Miguel de Carvalho
School of Mathematics, University of Edinburgh
Miguel.deCarvalho@ed.ac.uk

ESPELHO MEU, ESPELHO MEU: HÁ ALGUÉM QUE GASTE MAIS POR MÊS DO QUE EU?

As estatísticas de finanças pessoais são fundamentais tanto para bancos e consultores financeiros como para os seus clientes. Neste artigo faço um apanhado de um método estatístico que desenvolvi recentemente em colaboração com colegas e que foi motivado por uma questão colocada por uma empresa da fintech:

> *"Como é que comparo estatisticamente as despesas de um dos utilizadores da nossa plataforma com as de outros que ganham aproximadamente o mesmo?"*

A metodologia desenvolvida envolve uma extensão do conceito de regressão não paramétrica através de métodos de núcleo na qual o regressor é censurado.

1. INTRODUÇÃO E MOTIVAÇÃO

A Estatística desempenha um papel fundamental na resolução de problemas reais com os quais são confrontados diariamente várias entidades, tais como, empresas, reguladores, e outros institutos da área económica. O facto de que a mesma metodologia estatística (ex: regressão, teste de hipóteses, máxima verosimilhança) possa ser aplicada recorrentemente para resolver problemas de distintas áreas científicas – tais como climatologia, economia, geologia, medicina, etc. – demonstra a transversalidade e a riqueza da Estatística. Não é portanto surpreendente que a pesquisa científica por novas metodologias estatísticas seja assim uma área extremamente ativa de investigação.

Esta nota resume a análise de um problema de consultoria com o qual fui confrontado no âmbito de uma *comparação estatística indivíduo-grupo*; ou seja, um problema em que o objetivo é comparar estatisticamente um indíviduo *versus* um grupo de indivíduos "semelhantes" num sentido a detalhar em seguida. Uma versão ingénua da questão que motivou esta investigação é referida acima no resumo; a nossa solução para uma versão mais complexa do problema levou à publicação de um artigo no *Journal of the Operational Research Society* [1] e esta nota é baseada nesse artigo.

Uma das primeiras etapas em qualquer problema de consultoria é entender exatamente qual é o problema com que se confronta o cliente. Quando me reuni com o cliente pela primeira vez, o cliente não chegou com o problema formulado como um problema de comparação estatística indivíduo-grupo até porque essa é terminologia que eu próprio desenvolvi em colaboração com os meus coautores. O cliente chegou com uma base de dados enorme, recorria a jargão da *fintech* que não dominávamos e o meu primeiro objetivo foi converter as suas necessidades numa questão simples, como a questão que enuncio no resumo

desta nota. De modo grosseiro, a *fintech* é uma área emergente que recorre ao uso de tecnologias e técnicas inteligentes de automação para desenvolver novas ferramentas e serviços financeiros.

Há questões de consultoria que podem ser resolvidas recorrendo a metodologias estatísticas que já são conhecidas; outras questões requerem um tratamento mais sofisticado e podem levar ao desenvolvimento de novas técnicas. Em seguida, introduzo os métodos que foram desenvolvidos para dar uma resposta às questões do cliente; do ponto de vista conceptual, a nossa abordagem tem algumas ligações com o chamado *F-baricentro* do intervalo $(a, b]$ [4], o qual é definido como

$$b_F = \mathrm{E}\{X \mid X \in (a, b]\}, \tag{1.1}$$

onde se supõe que a variável aleatória X tem valor esperado $\mathrm{E}(X) < \infty$ e que a sua função de distribuição, $F(x) = P(X \leq x)$, é estritamente crescente.

2. METODOLOGIAS DESENVOLVIDAS

2.1 Valor médio comparativo - versão estática

Definição e contexto

O foco da abordagem é semelhante ao contexto de regressão, no sentido em que há que considerar uma resposta (Y) e um regressor (X); sejam $X_0 = x_0$ e $Y_0 = y_0$ os valores fixos de um indivíduo de referência. Embora definamos em seguida o *valor médio comparativo* como um conceito geral, no contexto aplicado de interesse:

- Y representa uma despesa e X denota um rendimento.
- o "indivíduo de referência" é um utilizador de uma plataforma financeira.

O objetivo é comparar o indivíduo de referência com indivíduos semelhantes, ou seja, indivíduos caracterizados por uma covariável X contida numa bola com raio $\delta > 0$ centrada em x_0, $B_\delta(x_0)$; para facilitar a exposição consideremos em seguida apenas um regressor por forma que $B_\delta(x_0) \equiv (x_0 - \delta, x_0 + \delta)$. O *valor médio comparativo* (de nível $\delta > 0$) é definido como

$$\mu_\delta = \mathrm{E}\{Y \mid X \in B_\delta(x_0)\}.^{1} \tag{2.1}$$

Se $\delta \to 0$, obtemos o valor esperado condicional $\mathrm{E}(Y|X = x_0)$ o qual é a base do modelo de regressão linear simples. Conforme se pode ver na equação (2.1), o valor médio comparativo é semelhante ao *F*-baricentro definido em (1.1), mas envolve duas variáveis aleatórias (X, Y).

Inferência

Em seguida discuto um estimador empírico para o valor médio comparativo. Suponhamos que existem $n + 1$ indivíduos na nossa plataforma e que estão disponíveis os seguintes dados $\{X_i, Y_i\}_{i=1}^{n}$, além dos dados do indivíduo de referência, (x_0, y_0). Definimos a *média amostral comparativa* (de nível $\delta > 0$) como

$$\hat{\mu}_\delta = \frac{1}{k} \sum_{i \in A_\delta} Y_i, \tag{2.2}$$

onde $A_\delta = \{i : X_i \in B_\delta(x_0)\}$ será designado por *janela de comparação* e $k \equiv k_{n,\delta} = |A_\delta|$ denota o seu cardinal. Suponhamos que

$$1. \lim_{\delta \to 0} k = |A_0| \quad 2. \lim_{\delta \to \infty} k = n + 1 \quad 3. \lim_{n \to \infty} k = \infty.$$

Com estas hipóteses, a média amostral comparativa tem as seguintes propriedades

$$\begin{aligned} &1. \lim_{\delta \to 0} \hat{\mu}_\delta = y_0, \text{ q.c.} \qquad 2. \lim_{\delta \to \infty} \hat{\mu}_\delta = \bar{Y}, \text{ q.c.} \\ &3. \ {}_{\delta} \xrightarrow[\mathrm{p}]{} \mu_\delta, \text{ com } n \to \infty, \text{ para } \delta > 0, \end{aligned} \tag{2.3}$$

onde "q.c." significa "quase certamente" (ou seja, com probabilidade 1) e "$\xrightarrow[\mathrm{p}]{}$" denota a convergência em probabilidade; a definição destes limites estocásticos pode ser encontrada, por exemplo, em [6].

Em seguida vamos considerar a extensão destas ferramentas para um contexto dinâmico.

2.2 Valor médio comparativo - versão dinâmica

Definição e contexto

Para captar a trajetória temporal do valor médio comparativo ao longo do tempo, desenvolvemos agora uma versão dinâmica. Sejam $\{X_t\}$ e $\{Y_t\}$ processos estocásticos e sejam $\{X_{0,t} = x_{0,t}\}$ e $\{Y_{0,t} = y_{0,t}\}$ os valores fixos do indivíduo de referência. O *valor médio comparativo dinâmico* é uma extensão trivial de (2.1) e é definido como

$$\mu_{\delta,t} = \mathrm{E}\{Y_t \mid X_t \in B_\delta(x_{0,t})\}. \tag{2.4}$$

Inferência

Seja $n_t + 1$ o número total de indivíduos no período t e sejam

$$\{X_{i,1}, Y_{i,1}\}_{i=0}^{n_1}, \ldots, \{X_{i,T}, Y_{i,T}\}_{i=0}^{n_T}$$

os dados de interesse (ou seja, rendimento e despesa). O estimador empírico de (2.4) é

$$_{\delta,t} = \frac{1}{k_t} \sum_{i \in A_{\delta,t}} Y_{i,t} \tag{2.5}$$

onde $A_{\delta,t} = \{i : X_{i,t} \in B_\delta(x_{0,t})\}$e $k_t = |A_{\delta,t}|$; o estimador definido na equação (2.4) será aqui designado por *média amostral comparativa dinâmica* (de nível $\delta > 0$). Trivialmente, o estimador $\hat{\mu}_{\delta,t}$ verifica as seguintes propriedades

$$\begin{aligned} &1.\ \lim_{\delta \to 0} \hat{\mu}_{\delta,t} = y_{0,t}, \text{ q.c.} \quad 2.\ \lim_{\delta \to \infty} \hat{\mu}_{\delta,t} = \bar{Y}_t, \text{ q.c.} \\ &3.\ _{\delta,t} \underset{p}{\to} \mu_{\delta,t}, \text{ com } n_t \to \infty, \text{ para } \delta > 0, \end{aligned} \tag{2.6}$$

onde $\bar{Y}_t = 1/n_t \sum_{i=1}^{n_t} Y_{i,t}$, se considerarmos uma versão dinâmica das hipóteses enunciadas na secção 2.1, ou seja,

$$1.\ \lim_{\delta \to 0} k_t = |A_{0,t}| \quad 2.\ \lim_{\delta \to \infty} k_t = n_t + 1 \quad 3.\ \lim_{n_t \to \infty} k_t = \infty.$$

Suavização

O estimador empírico introduzido na equação (2.5) não é "suave" ao longo do tempo e portanto nesta secção construímos um método para suavizar o comportamento de (2.5); para o efeito, recorremos a métodos de regressão polinomial local [3], os quais correspondem essencialmente a versões locais de métodos de regressão padrão. Recorremos a uma abordagem não paramétrica para visualizar a média amostral comparativa por forma a que seja mais fácil avaliar tendências ao longo do tempo; a suavização visa assim permitir que os resultados sejam mais intuitivos e fáceis de entender por parte de um utilizador.

O modelo de regressão não paramétrico é definido através da relação $\mu_{\delta,t} = m_t(\delta) + \varepsilon_t$, onde ε_t é variável erro independente e identicamente distribuída com $\mathrm{E}\{\varepsilon_t\} = 0$ e $\mathrm{var}\{\varepsilon_t\} = \sigma^2$. A função $m_t(\delta)$ pode ser estimada da seguinte forma

$$\hat{m}_t(\delta) = \frac{\sum_{i=1}^{T} K_h(t-i)\hat{\mu}_{\delta,i}}{\sum_{i=1}^{T} K_h(t-i)}. \tag{2.7}$$

Na equação (2.7), $K_h(\cdot) = K(\cdot/h)/h$, é K uma função de núcleo e $h > 0$ é um parâmetro de suavização (largura de banda) [8]. A largura de banda $h = h_T$ é uma sequência tal que $h \to 0$ e $hT \to \infty$, quando $T \to \infty$. Em termos práticos, a escolha do núcleo tem pouco impacto nas estimativas; no entanto, a escolha da largura de banda é importante pois uma escolha inapropriada pode levar ou a uma suavização excessiva ou insuficiente.

3. CASO REAL

3.1 Descrição do conjunto de dados

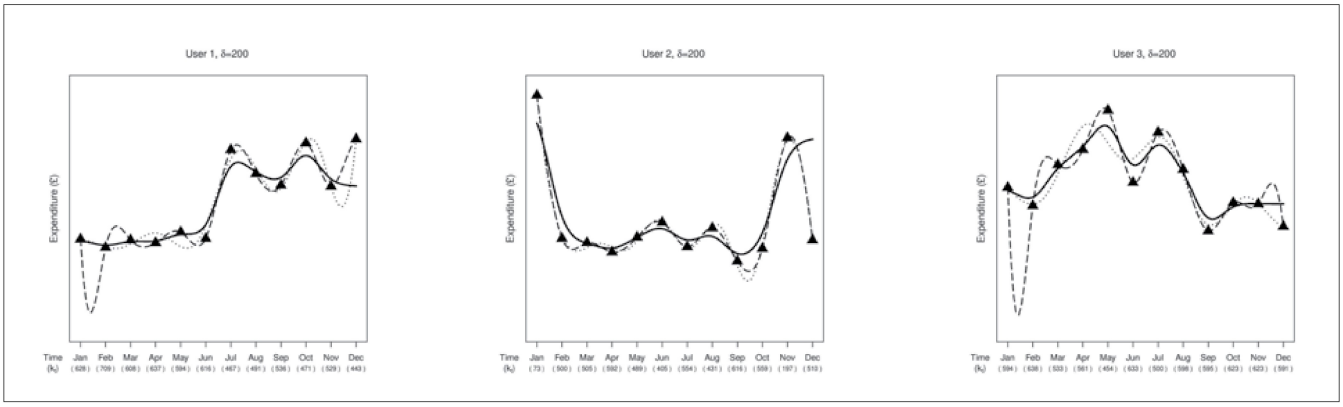
Vamos agora ilustrar a aplicação das metodologias desenvolvidas; para o efeito, vamos considerar três indivíduos de referência – ou seja, três utilizadores de uma plataforma financeira que implementa os métodos da secção 2 – e vamos comparar as suas despesas com as de indivíduos com rendimentos semelhantes. Os dados foram disponibilizados por um provedor de serviços financeiros (Money Dashboard, *www.moneydashboard.com*). Usando a metodologia da secção 2, um utilizador da plataforma tem a possibilidade de comparar as suas despesas ($\{y_{0t}\}$) com as de indivíduos que ganhem $\pm$£δ que esse mesmo utilizador por mês. Os dados a que tivemos acesso para implementar as ferramentas desenvolvidas consistem em rendimentos $X_{i,t}$ e despesas mensais $Y_{i,t}$, em libras (£), para o utilizador i no mês t durante o ano de 2017; tivemos acesso ao perfil de 10 689 utilizadores.

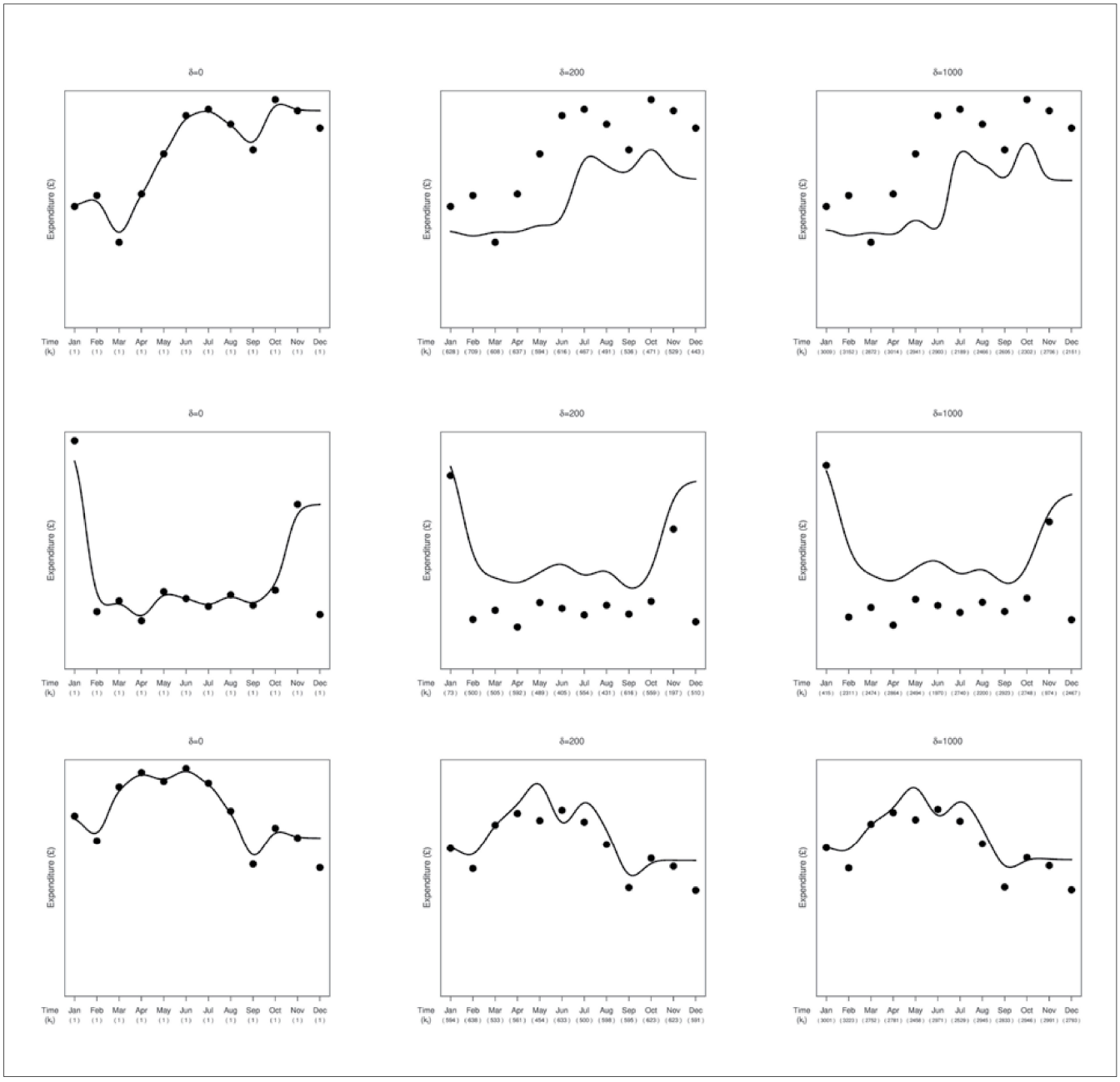
3.2 Análise de dados

Na figura 1 represento a despesa média amostral comparativa dinâmica (2.5) e a sua versão suavizada (2.7) para três utilizadores distintos, considerando $\delta =$ £200; devido a motivos de confidencialidade, não tenho permissão para revelar os valores de y na figura 1. Para comparação, apresento ainda na figura 1 a suavização média amostral comparativa dinâmica que seria obtida através de processos Gaussianos [5] e P-splines [2]; os resultos obtidos com os diferentes métodos de suavização são essencialmente equivalentes. O valor de δ foi escolhido por forma a permitir que os pares de comparação tivessem rendimentos semelhantes aos do indivíduo de referência. Conforme pode ser visto na figura 1, a suavização permite rastrear de uma forma intuitiva para um utilizador a dinâmica das despesas médias comparativas ao longo do tempo.

Na figura 2, contrastamos a despesa média comparativa dinâmica (2.7) com a despesa dos {indivíduos de referência para três valores diferentes de δ. Para ilustrar como podem ser interpretados os resultados obtidos, vamos concentrar a análise no caso $\delta =$ £200. A despesa média comparativa dinâmica (2.7) representa a despesa média dos utilizadores na faixa do rendimento do utilizador de referência $\pm$£200. Os pontos na figura 2 representam as despesas dos utilizadores de referência em cada mês de 2017. Podemos notar como os três utilizadors de referência diferem consideravelmente em termos das suas despesas mensais. Com exceção de março, o utilizador 1 gasta

[1] Todas as comparações nesta nota baseadas em (2.1) são para $\delta > 0$ com $f_X(x) = \mathrm{d}F_X/\mathrm{d}x > 0$ para todo o $x \in B_\delta(x_0)$; a extensão para o caso assimétrico, $B_{\delta_1,\delta_2}(x_0) = (x_0 - \delta_1, x_0 + \delta_2)$ é imediata mas torna a notação mais pesada.


User 1, δ=200
User 2, δ=200
User 3, δ=200
Expenditure (£)
Time
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec


δ=0
δ=200
δ=1000
Expenditure (£)
Time
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

◀Figura 1. Despesa média comparativa dinâmica suavizada: A despesa média comparativa dinâmica (2.5) é representada usando triângulos (▲) e a sua versão suavizada é representada através de uma linha sólida (núcleo, (2.7)), tracejada (processo Gaussiano) e pontilhada (P-spline). O eixo do x representa o tempo e contém ainda informação sobre o número de indivíduos na janela de comparação, k_t.

◀Figura 2. Despesa média comparativa dinâmica suavizada para diferentes janelas de comparação. A despesa média comparativa dinâmica suavizada (2.7) é representada através de uma linha sólida; os pontos (•) representam as despesas mensais de cada utilizador de referência, $y_{0,t}$. O eixo do x representa o tempo e contém ainda informação sobre o número de indivíduos na janela de comparação, k_t. Cada coluna corresponde a uma janela de comparação distinta.

consistentemente mais do que a média de utilizadors com o mesmo rendimento $\pm$£200. O caso oposto é o do utilizador 2, que tende a gastar menos do que a média de utilizadores com rendimentos semelhantes; as exceções são janeiro e novembro, quando os gastos deste utilizador estão alinhados com os dos seus pares. Finalmente as despesas do utilizador 3 estão alinhadas com as da média dos seus pares, com desvios excecionais em fevereiro, setembro e dezembro de 2017.

A análise da figura 2, permite fornecer conselho aos respetivos utilizadores sobre as suas finanças pessoais. Por exemplo, como o utilizador 1 tende a gastar consistentemente mais dinheiro do que os seus pares com rendimento semelhante, o utilizador pode ser considerado em risco de problemas financeiros imediatos ou futuros; pode-se assim fornecer conselhos sobre como reduzir gastos. O utilizador 2, que gasta consistentemente menos do que seus pares com um rendimento semelhante, poderia receber dicas sobre onde investir as suas poupanças. Finalmente, os consultores, ou a própria plataforma, podem ainda enviar avisos ou recomendações ao utilizador 3 se os seus gastos se desviarem substancialmente da média de outros clientes com rendimentos semelhantes.

COMENTÁRIOS FINAIS

A comparação de grupos é um assunto amplamente estudado do ponto de vista estatístico, por exemplo, através de um teste de hipóteses – o qual é introduzido em qualquer curso introdutório de Estatística. Esta nota aborda um problema relacionado mas muito menos estudado do ponto de vista estatístico: a comparação estatística entre um indivíduo *versus* um grupo de comparação.

O método desenvolvido permite comparar numa plataforma financeira as despesas de um utilizador com as de pares que auferem aproximadamente o mesmo rendimento; o método proposto permite aos clientes de uma empresa de *fintech* responderem a questões tais como

> *"Será que gasto muito por mês em comparação com indivídiduos que ganham $\pm\delta$ do que eu?"*

Por último, é importante destacar que ainda que o método tenha sido motivado por um problema no âmbito da *fintech*, a sua generalidade metodológica permite-lhe abordar questões de outras áreas, tais como por exemplo:

> *"Será que a minha frequência cardíaca é elevada em comparação com indivídiduos com $\pm\delta$ anos do que eu?"*

REFERÊNCIAS

[1] Svetloksak, de Carvalho, M., Calabrese R. "Subject--to-group statistical comparison for open banking-type data". *Journal of the Operational Research Society*, 2021, in press.

[2] Eilers, P. H., & Marx, B.D. "Flexible smoothing with B-splines and penalties." *Statistical Science*, 11(2), 89-102, 1996.

[3] Fan, J. *Local Polynomial Modelling and its Applications*. London: Chapman & Hall, 1996.

[4] Hill, T., & Monticino, M. (1998). "Constructions of random distributions via sequential barycenters". *The Annals of Statistics*, 26(4), 1242-1253, 1998.

[5] Rasmussen, C.E., & Williams, C. K. I. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[6] Van der Vaart, A.*Asymptotic Statistics*. Cambridge: Cambridge University Press, 1998.

[7] Ventura, C., Pestana, D., Ivette Gomes, M., Pestana, P., Glossário de Termos Estatísticos, Lisboa: Instituto Nacional de Estatística, 2013.

[8] Wand, M. P. *Kernel Smoothing*. London: Chapman & Hall, 1995.

Nota sobre a tradução de termos técnicos: Os termos técnicos são traduzidos seguindo sempre que possível o Glossário Estatístico Inglês - Português da Sociedade Portuguesa de Estatística (*https://www.spestatistica.pt/glossario)* e o Glossário de Termos Estatísticos da autoria de [7].

Miguel de Carvalho é *reader* em Estatística no Departamento de Matemática da Universidade de Edimburgo. Exerce neste momento as funções de Diretor do Centro de Estatística da Universidade de Edimburgo (*https://centreforstatistics.maths.ed.ac.uk/*) e de Presidente da Sociedade Portuguesa de Estatística (*https://spestatistica.pt*). A sua trajetória e investigação foram condecoradas com vários prémios de prestígio internacional na área da Estatística Matemática (ex: *Lindley Prize - International Society for Bayesian Analysis*) e é ainda editor associado de revistas científicas líderes nessa área, tais como o *Journal of the American Statistical Association* e o *Annals of Applied Statistics*. Para subscrever o canal YouTube de Miguel de Carvalho por favor clicar em *https://www.youtube.com/c/MigueldeCarvalho80*

Secção coordenada pela PT-MATHS-IN, Rede Portuguesa de Matemática para a Indústria e Inovação

pt-maths-in@spm.pt