suas perspetivas, mas também com outra tanta ousadia que aceitei o convite do Professor Fernando Rosado para escrever as linhas que aqui vos deixo e que embora rolando sobre uma perspetiva muito pessoal espero que deixem uma visão de alguma forma útil e motivadora e que possam também constituir, quando a outro que não o próprio autor se referirem, uma singela homenagem a alguns daqueles que de uma forma decisiva contribuíram e continuam a contribuir para o desenvolvimento desta interessante e estimulante área de investigação e aplicações.
Também com alguma ousadia, diria que se de facto a própria vida e o mundo que nos rodeia são multivariados, porque não também a Estatística? Se nos é quase sempre impossível pensar numa única razão, uma única consequência dos nossos atos, numa única variável que nos condiciona, ou numa única variável da qual gostaríamos de conhecer o valor, ou conseguir plenamente compreender, como pode a Estatística ser univariada?

2. Os trabalhos de Samuel Stanley Wilks

Embora os estudos e a investigação em Estatística envolvendo mais de uma variável tenham começado antes de Samuel Stanley Wilks (1906-1964) ter escrito os seus artigos, pode-se e deve-se bem considerar Wilks como o principal obreiro do desenvolvimento e da atenção que a Estatística Multivariada viria a conhecer.
Embora os testes de razão de verosimilhanças fossem já conhecidos, foi Wilks (1932, 1935, 1946, 1963) quem lhes acabou por dar o lugar de relevo que vieram a ter em Estatística Multivariada.
Wilks, como se tornou mundialmente conhecido, teve infelizmente uma vida demasiado curta para que pudéssemos ter usufruído plenamente do seu génio, mas tivemos a sorte de mesmo assim nos ter deixado obras onde se encontram ideias, conceitos e desenvolvimentos sem dúvida alguma dignos do nosso interesse e da nossa consideração. Foi Wilks (1932) quem lançou os fundamentos sólidos para a Análise de Variância Multivariada, hoje conhecida pelo seu acrónimo de raíz anglo-saxónica MANOVA (Multivariate ANalysis Of VAriance). Foi também Wilks (1935) quem desenvolveu o teste de razão de verosimilhanças para o teste de independência de vários grupos de variáveis, o qual pode de facto ser visto como um teste que abrange afinal muitos dos testes utilizados em modelos lineares ou com eles relacionados pois pode-se mostrar que tem como casos particulares o teste de ajustamento do modelo de Análise Canónica ou Regressão Multivariada, onde se testa a relação (de independência ou não) entre dois grupos de variáveis e onde se pode considerar um dos grupos como sendo o das variáveis resposta e o outro o das variáveis explicativas, mas também o teste de ajustamento para um

modelo MANOVA ou MANCOVA (Multivariate ANalysis Of COVAriance), onde o conjunto das variáveis explicativas além de incluir as variáveis indicatrizes dos fatores e eventuais interações também inclui algumas variáveis contínuas, e como tal também a Análise de Variância e de Covariância univariadas, onde temos apenas uma única variável resposta, a própria Regressão múltipla ou simples univariada (i.e., com uma única variável resposta) e mesmo até os testes T para amostras emparelhadas ou independentes (Knapp, 1978; Thompson 1991, 2000; Coelho, 1992; Vidal, 1997; Sherry e Henson, 2005).

Não se pode falar de Wilks sem falarmos também do teste de razão de verosimilhanças para 'outliers' que nos deixou (Wilks, 1963) e de tantos outros interessantíssimos artigos que felizmente podem ser encontrados juntos no livro "S. S. Wilks: collected papers – Contributions to Mathematical Statistics" editado pelo Professor T. W. Anderson (Anderson, 1967).

Samuel Stanley Wilks de facto não se dedicou apenas à área da Estatística Multivariada, tendo-nos também deixado importantes trabalhos noutras áreas da Estatística, como é o caso do seu livro em Estatística Matemática (Wilks, 1947, 1962), embora mesmo neste livro seja muito clara a sua tendência para passar muito rapidamente para a consideração de mais do que apenas uma variável aleatória, nomeadamente com a edição de 1962 a não poder deixar de incluir um último capítulo intitulado "Multivariate Statistical Theory", mas também já com a edição de 1947 a ter um último capítulo intitulado "An Introduction to Multivariate Statistical Analysis".

Os trabalhos de Wilks sobre estatísticas de razão de verosimilhanças e as estatísticas de razão de verosimilhanças por ele derivadas fizeram escola de tal forma que, e também devido ao facto de muitas destas estatísticas terem uma estrutura de alguma forma comum, o termo Lambda de Wilks acabou por se vulgarizar na área da Estatística Multivariada para designar uma estatística de razão de verosimilhanças, ou mais precisamente, para uma amostra de dimensão n (e para variáveis aleatórias reais com distribuição multivariada Normal), a potência $(2/n)$ da estatística de razão de verosimilhanças, a qual se designa habitualmente por Λ e que tem uma expressão da forma

$$\Lambda = \frac{|A|}{|A+B|} \tag{1}$$

onde A e B são duas matrizes, sendo que A tem uma distribuição Wishart (Wishart, 1928; Kshirsagar, 1972, Cap. 3; Anderson, 2003, Sec. 7.2; Muirhead, 2005, Sec. 3.2) digamos de dimensão p e graus de liberdade $n-q$, para por exemplo um teste envolvendo q amostras, e matriz de parâmetro Σ (definida-positiva), facto que denotaremos por $A \sim W_p(n-q,\Sigma)$, e B tem, sob a hipótese nula a ser testada, também uma distribuição Wishart de dimensão p e graus de liberdade $q-1$, também com matriz de parâmetro Σ, i.e., $B \sim W_p(q-1,\Sigma)$, sendo A e B independentes, de modo que $A+B \sim W_p(n-1,\Sigma)$ e sendo a distribuição de A válida quer sob a hipótese nula quer sob a hipótese alternativa. Na distribuição de A requer-se habitualmente $n-q>p$, enquanto que na distribuição de B podemos ter $p>q-1$, caso em que B terá uma distribuição Wishart legítima, ou $p \leq q-1$, caso em que B terá uma distribuição pseudo-Wishart (Kshirsagar, 1972, Sec. 3.6), em qualquer dos casos com $A+B$ a ter sempre uma distribuição Wishart legítima. Para variáveis aleatórias complexas com a distribuição multivariada Normal comummente mais utilizada (Wooding, 1956; Goodman, 1957, 1963a,b; James, 1964, Sec. 8; Khatri, 1965; Gupta, 1971; Krishnaiah et al., 1976; Fang et al., 1982; Brillinger, 2001, Sec. 4.2; Anderson, 2003, problema 2.64), uma estatística Λ do tipo da estatística em (1) será, para uma amostra aleatória de dimensão n, a potência $(1/n)$ da estatística de razão de verosimilhanças.

Demonstra-se que Λ tem a mesma distribuição da de um produto de variáveis aleatórias Beta independentes, facto a partir do qual se podem então elaborar vários estudos sobre a distribuição da estatística Λ, sendo que em algumas situações esta distribuição pode assumir formas relativamente simples enquanto noutros casos esta distribuição terá funções de densidade e de distribuição de probabilidade bem complicadas e mesmo impróprias para serem utilizadas em aplicações, facto pelo qual é comum o recurso a distribuições assintóticas ou quase-exatas. Veja-se a este propósito a Secção 6.

3. A Estatística Multivariada através de alguns dos mais importantes livros nesta área

Sem dúvida que os trabalhos de S. S. Wilks tiveram influência determinante naquele que viria a ser o surgimento de um grande número de obras de fundo na área da Estatística Multivariada quer em termos mais formais quer em termos mais aplicados, com uma enorme profusão de artigos nos anos

60, 70 e 80 do século XX. Mas, pretendo referir-me nesta secção a livros que marcaram e continuam a marcar não só a área em si mas também as vidas daqueles que tiveram a sorte, e diria mesmo, a bênção de os ler, pois tais livros acabam por de uma forma ou de outra marcar a vida daqueles que com eles se cruzam. São livros como os de Anderson (1958, 1984, 2003), Kshirsagar (1972) e Muirhead (1982, 2005) que são verdadeiros monumentos na área da Estatística Multivariada, cada um com o seu cunho muito próprio e com as suas facetas estimulantes. Estes livros são um manancial de conhecimentos e de informação e referências indispensáveis e impossíveis de ignorar para quem queira trabalhar na área da Estatística Multivariada, nomeadamente se tal trabalho envolver facetas mais relacionadas com a teoria. Posso dizer com toda a alegria e satisfação que tive a sorte de conhecer estes três autores, sendo somente de lamentar o facto de o Professor T. W. Anderson nos ter deixado há pouco mais de um ano, mas que na sua vida de quase um século nos bafejou com a sua presença. Embora cientificamente falando me tenha avidamente alimentado de cada uma destas obras, não posso deixar de frisar a grande influência que sobre mim teve o livro do Professor Anant M. Kshirsagar (Kshirsagar, 1972) que sem dúvida alguma foi o grande responsável pelo meu interesse pela área da Estatística Multivariada inferencial, quando pela primeira vez o li, ou talvez devesse antes dizer, tentei ler, na biblioteca da Universidade de Montpelier em 1985, vindo mais tarde a ter a magnífica oportunidade de ter o Professor Anant Kshirsagar como meu orientador, enquanto aluno de Doutoramento no Departamento de Bio-Estatística da Universidade de Michigan.

É claro que a história da Estatística Multivariada e a sua bibliografia, em termos de livros, não se faz apenas destes três livros, havendo muitas outras obras que não podemos deixar de mencionar, embora não seja de modo algum objetivo deste breve comentário exibir um conjunto bibliográfico exaustivo sobre a área, tarefa que aliás seria mais ou menos impossível. São exemplo de outras importantes obras na área da Estatística Multivariada os livros de Bilodeau e Brenner (1999) e um livro mais recente de Kollo e von Rosen (2005) que nos traz uma abordagem diferente, muito baseada na álgebra matricial.

Não podemos também esquecer outros livros que, com um cariz mais aplicado, nos fornecem um manancial de interessantes aplicações como sejam os livros de Morrison (1967, 1976, 1990, 2005), Johnson e Wichern (1982, 1988, 2007, 2014), Rencher (1995, 1998, 2002), Timm (2002) e Rencher e Christensen (2012).

O conjunto destes livros, bem como muitos outros aqui não referidos, teve ainda a grande virtude de ter contribuído para a 'democratização' da Estatística Multivariada, desmistificando uma área que perante algumas audiências, quer do ponto de vista teórico, quer do ponto de vista aplicado, estava afetada de um certo estigma de dificuldade e de quase impenetrabilidade, sendo apenas acessível a algumas mentes privilegiadas ou especialmente preparadas, embora seja verdade que é necessária uma preparação sólida em termos de conceitos básicos de Estatística Univariada, bem como alguns conhecimentos de Álgebra Linear e Matricial e de Análise Matemática, para que se possa plenamente entender os conceitos envolvidos e métodos utilizados e usufruir assim de tudo o que a Estatística Multivariada está pronta para nos proporcionar.

4.Amostras de pequena dimensão ou 'dados de grandes dimensões' ('high-dimensional data')

Mais recentemente e nomeadamente relacionado com estudos nas áreas da biologia e mais especificamente da genética, ou se quisermos, das chamadas 'ómicas' (genómica, transcriptómica, proteómica e metabolómica) surgiram em Estatística Multivariada os estudos na área de amostras de pequenas dimensões ou de 'dados de grandes dimensões' ('high-dimensional data'), onde as amostras, que tipicamente nos estudos clássicos têm sempre uma dimensão superior ao número de variáveis envolvidas, e que em alguns modelos clássicos que utilizam mais de uma amostra, têm mesmo de ter uma dimensão superior à soma do número de variáveis envolvidas com o número de amostras envolvidas, têm agora uma dimensão inferior ao número de variáveis analisadas ou observadas. Esta é uma questão que, embora não se coloque de uma forma direta em muitos dos modelos univariados mais simples pode também surgir sob formas algo diferentes em modelos como na Regressão e sobretudo na Análise de Covariância, mas que é hoje em dia de interesse fundamental em determinadas áreas como as acima referidas da genética e das 'ómicas', mas também noutras áreas como por exemplo em medicina, nomeadamente em estudos de doenças raras, ou em problemas de 'computer vision', mais precisamente em estudos sobre algoritmos de reconhecimento facial. São situações em que o número de variáveis medidas pode ser muito grande, ao mesmo tempo que existe

disponível apenas um número limitado e por vezes muito pequeno de indivíduos ou unidades de observação.
O desenvolvimento de procedimentos para a realização de testes a hipóteses de interesse e a obtenção das respetivas estatísticas de teste, na situação de 'high-dimensional data' (amostras pequenas), requer habitualmente o recurso a extensa formulação e à utilização de distribuições assintóticas e tem sido recentemente uma área de grande interesse e de intenso trabalho por parte de muitos investigadores, sendo possível encontrar na literatura um já grande número de artigos sobre o assunto como por exemplo os de Srivastava (2005, 2009), Srivastava e Du (2008), Srivastava e Yanagihara (2010), Chen e Qin (2010) e Srivastava, Katayama e Kano (2013).

5. "Big data"

"Big data" é, em termos simples, o termo recentemente cunhado para designar conjuntos de dados muito extensos que podem atingir vários milhões de observações realizadas sobre muitos milhares ou mesmo também milhões de variáveis, e que podem resultar de uma grande variedade de áreas de atividade, como por exemplo da recolha de dados sobre tráfego ou outras atividades na 'internet', ou mesmo de dados geográficos ou dados resultantes da digitalização de dados sismológicos ou ainda de estudos em marketing ou sociologia onde por exemplo podem ser estudadas preferências por determinados produtos ou o assumir de determinados comportamentos sociais complexos.
Cada vez mais, com a atual facilidade de recolha de dados as questões relacionadas com a análise e a extração de informação destes grandes conjuntos de dados que vão surgindo em quase todas as áreas do conhecimento se afiguram não só como problemas de interesse, mas em algumas áreas como as da segurança e do marketing, como problemas prementes.
Muitas das técnicas utilizadas, de alguma forma 'voltam' a estar relacionadas com técnicas e métodos de cariz essencialmente algébrico e geométrico muito próximos do que foi nos anos 70 e 80 do século XX designado por Métodos de Análise de Dados ou Métodos Fatoriais de Análise de Dados (Eisenbeis e Avery, 1972; Escoufier, 1973, 1975; Dagnelie, 1975; Cailliez e Pagés, 1976; Bouroche e Saporta, 1980; Sarbo, 1981; Volle, 1981, 1985; Dunn e Everitt, 1982; Coelho, 1986), sendo que muito frequentemente o que se pretende é mais o sumarizar de forma útil a informação nos dados ou procurar 'observações anómalas' (os chamados 'outliers'), mais do que efetivamente estabelecer modelos estruturais entre as variáveis envolvidas, embora frequentemente possa interessar o estudo das relações de associação ou antagonismo entre as variáveis. A literatura mais recente sobre o tópico 'Big Data' parece no entanto ir sendo construída mais à base de livros com os de Simon (2013), Foreman (2013), Mayer-Schönberger e Cukier (2013), Davenport (2014) e Baescus (2014) do que propriamente à base de artigos.

6. Uma perspetiva (demasiado) pessoal de alguma da investigação realizada em Portugal em Estatística Multivariada

O autor deste artigo de divulgação, tendo bebido da maioria das fontes referidas nas secções anteriores decidiu há vários anos tentar desenvolver aquilo a que chamou na altura de 'distribuições quase-exatas' e decidiu, com alguma imodéstia, que pede que lhe seja perdoada, aqui falar destas distribuições.
Em muitas situações as distribuições das estatísticas de razão de verosimilhanças utilizadas em Estatística Multivariada ou Análise Multivariada, conforme se goste mais de chamar, ou das suas potências (2/n) ou (1/n), consoante e trate de variáveis aleatórias reais ou complexas, mesmo para o caso de variáveis com distribuição Normal multivariada, ou com uma distribuição multivariada de contornos elípticos, têm frequentemente expressões demasiado complicadas para as suas funções de densidade e de distribuição de probabilidade, difíceis de implementar computacionalmente com a devida precisão e assim também difíceis de utilizar em aplicações, o que tem levado a uma utilização quase generalizada de aproximações assintóticas. Todavia sabe-se já desde há algum tempo que tais distribuições assintóticas não só podem, de uma forma geral, não exibir a precisão desejada, como têm um mau comportamento para amostras pequenas (i.e., situações em que a dimensão da amostra mal excede o número de variáveis envolvidas, ou a soma deste número com o número de amostras em questão) bem como para situações em que o número de variáveis envolvido é elevado (i.e., na ordem de algumas dezenas). Sabe-se aliás hoje em dia que algumas das usualmente mais utilizadas

distribuições assintóticas para estatísticas de razão de verosimilhanças utilizadas em Estatística Multivariada não são distribuições legítimas em situações em que o número de variáveis envolvidas na análise atinja algumas dezenas e a dimensão da amostra exceda este valor, eventualmente adicionado do número de amostras, só em algumas unidades (Coelho e Marques, 2012).

Foi com o objetivo de obter distribuições assintóticas que não enfermassem destes problemas e que desempenhassem muito bem mesmo para amostras de pequena dimensão que foram desenvolvidas as distribuições quase-exatas. Procurava-se ainda que estas distribuições fossem assintóticas não só em relação a valores crescentes da dimensão da amostra mas também em relação a valores crescentes do número de amostras envolvidas, o que algumas das usuais distribuições assintóticas ainda conseguem, e ainda em relação a valores crescentes do número de variáveis envolvidas, objetivo que as usuais distribuições assintóticas de forma alguma conseguem atingir, pretendendo-se ainda finalmente que, é claro, a distribuição obtida seja manejável, por forma a permitir um fácil e rápido cálculo de quantis e valores-de-p, i.e., valores da função de distribuição.

Embora à primeira vista a prossecução de todos estes objetivos em simultâneo possa parecer um tanto irrealista e o seu alcance um tanto impossível, tal não é o caso pois todos estes objetivos são possíveis de alcançar em simultâneo por deixar 'uma boa parte' da distribuição exata da estatística inalterada, aproximando a restante parte com uma aproximação assintótica com um comportamento asseguradamente bom para amostras de pequena dimensão e também com um bom comportamento assintótico em termos da dimensão da amostra, de forma a que depois de voltarmos a juntar as duas partes a distribuição resultante seja conhecida e manejável. Mas a questão então é: mas como podemos deixar 'uma boa parte' da distribuição original da estatística inalterada e o que é esta 'boa parte' e como a 'medimos' por forma a sabermos que corresponde de facto a 'uma boa parte' da distribuição original?

Chamemos Λ à estatística de razão de verosimilhanças em questão. Como em geral não é muito difícil obter a expressão para os seus momentos de ordem h e esta é válida para h pertencente a uma vizinhança de zero e ainda se mantem como uma expressão válida para h complexo, pode-se facilmente obter a expressão da função característica de $W = -\log \Lambda$, através da relação

$$\Phi_W(t) = E\left(e^{itW}\right) = E\left(e^{-it\log W}\right) = E\left(\Lambda^{-it}\right),$$

onde $i = \sqrt{-1}$ e $t \epsilon \mathbb{R}$. Em seguida fatorizamos $\Phi_W(t)$ e juntamos num fator, chamemos-lhe $\Phi_{W_1}(t)$, os fatores que vamos deixar intactos e noutro fator, chamemos-lhe $\Phi_{W_2}(t)$, os fatores que vamos aproximar assintoticamente, de forma que tanto $\Phi_{W_1}(t)$ como $\Phi_{W_2}(t)$ sejam funções características legítimas. De uma forma geral é possível juntar em $\Phi_{W_1}(t)$, a função característica que vamos deixar inalterada, 'a maior parte' dos termos em $\Phi_W(t)$, i.e., por forma que

$$\int_{-\infty}^{+\infty} \left|\Phi_W(t) - \Phi_{W_1}(t)\right| dt \ll \int_{-\infty}^{+\infty} \left|\Phi_W(t) - \Phi_{W_2}(t)\right| dt \,. \qquad (2)$$

Aproximamos então $\Phi_{W_2}(t)$ por $\Phi_2^*(t)$ por forma que

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty} \left|\Phi_{W_2}(t) - \Phi_2^*(t)\right| dt = 0 \,,$$

Onde *n* representa a dimensão da amostra, o que, tomando

$$\Phi^*(t) = \Phi_{W_1}(t)\, \Phi_2^*(t)$$

como a função característica quase-exata de W, assegurará que a distribuição quase-exata que corresponderá a $\Phi^*(t)$ será assintótica em termos de valores crescentes da dimensão da amostra, e acontecendo que, tendo a factorização de $\Phi_W(t)$ sido realizada de forma adequada, a presença da 'maioria dos termos' em $\Phi_{W_1}(t)$, no sentido de (2), assegurará que a distribuição quase-exata, i.e., a distribuição correspondente a $\Phi^*(t)$, será também assintótica em termos de valores crescentes do número de variáveis envolvidas e eventualmente também do número de amostras envolvidas, nos casos em que o teste em questão envolva mais do que uma amostra. Tudo isto terá de ser executado de modo que a distribuição a que $\Phi^*(t)$ corresponde seja uma distribuição conhecida e manejável, no sentido de que dela seja fácil obter quantis e valores-de-p, i.e., valores da função de distribuição.

A ideia foi lançada num artigo publicado no Journal of Multivariate Analysis (Coelho, 2004) e o processo tem sido eficazmente aplicado a um vasto leque de estatísticas de razão de verosimilhança com aplicações em Estatística Multivariada, sendo possível encontrar mais de 40 publicações sobre o tema, a maioria delas em revistas indexadas ou em livros publicados pela editora Springer (Marques e

Coelho, 2008; Coelho, Arnold e Marques, 2010; Marques, Coelho e Arnold, 2011; Coelho e Marques, 2012; Coelho e Arnold, 2014; Coelho, Marques e Arnold, 2015; Coelho e Roy, 2017; Marques, Coelho e Rodrigues, 2017; Coelho, 2017).

Outras áreas em que presentemente a investigação em Estatística Multivariada se realiza têm a ver com o desenvolvimento de expressões finitas, relativamente simples, para as funções densidade e distribuição de probabilidade de várias estatísticas de razão de verosimilhanças com aplicação em Estatística Multivariada, e consequentemente também para instâncias das funções G de Meijer (Meijer, 1946) e H de Fox (Fox, 1961), a partir das quais seja possível calcular quantis e valores-de-p em frações de segundo (Coelho e Arnold, 2018) e numa área algo diferente, mas que utiliza modelos de Estatística Multivariada, nomeadamente a Análise Canónica ou Regressão Multivariada que é a área de Controlo de Divulgação Estatística, mais conhecida pela sigla de origem anglo-saxónica SDC (Statistical Disclosure Control) (Moura, Klein, Coelho e Sinha, 2017; Moura, Sinha e Coelho, 2017).

Referências

Anderson, T. W. (1958, 1984, 2003) – *An Introduction to Multivariate Statistical Analysis*, 1ª, 2ª, 3ª ed. J. Wiley & Sons, New York.

Anderson, T. W. (ed.) (1967) - *S. S. Wilks: collected papers – Contributions to Mathematical Statistics*. J. Wiley & Sons, New York.

Baescus, B. (2014) - *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. J. Wiley & Sons, Hoboken, New Jersey.

Bilodeau, M., Brenner, D. (1999) - *Theory of Multivariate Statistics*. Springer, New York.

Bouroche, J. M., Saporta, G. (1980) - *L'Analyse des Données*. Presse Universitaire Française, Paris.

Brillinger, D. R. (2001) - *Time Series: Data Analysis and Theory*. SIAM, Philadelphia.

Cailliez, F., Pagés, J. P. (1976) - *Introduction à l'Analyse des Données*. SMASH, Paris.

Chen, S. X., Qin, Y. (2010) - A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* 38 808-835.

Coelho, C. A. (1986) - *Métodos Factoriais de Análise de Dados*. Trabalho de síntese apresentado nas Provas de Aptidão Pedagógica e Capacidade Científica, Instituto Superior de Agronomia, U.T.L.

Coelho, C. A. (1992) - *Generalized Canonical Analysis*. Ph.D. Thesis, The University of Michigan, Ann Arbor, MI, U.S.A.

Coelho, C. A. (2004) - The Generalized Near-Integer Gamma distribution: a basis for 'near-exact' approximations to the distribution of statistics which are the product of an odd number of independent Beta random variables. *Journal of Multivariate Analysis*, 89, 191-218.

Coelho, C. A. (2017) - The Likelihood Ratio Test for Equality of Mean Vectors with Compound Symmetric Covariance Matrices, in *Computational Science and Its Applications*, Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C. M., Rocha, A. M. A. C., Taniar, D., Apduhan, B. O., Stankova, E., Cuzzocrea, A. (eds.), Lecture Notes in Computer Science 10408, Vol. V, Springer, pp. 20-32 (ISBN: 978-3-319-62403-7, 978-3-319-62404-4 (eBook).

Coelho, C. A., Arnold, B. C. (2014) - On the exact and near-exact distributions of the product of generalized Gamma random variables and the generalized variance. *Communications in Statistics – Theory and Methods* 43, 2007-2033.

Coelho, C. A., Arnold, B. C. (2016) - Finite Form Representations of Instances of Meijer G and Fox H Functions – Applications: implementing likelihood ratio tests in Multivariate Analysis, Springer, 377+xv pp. (aceite para publicação na série Lecture Notes in Statistics, com prospetiva publicação em 2018).

Coelho, C. A., Arnold, B. C., Marques, F. J. (2010) - Near-exact distributions for certain likelihood ratio test statistics. *Journal of Statistical Theory and Practice* 4, 711-725.

Coelho, C. A., Marques, F. J. (2012) - Near-exact distributions for the likelihood ratio test statistic to test equality of several variance-covariance matrices in elliptically contoured distributions. *Computational Statistics* 27, 627-659.

Coelho, C. A., Marques, F. J., Arnold, B. C. (2015) - The exact and near-exact distributions of the main likelihood ratio test statistics used in the complex multivariate normal setting. *TEST* 24, 386-416 + supplementary material, 14pp.

Coelho, C. A., Roy, A. (2017) - Testing the hypothesis of a block compound symmetric covariance matrix for elliptically contoured distributions. *TEST* 26, 308-330.

Dagnelie, P. (1975) - *Analyse Statistique à Plusieurs Variables*. Les Presses Agronomiques de Gembloux, Gembloux.
Davenport, T. H. (2014) - *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press, U.S.A.
Dunn, G., Everitt, B. S. (1982) - *An Introduction to Mathematical Taxonomy*. Cambridge University Press.
Eisenbeis, R. A., Avery, R. B. (1972) - *Discriminant Analysis and Classification Procedures*. Lexington Books D. C., Heath & Co.
Escoufier, Y. (1973) - Le traitement des variables vectorielles. *Biometrics* 29, 751-760.
Escoufier, Y. (1975) - Le positionement multidimensionnel. *Revue de Statistique Appliquée* 24, 5-14.
Fang, C., Krishnaiah, P. R., Nagarsenker, B. N. (1982) - Asymptotic distributions of the likelihood ratio test statistics for covariance structures of the complex multivariate normal distributions. *Journal of Multivariate Analysis* 12, 597-611.
Foreman, J. W. (2013) - *Data Smart: Using Data Science to Transform Information into Insight*. J. Wiley & Sons, Hoboken, New Jersey.
Fox, C. (1961) - The G and H functions as symmetrical kernels. *Transactions of the American Mathematical Society* 98, 395-429.
Goodman, N. R. (1957) - On the Joint Estimation of the Spectra, Cospectrum and Quadrature Spectrum of a Two-Dimensional Stationary Gaussian Process. Scientific Paper No. 10, Engineering Statistics Laboratory, New York University / Ph.D. Dissertation, Princeton University.
Goodman, N. R. (1963a) - Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction). *Annals of Mathematical Statistics* 34, 152-177.
Goodman, N. R. (1963b) - The Distribution of the Determinant of a Complex Wishart Distributed Matrix. *Annals of Mathematical Statistics* 34, 178-180.
Gupta, A. K. (1971) - Distribution of Wilks' likelihood-ratio criterion in the complex case. *Annals of the Institute of Statistical Mathematics* 23, 77-87.
James, A. T. (1964) - Distributions of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics* 35, 475-501.
Johnson, R., Wichern, D. W. (1982, 1988) - *Applied Multivariate Statistical Analysis*. Pearson, 1ª, 2ª ed., Prentice Hall, Englewood Cliffs, New Jersey.
Johnson, R., Wichern, D. W. (2007, 2014) - *Applied Multivariate Statistical Analysis*. Pearson, 6ª ed., New International Edition. Prentice Hall, Upper Saddle River, New Jersey.
Khatri, C. G. (1965) - Classical Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution. *Annals of Mathematical Statistics* 36, 98-114.
Knapp, T. R. (1978) - Canonical correlation analysis: a general parametric significance-testing system. *Psychological Bulletin* 85, 410–416.
Kollo, T., von Rosen, D. (2005) - *Advanced Multivariate Statistics with Matrices*. Springer, New York.
Krishnaiah, P. R., Lee, J. C., Chang, T. C. (1976) - The distributions of the likelihood ratio statistics for tests of certain covariance structures of complex multivariate normal populations. *Biometrika* 63, 543-549.
Kshirsagar, A. M. (1972) - *Multivariate Analysis*. Marcel Dekker, New York.
Marques, F. J., Coelho, C. A. (2008) - Near-exact distributions for the sphericity likelihood ratio test statistic. *Journal of Statistical Planning and Inference*, 138, 726-741.
Marques, F. J., Coelho, C. A., Arnold, B. C. (2011) - A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. *TEST* 20, 180-203.
Marques, F. J., Coelho, C. A., Rodrigues, P. C. (2017) - Testing the equality of several linear regression models. *Computational Statistics* 32, 1453-1480.
Mayer-Schönberger, V., Cukier, K. (2013) - *Big Data: a Revolution that Will Transform How We Live, Work and Think*. Houghton Mifflin Harcourt, Boston.
Meijer, C. S. (1946) - On the G-function I–VIII. Proc. *Koninklijk Nederlandse Akademie van Weteenschappen* 49, 227-237, 344-356, 457-469, 632-641, 765-772, 936-943, 1063-1072, 1165-1175.
Morrison, D. F. (1967, 1976, 1990) – *Multivariate Statistical Methods*, 1ª, 2ª, 3ª ed. McGraw-Hill, New York.
Morrison, D. F. (2005) – *Multivariate Statistical Methods*, 4ª ed. Thomson Learning, London.

Moura, R., Klein, M., Coelho, C. A., Sinha, B. (2017) – Inference for Multivariate Regression Model based on Synthetic Data generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling. *REVSTAT* 15, 155-186.

Moura, R., Sinha, B., Coelho, C. A. (2017) - Inference for multivariate regression model based on multiply imputed synthetic data generated via posterior predictive sampling. *AIP Conference Proceedings* 1836, 020065.

Muirhead, R. J. (1982, 2005) - *Aspects of Multivariate Statistical Theory*, 1ª, 2ª ed., J. Wiley & Sons, New York.

Rencher, A. C. (1998) - *Multivariate Statistical Inference and Applications*. J. Wiley & Sons, New York.

Rencher, A.C. (1995, 2002) - *Methods of Multivariate Analysis*, 1ª, 2ª ed. J. Wiley & Sons, New York.

Rencher, A. C., Christensen, W. F. (2012) - *Methods of Multivariate Analysis*, 3ª ed. J. Wiley & Sons, New York.

Sarbo, W. De (1981) – Canonical/Redundancy factoring analysis. *Psychometrika*, 46, 307-329.

Sherry, A., Henson, R. K. (2005) - Conducting and interpreting Canonical Correlation Analysis in personality research: a user-friendly primer. *Journal of Personality Assessment* 84, 37-48.

Simon, P. (2013) - *Too Big to Ignore: The Business Case for Big Data*. Wiley & Sons, Hoboken, New Jersey.

Srivastava, M. S. (2005) - Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society* 35, 251-272.

Srivastava, M. S. (2009) - A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* 100, 518-532.

Srivastava, M. S., Du, M. (2008) - A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* 99, 386-402.

Srivastava, M. S., Katayama, S., Kano, Y. (2013) - A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114, 349-358.

Srivastava, M. S., Yanagihara, H. (2010) - Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis* 101, 1319-1329.

Thompson, B. (1991) - A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development* 24*,* 80–95.

Thompson, B. (2000) - Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 207–226). Washington, DC: American Psychological Association.

Timm, N. H. (2002) - Applied Multivariate Analysis. Springer, New York.

Vidal, S. (1997) - Canonical correlation analysis as the general linear model. Disponível online: https://files.eric.ed.gov/fulltext/ED408308.pdf ou https://eric.ed.gov/?id=ED408308

Volle, M. (1981, 1985) – *Analyse des Données*, 1ª, 2ª ed. Ed. Economica, Paris.

Wilks, S. S. (1932) - Certain generalizations in the analysis of variance. *Biometrika* 24, 471-494.

Wilks, S. S. (1935) - On the independence of k sets of normally distributed statistical variables. *Econometrica* 3, 309–326.

Wilks, S. S. (1946) - Sample criteria for testing equality of means, equality of variances, and equality of covariances in a Normal multivariate distribution. *Annals of Mathematical Statistics*, 17, 257–281.

Wilks, S. S. (1947) - *Mathematical Statistics*. Princeton University Press, Princeton, New Jersey.

Wilks, S. S. (1962) - *Mathematical Statistics*. J. Wiley & Sons, New York.

Wilks, S. S. (1963) - Multivariate statistical outliers. *Sankhya*, Ser. A 25, 407–426.

Wishart, J. (1928) - The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika* 20A, 32–52.

Wooding, R. A. (1956) - The Multivariate Distribution of Complex Normal Variables. *Biometrika* 43, 212-215.

Multivariada e Multidisciplinar. Caminhos divergentes. Uma Opinião!

Irene Oliveira, *ioliveir@utad.pt*

DM-UTAD, *Universidade de Trás-os-Montes e Alto Douro*
CEMAT*, Centro de Matemática Computacional e Estocástica*

Nos dias de hoje os termos multidisciplinar e interdisciplinar estão em voga no mundo académico. Por um lado são usados na descrição de candidaturas a projetos de investigação, por outro surgem nos objetivos de propostas a novos cursos, desde licenciaturas, mestrados a doutoramentos ditos interdisciplinares ou com uma componente forte de interdisciplinaridade. É habitual associarmos a investigação ou o ensino interdisciplinar à Estatística dada a necessidade de nesse mundo que interrelaciona as ciências aplicadas se exigir uma parte considerável de análise de dados e modelação estatística.

Pretende-se nesta reflexão evidenciar a ligação que a estatística multivariada tem com os novos caminhos da multidisciplinaridade e apresentar uma leitura da divergência de caminhos conforme se aborda o campo da investigação ou o campo do ensino e planos curriculares dos cursos de 2º e 3º ciclo no Ensino Superior.

A Investigação

Relativamente aos projetos de investigação é necessário ter em conta que as políticas de investigação científica, no atual contexto de investigação e Ensino Superior, levaram os Centros de Investigação a procurarem equipas de investigadores interdisciplinares, focados quase especificamente em realizar candidaturas vencedoras a grandes projetos organizados segundo os padrões internacionais, com vista a um retorno de avultadas verbas.

A investigação por exemplo nas áreas das ciências da vida e do ambiente, da saúde, animais e agrárias e das tecnologias de informação e computação, …, tem evoluído gradualmente ao longo dos últimos anos com o progresso de tecnologias de ponta, o rápido desenvolvimento dos instrumentos de recolha de dados e o desenvolvimento de aplicações informáticas e *software* de análise de dados, permitindo o processamento de grandes quantidades de dados.

Surgiram assim equipes de investigação colaborativas em projetos de investigação transversais a múltiplas áreas da ciência apresentando, frequentemente, alunos de mestrado, doutoramento e bolseiros trabalhando em simultâneo para vários orientadores de distintas áreas, e cientistas partilhando distintos conhecimentos e competência visando enriquecer o projeto com a exploração das potencialidades e dinâmica que ocorrem por interação dos investigadores.

Existe um provérbio africano que diz: "Se queres ir rápido, vai sozinho. Se queres ir longe, vai em grupo" e que nos remete para a ideia que nada existe isoladamente do resto do mundo.

A adoção dessa abordagem integrativa exige a interação e cooperação da equipe multidisciplinar envolvendo pesquisa científica, experimental e quase sempre análise de dados complementares realizada por especialistas em metodologias de estatística computacional, modelação e mineração de dados de grandes bases dados, ou de dados massivos frequentemente descritos pelo termo *Big Data*.

Mas se por um lado vai-se mais longe em grupo, não é de todo verdade que ir em grupo implique a união dos elementos do grupo. Atualmente os grandes projetos interdisciplinares apresentam alguns pontos fracos, nomeadamente para quem é de uma área transversal a todo o projeto, como seja a Estatística.

O caminho de avaliação destes projetos não seguiu a evolução esperada e a forma de pesar a contribuição dos membros das equipas individualmente levanta ainda algumas questões na

comunidade científica. Que crédito se deve atribuir a um investigador pela sua contribuição para uma investigação colaborativa?

O progresso das Ciências tem ocorrido de forma faseada, enquanto que a evolução da avaliação da produtividade científica dos investigadores ou de uma investigação está a acontecer mais lentamente e alvo constante de reflexão no meio da comunidade académica e científica. Atualmente o sucesso de uma investigação é medido, maioritariamente, com base em indicadores bibliométricos, que não captam o peso das contribuições individuais em equipas multidisciplinares. Muitos artigos associados a projetos multidisciplinares podem ter um conjunto de autores com grau de importância equivalente, quando bibliometricamente se continua a avaliar a importância dos investigadores principalmente pelo contributo do primeiro autor e o último autor (habitualmente um aluno ou bolseiro e o orientador), levando por isso, e a título de exemplo, que os estatísticos e analistas de dados deste tipo de projetos sejam remetidos para uma avaliação secundária.

A avaliação da produção científica usada atualmente é assim uma ameaça para a existência destes projetos multidisciplinares, onde o aumento da autoria múltipla pode levar à divisão e a alguma competição interna ao grupo de investigadores sujeitos constantemente a avaliações curriculares para bolsas, promoções na carreira, verbas e financiamento cientifico dependentes destas métricas, totalmente em divergência com a ideologia do que deveria representar uma equipa multidisciplinar.

No processo do sistema da avaliação de produção científica têm surgido algumas alterações, por exemplo atribuindo igual crédito à autoria, exigindo uma declaração/lista das contribuições individuais de cada autor no trabalho. Esta lista geralmente inclui os seguintes tipos de contribuição: conceção do estudo, realização de experiências, realização da análise de dados, desenvolvimento de ferramentas analíticas ou de modelação, escrita do artigo e discussão, fornecimento de reagentes e materiais e de suporte financeiro. Neste caso específico o papel dos estatísticos das equipas multidisciplinares é valorizado individualmente pelo seu trabalho de, por exemplo, gerar múltiplos modelos, criar ou desenvolver ferramentas de análise para a manipulação de grandes quantidades de dados os quais podemos considerar de extrema importância para a validação dos resultados dos projetos de investigação.

O Ensino Multidisciplinar

Há certamente um caminho a percorrer na reestruturação da avaliação dos investigadores relativamente à produção científica associada a projetos interdisciplinares mas, enquanto se exige que os investigadores trabalhem em equipe e partilhem conhecimentos, a nível do Ensino, principalmente a nível de doutoramento e mestrado, o termo multidisciplinar pode ter outra interpretação.

Nas duas últimas décadas as universidades portuguesas têm procurado dar resposta aos novos desafios da ciência criando vários mestrados, pós-graduações e doutoramentos frisando nos seus objetivos a mais-valia de se obter um curso com cariz multidisciplinar, por forma a satisfazer as exigências profissionais em áreas emergentes e com empregabilidade crescente. Nestes cursos é expectável os alunos estudarem em campos disciplinares de ponta onde é necessário aplicar conhecimentos de várias áreas em simultâneo, usando tecnologias de investigação avançada e linguagens programação, realizando análises de dados com base em metodologias estatísticas avançadas, surgindo assim uma janela de oportunidades para as ciências de dados introduzirem várias unidades curriculares nos seus planos de estudos. É necessário frisar que a regra geral nos tradicionais mestrados e doutoramentos, com uma componente forte de ciências, era a de exigir nos seus planos de estudos uma única unidade curricular de estatística avançada, considerando-se desta forma que os alunos estariam aptos para realizar uma dissertação com análises de dados. Mais recentemente observou-se em algumas universidades a "substituição" da unidade curricular de estatística avançada por metodologias de investigação cujos *curricula*, usualmente, engloba tópicos sobre a execução e apresentação de um trabalho científico, referências bibliográficas e citações, apresentação de trabalhos. Esta decisão levou a que muitos alunos sentissem a falta de conhecimentos estatísticos mais aprofundados na altura de execução das suas análises estatísticas. Por outro lado, ironicamente, nota-se que após a realização e discussão de uma tese, de mestrado ou doutoramento com apoio a metodologias estatísticas, estes mesmos alunos se julguem proficientes na área de estatística e aptos a se candidatarem a bolsas e projetos onde são exigidas competências em *software* estatístico e conhecimentos em estatística multivariada.

Para terminar, tem-se ainda observado o desenvolvimento a nível internacional de cursos de estatística e de *software* específico de estatística em plataformas *online* como por exemplo nas plataformas *Coursera*, *Udemy*,… Recentemente algumas universidades portuguesas começaram também a explorar as potencialidades do ensino massivo global, MOOC, Instituto Superior Técnico- *MOOC-Técnico*, na Universidade do Porto- *Miríada X*, e na Universidade Aberta- *EMMA*. Sendo a análise de dados uma área transversal e complementar com grande procura, a captação de alunos/formandos interessados em adquirir competências estatísticas por via de cursos *MOOC* é uma oportunidade a explorar para o ensino avançado da Estatística em Portugal. Ficando ainda a questão sobre se o uso da língua portuguesa nos MOOC é uma oportunidade ou uma ameaça.

Conclusão

Apesar de as universidades e os organismos de financiamento prestarem atenção ao conceito de interdisciplinaridade este tem vindo a ser abordado no mundo académico de forma distinta conforme abordamos a Investigação ou o Ensino.

Concluindo, há uma certa ironia e dualidade na forma como as entidades governamentais e de decisão de políticas de investigação pretendem um maior número de projetos e bolseiros e interdisciplinares, mas não alteram as regras de avaliação cientifica dos seus investigadores, nem os parâmetros que definem o impacto de uma investigação. O aumento da autoria múltipla levantou o problema da importância atribuída frequentemente ao primeiro e último autor de artigos, não permitindo, de um modo geral, avaliar as contribuições dos outros autores de uma forma justa. Publica-se principalmente em revistas científicas da área principal do projeto de investigação, onde está o investigador líder, levando a que os restantes investigadores e, quase sempre, os estatísticos de área aplicada a serem considerados pelos seus pares, como investigadores de segunda categoria. Daí que se exija uma reestruturação e renovação na avaliação dos trabalhos e artigos que permita um cálculo justa das contribuições reais de cada um. Essa renovação já está a acontecer mas ainda não é a regra, deve por isso o Estatístico ter esperança na caminhada do mundo da investigação interdisciplinar mas deve também estar atento, para que a equipa onde está inserido o valorize.

No campo do Ensino, tem-se vindo a observar uma grande procura de candidatos nos 2º e 3º ciclos das ofertas ditas multidisciplinares as quais, podemos acrescentar, exigem uma parte considerável de análise de dados e modelação estatística. Para os restantes cursos de doutoramento e de mestrado, ditos "não multidisciplinares"!, é consensual que um estudante não deva apenas entender e utilizar conceitos estatísticos nas suas pesquisas mas ter obrigatoriamente conhecimento de metodologias estatísticas mais avançadas contudo estes cursos continuam a considerar, nos seus planos de estudo, apenas uma unidade curricular de estatística ou, num cenário mais pessimista, nenhuma. Relativamente aos cursos *online*, estes abriram a oportunidade ao ensino massivo da estatística avançada e programação estatística permitindo a um elevado número de formandos e alunos procurar colmatar lacunas dos seus conhecimentos estatísticos, e este deveria ser um dos caminhos a considerar no ensino em Portugal e em português.

Encontram-se assim os Estatísticos "entre a espada e a parede". Por um lado, é verdade que há um crescimento de empregabilidade e de procura por estatísticos e que são necessários investigadores com conhecimentos avançados em metodologias estatísticas e multivariadas, mas por outro é muita das vezes assumido pelos investigadores de outras áreas que bastam conhecimentos básicos de estatística multivariada para se singrar no mundo multidisciplinar, sejam eles adquiridos por via de uma única unidade curricular num curso de 2º ou 3º Ciclo, ou por via de frequência de cursos *online*.

É esta a principal divergência entre o mundo da investigação e do ensino. Procura-se o interdisciplinar e o que complemente os conhecimentos numa investigação enquanto que no ensino considera-se frequentemente a estatística como uma área secundária de serviços, sem que haja a atribuição do reconhecimento merecido pelos investigadores das outras áreas.

Procurei com este texto deixar as minhas preocupações sobre o caminho da Estatística para quem trabalha maioritariamente com investigadores de áreas aplicadas, abordando as ameaças e oportunidades do ensino e investigação multidisciplinar. Certamente que uma perspetiva de século XXI para a Estatística Multivariada revelaria muito do trabalho multivariado que ocorre em Portugal, mas a minha preocupação não é "o quê?", mas o "como?" e o "porquê".

Sigamos então o caminho juntamente com os outros cientistas, para conseguirmos ir mais longe! …

Métodos Fatoriais de Análise de Dados e *Big Data*

Adelaide Figueiredo[1], *adelaide@fep.up.pt* e Fernanda Otília Figueiredo[2], *otília@fep.up.pt*

[1]*Faculdade de Economia da Universidade do Porto, LIAAD-INESC TEC Porto e CEAUL*
[2]*Faculdade de Economia da Universidade do Porto e CEAUL*

Introdução

Na literatura existem muitos métodos de redução de dimensão e de visualização de dados multidimensionais (*high dimensional data*). Estes métodos são importantes para transformar um grande número de variáveis correlacionadas num número menor de novas variáveis não correlacionadas entre si, e para visualizar a estrutura dos dados em espaços de dimensão menor. Em Análise de Dados, os métodos fatoriais de redução de dimensão mais populares são a Análise em Componentes Principais (ACP), a Análise das Correspondências Simples (ACS), a Análise das Correspondências Múltiplas (ACM) e a Análise Discriminante Linear (ADL). Estes métodos de estatística multidimensional têm muitas aplicações nas mais variadas áreas, como por exemplo, em saúde, engenharia, genética e ambiente, entre outras. A Análise em Componentes Principais e a Análise Discriminante Linear são também métodos muito usados nas áreas de data mining, *machine learning* e bioinformática.

Estes métodos de Análise de Dados aplicam-se a conjuntos de dados pequenos ou moderados, mas são difíceis de aplicar a *Big Data* devido a problemas de memória e armazenamento. Como se sabe hoje em dia, deparamo-nos com grandes volumes de dados, pelo que se torna necessário ver a adequabilidade dos métodos de redução de dimensão tradicionais, atrás referidos, para este tipo de dados (*big high dimensional data*). Em geral, considera-se um quadro de dados com n observações e p variáveis/caraterísticas (dimensões). Segundo Seng and Ang (2017) este quadro pode ser classificado em duas categorias: quadro com elevado número de observações (*Big Sample Data Set*) e quadro com elevado número de variáveis (*Big Feature Data Set*). Na primeira categoria ($n>>p$), a dimensão dos dados não é elevada, e à medida que o volume dos dados aumenta, o número de observações aumenta, mantendo-se o número de variáveis. Na segunda categoria ($p>>n$), o número de variáveis p é elevado, e pode ainda aumentar, tal como n pode aumentar com o crescimento do volume dos dados.

O enfoque deste trabalho é nos métodos fatoriais de redução de dimensão de Análise de Dados. Iremos referir, nas secções 2, 3 e 4, novas abordagens da Análise em Componentes Principais, da Análise das Correspondências e da Análise Discriminante Linear, respetivamente, desenvolvidas para enfrentar os desafios despoletados pela era do *Big Data*.

Análise em Componentes Principais e *Big Data*

A Análise em Componentes Principais (ACP) é um método de redução de dimensão clássico introduzido por Pearson (1901). Trata-se de um método que permite explicar a variabilidade subjacente a um conjunto de dados através de um número menor de novas variáveis não correlacionadas entre si. Estas novas variáveis, designadas por componentes principais, são combinações lineares das variáveis iniciais. O número de componentes principais que é possível determinar neste método é igual ao número de variáveis inicial, mas interessa-nos reter um número de componentes principais muito menor que o número inicial de variáveis, e que expliquem uma boa variabilidade dos dados. As representações gráficas dos indivíduos e das variáveis nos primeiros planos principais ajudam-nos a identificar a estrutura dos dados. Os resultados obtidos na Análise em Componentes Principais são em geral fáceis de interpretar, pelo que muitas vezes este método é utilizado antes de uma Análise Classificatória, de uma Regressão Linear, e de muitos outros métodos.

Na Análise em Componentes Principais começa-se por efetuar um pré-processamento da matriz dos dados, isto é, centram-se os dados de modo a que as variáveis tenham média zero; em seguida, diagonaliza-se a matriz de covariâncias para determinar as componentes principais. Em geral, as variáveis estão expressas em unidades de medida diferentes ou têm ordens de grandeza diferentes, pelo que é necessário ainda reduzir as variáveis de modo a que fiquem com desvio-padrão unitário, obtendo-se assim dados estandardizados; neste caso, diagonaliza-se em seguida a matriz de correlações entre as variáveis.

Zhang and Yang (2016) referem que as dificuldades em aplicar a Análise em Componentes Principais a *Big Data* estão relacionadas com problemas de memória e armazenamento, e propõem métodos e algoritmos para ultrapassar essas dificuldades, como iremos mencionar sucintamente em seguida. Em geral é impossível guardar a matriz de dados na memória de um computador e a estandardização das variáveis é um problema em *Big Data* porque é difícil guardar os resultados, ou na memória ou no disco duro de um computador. Ainda devido à quantidade massiva de dados que vão surgindo todos os dias, é frequente ser necessária a atualização de dados e a combinação de conjuntos de dados com os anteriores para voltarem a ser analisados. Numa abordagem clássica, temos de considerar o conjunto inteiro dos dados, o que é ineficiente num contexto de *Big Data*.

Assim, se um conjunto de dados grande pode ser guardado no disco duro de um computador, Zhang and Yang (2016) propõem um método para a Análise em Componentes Principais baseado num único processador, assumindo neste método que não pode haver estandardização. Frequentemente, não se está em condições de aplicar esta solução devido à enorme quantidade de dados que surgem diariamente, pelo que estes autores sugerem, em alternativa, recorrer a computação paralela na Análise em Componentes Principais.

Devido às dificuldades computacionais na aplicação da Análise em Componentes Principais a grandes conjuntos de dados, Halko *et al.* (2011) e Witten and Candes (2013) entre outros, aplicaram projeções de matrizes aleatórias. Estes últimos autores aproximam uma matriz de elevada dimensão pelo produto de duas matrizes de menor dimensão as quais podem ser tratadas de forma menos complicada.

Para lidar com dados não lineares, Schölkopf *et al.* (1998) propuseram a Análise em Componentes Principais Kernel (*Kernel PCA*). A matriz de covariâncias na Análise em Componentes Principais é substituída por uma matriz baseada numa função kernel.

Outros métodos de redução de dimensão têm sido propostos para visualizar *high dimensional data,* como por exemplo: o método de redução de dimensão não linear, Kernel Entropy Components Analysis (KECA), proposto por Jenssen (2010), no qual se maximiza a entropia quadrática de Renyi; técnicas de redes neuronais (*Deep neural network*) tais como *Deep Belief Network* (DBN), referido em Noulas and Krse (2008), ou *Staked Auto-encoders* (SAE), apresentado em Schmidhuber (2015).

Tsai (2011) usou a Análise em Componentes Principais para redução de dimensão, de forma a efetuar a visualização de outliers. Najim and Lim (2014) usaram a Análise em Componentes Principais como um método de redução de dimensão para avaliar a qualidade de visualização.

Zhan and El Ghaoui (2011) mostram que, na prática, a Análise em Componentes Principais esparsa (*sparse PCA*) pode ser mais simples que a ACP, e pode ser aplicada a conjuntos de dados muito grandes, como por exemplo, a dados textuais envolvendo milhões de documentos e com centenas de milhares de caraterísticas.

Análise das Correspondências e *Big Data*

A Análise das Correspondências permite estudar as relações entre duas variáveis qualitativas (Análise das Correspondências Simples) ou entre mais do que duas variáveis qualitativas (Análise das Correspondências Múltiplas).

A Análise das Correspondências de um número infinito de linhas ou observações por 1000 atributos foi abordada por Benzécri (1982, 1997). Murtagh (2015a) aplica a Análise das Correspondências a *Big Data*, isto é, a 30 milhões de palavras, apresentando os enormes outputs da análise de forma "inteligente". Murtagh (2015b) descreve propriedades úteis de espaços de dados com elevada dimensionalidade, as quais podem ter interesse na análise de *Big Data*.

Uma nova abordagem da Análise das Correspondências Múltiplas baseada em redes neuronais foi proposta por Tian and Chen (2017) para deteção de falhas em sistemas de gestão de informação.

Análise Discriminante Linear e *Big Data*

A Análise Discriminante Linear (ADL) é um método de redução de dimensão que tem por objetivo determinar as combinações lineares das variáveis iniciais que melhor discriminam os grupos de observações definidos à partida.

Este método também tem por objetivo classificar um novo indivíduo numa de várias classes com base nos valores observados das variáveis para esse indivíduo. Como a distribuição de probabilidade das variáveis é geralmente desconhecida, a regra de classificação é construída usando uma amostra treino.

Muitas vezes em problemas de classificação aplicam-se os dois métodos, ACP e ADL: começa-se por aplicar a ACP para reduzir a dimensionalidade dos dados, e em seguida, aplica-se a ADL. Contudo com *Big Feature Data Sets* pode não ser adequado aplicar a ACP antes da ADL, porque se pode perder poder discriminatório na ADL. A Análise Discriminante Linear para *Big Data*, em geral, ou especificamente para *Big Feature Data Sets*, não tem sido completamente explorada. No entanto, Seng and Ang (2017) propuseram uma abordagem designada por *Split and Combine Linear Discriminant Analysis* (SC-LDA) para *Big Feature Data Analytics*. Contrariamente às abordagens da Análise Discriminante Linear e suas extensões, em que o objetivo é essencialmente melhorar a velocidade e eficiência dos cálculos envolvidos no método, a abordagem SC-LDA tem por objetivo não só reduzir o custo de computação, como também dividir o problema da Análise Discriminante Linear em dois sub-problemas de dimensão menor, resolver os sub-problemas separadamente com um algoritmo base, e combinar os resultados dos dois sub-problemas para obter a solução final. Tal como a ACP, a Análise Discriminante Linear clássica requer a diagonalização de uma matriz que é muito dispendiosa computacionalmente. A abordagem SC-LDA substitui a diagonalização dessa matriz pela diagonalização de sub-matrizes mais pequenas que podem ser efetuadas em paralelo, e depois combina os resultados intermédios para obter os resultados da diagonalização da matriz global.

Existem outras abordagens recentes da Análise Discriminante Linear, tais como algoritmos de ADL de aprendizagem incremental (*Incremental Learning LDA algorithms*) e ADL para conjuntos de dados com poucas observações (*LDA for undersampled data sets*). Os algoritmos de ADL de aprendizagem incremental têm a vantagem de lidar com os novos dados que vão surgindo, i.e, são facilmente aplicados a *data streams*, e a aprendizagem de todos os dados desde o início não é requerida. Isto não exige elevada complexidade computacional e o sistema não necessita de muita capacidade de memória para armazenar os dados, quer aprendidos anteriormente, quer apresentados de novo. Algoritmos de ADL deste tipo foram propostos, por exemplo, por Uray *et al.* (2007), Kim *et al. (*2011) e Ghassabeh and Moghaddam (2013).

Os algoritmos de ADL para *undersampled data sets* pretendem resolver um problema bem conhecido na ADL clássica, designado por problema de singularidade, o qual ocorre quando o número de variáveis p na matriz de dados é elevado comparado com o número n de observações (*Big feature data set*). Outras abordagens para contornar o problema da singularidade têm sido propostas, as quais consistem em usar diferentes variantes da função objetivo da ADL clássica, como por exemplo, em Chu *et al.* (2011). Shao *et al.* (2011) propõem uma Análise Discriminante Linear esparsa (*sparse LDA*) para o caso em que o número de variáveis usada para a classificação é muito maior que a dimensão da amostra, uma vez que neste caso a ADL pode ter uma *performance* não adequada. Qiao *et al.* (2009) também desenvolvem um procedimento de *sparse LDA* eficaz para o caso de *high dimensional data* e amostras de dimensão pequena.

Ye and Wang (2006) apresentam um novo algoritmo para a Análise Discriminante Regularizada (ADR) no caso de *high dimensional data*. É de relembrar que a ADR foi proposta por Friedman (1989) como um compromisso entre a Análise Discriminante linear e quadrática, e tem mostrado ser flexível para lidar com várias classes de distribuições.

Referências

Benzécri, J. P. (1982). L'approximation stochastique en analyse des correspondances, *Les Cahiers de l'Analyse des Données*, **7**(4), 387-394.

Benzécri, J. P. (1997). Approximation stochastique, réseaux de neurones et analyse des données, *Les Cahiers de l' Analyse des Données*, **22**(2), 211-220.

Chu, D., Goh, S. T. and Hung, Y. S. (2011). Characterization of all solutions for undersampled uncorrelated linear discriminant analysis problems, *SIAM Journal on Matrix Analysis and Applications*, **32**(3), 820-844.

Friedman, J. H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**(405), 165-175.

Ghassabeh, Y. A. and Moghaddam, H. A. (2013). Adaptive linear discriminant analysis for online feature extraction, *Machine Vision and Applications*, **24**(4), 777-794.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review*, **53**(2), 217-288.

Jenssen, R. (2010). Kernel entropy component analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **32**(5), 847-860.

Qiao, Z., Zhou, L. and Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data, *IAENG International Journal of Applied Mathematics*, **39**(1), 48-60.

Murtagh, F. (2015a). Correspondence Factor Analysis of Big Data Sets: A Case Study of 30 Million Words; and Contrasting Analytics using Apache Solr and Correspondence Analysis in R, *Technical Report, Goldsmiths University of London*, 1-38.

Murtagh, F. (2015b). Big Data Scaling through Metric Mapping: Exploiting the Remarkable Simplicity of Very High Dimensional Spaces using Correspondence Analysis. *Technical Report, Goldsmiths University of London*, 1-13.

Najim, S. A. and Lim, I. S. (2014). Trustworthy dimension reduction for visualization different data sets, *Information Sciences*, 278, 206-220.

Noulas, A. K. and Krse, B. J. A. (2008). Deep Belief Networks for dimension reduction. In Proceedings of *Belgian-Dutch Conference on Artificial Intelligence*, Netherland, 20, 185-191.

Kim, T., Wong, S., Stenger, B., Kittler, J. and Cipolla, R. (2011). Incremental linear discriminant analysis using sufficient spanning sets and its applications, *International Journal of Computer Vision*, **9**(2), 216-232.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, Series 6, **2**(11), 559-572.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview, *Neural Networks*, 61, 85-117.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, July 1, **10**(5), 1299-1319.

Seng, J. K. P. and Ang, K. L.-M. (2017). Big Feature Data Analytics: Split and Combine Linear Discriminant Analysis (SC-LDA) for Integration Towards Decision Making Analytics, *IEEE Access*, 5, 14056-14065.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse Linear Discriminant Analysis by thresholding for high dimensional data*, The Annals of Statistics*, 39, 2, 1241-1265.

Tian, H. and Chen, S.-C. (2017). MCA-NN: Multiple Correspondence Analysis based Neural Network for Disaster Information Detection, *IEEE Third International Conference on Multimedia Big Data*, 268-275.

Tsai, F. S. (2011). Dimensionality reduction techniques for blog visualization, *Expert Systems with Applications*, **38**(3), 2766-2773.

Uray, M., Skocaj, D., Roth, P. M. and Bischof, A. L. H. (2007). Incremental LDA learning by combining reconstructive and discriminative approaches, in *Proceedings of British Machine Vision Conference (BMVC)*, 272-281.

Witten, R. and Candes, E. (2013). Randomized algorithms for low–rank matrix factorizations: Sharp performance bounds, *Algorithmica*, **72**(1), 264-281.

Ye, J. and Wang, T. (2006). Regularized Discriminant Analysis for High Dimensional Low Sample Size Data, *International Conference* on *Knowledge Discovery and Data Mining* (*KDD*), Philadelphia, PA, USA, 454-463.

Zhang, Y. and El Ghaoui, L. (2011). Large-Scale Sparse Principal Component Analysis with Application to Text Data, in *Proceedings Advances in Neural Information Processing Systems 24 (NIPS)*, 1-8.

Zhang, T. and Yang, B. (2016). Big Data Dimension Reducing using PCA, *IEEE International conference on Smart Cloud*, 152-157.

Ciência Estatística

• Capítulos de Livros

Braumann, C. A., Cortés, J.-C., Jódar, L. e Villafuerte, L. (Junho 2018, online Dec 6, 2017). On the random gamma function: Theory and computing. *Journal of Computational and Applied Mathematics* 335: 142-155. *https://doi.org/10.1016/j.cam.2017.11.045*

Carlos, C. e Braumann, C. A. (2017). General population growth models with Allee effects in a random environment. *Ecological Complexity* 30: 26-33. *http://dx.doi.org/10.1016/j.ecocom.2016.09.003*

Brites, N. M. e Braumann, C. A. (2017). Fisheries management in random environments: comparison of harvesting policies for the logistic model. *Fisheries Research* 195: 238-246. *http://dx.doi.org/10.1016/j.fishres.2017.07.016*

Aguiar, M., **Braumann, C. A.**, Kooi, B. W., Pugliese, A., Stollenwerk, N. (2017). Special issue "Dynamics in Bio-systems" of the journal *Ecological Complexity*, vol. 30: 1. *http://dx.doi.org/10.1016/j.ecocom.2017.01.001*

• Livros

Título: *Estratégia e Fundamentos Teóricos* - Tomo II
Autores: Carlos Manuel Mendes Dias e Jorge Manuel Dias Sequeira.
Ano: 2017. Editora: Letras Itinerantes, Edição e Distribuição de Livros Lda. ISBN: 978-989-99409-8-7. Depósito Legal: 394281/15.

• Teses de Mestrado

Título: Parameter Estimation of the Linear Phase Correction Model by Mixed-effects Models
Autor: Dominic Noy, *dominic.noy@googlemail.com*
Orientadora: Raquel Menezes.

Título: Desempenho dos Estudantes Portugueses: Modelos de Regressão Multinível
Autor: Manuel Castigo, *cmanueljoao@gmail.com*
Orientadora: Susana Faria.

Título: Modelos Longitudinais para Momentos de Inovação em Psicoterapia
Autora: Gina Voss, *gvoss16@gmail.com*
Orientadora: Inês Sousa.

Título: Estimação de Distribuições Multivariadas na Presença de Censura pela Direita
Autora: Beatriz Carneiro, *beatriz09sampaio@gmail.com*
Orientador: Luis Machado.

Título: Estimação em Pequenos Domínios para Obtenção de Estatísticas de Uso e Ocupação do Solo
Autora: Suelma Pina, *suelmadepina@hotmail.com*
Orientadores: A.Manuela Gonçalves e Pedro Campos.

Título: Modelação de séries temporais de dados oceanográficos
Autor: Bruno Vasconcelos Oliveira Monteiro, *farinellobvom@gmail.com*
Orientadores: Clara Cordeiro e Maria do Rosário Ramos.

PRÉMIOS "ESTATÍSTICO JÚNIOR 2018"

REGULAMENTO

Está aberto, **até 25 de Maio de 2018**, o concurso para atribuição de prémios **"Estatístico Júnior 2018"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2018"** 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 2018"** devem ser alunos do 3.º 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 2017-2018.
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 2018**" a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.º 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 2018"** 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 2018"** poderão não ser atribuídos.
15. O boletim de candidatura e este regulamento podem ser obtidos em
    http://www.spestatistica.pt/BoletimCandidaturaPEJ18.pdf
    http://www.spestatistica.pt/RegulamentoPEJ18.pdf

Apoio Porto Editora

Prémio Iniciação à Investigação

A Sociedade Portuguesa de Estatística institui o prémio Iniciação à Investigação, que premeia trabalho desenvolvido em Probabilidades e Estatística no âmbito de teses de mestrado.

O prémio é regido pelo seguinte regulamento.

1 - O **Prémio Iniciação à Investigação 2018** é constituído por uma quantia de 200 euros e a quota de sócio SPE relativa ao ano 2019.

2 - Ao **Prémio Iniciação à Investigação 2018** podem concorrer teses de mestrado submetidas e defendidas entre Setembro de 2017 e 31 de Julho de 2018, sobre temas de Probabilidades e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição.

3 - Os autores dos trabalhos candidatos ao **Prémio Iniciação à Investigação 2018** devem ser (ou ter sido) alunos de mestrado em alguma instituição portuguesa e não devem ter completado os 30 anos de idade até 31 de Dezembro de 2018.

4 - A candidatura, dirigida à Presidente da SPE, é constituída pelo trabalho concorrente em versão pdf, uma carta de motivação do candidato e uma carta do orientador, sumariando a importância do trabalho na área. Os trabalhos submetidos podem ser escritos em Português ou em Inglês.

5 - A entrega de candidaturas decorre até ao dia 31 de Agosto de 2018. Os elementos constituintes da candidatura devem ser enviados por correio electrónico para spe@spestatistica.pt e uma cópia da tese deve ser enviada por correio para

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 é da responsabilidade da Direcção da SPE.

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

8 - O júri seleccionará até 3 trabalhos vencedores. Os trabalhos galardoados com o **Prémio Iniciação à Investigação 2018** serão apresentados durante a cerimónia de comemoração do aniversário da SPE, em data e local a anunciar.

9 - O júri reserva-se o direito de não atribuir o **Prémio Iniciação à Investigação 2018.**

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

PRÉMIO SPE 2018

REGULAMENTO

Pretendendo estimular a atividade de estudo e investigação científica em Probabilidades e Estatística, a Sociedade Portuguesa de Estatística institui em 2018 o Prémio SPE, regido pelo seguinte regulamento.

Está aberto até 31 de Agosto de 2018 o concurso para atribuição do Prémio SPE 2018, doravante referido como prémio. O prémio é constituído por uma quantia de 1000 euros.

Ao prémio podem concorrer trabalhos originais sobre temas de Probabilidades e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição. O trabalho deverá ser apresentado em português ou em inglês e não poderá exceder 25 páginas A4.

Podem candidatar-se ao prémio sócios da SPE que não completem 35 anos de idade até 31 de Dezembro de 2018 e que não tenham recebido o prémio nas quatro edições anteriores.

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

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.

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

O Júri reserva-se o direito de não atribuir o Prémio SPE 2018.

As candidaturas ao prémio, dirigidas à Presidente da SPE, são constituídas pelos trabalhos concorrentes e pelo *curriculum vitae* dos autores. Podem ser enviadas por correio eletrónico para spe@spestatistica.pt ou, em carta registada, para a morada a seguir indicada. O carimbo do correio valida a data de entrega.

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

A entrega formal do Prémio SPE 2018, com apresentação do trabalho galardoado, terá lugar na cerimónia de comemoração do aniversário da SPE, em data e local a anunciar.

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR