André Filipe Pedrosa Estrada

ESTIMAÇÃO DE DISTRIBUIÇÕES DE PERDA ATRAVÉS DE MÉTODOS DO NÚCLEO

Dissertação de Mestrado em Métodos Quantitativos em Finanças, orientada pelo Professor Doutor Carlos Manuel Rebelo Tenreiro da Cruz e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia e à Faculdade de Economia da Universidade de Coimbra.

Fevereiro 2018

· U C ·

Universidade de Coimbra

Estimação de distribuições de perda através de métodos do núcleo

André Filipe Pedrosa Estrada

UNIVERSIDADE DE COIMBRA

Mestrado em Métodos Quantitativos em Finanças

Master in Quantitative Methods in Finance

Dissertação de Mestrado | MSc Dissertation

February 2018

Agradecimentos

Ao Professor Carlos Tenreiro pela orientação, disponibilidade e prontidão em todos os momentos,
Ao DMUC pela excelência em todos os momentos do meu percurso académico,
Ao CMUC pela oportunidade única de descobrir,
Ao NEMAT/AAC pelas memórias edificadas,
À AAC pela referência e irreverência,
Aos Amigos pelas palavras de encorajamento,
À Família pelo apoio incondicional,
A todos um grande OBRIGADO!

Resumo

Sendo a função de perda a densidade da variável aleatória real que representa o valor da indeminização paga ao cliente subscritor de um produto de uma seguradora, a sua estimação constitui um problema típico das ciências atuariais. Na sua estimação é pratica corrente a análise dos dados divididos em duas tranches consoante o valor dos mesmos, através da definição de um *threshold*. No entanto, a escolha da melhor combinação de distribuições a ajustar pelos métodos paramétricos revela-se um problema, dado o desconhecimento da verdadeira densidade. Nesta dissertação são apresentadas metodologias unificadas de estimação, onde os estimadores do núcleo surgem como ferramentas essenciais. Primeiramente, é feita a apresentação do estimador do núcleo usual, à qual se alia uma análise do desenvolvimento assintótico dos erros quadrático médio e quadrático médio integrado, bem como da janela ótima de estimação, cujo critério assenta na minimização do critério de erro global. Por sua vez, a existência de distribuições com problemas de regularidade na origem leva à introdução dos núcleos de fronteira. Nesta fase são introduzidas várias alternativas para núcleos de fronteira e respetiva análise dos aspetos mencionados para o primeiro estimador. Wand *et al.* (1991) concluem que a aplicação de uma transformação aos dados e a consecutiva estimação da densidade, baseada na densidade dos dados transformados por métodos do núcleo, se revela eficaz no aumento do desempenho da estimação. Assim, Buch-Larsen *et al.* (2005) introduzem a estimação da função de perda com métodos do núcleo com correção de fronteira bilateral pela aplicação da transformação pela função de distribuição da lei de Champernowne modificada, sendo esta a terceira alternativa apresentada para método do núcleo. Em último, é levado a cabo um estudo de simulação que contrapõe os três métodos do núcleo apresentados, bem como dois núcleos de fronteira distintos, para três distribuições e tamanhos de amostra. A precisão da estimação é avaliada de forma global tendo em conta os erros L_1 e L_2, sendo que o erro $WISE$ surge como medida associada ao desempenho do estimador na cauda da distribuição, dada a sua maior ponderação nesta zona do suporte da variável. Este estudo permite então concluir que os métodos do núcleo que assentam na estimação da densidade dos dados transformados produzem melhores estimativas para a densidade original.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A análise histórica das perdas que uma seguradora aufere face a um determinado tipo de produto segurado revela-se fundamental na sua forma de operar o negócio. A perceção do comportamento dos dados históricos permite não só adequar os preços dos produtos oferecidos aos clientes, mas também potenciar ferramentas para a análise da subsistência da própria seguradora, através de medidas de risco, como é o caso do *Value at Risk*.

Neste âmbito, uma das ferramentas primordiais analisada nas ciências atuariais é a função de perda. Esta função é caracterizada por ser a densidade da variável aleatória real (absolutamente contínua) representativa da indemnização paga ao cliente subscritor de um determinado produto da seguradora, a variável X. Sendo que, teoricamente, a função densidade de uma determinada variável aleatória real caracteriza totalmente a lei de probabilidade associada, esta ferramenta mostra-se bastante pertinente. Assim, o problema em estudo centra-se na proposta de metodologias de estimação da função densidade da variável em estudo, dado um determinado conjunto de observações históricas da mesma.

Geralmente, a este tipo de conjunto de dados, que tomam apenas valores positivos, está associada uma grande massa de probabilidade para valores diminutos, pois são bastante frequentes pequenos sinistros, sendo que as indemnizações de grande valor, embora em número muito inferior, atingem valores potencialmente muito elevados, o que resulta em distribuições de dados com caudas pesadas.

Uma abordagem intuitiva possível seria ajustar diretamente aos dados distribuições de famílias que verificam este tipo de comportamento, como é o caso das leis Lognormal, de Weibull, de Pareto ou Gamma, recorrendo aos métodos usuais de estimação de parâmetros a elas associadas, método da máxima verosimilhança ou método dos momentos. Foi este o ponto de partida para o estudo apresentado por Käärik e Umbleja (2010) e (2011).

As limitações destas metodologias prendem-se com o seu desajuste em certas zonas do suporte. Exemplo disso é o caso da lei Lognormal cuja cauda tende rapidamente para zero, não modelando convenientemente as indeminizações de valores mais elevados. Por outro lado, leis como a de Pareto e a sua versão generalizada apresentam um comportamento monótono decrescente que pode resultar em falhas na modelação dos dados de valor menor. Assim, é prática corrente a análise dos dados divididos em duas tranches - valores de indemnizações reduzidos e substanciais - através do estabelecimento de um limiar divisório dos dados, o *threshold*, permitindo fazer a modelação dos dois tipos de dados de forma independente. A introdução desta metodologia, por Cooray e Ananda (2005), sugere a modelação da densidade de probabilidade das indeminizações através de

uma mistura da lei Lognormal - para as indemnizações de valor reduzido - com a lei de Pareto ajustada à cauda da distribuição. Este novo método viria a ser explorado no trabalho de Käärik e Umbleja (2012), produzindo melhores resultados que os anteriormente obtidos pelos mesmos autores. Adicionalmente, foram estudadas diversas combinações de leis para a estimação da densidade de probabilidade pretendida. Preda e Ciumara (2006) comparam a utilização dos modelos lognormal-Pareto e Weibull-Pareto, obtendo resultados similares. Por outro lado, Teodorescu e Vernic (2006) exploram o modelo exponencial-Pareto e Abu Bakar *et al.* (2015) o modelo Weibull-Burr.

Sendo que à partida não temos conhecimento de qual a melhor lei (ou mistura de leis) de probabilidade a considerar, torna-se pertinente a implementação de um método unificado de estimação da função de perda que não pressuponha qualquer tipo de comportamento aos dados, isto é, um método não-paramétrico. Desta forma, os estimadores do núcleo surgem naturalmente como resposta, sendo que a base do seu funcionamento assenta na análise local das observações contidas num intervalo centrado no ponto em que se calcula o valor da densidade. Assim, ao ser implementado este método em todos os pontos do suporte da variável aleatória real, obtemos uma estimativa da função de perda.

Neste sentido, começaremos por introduzir uma metodologia de estimação de f não paramétrica baseada nos estimadores do núcleo introduzidos por Rosenblatt (1956) e Parzen (1962). Segue-se depois a análise assintótica dos erros local e global obtidos na estimação, através dos erros quadrátrico médio (MSE) e quadrático médio integrado ($MISE$). Desta forma, estaremos em condições de estabelecer limites para os dados a considerar na estimação do valor da densidade num determinado ponto x do suporte, isto é, limitar o intervalo centrado nesse mesmo ponto objeto de análise. Esta limitação, denominada escolha da janela (semi-amplitude do intervalo), é motivada pelo aumento do desempenho da estimação, tentando definir um equilíbrio entre a diminuição do viés e da variância globais do estimador.

Por sua vez, notando problemas de não regularidade de f e das suas derivadas na origem (ponto fronteiro do suporte), patentes em algumas distribuições associadas a problemas de origem atuarial, insurge uma nova ponderação dos dados na região de fronteira. Esta alteração terá como base a introdução de métodos do núcleo com correção de fronteira, através dos núcleos de fronteira que dependem da distância do ponto onde se efetua a estimação à origem. Com a introdução desta segunda metodologia, procede-se novamente a análise assintótica do MSE do estimador. Este estudo permite então notar que poderá ser vantajosa a consideração de núcleos de fronteira de segunda ordem (média nula e variância não nula), pelo que se efetua a análise do erro quadrático médio. Por fim, o cálculo assintótico do $MISE$ nas duas situações potencia de novo um critério de escolha da janela de estimação.

Wand *et al.* (1991) identificam problemas na estimação da densidade com métodos do núcleo, quando a lei de probabilidade apresenta comportamentos muito dispares da lei gaussiana, e propõem como alternativa a aplicação de transformações às observações da variável em estudo e consequente estimação da densidade por métodos do núcleo dos dados transformados. Aplicando uma transformação inversa à densidade estimada, os autores obtêm, por fim, um estimador para a verdadeira densidade dos dados e concluem "*We have found that these transformations can substancially increase the accuracy of kernel estimates of Cauchy and other heavy-tailed densities.*".

Baseados no trabalho de Wand *et al.* (1991), Buch-Larsen *et al.* (2005) sugerem a estimação de f utilizando como tranformação a função de repartição associada à lei de Champernowne modificada,

bem como a utilização de núcleos de fronteira aplicados a ambos os pontos extremos do suporte da variável transformada. Assim, é levado a cabo um breve estudo de algumas características desta lei de probabilidade que justificam a sua utilização e é feita a apresentação do estimador da densidade dos dados transformados com correção de fronteira bilateral. O facto dos dados transformados não apresentarem independência nem distribuição idêntica impossibilita o estudo análogo do *MSE* e do *MISE* do novo estimador, pelo que a escolha da janela surge como adaptação empírica dos critérios apresentados para os estimadores com correção de fronteira. Este novo método acaba por se tornar semi-paramétrico, na medida em que se ajusta uma distribuição aos dados para obtenção dos parâmetros da transformação.

Em última instância, torna-se necessária a implementação de um estudo de simulação para comparar os vários estimadores apresentados, bem como núcleos de fronteira. Nesta fase, o desempenho dos estimadores é analisado face a dados simulados por leis de probabilidade com diversos comportamentos na origem e respetiva cauda. Mais ainda, o tamanho amostral é também um fator a ter em conta pelo que se consideram tamanhos da amostra díspares.

Capítulo 2

Estimador do núcleo da densidade de probabilidade

2.1 Noções básicas

Numa primeira fase, serão introduzidos conceitos e resultados teóricos relevantes para a estimação não paramétrica da função de densidade através do estimador do núcleo. Sejam então f a densidade de probabilidade da variável aleatória real absolutamente contínua e não negativa X, $X_1, X_2, ..., X_n$, com $n \in \mathbb{N}$, observações de X independentes e identicamente distribuídas com essa mesma variável e F a função de distribuição associada. De uma forma inicial, baseado no trabalho de Rosenblatt (1956), Parzen (1962) motiva o estimador do núcleo, começando por notar que

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx\,,$$

uma vez que, através de estimadores da função de distribuição, poder-se-ão obter estimadores para a densidade de probabilidade. Note-se que o integral é tomado no sentido de Lebesgue. De facto, daqui em diante todos os integrais deverão ser analisados na medida referida. Naturalmente, a função de distribuição empírica, $F_n(x)$, dada por

$$F_n(x) = \frac{1}{n} \#\{X_i \leq x, i = 1, ..., n\}$$

surge como estimador para $F(x)$. Adicionalmente, notando é a derivada da função de distribuição através de derivação, é também uma escolha intuitiva

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \tag{2.1}$$

para estimador da densidade f, em x, sendo esta a ideia inicialmente introduzida por Rosenblatt (1956). De facto, a quantidade h deverá ser um função positiva de tal forma que, à medida que o número de observações aumente, a mesma se aproxime de zero, isto é, h deve ser uma função de n que satisfaça

$$\lim_{n \to \infty} h_n = 0\,.$$

Analisando agora a formulação exposta em (2.1), é de notar que

$$f_n(x) = \frac{1}{n}\frac{\sharp\{x-h \leq X_i \leq x+h, i=1,...,n\}}{2h} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),$$

com K dado por

$$K(z) = \begin{cases} \frac{1}{2}, & |z| \leq 1, \\ 0, & |z| > 1. \end{cases}.$$

Note-se agora que esta primeira versão do estimador dito *naive* ou da janela móvel pondera de forma igual todas as observações que caem no intervalo $[x-h, x+h]$. Neste sentido, a consideração de funções K diferentes produz estimadores distintos em que a ponderação das observações é díspar, pelo que estamos na presença de uma classe de estimadores mais abrangente. Naturalmente, são escolhas usuais de funções peso funções positivas e simétricas relativamente à origem, uma vez que ponderam de igual forma obervações equidistantes do ponto onde se pretende estimar a densidade. Definamos agora a família de funções peso, núcleos, com que trabalharemos ao longo desta dissertação.

Definição 1. *Seja $K : \mathbb{R} \longrightarrow \mathbb{R}$ uma função não negativa, simétrica e integrável. Diz-se que K é núcleo se verifica*

$$\int K(z)\, dz = 1\,.$$

É do nosso interesse considerar núcleos que atinjam o seu máximo em zero e que decresçam à medida que nos afastamos deste mesmo ponto, pois as observações mais próximas de x apresentam uma maior ponderação ao invés de observações mais distantes. Mais ainda, dada a formulação do estimador da janela móvel, concentrar-nos-emos em núcleos de suporte $[-1,1]$. Apresentam-se de seguida alguns núcleos patentes na literatura que verificam as condições discutidas.

Tabela 2.1 Exemplos de núcleos

Núcleo	$K(z)$
Uniforme	$\frac{1}{2}\mathbb{1}_{\{\lvert z\rvert \leq 1\}}$
Triangular	$(1-\lvert z\rvert)\mathbb{1}_{\{\lvert z\rvert \leq 1\}}$
Epanechnikov	$\frac{3}{4}(1-z^2)\mathbb{1}_{\{\lvert z\rvert \leq 1\}}$
Quartic	$\frac{15}{16}(1-z^2)^2\mathbb{1}_{\{\lvert z\rvert \leq 1\}}$

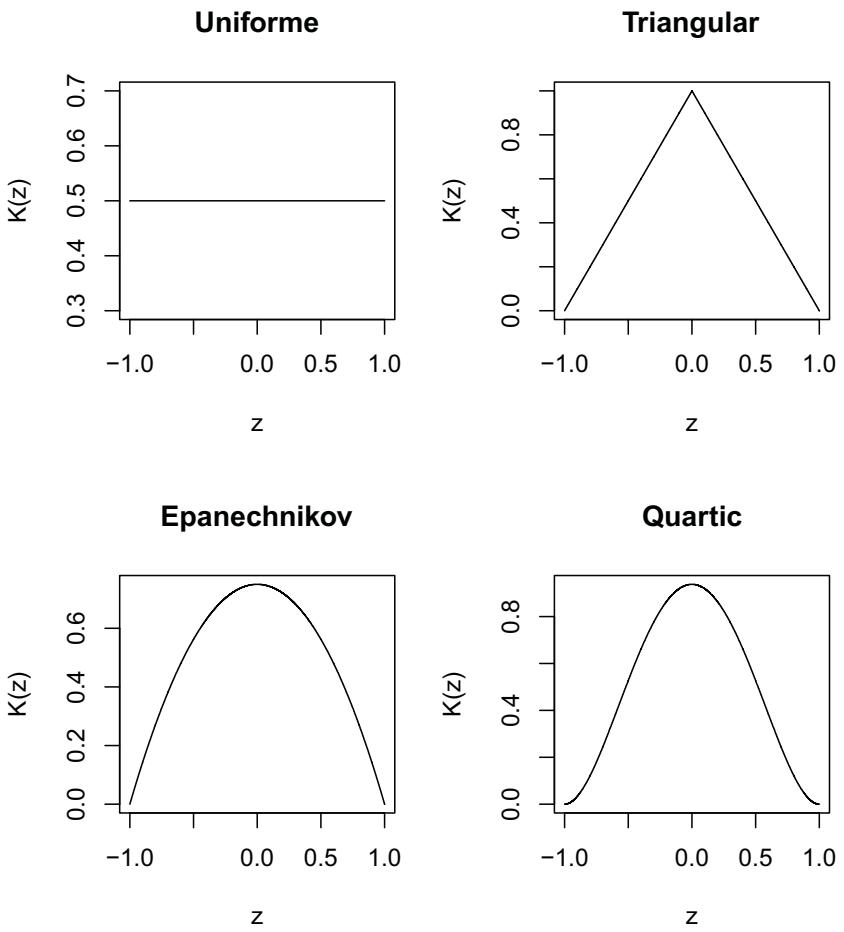


Fig. 2.1 Representações gráficas dos núcleos apresentados

É de salientar que existem outras formas de ponderação dos dados que assentam em núcleos de suporte ilimitado, como é o caso do núcleo gaussiano dado por $K(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}z^2)$. No entanto, a existência de dados de magnitude elevada patente nos problemas em estudo, levam-nos à exclusão deste tipo de núcleos de forma a garantir que a cauda pesada da distribuição não se reflita na estimação da densidade nos pontos do suporte mais próximos da origem.

Adicionalmente, para efeitos demonstrativos posteriores, torna-se necessário definir núcleos de segunda ordem.

Definição 2. *Seja K um núcleo. Considerando, para* $j \in \mathbb{N}$,

$$\mu_j(K) = \int z^j K(z)\, dz,$$

diz-se que K é um núcleo de segunda ordem, se $\mu_2(K)| \neq 0$.

Por fim, introduzamos agora a formulação final do estimador do núcleo a utilizar. Seja K um núcleo de segunda ordem e h a janela considerada, o estimador do núcleo da densidade de probabilidade, $\hat{f}_h$, é dado, para $x \in \mathbb{R}$, por

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) . \tag{2.2}$$

2.2 Medidas de desempenho do estimador

Porquanto estarmos focados no desenvolvimento de metodologias de estimação, torna-se relevante o estabelecimento de critérios de erro que permitam efectuar uma análise coerente do desempenho do estimador. Assim, com o intuito de estudar a precisão do estimador num ponto x do suporte, focar-nos-emos focar no erro quadrático médio, uma vez que esta é a medida mais usual na avaliação do desempenho local de um estimador da densidade. Pode-se definir o erro quadrático médio do estimador $\hat{f}$ da densidade f calculado num ponto x como sendo

$$MSE(\hat{f}_h(x)) = E(\hat{f}_h(x) - f(x))^2 \, .$$

Proceda-se então ao cálculo assintótico do erro quadrático médio do estimador definido em (2.2), na medida em que tomaremos a sucessão h_n a tender para 0. Para tal, introduza-se a seguinte notação. Seja $g : \mathbb{R} \longrightarrow \mathbb{R}$ uma função integrável no seu domínio, denota-se por $R(g)$ o integral do quadrado de g, isto é

$$R(g) = \int g^2(x)\, dx \, .$$

Note-se ainda que neste capítulo, apesar do suporte da variável ser $\mathbb{R}^+$, assumiremos sempre condições de regularidade na reta real.

Teorema 1. *Sejam f com derivadas contínuas até à segunda ordem numa vizinhança de x e $\hat{f}_h$ o estimador de f dado por (2.2). Se $h_n \to 0$ e $nh \to \infty$, então*

$$MSE(\hat{f}_h(x)) = \frac{1}{nh} f(x) R(K) + \frac{h^4}{4} (f''(x))^2 \mu_2^2(K) + o(h^4) + o\left(\frac{1}{nh}\right) \, .$$

Demonstração. Primeiramente, notemos que, para $x \in \mathbb{R}$, podemos escrever

$$MSE(\hat{f}_h(x)) = V(\hat{f}_h(x)) + (E(\hat{f}_h(x)) - f(x))^2 \, , \tag{2.3}$$

ou seja, o erro quadrático médio pode ser decomposto em dois termos, um de viés e um de variância. Assim, para a determinação desta medida de erro, comecemos por calcular a variância do estimador $\hat{f}(x)$. Note-se que $X_1, X_2, ..., X_n$ são independentes e identicamente distribuídas com X. Logo,

$$V(\hat{f}_h(x)) = \frac{1}{nh^2} Var\left(K\left(\frac{x-X}{h}\right)\right) = \frac{1}{nh^2}\left[E\left(K^2\left(\frac{x-X}{h}\right)\right) - E^2\left(K\left(\frac{x-X}{h}\right)\right)\right] \, ,$$

ou seja,

$$V(\hat{f}_h(x)) = \frac{1}{nh^2}\left[\int K^2\left(\frac{x-y}{h}\right) f(y) dy - \left(\int K\left(\frac{x-y}{h}\right) f(y) dy\right)^2\right] \, .$$

Tomemos agora a mudança de variável definida por $z = \frac{x-y}{h}$. Tem-se

$$V(\hat{f}_h(x)) = \frac{1}{nh} \int_{-1}^{1} K^2(z) f(x - zh) dz - \frac{1}{n}\left(\int_{-1}^{1} K(z) f(x - zh) dz\right)^2 \, . \tag{2.4}$$

Nesta fase, apliquemos o Teorema da Convergência Dominada de Lebesgue (TCD) (ver no Anexo A - Teorema 10) à sucessão de funções $g_n^i(z) = K^i(z)f(x-zh)$, $n \in \mathbb{N}$, $i = 1,2$. Primeiramente, note-se que, atendendo à continuidade de f em x, podemos estabelecer

$$\lim_n g_n^i(z) = K^i(z)f(x)\,.$$

Adicionalmente, atendendo às condições de regularidade de f, esta função é contínua numa vizinhança de x, V_x, e, portanto, é limitada nessa mesma vizinhança. Desta forma,

$$|g_n^i(z)| \leq \sup_{y\in V_x} |f(y)||K^i(z)| = r(z)\,,$$

com r integrável não negativa. Verificadas as hipóteses do TCD, podemos concluir que

$$\lim_{n\to\infty} \int K^i(z)f(x-zh)\,dz = f(x)\int K^i(z)\,dz\,,$$

de onde deduzimos que

$$V(\hat{f}_h(x)) = \frac{1}{nh}R(K)f(x) + o\left(\frac{1}{nh}\right)\,, \tag{2.5}$$

à medida que $n \to \infty$.

Foquemos agora a nossa atenção para o termo de viés presente no desenvolvimento do erro quadrático médio, calculando a esperança do nosso estimador. Analogamente ao caso anterior, pelas propriedades das observações $X_1, X_2, ..., X_n$ e tomando a mesma mudança de variável, vem

$$E(\hat{f}_h(x)) = \int_{-1}^{1} K(z)f(x-zh)\,dz\,. \tag{2.6}$$

Agora, dada a continuidade das derivadas de f em x, pode-se desenvolver a fórmula de Taylor (ver Anexo A - Teorema 11) até à ordem 1 em torno do ponto x, tendo-se

$$f(x-zh) = f(x) - zhf'(x) + z^2h^2\int_0^1 (1-t)f''(x-tzh)\,dt\,.$$

Assim, substituindo $f(x-zh)$ em (2.6), pelo desenvolvimento acima e notando que K é núcleo de segunda ordem, vem

$$E(\hat{f}_h(x)) = f(x) + h^2\int_{-1}^{1}\int_0^1 z^2K(z)f''(x-tzh)(1-t)\,dt\,dz\,. \tag{2.7}$$

Desta vez, considerando a sucessão de funções $g_n(z,t) = z^2K(z)f''(x-tzh)(1-t)$, pela continuidade da segunda derivada de f em x, podemos estabelecer

$$lim_n g_n(z,t) = z^2K(z)f''(x)(1-t)\,.$$

Desta forma, dada a continuidade de f'' numa vizinhança de x, V_x, a majoração de $|g_n(z,t)|$ por

$$z^2|K(z)|\sup_{y\in V_x}|f''(y)|(1-t) = r(z,t)$$

- função integrável não negativa - permite-nos aplicar o TCD e concluir que

$$E(\hat{f}_n(x)) = f(x) + \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2) \tag{2.8}$$

de onde se deduz que

$$\left(E(\hat{f}_h(x)) - f(x)\right)^2 = \frac{h^4}{4}(f''(x))^2\mu_2^2(K) + o(h^4) \tag{2.9}$$

De facto, de (2.3), (2.5) e (2.9), conclui-se a demonstração do teorema. □

O erro quadrático médio revela-se uma ferramenta essencial na análise da precisão pontual do estimador considerado. No entanto, torna-se especialmente necessário o estabelecimento de um critério que avalie de forma global o desempenho do mesmo, o que não acontece com o *MSE*, na medida em que se revela um critério de erro local. Nesta linha, o erro quadrático médio integrado, *MISE*, surge como alternativa, uma vez que resulta da integração do erro quadrático médio de $\hat{f}$ sobre o suporte da variável, funcionando como uma ferramente de análise global da precisão do estimador. Isto é,

$$MISE(\hat{f}_h) = \int MSE(\hat{f}_h(x))\, dx\,.$$

Temos agora como objetivo a obtenção de expressões assintóticas para o *MISE* do nosso estimador. De forma a cumprir o exposto, surge no seguinte resultado.

Teorema 2. *Sejam $f \in C^2(\mathbb{R})$ com f'' limitada e integrável, e $\hat{f}_h$ estimador de f dado por (2.2). Se $h_n \to 0$ e $nh \to \infty$, então*

$$MISE(\hat{f}_h) = \frac{1}{nh}R(K) + \frac{h^4}{4}R(f'')\mu_2^2(K) + o(h^4) + o\left(\frac{1}{nh}\right)$$

Demonstração. Primeiramente, atendendo à decomposição do *MSE*, em (2.3), é válida a seguinte expressão

$$MISE(\hat{f}_h) = \int V(\hat{f}_h(x))\, dx + \int (E(\hat{f}_h(x)) - f(x))^2\, dx\,. \tag{2.10}$$

Comecemos por calcular $\int V(\hat{f}_h(x))\, dx$. De (2.4), podemos estabelecer

$$\int V(\hat{f}_h(x))\, dx = \frac{1}{nh}\int\int_{-1}^{1} K^2(z) f(x - zh) dz\, dx - \frac{1}{n}\int \left(\int_{-1}^{1} K(z) f(x - zh) dz\right)^2 dx\,.$$

Trocando a ordem de integração da primeira parcela, à luz do Teorema de Fubini (ver Anexo A, Teorema 12), vem

$$\int V(\hat{f}_h(x))\, dx = \frac{1}{nh}R(K) - \frac{1}{n}\int \left(\int_{-1}^{1} K(z) f(x - zh) dz\right)^2 dx\,.$$

Desta, convertendo a última parcela num integral triplo, tem-se

$$\int V(\hat{f}_h(x))\,dx = \frac{1}{nh}R(K) - \frac{1}{n}\int\int_{-1}^{1}\int_{-1}^{1} K(z)K(y)f(x-zh)f(x-yh)\,dy\,dz\,dx\,.$$

Efectuando a mudança de variável $u = x - zh$, obtemos

$$\int V(\hat{f}_h(x))\,dx = \frac{1}{nh}R(K) - \frac{1}{n}\int\int_{-1}^{1}\int_{-1}^{1} K(z)K(y)f(u)f(u+zh-yh)\,dy\,dz\,du\,.$$

Considerando $g_n(z,y,u) = K(z)K(y)f(u)f(u+zh-yh)$, tem-se $\lim_n g_n(z,y,u) = K(z)K(y)f^2(u)$. Mais ainda, sendo f uma função contínua em $\mathbb{R}$, é também limitada numa vizinhança de $u+zh-yh$, $V_{u+zh-yh,\infty}$. Logo

$$|g_n(z,p,u)| \leq \sup_{y\in V_{u+zh-yh}} |f(y)||K(z)||K(p)|f(u) = r(z)\,,$$

com r integrável não negativa. Portanto, aplicando o TCD, tem-se

$$\int V(\hat{f}_h(x))\,dx = \frac{1}{nh}R(K) - \frac{1}{n}R(f) + o\left(\frac{1}{nh}\right) = \frac{1}{nh}R(K) + o\left(\frac{1}{nh}\right)\,. \tag{2.11}$$

Calculemos agora a parcela respeitante à integração do termo de viés. De (2.7), vem

$$\int (E(\hat{f}_h(x)) - f(x))^2\,dx = h^4 \int \left(\int_{-1}^{1}\int_{0}^{1} z^2K(z)(1-t)f''(x-tzh)\,dt\,dz\right)^2 dx\,.$$

Recorrendo novamente ao desdobramento em integral múltiplo, obtemos

$$\begin{aligned}&\int (E(\hat{f}_h(x)) - f(x))^2\,dx = \\ &h^4 \int\int_{-1}^{1}\int_{-1}^{1}\int_{0}^{1}\int_{0}^{1} z^2y^2K(z)K(y)(1-t)(1-s)f''(x-tzh)f''(x-syh)\,ds\,dt\,dy\,dz\,dx\,.\end{aligned} \tag{2.12}$$

Tome-se a mudança de variável definida por $u = x - tzh$. Ora,

$$\begin{aligned}&\int (E(\hat{f}_h(x)) - f(x))^2\,dx = \\ &h^4 \int\int_{-1}^{1}\int_{-1}^{1}\int_{0}^{1}\int_{0}^{1} z^2y^2K(z)K(y)(1-t)(1-s)f''(u)f''(u+tzh-syh)\,ds\,dt\,dy\,dz\,du\,.\end{aligned}$$

Com vista à utilização do TCD, considere-se a sucessão de funções

$$g_n(z,y,t,s,u) = z^2y^2K(z)K(y)(1-t)(1-s)f''(u)f''(u+tzh-syh)\,,$$

com $n \in \mathbb{N}$. Notemos que, dada a continuidade de f'' em $\mathbb{R}$,

$$\lim_n g_n(z,p,t,s,u) = z^2y^2K(z)K(y)(1-t)(1-s)(f''(u))^2$$

e que, uma vez que f'' é limitada, podemos estabelecer

$$|g_n(z,p,t,s,u)| \leq 2z^2p^2K(z)K(p)|f''(u)|||f''||_\infty = r(z,p,t,s,u)\,,$$

com r integrável não negativa. Desta forma, vale a seguinte expressão

$$\int (E(\hat{f}_h(x)) - f(x))^2\,dx = \frac{h^4}{4}\mu_2^2(K)R(f'') + o(h^4) \tag{2.13}$$

Por fim, de (2.10), (2.11) e (2.13), sai imediatamente o resultado. □

2.3 Escolha da janela

Com a obtenção de uma expressão para o *MISE* do estimador, medida de erro global de estimação considerada, surge a oportunidade de definir um critério de escolha da janela. A determinação de h revela-se particularmente essencial, uma vez que desta quantidade dependerá a qualidade da estimação. Nesta ótica, a escolha de uma janela demasiado elevada pode levar a uma estimação grosseira da função de perda, ou seja, um estimador com viés demasiado elevado. Neste caso, a cauda pesada dos dados poderá provocar mau ajustamento aos dados de menor valor. Por outro lado, a escolha de uma janela significativamente menor provocará eventualmente *oversmoothing* no ajustamento ao conjunto de dados e consecutivo aumento da variabilidade do estimador, não fornecendo uma estimação realista da função pretendendida. Neste sentido, o critério de escolha de h deverá ser fruto de um critério rigoroso com objetivo de melhorar a precisão da estimação.

Considere-se então o erro quadrático médio integrado assintótico, *AMISE*, definido como a soma dos dois termos mais significativos do *MISE* do estimador, isto é,

$$AMISE(\hat{f}_h) = \frac{1}{nh}R(K) + \frac{h^4}{4}\mu_2^2(K)R(f'')\,.$$

Definida esta medida, podemos tomar como critério de escolha da janela assintoticamente ótima o valor de h que minimiza o *AMISE*, ou seja, que assintoticamente diminui tanto o viés como a variância do estimador. Desta forma, surge o seguinte resultado.

Teorema 3. *Nas condições do Teorema 2, a janela assintoticamente ótima, h_{opt}, é dada por*

$$h_{opt} = \left(\frac{R(K)}{n\mu_2^2(K)R(f'')}\right)^{1/5}.$$

Note-se que o critério obtido pressupõe o conhecimento de $R(f'')$, ou seja, da verdadeira densidade. Este impedimento pode ser contornado ao obtermos uma aproximação. Desta forma, poderemos ajustar uma distribuição com comportamento semelhante, como é o caso das lei lognormais, pois através dos dados é relativamente simples calcular os parâmetros que a definem, através do método da máxima verosimilhança. Mais ainda, torna-se desaconselhável o uso de distribuições de referência em que o cálculo dos parâmetros resulte de métodos numéricos, através da resolução de sistemas de equações não lineares, como é o caso da de Weibull, uma vez que o erro associado à determinação dos parâmetros da lei considerada será transportado para a janela assintoticamente ótima.

Capítulo 3

Estimador do núcleo da densidade de probabilidade com correção de fronteira

3.1 Motivação

Estabelecidos os resultados do capítulo anterior, é de notar que algumas densidades típicas nos problemas desta índole revelam problemas de continuidade na origem - o caso da lei de Pareto ou da lei de Weibull para alguns valores dos seus parâmetros - ou até de derivadas não limitadas uma vez que, por exemplo, no caso da lei de Weibull, determinadas escolhas de parâmetros fazem com que $\lim_{x\to 0} f'(x) = -\infty$. Estas situações constituem, portanto, uma violação às hipóteses estabelecidas nos Teoremas 1 e 2. Assim, torna-se especialmente interessante a análise do critério de erro local estabelecido (MSE) na origem, uma vez que é o ponto extremo do suporte e são recorrentes as funções densidade pouco regulares nestes pontos. Comecemos por analisar a esperança matemática de $\hat{f}_h$ em 0. Ora, atendendo à simetria do núcleo K e estabelecendo a mudança de variável $z = y/h$, tem-se

$$E(\hat{f}_h(0)) = \frac{1}{h}\int_0^{+\infty} K\left(\frac{-y}{h}\right) f(y)\,dy = \int_0^1 K(z) f(zh)\,dz\,.$$

Logo, fazendo agora a expansão de Taylor de ordem 1 em torno da origem com resto integral, sai

$$E(\hat{f}_h(0)) = \frac{1}{2} f(0) + h f'(0)\int_0^1 zK(z) + h^2 \int_0^1 \int_0^1 z^2 K(z) f''(tzh)(1-t)\,dt\,dz\,.$$

De facto, torna-se evidente que a existência de densidades típicas nas ciências atuariais pouco regulares na origem inviabiliza, em alguns casos, o desenvolvimento obtido para o erro quadrático médio na origem. Na verdade, Tenreiro ([15], pg. 96) aponta a pouca concentração de observações em $[x-h, x+h]$, para $x \in]0,h]$, como fundamento para a fraca estimação da densidadde neste ponto. Logo, surge a necessidade de repensar a ponderação que é feita no intervalo $]0,h]$, uma vez que para valores de x superiores a h os desenvolvimentos obtidos permanecem válidos. Neste sentido, os estimadores do núcleo com correção de fronteira manifestam-se especialmente úteis, uma vez que dependem da distância do ponto em estudo ao extremo do suporte de X. Introduza-se então a noção de núcleo de fronteira.

Definição 3. *Seja K^L uma função real de variáveis (z,α) reais definida sobre $\mathbb{R}\times]0,1]$. Diz-se que, para todo o $\alpha\in]0,1]$, K^L é* ***núcleo de fronteira*** *se é limitada e verifica*

$$\int K^L(z,\alpha)\,dz = 1\,.$$

Notemos que, para $x=\alpha h$, com $\alpha\in]0,1]$, pretendemos efetuar uma ponderação específica em $]0,x+h]$. Notando que os estimadores do núcleo assentam na soma do núcleo em $\frac{x-X_i}{h}$, $i\in\{1,...,n\}$ é natural que os núcleos de fronteira sejam de suporte contido em $[-1,\alpha]$. Assim, introduz-se o estimador do núcleo com correção de fronteira da função densidade da variável X, com base em $X_1,X_2,...,X_n$,

$$\hat{f}_h^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K_{x,h}\left(\frac{x-X_i}{h}\right)\,, \tag{3.1}$$

com

$$K_{x,h}(z) = \begin{cases} K^L(z,x/h), & 0<x\le h,\\ K(z), & x>h. \end{cases}$$

onde $K(\cdot)$ e $K^L(\cdot,\alpha)$ são núcleo de segunda ordem e núcleo de fronteira sobre $[-1,\alpha]$, respetivamente.

Definição 4. *Para cada $\alpha\in]0,1]$, K^L é um núcleo de segunda ordem com suporte contido no intervalo $[-1,\alpha]$ se*

$$\int K^L(z,\alpha)\,dz = 1,\ \int zK^L(z,\alpha)\,dz = 0\ e\ \int z^2K^L(z,\alpha)\,dz \neq 0\,.$$

São várias as metodologias possíveis para a construção dos núcleos de fronteira sobre $[-1,\alpha]$. Por um lado, Gasser e Müller (1979) determinam um núcleo de fronteira de segunda ordem que responde à minimização do termo de variância no *MSE* do estimador com correção de fronteira a ser calculado adiante. Surge como solução o núcleo de fronteira sobre $[-1,\alpha]$ e respetiva representação gráfica.

$$K_1^L(z,\alpha) = \frac{1}{\alpha+1}\left(1+3\left(\frac{1-\alpha}{1+\alpha}\right)^2+6\frac{1-\alpha}{(1+\alpha)^2}z\right)\mathbb{1}_{\{-1\le z\le\alpha\}}$$

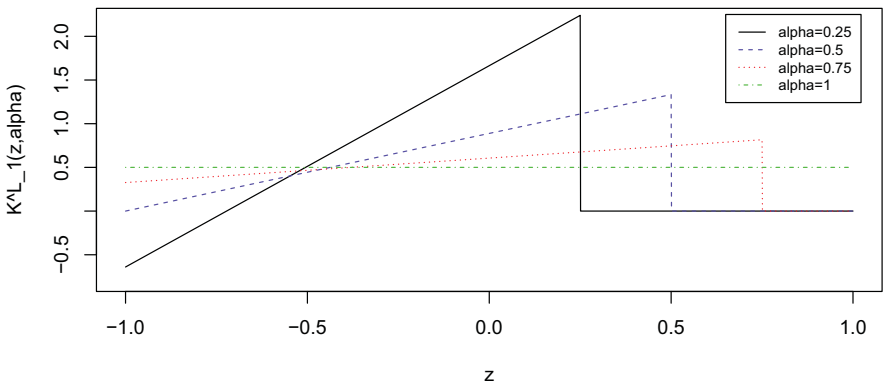


Fig. 3.1 K_1^L para diferentes valor de α

Por outro, os autores apresentam uma metodologia de construção de núcleos de fronteira que tem como base o núcleo utilizado, K. Surge então um núcleo de fronteira de segunda ordem sobre $[-1,\alpha]$ resultante da multiplicação do núcleo por uma função linear. A solução encontrada é então

$$K_2^L(z,\alpha) = (A_\alpha(K) + B_\alpha(K)z)K(z)\mathbb{1}_{\{-1\leq z\leq\alpha\}}$$

em que as quantidades A_α e B_α são determinadas, definindo

$$\mu_{k,\alpha}(K) = \int_{-1}^{\alpha} z^k K(z)\,dz,$$

por

$$A_\alpha(K) = \frac{\mu_{2,\alpha}(K)}{\mu_{0,\alpha}(K)\mu_{2,\alpha}(K) - \mu_{1,\alpha}^2(K)} \text{ e } B_\alpha(K) = \frac{-\mu_{1,\alpha}(K)}{\mu_{0,\alpha}(K)\mu_{2,\alpha}(K) - \mu_{1,\alpha}^2(K)},$$

garantindo desta forma que K_2^L é de segunda ordem. Na figura seguinte, pode-se notar o comportamento do núcleo de fronteira sobre $[-1,\alpha]$ com base no núcleo de Epanechnikov para os diferentes valores de α considerados.

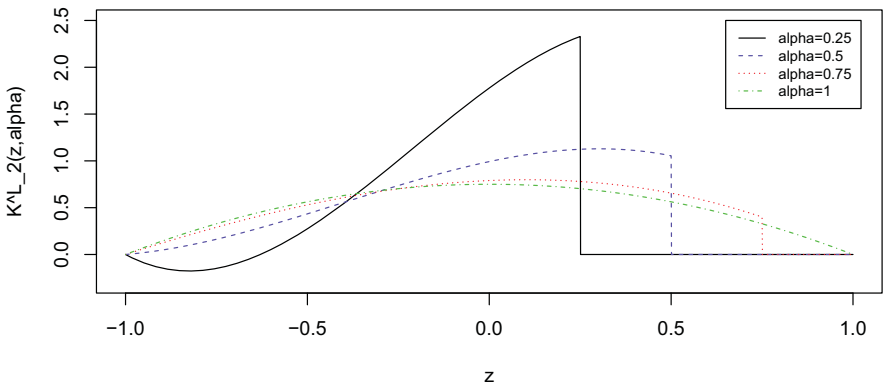


Fig. 3.2 K_2^L para diferentes valor de α

Analisando as representação do núcleos de fronteira sobre $[-1,\alpha]$ K_1^L e K_2^L para $\alpha = 0.25$, é evidente que para alguns valores de x em $]0,h]$ é possível que sejam produzidas estimativas negativas para $f(x)$. Nestes casos, notando que $f(x) \geq 0$, para todo o x em $\mathbb{R}^+$, dever-se-á considerar $\hat{f}_h^*(x)$ nulo.

Por sua vez, Buch-Larsen *et al.* (2005) introduz como núcleo de fronteira o núcleo usual corrigido através da divisão pelo integral do núcleo considerado, entre $[-1,\alpha]$, assegurando assim que K^L integra 1 sobre o seu domínio. Este núcleo de fronteira sobre $[-1,\alpha]$ assume a seguinte formulação:

$$K_3^L(z,\alpha) = \frac{K(z)}{\int_{-1}^{\alpha} K(z)\,dz}\mathbb{1}_{\{-1\leq z\leq\alpha\}}.$$

Por fim, considerando K o núcleo de Epanechnikov, resta notar a expressão gráfica de K_3^L abaixo, face aos diferentes valores da variável α considerados.

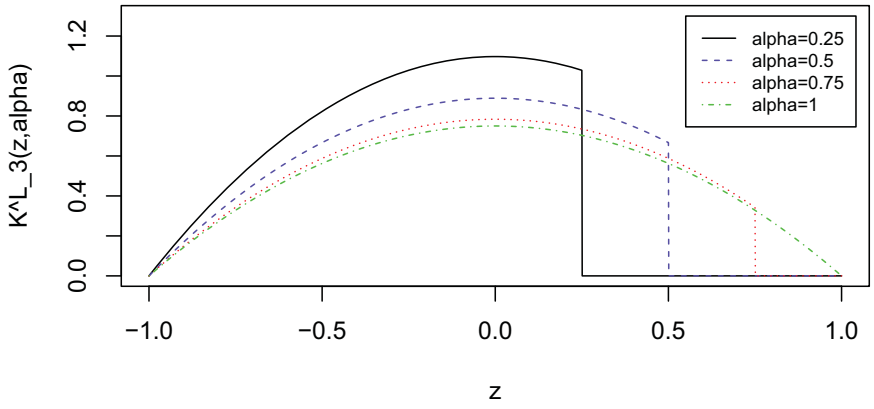


Fig. 3.3 K_3^L para diferentes valor de α

3.2 Medidas de desempenho do estimador com correção de fronteira

A introdução de uma nova metodologia de estimação da densidade faz com que a obtenção de expressões para os critérios de erro volte a ser necessária por forma a analisar os erros de estimação obtidos. Assim, comecemos, à semelhança do capítulo anterior, por apresentar o resultado referente ao erro quadrático médio.

Teorema 4. *Sejam a restrição de f ao intervalo $[0,+\infty[$ com derivada contínua numa vizinhança de x, e $\hat{f}_h^*$ o estimador de f dado por (3.1). Se $h_n \to 0$ e $nh \to \infty$, então, para $x = \alpha h$, com $\alpha \in]0,1]$,*

$$MSE(\hat{f}_h^*(x)) = \frac{1}{nh} f(x) R(K^L(\cdot,\alpha)) + h^2 (f'(x))^2 \mu_1^2(K^L(\cdot,\alpha)) + o\left(\frac{1}{nh}\right) + o(h^2),$$

com $\mu_1(K^L(\cdot,\alpha) = \int z K^L(z,\alpha)\, dz$ e $R(K^L(\cdot,\alpha)) = \int (K^L(z,\alpha))^2\, dz$.

Demonstração. Dada a decomposição do MSE, em (2.3), comecemos pelo cálculo da variância do estimador. Atendendo às propriedades das observações $X_1, X_2, ..., X_n$ e ao facto de f ser uma densidade sobre $\mathbb{R}^+$, sai

$$\begin{aligned} V(\hat{f}_h^*) &= \frac{1}{nh^2} V\left(K^L\left(\frac{x-X}{h}, \frac{x}{h}\right)\right) \\ &= \frac{1}{nh^2}\left[\int_0^\infty K^L\left(\frac{x-y}{h}, \frac{x}{h}\right)^2 f(y)\, dy - \left(\int_0^\infty K^L\left(\frac{x-y}{h}, \frac{x}{h}\right) f(y)\, dy\right)^2\right]. \end{aligned}$$

Tomando a mudança de variável definida por $z = \frac{x-y}{h}$, tem-se

$$V(\hat{f}_h^*) = \frac{1}{nh}\int_{-\infty}^{\frac{x}{h}} K^L\left(z, \frac{x}{h}\right)^2 f(x-zh)\,dz - \frac{1}{n}\left(\int_{-\infty}^{\frac{x}{h}} K^L\left(z, \frac{x}{h}\right) f(x-zh)\,dz\right)^2 . \tag{3.2}$$

Adicionalmente, x pode-se escrever na forma αh, com $\alpha \in]0,1]$. Ora, retomando o suporte de K^L, $[-1,\alpha]$, sai

$$V(\hat{f}_h^*) = \int_{-1}^{\alpha} K^L(z,\alpha)^2 f(x-zh)\,dz - \frac{1}{n}\left(\int_{-1}^{\alpha} K^L(z,\alpha) f(x-zh)\,dz\right)^2 .$$

Agora, a integração sobre $[-1,\alpha]$, em relação a z, permite-nos estabelecer $0 \leq x - zh \leq (\alpha+1)h$. Desta forma, através da continuidade de f em x e notando que $h_n \to 0$, da aplicação do TCD à sucessão de funções $g_n^i(z) = (K^L)^i(z,\alpha) f(x-zh)$, $i = 1,2$, vem

$$V(\hat{f}_h^*(x)) = \frac{1}{nh} f(x) R(K^L(\cdot,\alpha)) + o\left(\frac{1}{nh}\right) . \tag{3.3}$$

Atente-se agora ao termo de viés do *MSE*, calculando primeiramente a esperança do estimador. Ora, as propriedades das observações e o suporte de f permitem-nos estabelecer

$$E(\hat{f}_h^*(x)) = \frac{1}{h}\int_0^{\infty} K^L\left(\frac{x-y}{h}, \frac{x}{h}\right) f(y)\,dy.$$

De novo, considere-se $z = \frac{x-y}{h}$. Logo,

$$E(\hat{f}_h^*(x)) = \int_{-\infty}^{\frac{x}{h}} K^L\left(z, \frac{x}{h}\right) f(x-zh)\,dz.$$

Notando que $x = \alpha h$, com $\alpha \in]0,1]$, e que $K^L(z,\alpha)$ tem suporte contido em $[-1,\alpha]$, podemos estabelecer

$$E(\hat{f}_h^*(x)) = \int_{-1}^{\alpha} K^L(z,\alpha) f(x-zh)\,dz. \tag{3.4}$$

Sendo f' contínua em x, vale o seguinte desenvolvimento de $f(x-zh)$ pela Fórmula de Taylor

$$f(x-zh) = f(x) - \int_0^1 zhf'(x-tzh)\,dt\,. \tag{3.5}$$

Substituindo o desenvolvimento acima em (3.4) e notando que $\int_{-1}^{\alpha} K^L(z,\alpha)\,dz = 1$, vem

$$E(\hat{f}_h^*(x)) = f(x) - h\int_{-1}^{\alpha}\int_0^1 zK^L(z,\alpha) f'(x-tzh)\,dt\,dz\,.$$

Agora, a utilização do TCD na sucessão de funções $g_n(t,z) = zK^L(z,\alpha) f'(x-tzh)$ permite estabelecer que

$$\int_{-1}^{\alpha}\int_0^1 zK^L(z,\alpha) f'(x-tzh)\,dt\,dz = f'(x)\mu_1(K^L(\cdot,(\alpha)) + o(1)\,,$$

uma vez que $0 \leq x - tzh \leq (\alpha+1)h$ e, portanto, estamos numa região de continuidade de f', pois $h_n \to 0$. De facto,

$$E(\hat{f}_h^*(x)) = f(x) - hf'(x)\mu_1(K^L(\cdot,\alpha)) + o(h)\,,$$

ou seja,

$$\left(E(\hat{f}_h^*(x)) - f(x)\right)^2 = h^2(f'(x))^2\mu_1^2(K^L(\cdot,\alpha)) + o(h^2)\,. \tag{3.6}$$

Por fim, de (2.3), (3.3) e (3.6), podemos concluir a demonstração. □

Resta ainda ressalvar que, para $x > h$, como o estimador com correção de fronteira coincide com o estimador do núcleo na sua primeira versão, neste caso, voltam a valer os desenvolvimentos do *MSE* estabelecidos no Capítulo 2.

Face ao desenvolvimento obtido, apresenta-se agora uma representação gráfica da evolução do termo $R(K^L(\cdot,\alpha))$ referente à variância do estimador com correção de fronteira, para $x = \alpha h$, com $\alpha \in]0,1]$.

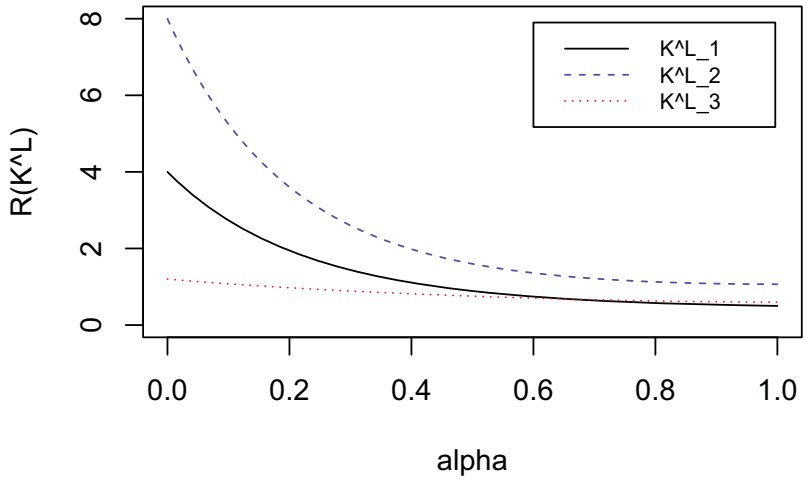


Fig. 3.4 Evolução de $R(K^L(\cdot,\alpha))$, para K_i^L, $i = 1,2,3$, em função de α

Acerca do comportamento do termo de variância acima exposto para os núcleos de fronteira sobre $[-1,\alpha]$ apresentados, é de notar que $R(K_1^L) < R(K_2^L)$, para qualquer valor de α em $]0,1]$. Esta observação é facilmente justificada, uma vez que K_1^L foi determinado por forma a minimizar o termo de variância do *MSE*, na classe dos núcleos de fronteira sobre $[-1,\alpha]$ de segunda ordem. Mais ainda, a consideração do núcleo de fronteira sobre $[-1,\alpha]$ proposto por Buch-Larsen *et al.* (2005), de uma forma geral, apresenta um termo associado à variabilidade do estimador menor do que os restantes. No entanto, uma análise semelhante para o termo $\mu_1(K^L(\cdot,\alpha))$ produziria resultados contrários, uma vez que os dois primeiros núcleos de fronteira são de segunda ordem, o que não acontece com K_3^L.

Estabelecido o resultado referente ao *MSE* do estimador com correção de fronteira, surge novamente a necessidade de estabelecer um resultado análogo ao teorema 2 para o novo estimador. Neste sentido, segue o seguinte resultado.

Teorema 5. *Sejam a restrição de f ao intervalo $[0,+\infty[$ com derivadas contínuas até à segunda ordem, f' limitada e $f'(0)\neq 0$ e $\hat{f}_h^*$ estimador de f dado por (3.1) com K^L a verificar*

$$\int_0^1 \left(\int_{-1}^{\alpha} |K^L(z,\alpha)|\,dz\right)^2 d\alpha < \infty. \tag{3.7}$$

Se $h_n \to 0$ e $nh \to \infty$, então

$$MISE\left(\hat{f}_h^*\right) = \frac{1}{nh}R(K) + h^3(f'(0))^2 \int_0^1 \mu_1^2(K^L(\cdot,\alpha))\,d\alpha + o\left(\frac{1}{nh}\right) + o(h^3).$$

Demonstração. Dada a expressão do estimador em causa, iremos analisar em separado os integrais que vigoram nos intervalos $]0,h]$ e $]h,+\infty[$. Comece-se por analisar o integral sobre $]0,h]$ referente à variância do estimador. Por (3.2), tem-se

$$\int_0^h Var(\hat{f}_h^*(x))\,dx = \frac{1}{nh}\int_0^h \int_{-\infty}^{\frac{x}{h}} K^L\left(z;\frac{x}{h}\right)^2 f(x-zh)\,dz\,dx - \frac{1}{n}\int_0^h \left(\int_{-\infty}^{\frac{x}{h}} K^L\left(z;\frac{x}{h}\right) f(x-zh)\,dz\right)^2 dx.$$

Considere-se a mudança de variável definida por $\alpha = \frac{x}{h}$. Ora, atendendo ao domínio de K^L, vem

$$\int_0^h Var(\hat{f}_h^*(x))\,dx = \frac{1}{n}\int_0^1 \int_{-1}^{\alpha} K^L(z;\alpha)^2 f(x-zh)\,dz\,d\alpha - \frac{h}{n}\int_0^1 \left(\int_{-1}^{\alpha} K^L(z;\alpha) f(x-zh)\,dz\right)^2 d\alpha.$$

De facto, a condição (3.7) aliada à regularidade de f permite estabelecer, pelo TCD, que

$$\int_0^h Var(\hat{f}_h^*(x))\,dx = o\left(\frac{1}{n}\right) + o\left(\frac{h}{n}\right). \tag{3.8}$$

Trabalhemos agora no termo correspondente à integração do viés do estimador sobre $]0,h]$. Ora, usando a mudança de variável $z = \frac{x-y}{h}$, tem-se

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = \int_0^h \left(\int_{-1}^{\frac{x}{h}} K^L\left(z,\frac{x}{h}\right) f(x-zh)\,dz - f(x)\right)^2 dx.$$

Utilizando agora o desenvolvimento de Taylor exposto em (3.5) e notando que $\int K^L(z,\alpha)\,dz = 1$, vem

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = h^2 \int_0^h \left(\int_{-1}^{\frac{x}{h}} \int_0^1 zK^L\left(z,\frac{x}{h}\right) f'(x-tzh)\,dt\,dz\right)^2 dx,$$

isto é,

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^2 \int_0^h \int_{-1}^{\frac{x}{h}} \int_{-1}^{\frac{x}{h}} \int_0^1 \int_0^1 zK^L\left(z, \frac{x}{h}\right) yK^L\left(y, \frac{x}{h}\right) f'(x-tzh) f'(x-syh)\, ds\, dt\, dy\, dz\, dx\,.$$

Tomando a mudança de variável $\alpha = \frac{x}{h}$, vem

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^3 \int_0^1 \int_{-1}^{\alpha} \int_{-1}^{\alpha} \int_0^1 \int_0^1 zK^L(z,\alpha)\, yK^L(y,\alpha)\, f'((\alpha - tz)h) f'((\alpha - sy)h)\, ds\, dt\, dy\, dz\, d\alpha\,.$$

Na medida em que $h_n \to 0$, a função integranda converge para $zK^L(z,\alpha)yK^L(y,\alpha)f'(0)^2$ e o seu módulo encontra-se majorado por $|K^L(z,\alpha)||K^L(y,\alpha)| \sup_{y\in\mathbb{R}_0^+} |f'(y)|^2$, função integrável, a aplicação do TCD permite concluir que

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = h^3 (f'(0))^2 \int_0^1 \mu_1^2(K^L(\cdot,\alpha))\, d\alpha. + o(h^3) \tag{3.9}$$

Numa segunda fase, focar-nos-emos na obtenção de expressões correspondentes ao intervalo $]h, +\infty[$. Usando o mesmo tipo de argumentos do capítulo anterior, vigora

$$\int_h^{+\infty} Var(\hat{f}_h^*(x))\, dx = \frac{1}{nh^2} \int_h^{+\infty} \int K\left(\frac{x-y}{h}\right)^2 f(y)\, dy\, dx - \frac{1}{nh^2} \int_h^{+\infty} \left(\int K\left(\frac{x-y}{h}\right) f(y)\, dy\right)^2 dx\,.$$

Através da mudança de variável de y para z já utilizada, obtem-se

$$\int_h^{+\infty} Var(\hat{f}_h^*)\, dx = \frac{1}{nh} \int_h^{+\infty} \int_{-1}^1 K(z)^2 f(x-zh) dz dx - \frac{1}{n} \int_h^{+\infty} \left(\int_{-1}^1 K(z) f(x-zh) dz\right)^2 dx\,.$$

De forma equivalente, usando o Teorema de Fubini, vem

$$\int_h^{+\infty} Var(\hat{f}_h^*)\, dx = \frac{1}{nh} \int_{-1}^1 \int_h^{\infty} K^2(z) f(x-zh)\, dx\, dz - \frac{1}{n} \int_h^{+\infty} \left(\int_{-1}^1 K(z) f(x-zh) dz\right)^2 dx$$

e, por sua vez,

$$\int_h^{+\infty} Var(\hat{f}_h^*)\, dx = \frac{1}{nh} \int_{-1}^1 \int_{h+zh}^{\infty} K^2(z) f(x)\, dx\, dz - \frac{1}{n} \int_h^{+\infty} \left(\int_{-1}^1 K(z) f(x-zh) dz\right)^2 dx\,,$$

ou ainda,

$$\int_h^{+\infty} Var(\hat{f}_h^*)\, dx = \frac{1}{nh} \int_{-1}^1 \int_0^{\infty} K^2(z) f(x) \mathbb{1}_{\{x \geq h-zh\}}\, dx\, dz + -\frac{1}{n} \int_h^{+\infty} \left(\int_{-1}^1 K(z) f(x-zh) dz\right)^2 dx\,.$$

Aplicando agora o TCD à primeira parcela, tem-se

$$\int_h^{+\infty} Var(\hat{f}_h^*)\,dx = \frac{1}{nh}R(K) + o\left(\frac{1}{nh}\right) - \frac{1}{n}\int_h^{+\infty}\left(\int_{-1}^{1} K(z)f(x-zh)dz\right)^2 dx.$$

Agora, utilizando o desdobramento do quadrado do integral em z em integral múltiplo, podemos escrever

$$\int_h^{+\infty}\left(\int_{-1}^{1} K(z)f(x-zh)dz\right)^2 dx = \int_h^{+\infty}\int_{-1}^{1}\int_{-1}^{1} K(z)K(y)f(x-zh)f(x-yh)\,dy\,dz\,dx,$$

isto é, utilizando o Teorema de Fubini e a mudança de variável $u = x - zh$,

$$\int_h^{+\infty}\left(\int_{-1}^{1} K(z)f(x-zh)dz\right)^2 dx = \int_{-1}^{1}\int_{-1}^{1}\int_{h-zh}^{+\infty} K(z)K(y)f(u)f(u+zh-yh)\,du\,dy\,dz.$$

Ora, tendo

$$\int_h^{+\infty}\left(\int_{-1}^{1} K(z)f(x-zh)dz\right)^2 dx = \int_{-1}^{1}\int_{-1}^{1}\int_{0}^{+\infty} K(z)K(y)f(u)f(u+zh-yh)\mathbb{1}_{\{u\geq h-zh\}}\,du\,dy\,dz,$$

podemos utilizaor o TCD e estabelecer

$$\int_h^{+\infty}\left(\int_{-1}^{1} K(z)f(x-zh)dz\right)^2 dx = R(f) + o(1).$$

Logo,

$$\int_h^{+\infty} Var(\hat{f}_h^*)\,dx = \frac{1}{nh}R(K) + o\left(\frac{1}{nh}\right) + o\left(\frac{1}{n}\right). \tag{3.10}$$

Desta, calculemos o termo do *MISE* referente à integração do termo de viés sobre o intervalo $]h,+\infty[$. De facto, de (2.8), temos

$$\int_h^{+\infty}\left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = h^4\int_h^{\infty}\left(\int_{-1}^{1}\int_{0}^{1} z^2K(z)(1-t)f''(x-tzh)\,dt\,dz\right)^2 dx.$$

Agora o desdobramento em integral múltiplo permite estabelecer

$$\int_h^{+\infty}\left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^4\int_h^{\infty}\int\int\int\int_D z^2p^2K(z)K(y)(1-t)(1-s)f''(x-tzh)f''(x-syh)\,ds\,dt\,dy\,dz\,dx,$$

com $D = [-1,1]^2\times[0,1]^2$. Alterando a ordem de integração, recorrendo ao Teorema de Fubini, e com a mudança de variável $u = x - tz$ segue que

$$\int_h^{+\infty}\left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^4\int\int\int\int_D\int_{h-tzh}^{\infty} z^2p^2K(z)K(y)(1-t)(1-s)f''(u)f''(x+tzh-syh)\,du\,ds\,dt\,dy\,dz,$$

ou seja,

$$\int_h^{+\infty} \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^4 \int\int\int\int_D \int_0^{\infty} z^2 p^2 K(z) K(y)(1-t)(1-s) f''(u) f''(x+tzh-syh) \mathbb{1}_{\{u \geq h-tzh\}} \, du\, ds\, dt\, dy\, dz\,.$$

Por fim, o TCD determina que

$$\int_h^{+\infty} \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = o(h^4)\,. \tag{3.11}$$

Assim, o resultado pretendido sai de (2.10), (3.8), (3.9), (3.10) e (3.11). □

Note-se que a condição $f'(0) \neq 0$ pode não ser verificada em algumas densidades típicas nas ciências atuariais, como é o caso da densidade logormal. Nestas a demonstração da expansão assintótica do *MISE* será em tudo semelhante à que se apresenta abaixo para núcleos de fronteira de segunda ordem.

Façamos agora um paralelismo entre os Teoremas 2 e 5. De facto, os termos referentes à integração da variância dos estimadores em estudo apresentam a mesma ordem de convergência. Por outro lado, os termos referentes ao viés apresentam-se com ordem distintas de convergência, traduzindo-se num *MISE* que tenderá mais lentamente para 0 no Teorema 5. Assim, a apresentação prévia de núcleos que satisfazem a condição de segunda ordem, motiva a particularização do *MSE* e *MISE* para estimadores com núcleos de fronteira de segunda ordem sobre $[-1, \alpha]$, pelo que se apresenta o seguinte resultado.

Teorema 6. *Sejam a restrição de f ao intervalo $[0, +\infty[$ com derivadas contínuas até à segunda ordem numa vizinhança de x, $\hat{f}_h^*$ o estimador de f dado por (3.1) e K^L núcleo de fronteira de segunda ordem. Se $h_n \to 0$ e $nh \to \infty$, então, para $x = \alpha h$, com $\alpha \in]0,1]$,*

$$MSE(\hat{f}_h^*(x)) = \frac{1}{nh} f(x) R(K^L(\cdot, \alpha)) + \frac{h^4}{4} (f''(x))^2 \mu_2^2(K^L)(\alpha) + o\left(\frac{1}{nh}\right) + o\left(\frac{1}{n}\right) + o(h^4)\,,$$

Demonstração. Notando que, na obtenção da expressão (3.3) para a variância de $\hat{f}_h^*(x)$, a segunda ordem do núcleo de fronteira sobre $[-1, \alpha]$ em nada toma parte, podemos atestar imediatamente que

$$V(\hat{f}_h^*(x)) = \frac{1}{nh} f(x) R(K^L(\cdot, \alpha)) + o\left(\frac{1}{nh}\right)\,. \tag{3.12}$$

Para o cálculo do quadrado do viés do estimador, comecemos por notar que, de (3.4), obtemos

$$E(\hat{f}_h^*(x)) = \int_{-1}^{\alpha} K^L(z, \alpha) f(x - zh)\, dz.$$

Retomando a regularidade admitida para f e as suas duas primeiras derivadas e o desenvolvimento de Taylor associado, atente-se a que K^L intregra 1 sobre $[-1, \alpha]$ e que o mesmo é de segunda ordem. Logo,

$$E(\hat{f}_h^*(x)) = f(x) + h^2 \int_{-1}^{\alpha} \int_0^1 z^2 K^L(z, \alpha)(1-t) f''(x - tzh)\, dt\, dz\,.$$

Note-se que, por $t \in [0,1]$ e $\alpha \in]0,1]$, se tem $0 \leq x - tzh \leq (\alpha+1)h$ e, portanto, $x - tzh$ é um ponto onde f e as suas duas primeiras derivadas são contínuas e limitadas numa sua vizinhança, relembrando que a sucessão h_n tende para 0. Considerando então a sucessão de funções

$$g_n(z,t) = z^2 K^L(z,\alpha)(1-t)f''(x-tzh)$$

, a aplicação do TCD resulta em

$$\int_{-1}^{\alpha}\int_{0}^{1} z^2 K^L(z,\alpha)(1-t)(f''(x-tzh)-f''(x))\,dt\,dz = \frac{f''(x)}{2}R(K^L(\cdot,\alpha)) + o(1)$$

Logo,

$$\left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 = \frac{h^4}{4}(f''(x))^2\mu_2^2(K^L)(\alpha) + o(h^4)\,. \tag{3.13}$$

Em última análise, de (2.3), (3.12) e (3.13), tem-se o resultado pretendido. □

De seguida, apresenta-se a evolução gráfica do termo associado ao viés do estimador com correção de fronteira, $\mu_2(K^L)$, função da distância à origem em h unidades. Consideram-se apenas os núcleos de fronteira sobre $[-1,\alpha]$ K_1^L e K_2^L, uma vez que K_3^L não é núcleo de fronteira de segunda ordem.

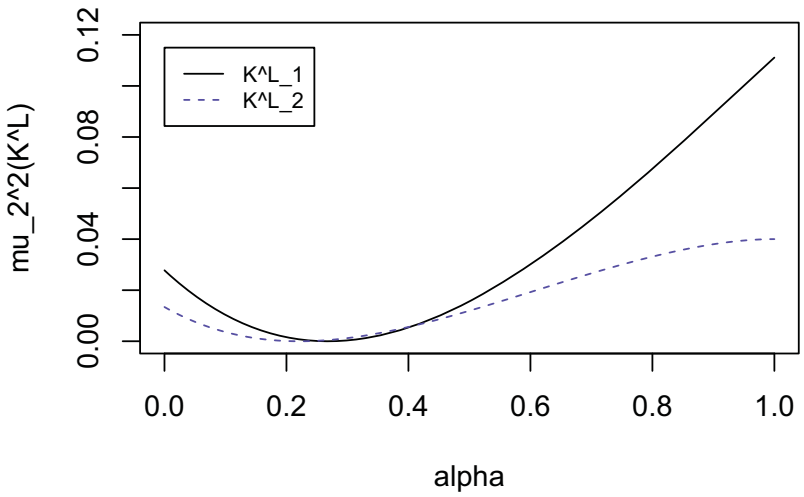


Fig. 3.5 Evolução de $\mu_2^2(K^L)$, para K_i^L, $i = 1,2$, em função de α

No que diz respeito à análise do termo associado ao viés do estimador, a comparação dos núcleos de fronteira de segunda ordem sobre $[-1,\alpha]$ apresentados permite atestar que K_2^L apresenta, de uma forma geral, melhor desempenho na minimização de $\mu_2^2(K^L(\cdot,\alpha))$. Isto revela que a determinação de um núcleo de fronteira que minimiza a variância do estimador não produz núcleos de fronteira que minimizem de forma global o *MSE* do estimador com correção de fronteira em pontos x tais que $0 < x \leq h$.

Com o intuito de definir, tal como no caso anterior, um critério de erro global, apresenta-se agora o resultado referente ao *MISE* do novo estimador com correção de fronteira.

Teorema 7. *Sejam a restrição de f ao intervalo $[0,+\infty[$ com derivadas contínuas até à segunda ordem e f'' limitada e $\hat{f}_h^*$ estimador de f dado por (3.1) com K^L núcleo de segunda ordem a verificar*

$$\int_0^1 \left(\int_{-1}^{\alpha} |K^L(z,\alpha)|\, dz \right)^2 d\alpha < \infty \tag{3.14}$$

. Então

$$MISE\left(\hat{f}_h^*\right) = \frac{1}{nh} R(K) + \frac{h^4}{4} \mu_2^2(K) R(f'') + o(h^4) + o\left(\frac{1}{nh}\right),$$

com $h \to 0$ e $nh \to \infty$.

Demonstração. Comecemos por notar que a alteração das condições face ao teorema 5, não se traduz em qualquer mudança na obtenção do integral da variância de $\hat{f}_h^*$ sobre $]0,+\infty[$. Assim, vale

$$\int V(\hat{f}_h^*(x))\, dx = \frac{1}{nh} R(K) + o\left(\frac{1}{nh}\right). \tag{3.15}$$

Por outro lado, a formulação do estimador encoraja a análise independente dos integrais do viés a calcular nos intervalos $]0,h]$ e $]h,+\infty[$. Desta forma, incitemos o cálculo de $\int (E(\hat{f}_h^*(x)) - f(x))^2\, dx$, no intervalo $]0,h]$. Ora

$$E(\hat{f}_h^*(x)) = \int_{-\infty}^{\frac{x}{h}} K^L\left(z, \frac{x}{h}\right) f(x - zh)\, dy,$$

pelo que, atendendo à definição do domínio de K^L,

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = \int_0^h \left(\int_{-1}^{\frac{x}{h}} K^L\left(z, \frac{x}{h}\right) f(x - zh)\, dz - f(x) \right)^2 dx.$$

Agora, graças ao desenvolvimento (2.12) e à segunda ordem do núcleo de fronteira sobre $[-1,\alpha]$, tem-se

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = h^4 \int_0^h \left(\int_{-1}^{\frac{x}{h}} z^2 K^L\left(z, \frac{x}{h}\right) (1-t) f''(x - tzh)\, dt\, dz \right)^2 dx.$$

Utilizemos agora uma nova variável definida por $\alpha = \frac{x}{h}$. Assim,

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = h^5 \int_0^1 \left(\int_{-1}^{\alpha} z^2 K^L(z,\alpha)(1-t) f''((\alpha - tz)h)\, dt\, dz \right)^2 dx.$$

Por sua vez, a aplicação do TCD aliada à condição (3.14) permite determinar que

$$\int_0^h \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = o(h^5). \tag{3.16}$$

Por fim, a inalteração do estimador para $x > h$, de (2.12), atesta-se que

$$\int_h^{+\infty} \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx =$$
$$h^4 \int\int\int\int_D \int_0^{\infty} z^2 p^2 K(z) K(y)(1-t)(1-s) f''(u) f''(x+tzh-syh) \mathbb{1}_{\{u \geq h-tzh\}} \, du\, ds\, dt\, dy\, dz,$$

com $D = [-1,1]^2 \times [0,1]^2$, que, pelo TCD, equivale a

$$\int_h^{+\infty} \left(E\left(\hat{f}_h^*(x)\right) - f(x)\right)^2 dx = \frac{h^4}{4} R(f'') \mu_2^2(K). \tag{3.17}$$

Em última análise, por (2.10), (3.15), (3.16) e (3.17), sai o pretendido. □

3.3 Escolha da janela

O estudo dos estimadores do núcleo com correção de fronteira, culmina então com a determinação de critérios de escolha da janela. Comecemos então por apresentar o *AMISE* do estimador nas condições do Teorema 5.

$$AMISE(\hat{f}_h^*) = \frac{1}{nh} R(K) + h^3 (f'(0))^2 \int_0^1 \mu_1^2(K^L(\cdot, \alpha))\, d\alpha\,.$$

Na verdade, o *AMISE* do estimador com correção de fronteira permite a obtenção de uma janela assintoticamente ótima na minimização desta medida de erro.

Teorema 8. *Nas condições do Teorema 5, a janela assintoticamente ótima, h_{opt}, é dada por*

$$h_{opt} = \left(\frac{R(K)}{3n(f'(0))^2 \int_0^1 \mu_1^2(K^L(\cdot,\alpha))\, d\alpha} \right)^{1/4}$$

Por sua vez, considerando o estimador com correção de fronteira, com K^L de segunda ordem, do Teorema 7, basta notar que o *AMISE* deste estimador é igual ao do estimador sem correção de fronteira, pelo que voltamos a motivar a escolha da janela pela minimização do *MISE* assintótico. Logo,

Teorema 9. *Nas condições do Teorema 7, a janela assintoticamente ótima, h_{opt}, é dada por*

$$h_{opt} = \left(\frac{R(K)}{nR(f'')\mu_2^2(K)} \right)^{1/5}$$

Capítulo 4

Distribuição de Champernowne modificada e estimação dos dados transformados

4.1 Motivação

Nas ciências atuariais, são frequentes os conjuntos de dados de indeminizações de um determinado produto segurado que apresentam uma grande assimetria (positiva). Se, por um lado, são frequentes indemnizações de valores reduzidos, por outro, as indemnizações de grande magnitude são raras, mas devido às responsabilidades assumidas pelas seguradoras, são também de especial interesse. De facto, a pouca frequência dos dados elevados aliada à aleatoriedade da variável que descreve as indemnizações, poder-se-á traduzir numa estimação desadequada da cauda da distribuição e consequentes perdas para a empresa seguradora. Assim, torna-se ainda mais fulcral a boa estimação da função de perda associada.

Tendo como mote o trabalho de Wand, Marron e Ruppert (1991) que oferece melhorias à estimação de funções densidade por métodos do núcleo através da aplicação de uma transformação aos dados recolhidos (*shift power transformation*, no caso), Buch-Larsen *et al.* (2005) introduzem a transformação dos dados através da distribuição de Champernowne modificada. Começando por escrutinar a distribuição de Champernowne, cuja função densidade se assume

$$f(x) = \frac{\alpha M^{\alpha} x^{\alpha-1}}{(x^{\alpha} + M^{\alpha})^2},$$

com α e M parâmetros da distribuição, os autores identificam que a distribuição apresentada converge para a lei de Pareto na cauda. Mais ainda, assemelha-se a uma lei lognormal na proximidade da origem, para valores de α superiores à unidade. De facto, estas duas distribuições são bastantes referenciadas nos métodos paramétricos de estimação de funções de perda, basta para isso notar Cooray e Ananda (2005), o que poderia induzir diretamente a transformação dos dados, através da aplicação da função de repartição associada para o intervalo $[0,1]$.

No entanto, Buch-Larsen *et al.* (2005) identificam problemas na origem, graças à inflexibilidade da distribuição perto de 0. É de notar que a distribuição de Champernowne permite, para $\alpha > 1$, que $t(0) = 0$ e que, por sua vez, se considerarmos $\alpha < 1$, $lim_{x\to 0} t(x) = \infty$. No entanto, para conseguir $t(0) > 0$ será necessário impor taxativamente $\alpha = 1$, o que se torna limitativo, uma vez que nos problemas em estudo são frequentes as densidades não contínuas na origem.

Por fim, através de estudos de simulação, os autores concluíram que a consideração da transformação de Champernowne modificada se traduz num estimador para a densidade com melhor desempenho.

4.2 Distribuição de Champernowne modificada

Nesta secção, será apresentado um estudo sucinto relativo à distribuição proposta por Buch-Larsen *et al.* (2005) para a transformação dos dados.

Definição 5. *A função de distribuição da Lei de Champernowne modificada é definida, para* $x \geq 0$*, como sendo*

$$T_{\alpha,M,c}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha}, \tag{4.1}$$

com parâmetros $\alpha > 0$*,* $M > 0$ *e* $c \geq 0$ *e densidade*

$$t_{\alpha,M,c}(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^\alpha - c^\alpha)}{((x+c)^\alpha + (M+c)^\alpha - 2c^\alpha)^2}.$$

De seguida, apresentam-se representações gráficas alusivas à Lei de Champernowne modificada para vários valores dos parâmetros. Foram utilizados os mesmos valores do artigo original por se considerarem explicativos no estudo da distribuição em causa.

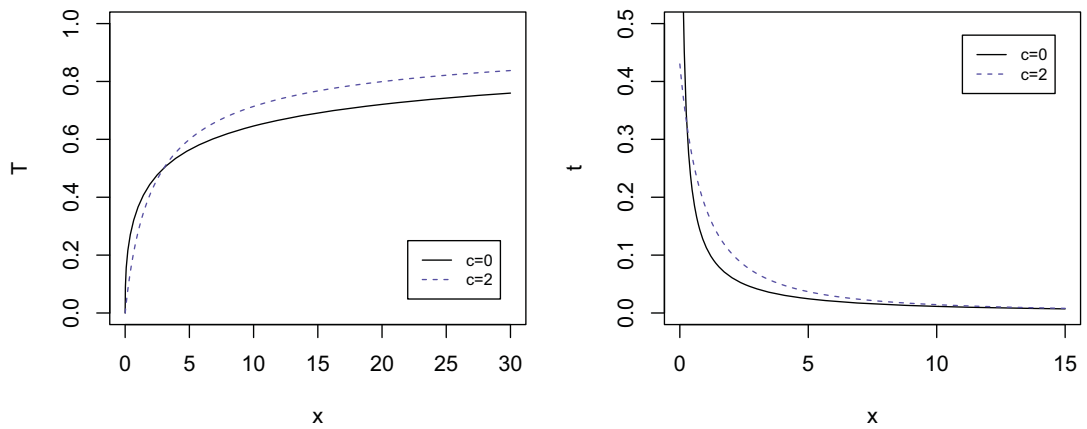


Fig. 4.1 Representações gráficas das funções de distribuição e densidade da Lei de Champernowne modificada para $\alpha = 0.5, M = 3$

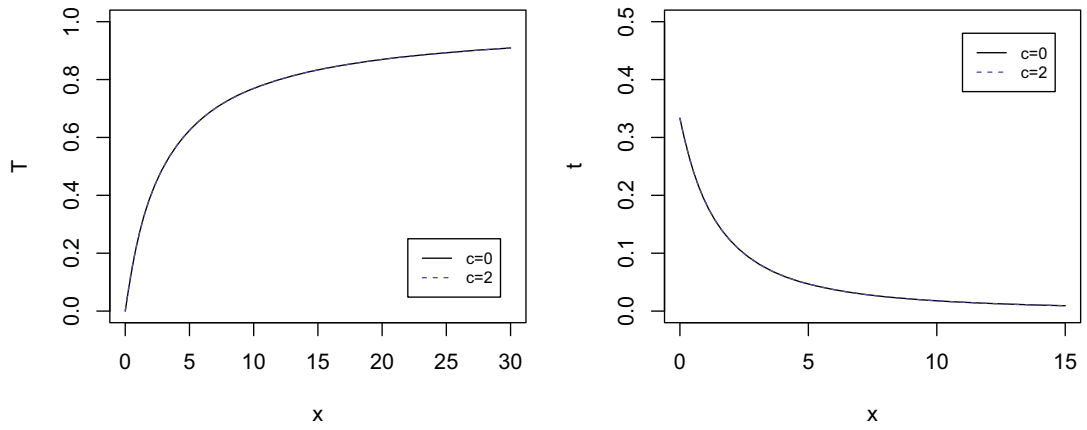


Fig. 4.2 Representações gráficas das funções de distribuição e densidade da Lei de Champernowne modificada para $\alpha = 1, M = 3$

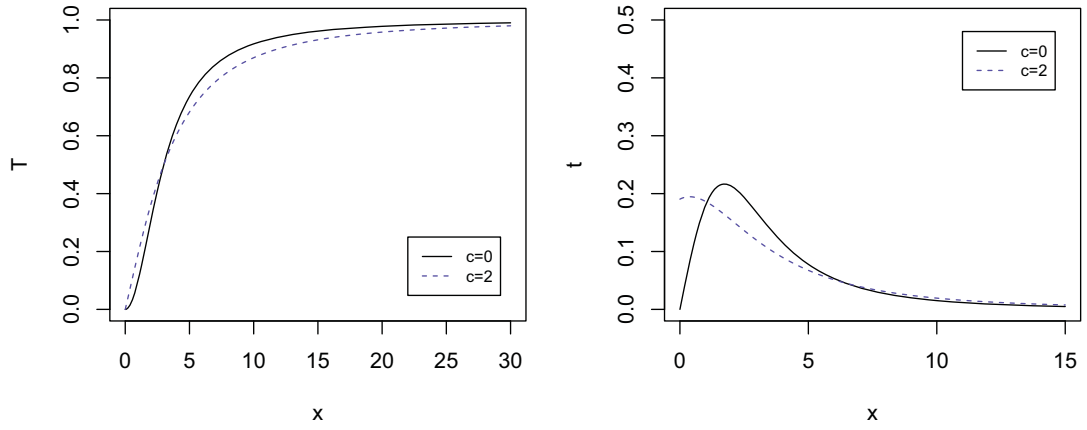


Fig. 4.3 Representações gráficas das funções de distribuição e densidade da Lei de Champernowne modificada para $\alpha = 2, M = 3$

Façamos uma análise mais detalhada ao comportamento da Lei de Champernowne modificada, notando as alterações que a consideração do parâmetro c produz face à distribuição de Champernowne.

Primeiramente, notemos que se $c \neq 0$ então, obtemos incondicionalmente uma densidade que se revela positiva e finita na origem, tal como havia sido motivado.

Por outro lado, quando $\alpha < 1$, a consideração do parâmetro c traduz-se numa distribuição com uma cauda mais leve, como se pode observar nas funções de distribuição presente na Figura 4.1, pois, a partir de um determinado valor, esta função da nova lei é superior à original. O inverso acontece na Figura 4.3 e para qualquer $\alpha > 1$, pois a cauda da distribuição torna-se mais pesada para escolhas de c positivo, tendo por base uma análise inversa à anterior.

É ainda relevante salientar que a introdução do parâmetro c em nada influencia a nova lei de probabilidade, para $\alpha = 1$, pois neste caso a distribuição modificada e a original coincidem, como exposto na Figura 4.2

Apresentada a distribuição que motiva a transformação dos dados, torna-se agora essencial a proposta de metodologias de estimação dos parâmetros da Lei de Champernowne modificada. Primeiramente, notando que $T_{\alpha,M,c}(M) = 0.5$, o parâmetro M pode ser estimado pela mediana empírica das observações de X. Os autores, suportam a escolha deste estimador pelo aumento da sua eficácia do estimador à medida em que a cauda fica mais pesada. Por fim, sugere-se a utilização do método da máxima verosimilhança com o intuito da determinação de estimativas para o par de parâmetros (α, c).

4.3 Estimação da densidade dos dados transformados

Nesta etapa, assumimos como objetivo a estimação da densidade dos dados transformados, g, pois a partir desta, aplicando a transformação inversa, podemos inferir sobre a densidade de X. Ora, obtidas estimativas para os parâmetros da Lei de Champernowne modificada, $(\hat{\alpha}, \hat{M}, \hat{c})$, trabalhemos com as observações transformadas

$$Y_i = T_{\hat{\alpha},\hat{M},\hat{c}}(X_i), i = 1, ..., n,$$

com T definida por (4.1).

De uma forma geral, se tomarmos X uma v.a.r. absolutamente contínua, com função de distribuição F_X, ao definir uma nova variável $Y = F_X(X)$, obtemos $0 \leq Y \leq 1$. Logo, para $y \in [0,1]$,

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

Desta forma, uma vez que a função de distribuição caracteriza a lei de probabilidade, Y será uma variável distribuída uniformemente no intervalo $[0,1]$. Assim, o correto ajustamento da distribuição de Champernowne modificada aos dados e consequente transformação dos mesmos pela função de distribuição resulta em $Y_i, i = 1, ..., n$ com uma distribuição aproximadamente uniforme sobre $[0,1]$.

Tal como no capítulo anterior, aqui os núcleos de fronteira assumem um papel preponderante enquanto ferramenta de estimação da densidade dos dados transformados, pois quer na origem, quer na unidade, a densidade g não é necessariamente contínua. Na verdade, a existência de um ponto extremo direito do suporte dos dados tranformados, promove a extensão da utilização dos núcleos de fronteira de suporte contido no intervalo $[-\alpha, 1]$.

Analisando o domínio dos núcleos de fronteira previamente utilizados, para a construção de núcleos de fronteira sobre $[-\alpha, 1]$, K^R, basta considerar $K^R(z, \alpha) = K^L(-z, \alpha)$. Assim, estamos em condições de definir um estimador para a densidade transformada. Dadas as observações transformadas $Y_1, Y_2, ..., Y_n$, o estimador de $g(y)$, $y \in]0,1[$, com correção de fronteira bilateral assume a seguinte formulação:

$$\hat{g}_h(y) = \frac{1}{nh} \sum_{i=1}^{n} K_{y,h}\left(\frac{y - Y_i}{h}\right), \tag{4.2}$$

com

$$K_{y,h}(z) = \begin{cases} K^L(z, y/h), & 0 < y \leq h, \\ K(z), & h < y < 1-h, \\ K^R(z, (1-y)/h), & 1-h \leq y < 1. \end{cases} \tag{4.3}$$

onde $K(\cdot)$ é núcleo de segunda ordem e $K^L(\cdot, \alpha) = K^R(-\cdot, \alpha)$ são núcleos de fronteira sobre $[-1, \alpha]$. Contudo, para que estimador esteja bem definido é preciso estabelecer que $h \leq \frac{1}{2}$, por forma a anular qualquer ambiguidade na sua implementação.

4.4 Escolha da janela

Buch-Larsen *et al.* (2005) sugerem a utilização da *rule of thumb*, proposta por Silverman ([14], pg.45), para a escolha da janela. No entanto, esta regra assenta na minimização do MISE do estimador sem considerar as correções de fronteira e, por outro lado, baseia-se numa distribuição de referência normal. É esta desadequação que motiva o estabelecimento de outro critério de escolha da janela de estimação.

Notemos que $Y_1, Y_2, ..., Y_n$ são observações não independentes nem identicamente distribuídas pelo que a análise dos erros quadrático médio e quadrático médio integrado não se pode desenvolver conforme exposto nos Capítulos 2 e 3. Apesar disso, para os estimadores que assentam na estimação da densidade dos dados transformados pela distribuição de Champernowne modificada, sugere-se a utilização de

$$h = \left(\frac{R(K)}{nR(g'')\mu_2^2(K)} \right)^{1/5}, \tag{4.4}$$

para a janela de estimação, uma vez que surge como adaptação empírica do estabelecido no capítulo anterior. Agora resta-nos estabelecer metodologias para a obtenção de estimativas para $R(g'')$. Tal como referido, as observações Y_i, $i = 1, ..., n$ deverão apresentar uma distribuição aproximadamente uniforme. No entanto, a consideração desta distribuição de derivadas nulas como referência, não se revela pertinente na determinação de h, uma vez que resultaria em janelas infinitamente grandes. Outra aboragem é tomada por Tenreiro (2013), sugerindo a utilização da distribuição beta sobre $[0, 1]$ de parâmetros p e q superiores a 2, como distribuição de referência. No entanto, Tenreiro centra-se na estimação da função de distribuição pelo que a limitação dos parâmetros garante a integrabilidade do quadrado da primeira derivada de g. No nosso caso, sendo que estaremos a trabalhar com g'' é necessário estabelecer $p, q > 2.5$, para garantir que $R(g'')$ se encontra definida. Por fim, o autor sugere o método dos momentos na estimação dos parâmetros da lei beta que resulta nos estimadores

$$\hat{p} = \bar{Y}(\bar{Y}(1-\bar{Y})\hat{S}^{-2} - 1) \text{ e } \hat{q} = (1-\bar{Y})(\bar{Y}(1-\bar{Y})\hat{S}^{-2} - 1)$$

com $\bar{Y}$ a média dos dados transformados e $\hat{S}$ a respetiva variância.

4.5 Estimação da densidade

Em virtude do nosso objetivo principal ser a obtenção de estimativas da verdadeira densidade dos dados, surge a necessidade de estabelecer uma relação entre esta e a densidade dos dados transformados. Ora,

atentemos a que se X é uma v.a.r. absolutamente contínua e $Y = T(X)$, com T função estritamente crescente no suporte de X, então

$$F_X(x) = P(X \leq x) = P(T^{-1}(Y) \leq x) = P(Y \leq T(x)) = F_Y(T(X))$$

e assim

$$f_X(x) = F_X'(x) = (F_Y(T(x)))' = f_Y(T(x))T'(x).$$

De forma natural, estabelece-se como estimador da densidade f, em $x \in \mathbb{R}^+$,

$$\hat{f}_h^{**}(x) = \frac{1}{nh} \sum_{i=1}^{n} K_{T_{\hat{\alpha},\hat{M},\hat{c}}(x),h} \left(\frac{T_{\hat{\alpha},\hat{M},\hat{c}}(x) - T_{\hat{\alpha},\hat{M},\hat{c}}(X_i)}{h} \right) T'_{\hat{\alpha},\hat{M},\hat{c}}(x), \tag{4.5}$$

com $K_{y,h}$ definido por (4.3).

Capítulo 5

Estudo de simulação

Este capítulo tem por base a implementação, através do *software* R, de um estudo de simulação comparativo dos vários métodos do núcleo apresentados, bem como a análise posterior dos resultados obtidos. Foram simulados conjuntos amostrais de diferentes tamanhos, tendo por base algumas distribuições distintas no que toca ao comportamento na origem, bem como ao peso das respetivas caudas. É de ressalvar que, para cada combinação de cardinalidade da amostra, distribuição e método do núcleo, foram implementadas 500 repetições por forma a garantir credibilidade ao estudo apresentado. O código em uso pode ser consultado no Anexo B

5.1 Tamanho da amostra

Neste estudo de simulação, o tamanho da amostra será também objeto de análise, por forma a avaliar o desempenho dos vários estimadores face à reduzida ou elevada informação acerca do comportamento da variável X. Para tal, consideram-se amostras com 100, 1000 e 10000 observações.

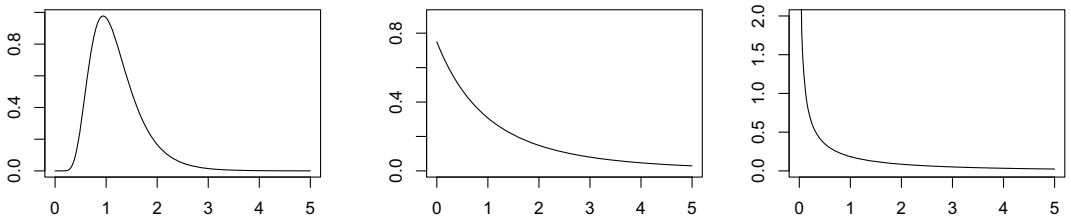
5.2 Distribuições

Foram simuladas amostras de três distribuições distintas: Lognormal, de Pareto e de Weibull. A primeira é uma distribuição cuja cauda não é muito pesada e que apresenta continuidade na origem, ou seja, $\lim_{x\to 0} f(x) = 0$. A distribuição de Pareto apresenta uma cauda mais pesada que as anteriormente referidas e o seu comportamento na origem é caracterizado pela descontinuidade em 0, pois apresenta um valor finito e positivo para a densidade neste ponto. Sempre que nos referirmos à lei de Pareto, estaremos a considerar a sua formulação no tipo 2, com o parâmetro de localização nulo. Por fim, a distribuição de Weibull apresenta um comportamento semelhante à distribuição lognormal na cauda, mas o seu comportamento na origem é distinto, uma vez que a escolha de parâmetros permite alternar entre $f(0)$ nulo, finito positivo ou infinito, sendo este último caso o considerado no presente trabalho. As densidades consideradas, bem como os respetivos parâmetros, podem ser observados na tabela abaixo.

Tabela 5.1 Distribuições e parâmetros utilizados no estudo de simulação

Distribuição	Dendidade para $x > 0$	Parâmetros
Lognormal (μ, σ)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$	$(\mu, \sigma) = (0.1, 0.4)$
Pareto (a, b)	$f(x) = \frac{ab^a}{(x+b)^{a+1}}$	$(a, b) = (3, 4)$
Weibull (k, λ)	$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$	$(k, \lambda) = (0.5, 1)$

A Figura 5.1 é então representativa das densidades e parâmetros considerados.



(a) Lognormal (b) Pareto (c) Weibull

Fig. 5.1 Representações gráficas das densidades apresentadas

5.3 Métodos do núcleo

Com o intuito de comparar os vários métodos do núcleo apresentados, foram considerados cinco estimadores da densidade.

O primeiro, $\hat{f}_1$ é o estimador definido por (2.2). Neste foi considerado como função peso o núcleo de Epanechnikov e a janela foi determinada aplicando o Teorema 3. A escolha do núcleo de Epanechnikov é usual neste tipo de problemas, uma vez que surge como núcleo ótimo na minimização do erro quadrático médio integrado assintótico. Este resultado pode ser consultado em [3]. Como sugerido anterioremente, a aproximação da quantidade $R(f'')$ é obtida ajustando aos dados uma distribuição lognormal.

No campo dos estimadores com correção de fronteira, serão utilizados dois métodos do núcleo, $\hat{f}_2$ e $\hat{f}_3$, ambos definidos por (3.1). Será de novo tomado como núcleo o de Epanechnikov, sendo que o que difere os dois métodos é a escolha do núcleo de fronteira utilizado. Para $\hat{f}_2$ considerar-se-á para o núcleo de fronteira sobre $[-1, \alpha]$ a funcional introduzida como K_1^L. Esta escolha, face à alternativa K_2^L, é justificada pela minimização que K_1^L oferece ao termo de variância do $MISE$. Mais ainda, apesar de não o fazer no termo do viés, a diferença entre os dois núcleos de fronteira sobre $[-1, \alpha]$ é bastante menor do que a verificada na variância. Já $\hat{f}_3$ apresenta como núcleo de fronteira a funcional sugerida por Buch-Larsen *et al.* (2005). Para os estimadores em causa, a escolha da

janela de estimação procede-se à luz do Teorema 9, uma vez que iremos de novo considerar como distribuição de referência uma lei lognormal. Tal como explicitado no Capítulo 3, esta densidade apresenta derivada nula na origem pelo que as janelas assintoticamente ótimas coincidem.

Os últimos dois estimadores terão como base a estimação dos dados transformados pela função de distribuição da lei de Champernowne modificiada, sendo definidos de acordo com (4.5). Assim, renovando a escolha do núcleo de Epanechnikov para K, $\hat{f}_4$ apresenta como núcleo de fronteira sobre $[-1,\alpha]$ a função K_1^L, enquanto $\hat{f}_5$ apresenta K_3^L para o desempenho desse papel. Resta ressalvar que as janelas de estimação são determinadas atendendo a (4.4), em que a densidade dos dados transformados será aproximada por uma distribuição beta no intervalo $[0,1]$ de parâmetros superiores a 2.5, para a determinação da constante $R(g'')$.

5.4 Medidas de erro

Por fim, resta apresentar as medidas utilizadas para a determinação do desempenho do estimador. Neste sentido, serão considerados os erros L_1, L_2 e $WISE$, como objeto de avaliação da precisão da estimação. O primeiro avalia a distância entre as densidades (estimada e verdadeira) sobre todo o suporte, isto é,

$$L_1 = \int_0^\infty |\hat{f}(x) - f(x)|\, dx\,.$$

Por sua vez, o erro L_2 apresenta-se sendo a norma L_2 da diferença entre a densidade estimada e a verdadeira. Ora,

$$L_2 = \left(\int_0^\infty (\hat{f}(x) - f(x))^2\, dx \right)^{\frac{1}{2}}\,.$$

Sendo que a boa estimação da cauda é um argumento central para a obtenção de estimadores precisos, o erro quadrático integrado ponderado, $WISE$, surge como ferramenta de medida que pondera de maior forma os erros na cauda, uma vez que apresenta a seguinte formulação

$$WISE = \left(\int_0^\infty (\hat{f}(x) - f(x))^2 x^2\, dx \right)^{\frac{1}{2}}\,.$$

À semelhança do que se fez para a ponderação dos erros na cauda, poder-se-ia pensar numa alternativa semelhante para ponderar de maior forma a estimação efetuada nas proximidades do extremo do suporte e assim avaliar o desempenho dos núcleos de fronteira. No entanto, caso procedessemos à divisão de $(\hat{f}(x) - f(x))^2$ por x^2 iriam surgir problemas de indefinição do integral em causa, conforme testado em simulação.

Por outro lado, poderia ser pensada uma medida de erro que avaliasse apenas a estimação feita no intervalo $]0,h]$. Na verdade, esta opção também não se revela benéfica na medida em que a janela h não é igual para todos os estimadores considerados e, mesmo avaliando esta medida no intevalo $]0,h_{max}]$, com h_{max} a maior janela definida entre os vários métodos propostos, as ilações a tirar não seriam conclusivas, na medida em que em certos pontos do intervalo estaríamos a comparar o desempenho do núcleo contra o núcleo de fronteira considerado. Se considerarmos $]0,h_{min}]$, com h_{min} a menor janela definida entre os vários métodos propostos, não estaríamos a avaliar todo o desempenho na correção de fronteira.

5.5 Análise dos Resultados

Apresenta-se agora as tabelas das médias dos erros obtidos na estimação da densidade lognormal para cada combinação de tamanho da amostra e estimador considerada.

Tabela 5.2 Médias dos erros na estimação da densidade de probabilidade lognormal considerada

		$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$
$n = 100$	L_1	0.16317903	0.16494140	0.16370892	0.13560967	0.13552768
	L_2	0.11996363	0.12133401	0.12010719	0.10563149	0.10562234
	$WISE$	0.15268501	0.15277938	0.15268998	0.11614908	0.11613008
$n = 1000$	L_1	0.07064163	0.07085649	0.07065526	0.05743315	0.05743315
	L_2	0.05312868	0.05316569	0.05312958	0.04646028	0.04646028
	$WISE$	0.06583769	0.06583874	0.06583771	0.05117734	0.05117734
$n = 10000$	L_1	0.02878066	0.02878528	0.02878078	0.02384875	0.02384875
	L_2	0.02219474	0.02219483	0.02219474	0.01983457	0.01983457
	$WISE$	0.02711381	0.02711381	0.02711381	0.02176807	0.02176807

Comece-se por observar que, tal como seria expectável, em média, o aumento do tamanho da amostra têm consequências na diminuição de todos os erros associados a cada estimador. Por sua vez, quando considerada a densidade de probabilidade lognormal, contínua em $\mathbb{R}$, a introdução dos estimadores com correção de fronteira, $\hat{f}_2$ e $\hat{f}_3$, em nada melhora a estimação da densidade, uma vez que os erros associados a estes dois estimadores são, no máximo, iguais aos erros do primeiro método do núcleo apresentado, em termos médios. Mais ainda, é bastante recorrente que estes sejam superiores aos associados ao estimador sem correção de fronteira. A análise da quarta e quinta coluna permite-nos estabelecer que o desempenho médio dos estimadores que recorrem à transformação dos dados simulados pela aplicação da função de distribuição associada à lei de Champernowne modificada é, de facto, melhor face aos restantes métodos do núcleo. No entanto, o aumento da cardinalidade da amostra faz com que estas diferenças se tornem menores. Analisando a tabela acima, é notória a grande diferença entre a média do $WISE$ dos três primeiros estimadores face aos dois últimos, em amostras de cardinalidade 100, revelando um bom desempenho destes últimos na estimação da cauda da distribuição.

No que diz respeito aos dois núcleos de fronteira utilizados, a análise comparativa da média dos erros dos pares de estimadores $(\hat{f}_2, \hat{f}_3)$ e $(\hat{f}_4, \hat{f}_5)$ induz ao melhor desempenho dos métodos que recorrem ao núcleo de fronteira apresentado por Buch-Larsen *et al.* (2005), porque os erros associados aos segundo e quarto métodos são sempre superiores ou iguais aos relacionados com os terceiro e quinto método, respetivamente. É ainda relevante salientar que, em média, para amostras com 1000 ou 10000 observações, os dois últimos métodos considerados revelam-se semelhantes, pois os erros associados apresentam-se iguais.

A análise da amplitude interquartil nos *boxplots* presentes no Anexo C permite atestar que a variabilidade dos erros dos cinco estimadores é semelhante, com exceção do erro L_1 para amostras com 100 observações, na medida em que os métodos assentes na transformação de dados apresentam

uma variabilidade ligeiramente superior. Mais ainda, está patente um decréscimo desta medida, com o aumento do tamanho da amostra, que se traduz numa menor dispersão dos dados. Por último, resta ressalvar que para $n = 100$ a ocorrência de *outliers* é mais severa nos estimadores que tratam os dados sem transformação, sendo que o inverso acontece quando o tamanho da amostra é dez vezes superior.

Tabela 5.3 Médias dos erros na estimação da densidade de probabilidade de Pareto considerada

		$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$
$n = 100$	L_1	0.48598953	0.48628193	0.48296513	0.13887790	0.13462787
	L_2	0.30182392	0.30459475	0.29843034	0.11449099	0.10232126
	$WISE$	0.52151702	0.52151594	0.52151376	0.09092263	0.09199779
$n = 1000$	L_1	0.19286810	0.19067074	0.18977448	0.05559002	0.05428802
	L_2	0.12285378	0.11924502	0.11763390	0.04555935	0.04207779
	$WISE$	0.20446025	0.20445991	0.20445985	0.03730548	0.03731946
$n = 10000$	L_1	0.07817725	0.07557446	0.07539929	0.02265668	0.02226697
	L_2	0.05485233	0.04696679	0.04663734	0.01809401	0.01712135
	$WISE$	0.08118316	0.08118296	0.08118296	0.01601341	0.01601330

Na estimação da densidade, através da simulação de dados pela lei de Pareto de parâmetros $(3,4)$, o aumento do tamanho da amostra volta a produzir a diminuição, de forma absoluta, da média dos erros considerados. Neste caso, a introdução dos núcleos de fronteira já se revela benéfica uma vez que, em média, os erros associados aos segundo e terceiro método do núcleo são inferiores aos de $\hat{f}_1$, com execção dos erros L_1 e L_2 para amostras com 100 observações. Novamente, a comparação das médias do erros L_1, L_2 e $WISE$ dos estimadores que se baseiam na estimação dos dados transformados com os restantes, permite inferir um melhor desempenho dos primeiros quer na estimação global, quer na estimação da cauda. De facto, para tamanhos de amostra menores, a diferença no desempenho destes verifica-se bastante superior.

Analisando as diferenças do desempenho médio dos estimadores com correção de fronteira unilateral ou bilateral é notório o melhor desempenho global do estimador que considera como núcleo de fronteira K_3^L. No entanto, a análise do $WISE$ associado a $\hat{f}_4$ e $\hat{f}_5$ revela K_1^L como núcleo mais adequado na estimação da cauda da distribuição após a transformação dos dados para amostras de tamanho 1000 e 10000. Mais ainda, a diferença entre os dois mostra-se residual quando se considera o maior tamanho da amostra.

A análise dos *boxplots* do Anexo C demonstra um maior distânciamento das várias medidas de localização entre os três primeiros métodos e os restantes, face às apresentadas na estimação da densidade lognormal. Geralmente, o mínimo dos erros associados aos três primeiros estimadores é maior que o máximo dos erros respetivos aos restantes estimadores. Esta análise permite ainda notar uma menor variabilidade dos erros produzidos pela estimação da densidade com o quarto e quinto estimador, face aos restantes, bem como a ocorrência de *outliers* de grande magnitude para os erros associados a $\hat{f}_1$, $\hat{f}_2$ e $\hat{f}_3$ quando $n = 100$.

Tabela 5.4 Médias dos erros na estimação da densidade de probabilidade de Weibull considerada

		$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$
$n = 100$	L_1	0.59995942	0.59845483	0.59551206	0.15848918	0.15247908
	L_2	0.44166958	0.44191155	0.43346726	0.15995754	0.14392687
	$WISE$	0.54989834	0.54989708	0.54989625	0.08548387	0.08371968
$n = 1000$	L_1	0.24523441	0.24124778	0.24087305	0.06199578	0.06143031
	L_2	0.18011193	0.16801213	0.16680993	0.06486620	0.06404545
	$WISE$	0.21329229	0.21329197	0.21329200	0.03210136	0.03207547
$n = 10000$	L_1	0.10018689	0.09677818	0.09671189	0.02601998	0.02589290
	L_2	0.08669514	0.06720594	0.06723493	0.02918816	0.03256246
	$WISE$	0.08570827	0.08570816	0.08570814	0.01359378	0.01362449

Tal como nos dois casos anteriores, em média, está patente na tabela acima a diminuição absoluta dos erros com o aumento do tamanho da amostra. Comparando os vários estimadores é também notório o melhor desempenho médio dos estimadores com correção de fronteira avaliado pelos erros considerados. De uma forma geral, podemos observar que os erros médios associados às estimativas obtidas por $\hat{f}_2$ e $\hat{f}_3$ são menores que os respetivos quando consideramos o primeiro método do núcleo implementado. Na estimação da densidade através de simulação de observações da distribuição de Weibull de parâmetros $k = 0.5$ e $\lambda = 1$ torna-se ainda evidente que os quarto e quinto estimadores aumentam de forma significativa o desempenho da estimação, em termos médios, com a diminuição drástica das medidas de erro, sendo mais notório quanto menor o tamanho da amostra considerado. É ainda importante ressalvar que tanto $\hat{f}_4$ como $\hat{f}_5$ se revelam ótimas alternativas na estimação da cauda da densidade, uma vez que a diminuição do $WISE$ é bastante maior nestes métodos do núcleo quando comparada com a respetiva diminuição das medidas L_1 e L_2.

Por outro lado, a comparação da média dos erros associados a $\hat{f}_2$ com os respetivos de $\hat{f}_3$, bem como no caso dos estimadores $\hat{f}_4$ e $\hat{f}_5$, permite inferir que os estimadores que consideram como núcleo de fronteira K_3^L têm um melhor desempenho que os estimadores que tomam K_1^L para esse papel, na maior parte dos casos.

Quanto à variabilidade dos erros considerados, patente nos *boxplots* do Anexo C, de uma forma global, os segundo e terceiro estimadores demonstram-se menos consistentes na estimação da densidade, dada a sua maior variação. Por sua vez, os métodos do núcleo que recorrem à transformação dos dados demonstram erros de estimação com menor amplitude interquartil, ou seja, menos variáveis. Na verdade, o aumento do número de observações têm um efeito na diminuição da dispersão dos erros dos vários estimadores, dada a diminuição da amplitude interquartil. Finalmente, para $n = 100$, está patente uma ocorrência de *outliers* mais severos nos erros associados aos três primeiros estimadores, face aos associados a $\hat{f}_4$ e $\hat{f}_5$.

Por fim, a análise comparativa das três tabelas, para $n = 100$ permite notar que a estimação da primeira densidade apresenta médias de erros muito menores que as restantes, sendo que a estimação da densidade à custa de observações da lei de Weibull apresenta os erros L_1 e L_2 maiores. Quanto à

estimação da cauda, para amostras do menor tamanho os estimadores mostram menor desempenho na lei com a cauda mais pesada, a lei de Pareto.

No que diz respeito à estimação da densidade com amostras de tamanho 1000 e 10000, podemos notar que para os três primeiros métodos do núcleo, os erros produzidos na estimação são sempre menores para a primeira densidade considerada, sendo que estes métodos apresentam um desempenho mais fraco na estimação da lei de Weibull, em termos médios. Por sua vez, a análise das medidas L_1 e L_2 associadas aos estimadores $\hat{f}_4$ e $\hat{f}_5$ permitem auferir um estimação mais precisa para a densidade da lei de Pareto, seguindo-se a lognormal e, por último, a de Weibull. Quanto à estimação da cauda, a medida $WISE$ sugere que, neste caso, a estimação da cauda é mais precisa na lei de Weibull, tendo menor desempenho na lei lognormal.

É ainda importante ressalvar a maior variabilidade nas médias dos erros associados aos três primeiros estimadores, face aos restantes. De facto, nos primeiros, para as várias densidades obtemos valores bastante díspares, ao passo em que os valores associados aos últimos dois estimadores apresentam magnitudes semelhantes quando alteradas as densidades.

Em suma, na maior parte dos casos, podemos notar que quanto mais regular a densidade de probabilidade, melhor o desempenho dos vários métodos do núcleo. Paralelamente, para todas as densidades, os métodos do núcleo com correção de fronteira mostram-se mais precisos que o estimador usual, sendo que $\hat{f}_4$ e $\hat{f}_5$ constituem de forma absoluta melhores alternativas na estimação da função de perda.

Capítulo 6

Conclusão

Esta dissertação tinha como objetivo a apresentação de metodologias de estimação da função de perda que fossem alternativas ao ajustamento paramétrico de densidades de probabilidade (ou misturas destas) aos dados. Tal foi alcançado pela exposição de métodos do núcleo.

No Capítulo 2, foi apresentado o estimador do núcleo usual e discutidos o erros quadrático médio e quadrático médio integrado assumindo a continuidade da verdadeira densidade e das suas duas primeiras derivadas. Por sua vez, no capítulo seguinte surgiu a necessidade de introduzir os núcleos de fronteira que ponderam os dados com maior cuidado na região fronteira, pois dependem não só do ponto estimado como também da sua distância ao extremo do suporte. Esta novidade prendeu-se com a existência de alguns problemas de regularidade na origem em algumas densidades estudadas nas ciências atuariais. O estudo dos erros considerados foi também levado a cabo assumindo as condições de regularidades para a restrição de f ao suporte da variável. De uma forma geral, o critério de escolha baseou-se na minimização do *MISE* para garantir a diminuição tanto do viés como da variância globais do estimador. O Capítulo 4 constituiu uma alternativa às metodologias de estimação anteriores, uma vez que se baseia na estimação dos dados transformados pela função de distribuição da lei de Champernowne modificada ajustada aos dados originais. Aqui a escolha da janela determinou-se pela adaptação empírica dos critérios anteriormente obtidos.

Por fim, foi levado a cabo um estudo de simulação que visava contrapôr as metodologias introduzidas ao longo do trabalho e sobre o qual podemos retirar alguns resultados.

No que diz respeito à análise do desempenho dos estimadores face ao tamanho da amostra, é possível concluir que o aumento do mesmo se traduz em melhorias na estimação da função de perda, dada as diminuições da média dos erros e da dispersão dos erros obtidos. Este resultado era expectável, pois é natural que, quanto mais informação se tem sobre o comportamento da distribuição, mais precisa será a estimação da sua densidade.

Por sua vez, a análise comparativa das três distribuições em estudo permite estabelecer uma correlação direita entre o aumento do desempenho do estimador com a regularidade da verdadeira densidade. De facto, na densidade mais regular, lognormal, os erros de estimação produzidos foram bastante inferiores àqueles obtidos nas restantes distribuições. Contrariamente, na estimação da densidade de Weibull, foi notório algum desajuste dos três primeiros métodos do núcleo testado. Esta irregularidade pode ser justificada pelo facto da densidade tender para infinito à medida que nos aproximamos da origem, bem como pela existência de derivadas não limitadas para a escolha de

parâmetros tomada, na medida em que esta realidade constitui uma violação nos desenvolvimentos assintóticos do *MISE*, traduzindo-se numa escolha desaquada da janela. É também importante referir que a conformidade na estimação da lei lognormal nestes métodos também advirá do facto da distribuição de referência no cálculo da janela de estimação ser a própria. No que diz respeito à estimação das outras duas densidades, esta escolha da distribuição de referência terá sido prejudicial, pois a mesma se revela inflexível na modelação do comportamento na zona fronteira do suporte, sendo contínua no extremo deste e podendo resultar numa escolha de janela imprópria e consequente imprecisão na estimação. Uma possível solução para este problema poderia passar pela escolha de uma lei de referência mais flexível na origem, como é o caso da lei de Weibull. No entanto, tal como explanado no segundo capítulo, esta escolha poderia acarretar novos problemas, na medida em que a estimação dos parâmetros da lei de Weibull passa pela resolução de um sistema não linear de equações que poderia acarretar os respetivos erros para a estimação da função de perda.

Comparem-se agora os diversos métodos do núcleo apresentados. Analisando o desempenho geral do estimador do núcleo usual com os métodos que tomam correção de fronteira na zona extrema do suporte, podemos concluir que a sua introdução não é apenas benéfica em densidades que não revelem problemas de regularidade na origem. Na verdade, para a densidade de probabilidade mais regular a introdução dos núcleos de fronteira nunca se repercutiu em melhorias na precisão da estimação. Já no caso das densidades de Pareto e de Weibull, a análise das tabelas das médias dos erros considerados, bem como dos respetivos *boxplots*, permite-nos inferir melhorias na estimação com a introdução da nova ponderação no intervalo $]0,h]$, o que vai de encontro à motivação teórica dos estimadores em causa. Contrapondo o segundo e terceiro método do núcleo torna-se claro que o núcleo introduzido por Buch-Larsen *et al.* (2005) é mais adequado na estimação da densidade nos pontos onde é aplicado. Tal pode ser explicado pela minimização do termo de variância que oferece, patente na Figura 3.4, compensando assim a menor ordem de convergência no critério de erro global.

No que diz respeito aos estimadores que assentam na estimação da densidade dos dados transformados, a análise do seu desempenho revela que estes são preferíveis na estimação da função de perda, porque são absolutamente mais precisos que os três métodos do núcleo anteriormente escrutinados. Isto revela também que a distribuição de Champernowne modificada pode constituir uma boa ferramenta, caso o nosso objetivo fosse a implementação de métodos paramétricos para a obtenção de estimativas da densidade de probabilidade da variável X. Torna-se necessário ressalvar que a diferença gritante nas médias do $WISE$ induz que estes métodos, para além de serem eficazes na estimação global da densidade produzem muito boas estimativas também na cauda da mesma. Tal poder-se-á dever à transformação dos dados, na medida em que a estimação da densidade dos dados transformados permite uma análise mais cuidada da cauda da distribuição, dada a limitação do suporte da variável transformada. A pouca variação das médias dos erros, nas diferentes densidades, mostram-nos que estes dois métodos do núcleo não se revelam particularmente sensíveis à pouca regularidade de algumas densidades no ponto extremo do suporte. A comparação de $\hat{f}_4$ com $\hat{f}_5$ segue a mesma linha de raciocínio do confronto entre $\hat{f}_2$ e $\hat{f}_3$, em virtude do método que apresenta K_3^L como núcleo de fronteira gerar melhores estimativas, em termos médios.

Em suma, podemos concluir que os métodos do núcleo são alternativas viáveis na estimação da função de perda, na medida em que pressupõe pouco conhecimento do conjunto de dados, ao invés dos métodos paramétricos. Segue também que, neste tipo de problemas, a metodologia de

estimação introduzida por Buch-Larsen *et al.* (2005) é, de facto, uma melhoria aos métodos do núcleo usuais, pois não só expõe melhores estimativas globais da densidade, como aumenta o desempenho na estimação da cauda, fator bastante relevante no problema em estudo. Por fim, não se revela particularmente necessária a consideração de núcleos de fronteira de segunda ordem, uma vez que o núcleo de fronteira K_3^L foi o que se mostrou melhor na ponderação dos dados na zona fronteira do suporte.

Anexo A

Resultados auxiliares

Nesta secção, apresentam-se resultados teóricos necessários para a obtenção das expressões relativas aos *MSE* e *MISE* dos estimadores apesentados aos longo dos Capítulos 2 e 3.

Teorema da Convergência Dominada de Lebesgue - TCD

Teorema 10. *(Cohn, [5], pg. 63) Sejam r uma função não negativa definida em $\mathbb{R}^d$, $d \in \mathbb{N}$, e integrável à Lebesgue e $g, g_1, g_2, \ldots$ funções mensuráveis reais mensuráveis definidas em $\mathbb{R}^p$, com $p \in \mathbb{N}$ e $p \leq d$, verificando as relações*

- $g(x) = \lim_n g_n(x)$,

- $|g_n(x)| \leq r(x), n = 1, 2, \ldots$.

para todo o $x \in \mathbb{R}^p$ com execção num conjunto de medida de Lebesgue nula. Então, g e $g_1, g_2, \ldots$ são funções integráveis e

$$\lim_{n\to\infty} \int g_n(x)\,dx = \int g(x)\,dx.$$

Fórmula de Taylor com Resto Integral

Teorema 11. *(Apostol, [2], pg. 279) Sejam $k \in \mathbb{N}$ e f com derivadas contínuas até à ordem $k+1$ num certo intervalo que contenha a. Então, para todo o x nesse intervalo,*

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(a)}{i!}(x-a)^i + R_k(x),$$

$$R_k(x) = \int_a^x \frac{f^{(k+1)}(s)}{k!}(x-s)^k\,ds.$$

Sendo que nos Capítulos 2 e 3 se torna necessário o desenvolvimento de $f(x-zh)$ em torno de x, particularizemos o resultado acima para o pretendido. Ora, podemos escrever

$$f(x-zh) = f(x) + \sum_{i=1}^{k} \frac{f^{(k)}(x)}{k!}(-zh)^k + R_k(x-zh)\,ds\,,$$

com

$$R_k(x) = \int_x^{x-zh} \frac{f^{(k+1)}(s)}{k!}(x-zh-s)^k\,ds\,.$$

De facto, introduzindo a mudança de variável definida por $s = x - tzh$, tem-se

$$R_k(x) = \int_0^1 \frac{f^{(k+1)}(x-tzh)}{k!}(-zh+tzh)^k(-zh)\,dt\,.$$

De forma equivalente, sai

$$R_k(x) = \frac{(-zh)^{k+1}}{k!}\int_0^1 f^{(k+1)}(x-tzh)(1-t)^k\,dt\,.$$

Teorema de Fubini

Teorema 12. *(Cohn, [5], pg. 147) Se a funão g com valor em* $\mathbb{R}$ *é integrável à Lebesgue, então*

$$\int_{\mathbb{R}^p\times\mathbb{R}^q} g(x,y)\,d(x,y) = \int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^q} g(x,y)\,dy\right)dx = \int_{\mathbb{R}^q}\left(\int_{\mathbb{R}^p} g(x,y)\,dx\right)dy\,.$$

Anexo B

Códigos

```
integravector = function(yy,inf,sup) \#length(yy)=impar
{
  num <- length(yy)
  passo <- (sup-inf)/(num-1)
  yy1 <- yy[1:(num-1)]
  oo1 <- rep(c(0,1),times=(num-1)/2)
  s4 <- sum(na.omit(oo1*yy1))
  yy2 <- yy[2:(num-2)]
  oo2 <- rep(c(0,1),times=(num-1)/2-1)
  s2 <- sum(na.omit(oo2*yy2))
  if (is.infinite(yy[1])|is.na(yy[1])) {yy[1]=yy[2]}
  passo*(yy[1]+4*s4+2*s2+yy[num])/3
}

K=function(z){
  3*(1-z^2)*(abs(z)<1)/4
}
KL1=function(z,w){
  ((1+3*((1-w)/(1+w))^2+6*(1-w)*z/((1+w)^2))/(w+1))*(-1<z)*(z<w)
}
KL3=function(z,w){
  (3*(1-z^2)/(3*w-w^3+2))*(-1<z)*(z<w)
}

tam=c(100,1000,10000)
y=seq(0,5,0.001)
erroslog=array(,c(5,3,500,3))
#erroswei=array(,c(5,3,500,3))
#errospar=array(,c(5,3,500,3))

for (j in 1:500){
for (l in 1:3) {
  n=tam[l]
  sdata=rlnorm(n,0.1,0.4)
  #sdata=rweibull(n,0.9,0.9)
  #sdata=rpareto(n,2,3)


  #\hat{f}_1
  mlmu=sum(log(sdata))/n
  mlsigma=sqrt(sum((log(sdata)-mlmu)^2)/n)
  h1=(15/(n*(exp(-5*mlmu+25*(mlsigma^2)/4)*(12+20*(mlsigma^2)+9*(mlsigma)^4)/(32*sqrt(pi)*
```

```
(mlsigma^5)))))^(1/5)
fest1=function(x){
  sum(K((x-sdata)/h1))/(n*h1)
}
f1=vector()
for (i in 1:5001)
{
  f1[i]=fest1((i-1)/1000)
}

#\hat{f}_2 h_2=h_1
fest2=function(x){
  if (x>h1){
    result=sum(K((x-sdata)/h1))/(n*h1)
    return(result)
  } else {
    alpha=x/h1
    result=sum(KL1((x-sdata)/h1,alpha))/(n*h1)
    return(max(0,result))
  }
}
f2=vector()
for (i in 1:5001)
{
  f2[i]=fest2((i-1)/1000)
}

#\hat{f}_3 h3=h1
fest3=function(x){
  if (x>h1){
    result=sum(K((x-sdata)/h1))/(n*h1)
    return(result)
  } else {
    alpha=x/h1
    result=sum(KL3((x-sdata)/h1,alpha))/(n*h1)
    return(max(0,result))
  }
}
f3=vector()
for (i in 1:5001)
{
  f3[i]=fest3((i-1)/1000)
}

#Champernowne
m=median(sdata)
maxv=function(j){
  k=numeric(2)
  k[1]=n/j[1]   +n*(log(m+j[2])*((m+j[2])^j[1])-log(j[2])*(j[2]^j[1]))/((m+j[2])^j[1]
  -(j[2])^j[1])+sum(log(sdata+j[2]))-2*sum(((log(sdata+j[2])*((sdata+j[2])^j[1])
  +log(m+j[2])*((m+j[2])^j[1])-2*log(j[2])*(j[2]^j[1])))/((((sdata+j[2])^j[1]
                +(m+j[2])^j[1]-2*(j[2]^j[1])))))
  k[2]=n*j[1]*((m+j[2])^(j[1]-1)-(j[2])^(j[1]-1))/((m+j[2])^j[1]-(j[2])^j[1])+(j[1]-1)
  *sum(1/(sdata+j[2]))-2*j[1]*sum((((sdata+j[2])^(j[1]-1)+(m+j[2])^(j[1]-1)-2*j[2]
                ^(j[1]-1)))/((((sdata+j[2])^j[1]+(m+j[2])^j[1]-2*(j[2]^j[1])))))
  return(k)
}
result=vector()
result=nleqslv(c(2,2),maxv)\$x
a=result[1]
```

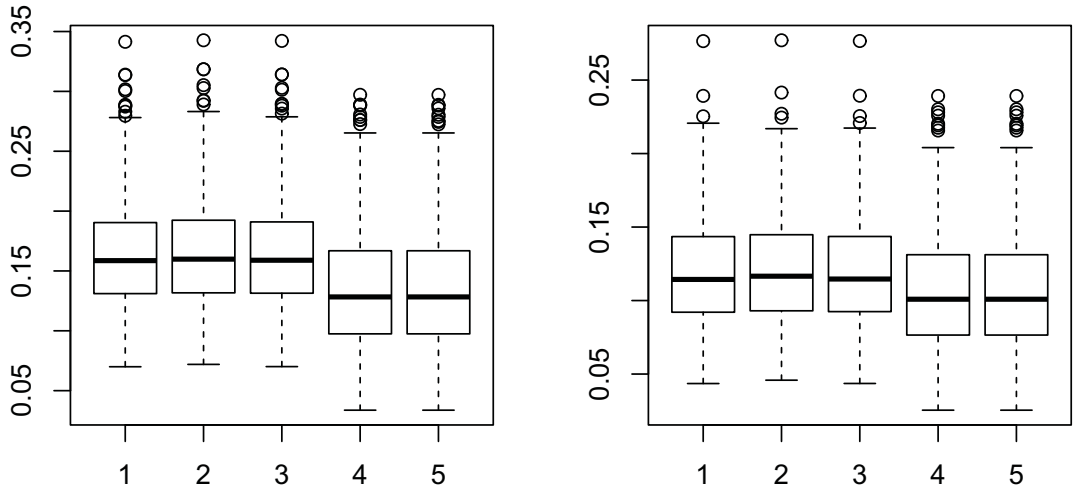
```
c=result[2]
champ=function(y,a,m,c){
  (((y+c)/50)^a-(c/50)^a)/(((y+c)/50)^a+((m+c)/50)^a-2*(c/50)^a)
}
champ2=function(y,a,m,c){
  (((a*((y+c)/10)^(a-1))/(((y+c)/10)^a+((m+c)/10)^a-2*(c/10)^a))*(((m+c)/10)^a-(c/10)^a)
  /(((y+c)/10)^a+((m+c)/10)^a-2*(c/10)^a))/10
}
sdatat=champ(sdata,a,m,c)

#\hat{f}_4
med=mean(sdatat)
v=var(sdatat)
p=max(2.55,med*((med*(1-med))/v-1))
q=max(2.55,(p*(1-med)/med))
h4=min(0.5,(15*(beta(p,q)^2)/(n*((3*(p-1)*(q-2)*(q-1)*gamma(2*p-3)*gamma(2*q-5))
/(2*(2*p-5)*(2*p+2*q-9)*(2*p+2*q-7)*gamma(2*(p+q-5))))))^(1/5))
fest4=function(x){
  if (h4<champ(x,a,m,c) & champ(x,a,m,c)<(1-h4)){
    result=sum(K((champ(x,a,m,c)-sdatat)/h4))*champ2(x,a,m,c)/(n*h4)
    return(result)
  } else {
    if (champ(x,a,m,c)<=h4){
      alpha=champ(x,a,m,c)/h4
      result=sum(KL1((champ(x,a,m,c)-sdatat)/h4,alpha))*champ2(x,a,m,c)/(n*h4)
      return(max(0,result))
    } else {
      alpha=(1-champ(x,a,m,c))/h4
      result=sum(KL1(-(champ(x,a,m,c)-sdatat)/h4,alpha))*champ2(x,a,m,c)/(n*h4)
      return(max(0,result))
    }
  }
}
f4=vector()
for (i in 1:5001)
{
  f4[i]=fest4((i-1)/1000)
}

#\hat{f}_5 h_5=h_4
fest5=function(x){
  if (h4<champ(x,a,m,c) & champ(x,a,m,c)<(1-h4)){
    result=sum(K((champ(x,a,m,c)-sdatat)/h4))*champ2(x,a,m,c)/(n*h4)
    return(result)
  } else {
    if (champ(x,a,m,c)<=h4){
      alpha=champ(x,a,m,c)/h4
      result=sum(KL3((champ(x,a,m,c)-sdatat)/h4,alpha))*champ2(x,a,m,c)/(n*h4)
      return(max(0,result))
    } else {
      alpha=(1-champ(x,a,m,c))/h4
      result=sum(KL3(-(champ(x,a,m,c)-sdatat)/h4,alpha))*champ2(x,a,m,c)/(n*h4)
      return(max(0,result))
    }
  }
}
f5=vector()
for (i in 1:5001)
{
  f5[i]=fest5((i-1)/1000)
```

```
}
erroslog[1,l,j,1]=integravector(abs(f1-dlnorm(y,0.1,0.4)),0,5)
erroslog[1,l,j,2]=(integravector((f1-dlnorm(y,0.1,0.4))^2,0,5))^(1/2)
erroslog[1,l,j,3]=(integravector(((f1-dlnorm(y,0.1,0.4))^2)*y^2,0,5))^(1/2)
erroslog[2,l,j,1]=integravector(abs(f2-dlnorm(y,0.1,0.4)),0,5)
erroslog[2,l,j,2]=(integravector((f2-dlnorm(y,0.1,0.4))^2,0,5))^(1/2)
erroslog[2,l,j,3]=(integravector(((f2-dlnorm(y,0.1,0.4))^2)*y^2,0,5))^(1/2)
erroslog[3,l,j,1]=integravector(abs(f3-dlnorm(y,0.1,0.4)),0,5)
erroslog[3,l,j,2]=(integravector((f3-dlnorm(y,0.1,0.4))^2,0,5))^(1/2)
erroslog[3,l,j,3]=(integravector(((f3-dlnorm(y,0.1,0.4))^2)*y^2,0,5))^(1/2)
erroslog[4,l,j,1]=integravector(abs(f4-dlnorm(y,0.1,0.4)),0,5)
erroslog[4,l,j,2]=(integravector((f4-dlnorm(y,0.1,0.4))^2,0,5))^(1/2)
erroslog[4,l,j,3]=(integravector(((f4-dlnorm(y,0.1,0.4))^2)*y^2,0,5))^(1/2)
erroslog[5,l,j,1]=integravector(abs(f5-dlnorm(y,0.1,0.4)),0,5)
erroslog[5,l,j,2]=(integravector((f5-dlnorm(y,0.1,0.4))^2,0,5))^(1/2)
erroslog[5,l,j,3]=(integravector(((f5-dlnorm(y,0.1,0.4))^2)*y^2,0,5))^(1/2)
#erroswei[1,l,j,1]=integravector(abs(f1-dweibull(y,0.9,0.9)),0,5)
#erroswei[1,l,j,2]=(integravector((f1-dweibull(y,0.9,0.9))^2,0,5))^(1/2)
#erroswei[1,l,j,3]=(integravector(((f1-dweibull(y,0.9,0.9))^2)*y^2,0,5))^(1/2)
#erroswei[2,l,j,1]=integravector(abs(f2-dweibull(y,0.9,0.9)),0,5)
#erroswei[2,l,j,2]=(integravector((f2-dweibull(y,0.9,0.9))^2,0,5))^(1/2)
#erroswei[2,l,j,3]=(integravector(((f2-dweibull(y,0.9,0.9))^2)*y^2,0,5))^(1/2)
#erroswei[3,l,j,1]=integravector(abs(f3-dweibull(y,0.9,0.9)),0,5)
#erroswei[3,l,j,2]=(integravector((f3-dweibull(y,0.9,0.9))^2,0,5))^(1/2)
#erroswei[3,l,j,3]=(integravector(((f3-dweibull(y,0.9,0.9))^2)*y^2,0,5))^(1/2)
#erroswei[4,l,j,1]=integravector(abs(f4-dweibull(y,0.9,0.9)),0,5)
#erroswei[4,l,j,2]=(integravector((f4-dweibull(y,0.9,0.9))^2,0,5))^(1/2)
#erroswei[4,l,j,3]=(integravector(((f4-dweibull(y,0.9,0.9))^2)*y^2,0,5))^(1/2)
#erroswei[5,l,j,1]=integravector(abs(f5-dweibull(y,0.9,0.9)),0,5)
#erroswei[5,l,j,2]=(integravector((f5-dweibull(y,0.9,0.9))^2,0,5))^(1/2)
#erroswei[5,l,j,3]=(integravector(((f5-dweibull(y,0.9,0.9))^2)*y^2,0,5))^(1/2)
#errospar[1,l,j,1]=integravector(abs(f1-dpareto(y,2,3)),0,5)
#errospar[1,l,j,2]=(integravector((f1-dpareto(y,2,3))^2,0,5))^(1/2)
#errospar[1,l,j,3]=(integravector(((f1-dpareto(y,2,3))^2)*y^2,0,5))^(1/2)
#errospar[2,l,j,1]=integravector(abs(f2-dpareto(y,2,3)),0,5)
#errospar[2,l,j,2]=(integravector((f2-dpareto(y,2,3))^2,0,5))^(1/2)
#errospar[2,l,j,3]=(integravector(((f2-dpareto(y,2,3))^2)*y^2,0,5))^(1/2)
#errospar[3,l,j,1]=integravector(abs(f3-dpareto(y,2,3)),0,5)
#errospar[3,l,j,2]=(integravector((f3-dpareto(y,2,3))^2,0,5))^(1/2)
#errospar[3,l,j,3]=(integravector(((f3-dpareto(y,2,3))^2)*y^2,0,5))^(1/2)
#errospar[4,l,j,1]=integravector(abs(f4-dpareto(y,2,3)),0,5)
#errospar[4,l,j,2]=(integravector((f4-dpareto(y,2,3))^2,0,5))^(1/2)
#errospar[4,l,j,3]=(integravector(((f4-dpareto(y,2,3))^2)*y^2,0,5))^(1/2)
#errospar[5,l,j,1]=integravector(abs(f5-dpareto(y,2,3)),0,5)
#errospar[5,l,j,2]=(integravector((f5-dpareto(y,2,3))^2,0,5))^(1/2)
#errospar[5,l,j,3]=(integravector(((f5-dpareto(y,2,3))^2)*y^2,0,5))^(1/2)
      }}
```

Anexo C

Resultados do estudo de simulação

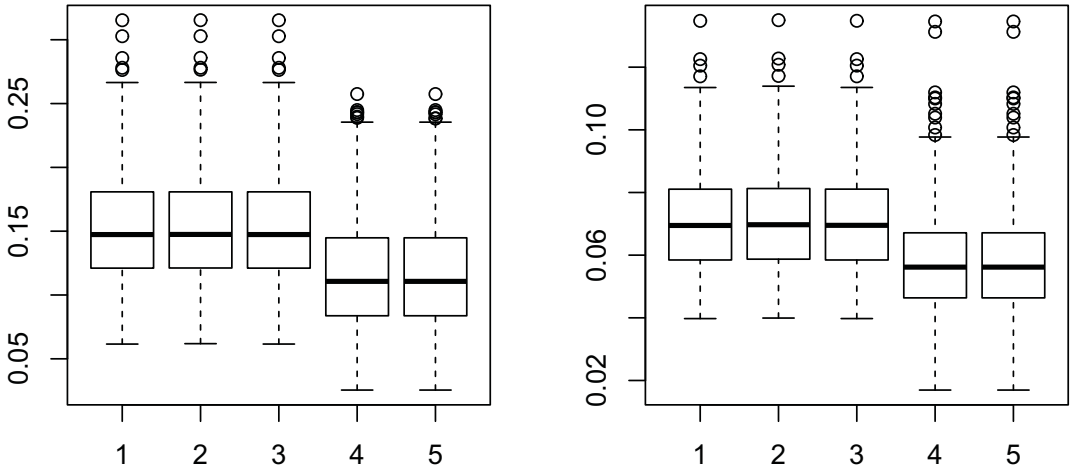
Este anexo tem como objetivo a apresentação dos boxplots dos erros obtidos na estimação da densidade pelos cinco métodos do núcleo e para cada combinação de lei de probabilidade e tamanho da amostra. Em todos as representações gráficas, a indicação $i \in \{1,2,3,4,5\}$ refere-se ao estimador $\hat{f}_i$.



(a) $L_1 \mid n = 100 \mid$ lognormal

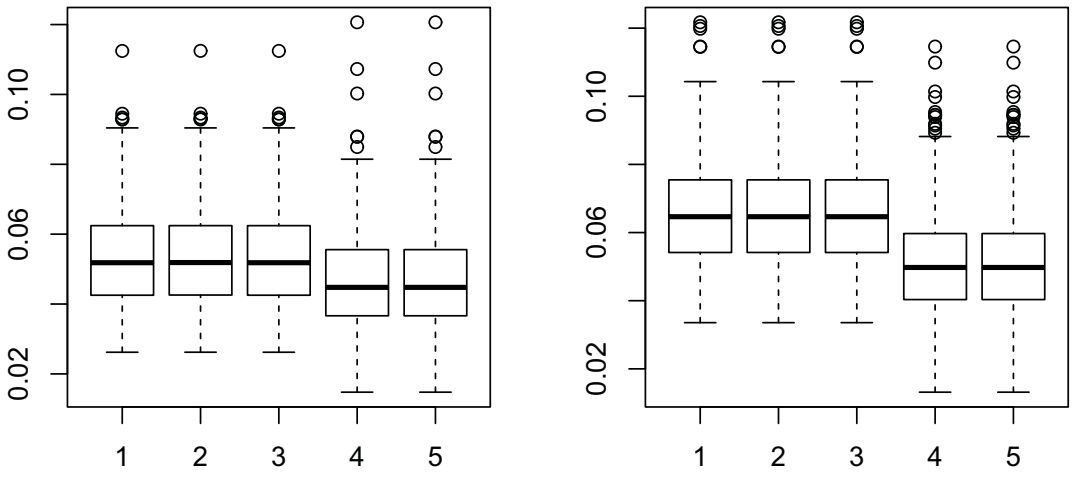
(b) $L_2 \mid n = 100 \mid$ lognormal

Fig. C.1 Boxplots dos erros L_1 e L_2 obtidos na estimação da densidade lognormal considerada com base em 100 observações



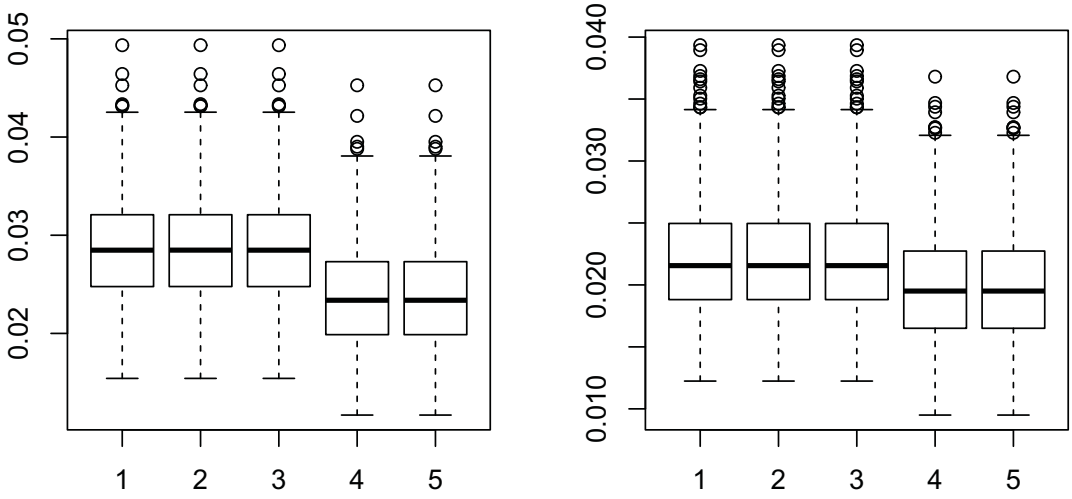
(a) $WISE \mid n = 100 \mid$ lognormal (b) $L_1 \mid n = 1000 \mid$ lognormal

Fig. C.2 Boxplots dos erros $WISE$ e L_1 obtidos na estimação da densidade lognormal considerada com base em 100 e 1000 observações, respetivamente



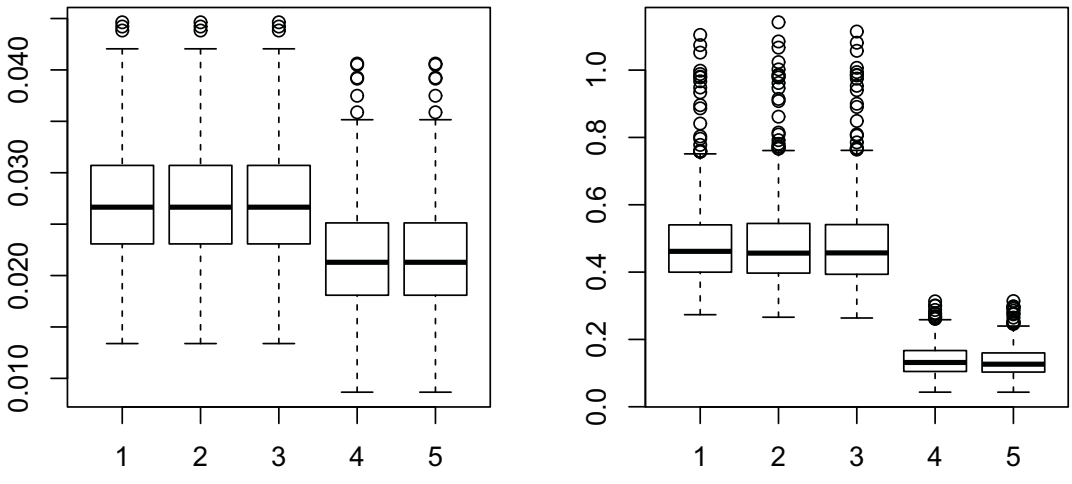
(a) $L_2 \mid n = 1000 \mid$ lognormal (b) $WISE \mid n = 1000 \mid$ lognormal

Fig. C.3 Boxplots dos erros L_2 e $WISE$ obtidos na estimação da densidade lognormal considerada com base em 1000 observações



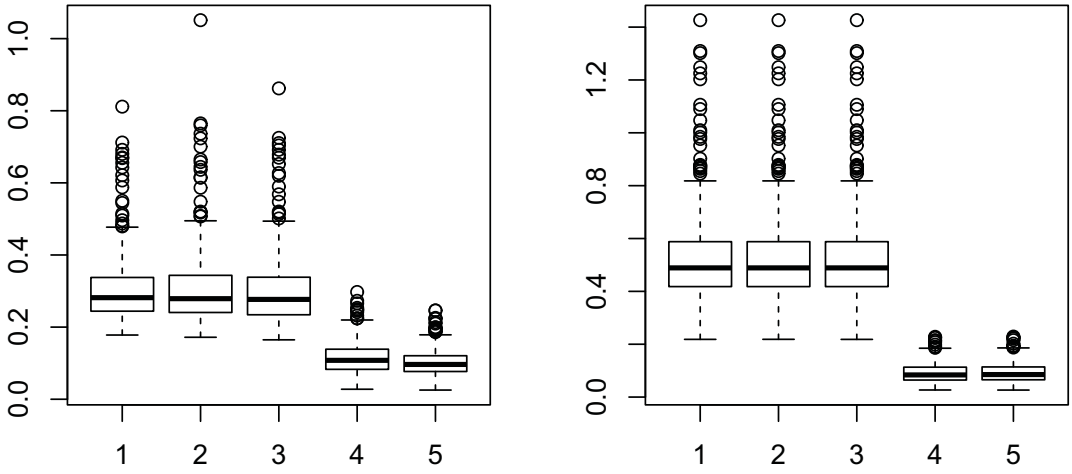
(a) L_1 | $n = 10000$ | lognormal (b) L_2 | $n = 10000$ | lognormal

Fig. C.4 Boxplots dos erros L_1 e L_2 obtidos na estimação da densidade lognormal considerada com base em 10000 observações



(a) $WISE$ | $n = 10000$ | lognormal (b) L_1 | $n = 100$ | Pareto

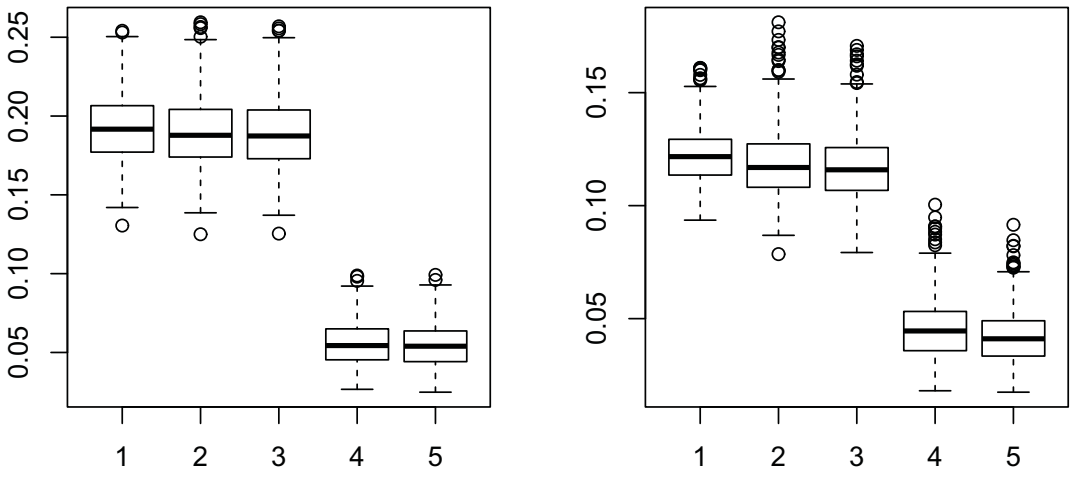
Fig. C.5 Boxplots dos erros $WISE$ e L_1 obtidos na estimação da densidade lognormal e de Pareto considerada com base em 10000 e 100 observações, respetivamente



(a) L_2 | $n = 100$ | Pareto

(b) $WISE$ | $n = 100$ | Pareto

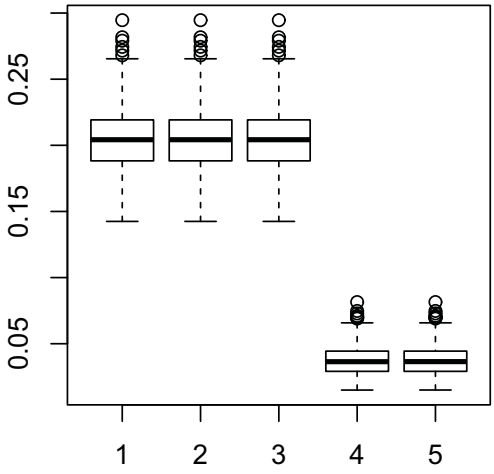
Fig. C.6 Boxplots dos erros L_2 e $WISE$ obtidos na estimação da densidade de Pareto considerada com base em 100 observações



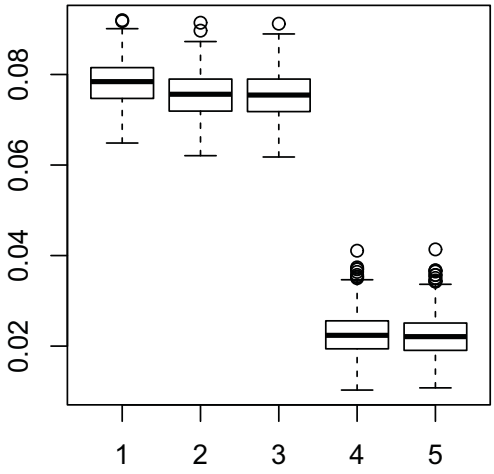
(a) L_1 | $n = 1000$ | Pareto

(b) L_2 | $n = 1000$ | Pareto

Fig. C.7 Boxplots dos erros L_1 e L_2 obtidos na estimação da densidade de Pareto considerada com base em 1000 observações

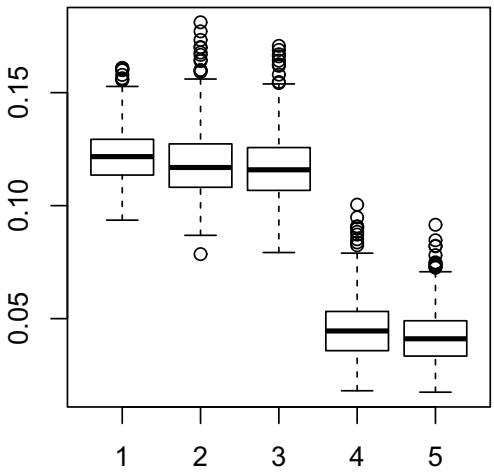


(a) $WISE$ | $n = 1000$ | Pareto

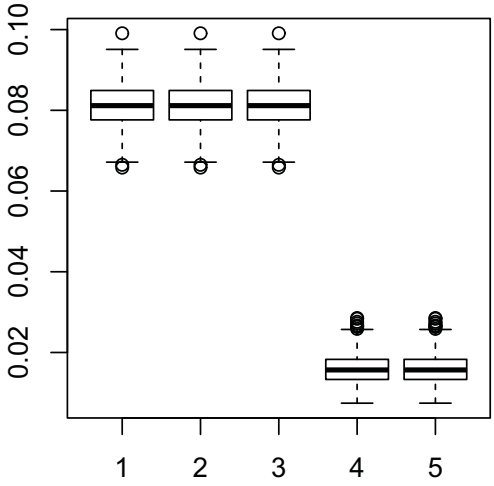


(b) L_1 | $n = 10000$ | Pareto

Fig. C.8 Boxplots dos erros $WISE$ e L_1 obtidos na estimação da densidade de Pareto considerada com base em 1000 e 10000 observações, respetivamente

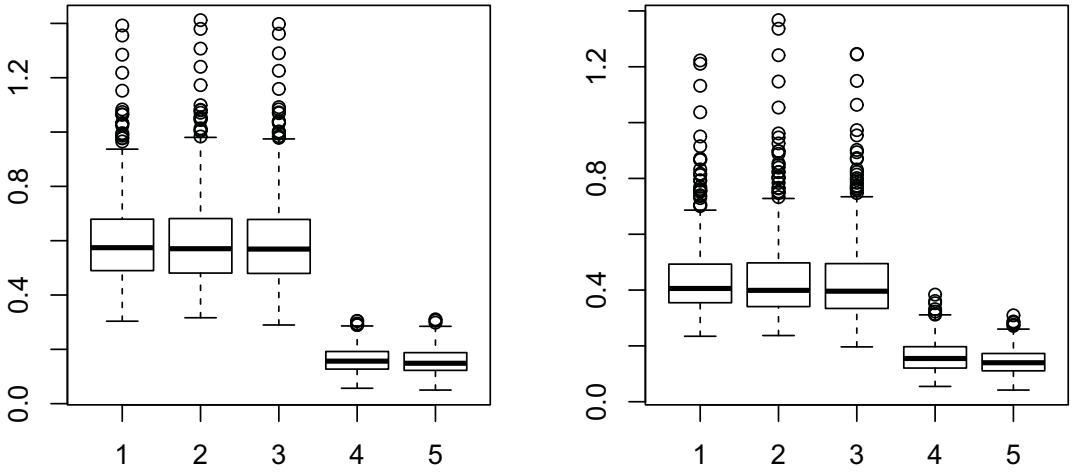


(a) L_2 | $n = 10000$ | Pareto



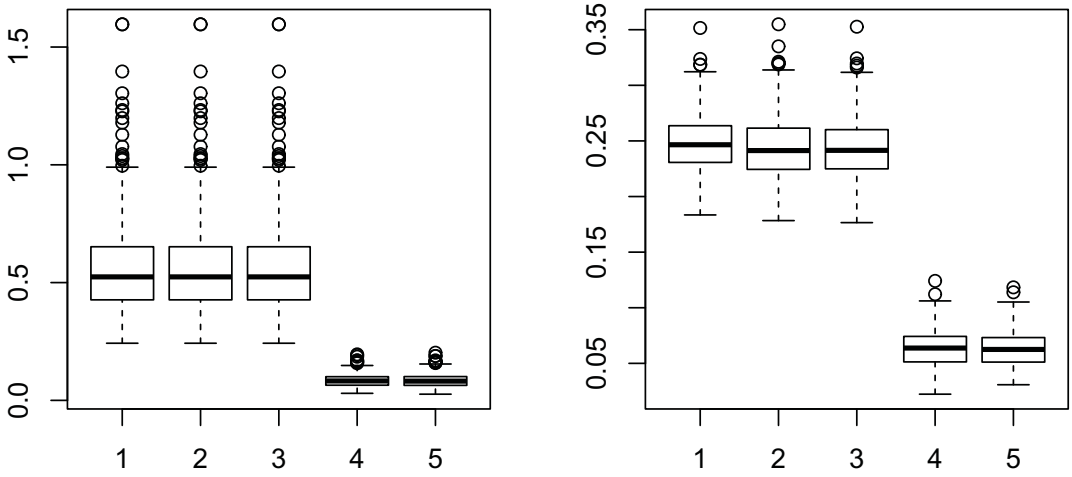
(b) $WISE$ | $n = 10000$ | Pareto

Fig. C.9 Boxplots dos erros L_2 e $WISE$ obtidos na estimação da densidade de Pareto considerada com base em 10000 observações



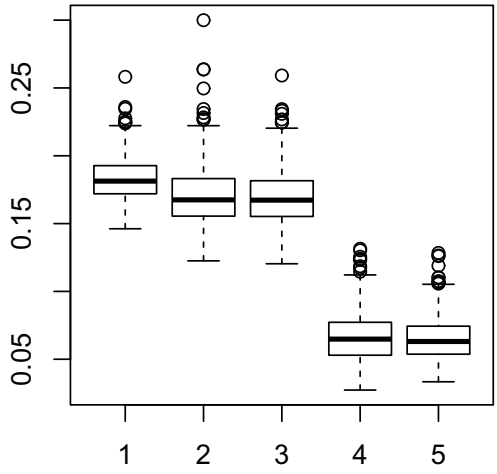
(a) L_1 | $n = 100$ | Weibull (b) L_2 | $n = 100$ | Weibull

Fig. C.10 Boxplots dos erros L_1 e L_2 obtidos na estimação da densidade de Weibull considerada com base em 100 observações

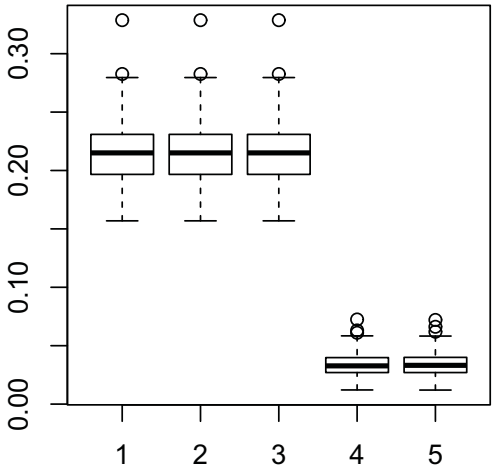


(a) $WISE$ | $n = 100$ | Weibull (b) L_1 | $n = 1000$ | Weibull

Fig. C.11 Boxplots dos erros $WISE$ e L_1 obtidos na estimação da densidade de Weibull considerada com base em 100 e 1000 observações, respetivamente

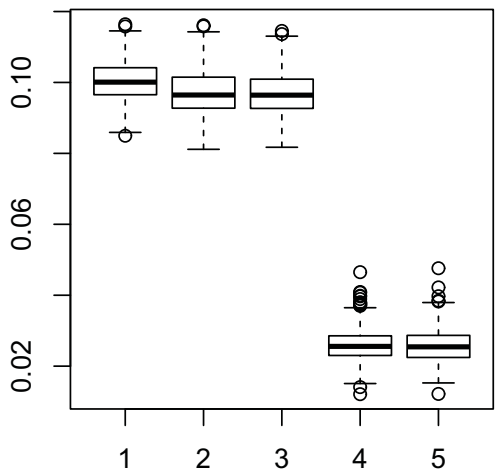


(a) L_2 | $n = 1000$ | Weibull

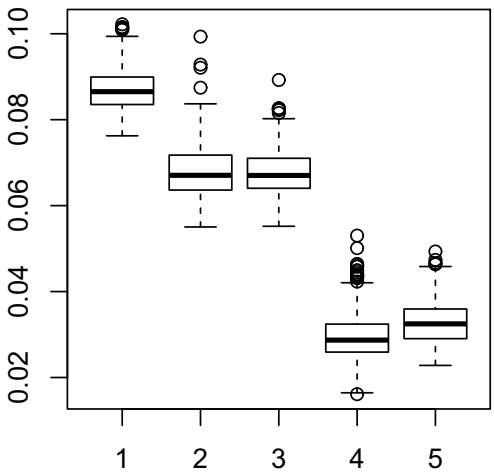


(b) $WISE$ | $n = 1000$ | Weibull

Fig. C.12 Boxplots dos erros L_2 e $WISE$ obtidos na estimação da densidade de Weibull considerada com base em 1000 observações

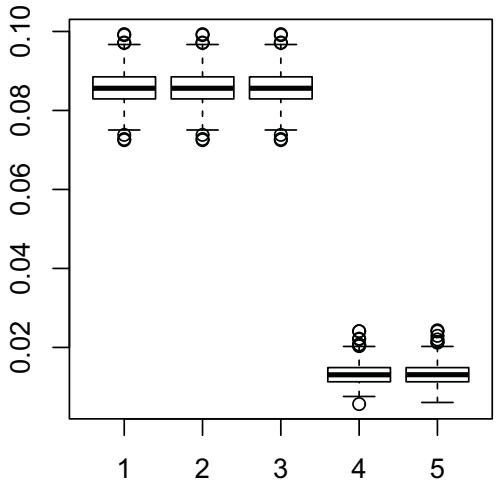


(a) L_1 | $n = 10000$ | Weibull



(b) L_2 | $n = 10000$ | Weibull

Fig. C.13 Boxplots dos erros L_1 e L_2 obtidos na estimação da densidade de Weibull considerada com base em 10000 observações



(a) $WISE \mid n = 10000 \mid$ Weibull

Fig. C.14 Boxplot dos erros $WISE$ obtidos na estimação da densidade de Weibull considerada com base em 10000 observações

Bibliografia

Abu Bakar, S., Hamzah, N., Maghsoudi, M., and Nadarajah, S. (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics*, (16):146–154.

Apostol, T. M. (1967). *Calculus - One-Variable Calculus, with an Introduction to Linear Algebra*. John Wiley & Sons, Inc., United States of America, second edition.

Bosq, D. and Lecoutre, J. (1987). *Theorie de L'estimation Fonctionnelle*. Economica, Paris.

Buch-Larsen, T., Nielsen, J. P., Guillém, M., and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–518.

Cohn, D. L. (1980). *Measure Theory*. Birkhäuser, Boston.

Cooray, K. and Ananda, M. (2005). Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal*, 2005(5):321–334.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression funcions. in: Smoothing techniques for curve estimation, gasser, t., rosenblatt, m. *Lecture Notes in Mathematics*, (757):23–68.

Käärik, M. and Umbleja, M. (2010). Estimation on claim size distributions in estonia traffic insurance. *Selected Topics in Applied Computing*, pages 28–32.

Käärik, M. and Umbleja, M. (2011). On claim size fitting and rough estimation of risk premiums based on estonian traffic insurance example. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(5):17–24.

Käärik, M. and Umbleja, M. (2012). Estimation on claim size distributions in estonia traffic insurance. *Acta et Commentationes Universitatis Tartuenseis de Mathematica*, 16(1):53–67.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076.

Preda, V. and Ciumara, R. (2006). On composite models: Weibull-pareto and lognormal-pareto. a comparative study. *Romanian Journal of Economic Forecasting*, 3(2):32–46.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Tenreiro, C. (2010). *Uma introdução não paramétrica à estimação não-paramétrica da densidade*. Coimbra.

Tenreiro, C. (2013). Boundary kernel for distribution function estimation. *REVSTAT - Statistical Journal*, 11(2):169–190.

Teodorescu, S. and Vernic, R. (2006). A composite exponential-pareto distribution. *The Annals of the "Ovidius" University of Constanta, Mathematics Series*, 14(1):99–108.

Wand, P., Marron, J., and Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, (86):343–361.