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.