

DM

Estudo Sobre a Escolha Estatística de Modelos Extremais na Metodologia POT

DISSERTAÇÃO DE MESTRADO

António Miguel Nunes Gomes

MESTRADO EM MATEMÁTICA



UNIVERSIDADE da MADEIRA

A Nossa Universidade

www.uma.pt

fevereiro | 2018

Estudo Sobre a Escolha Estatística de Modelos Extremos na Metodologia POT

DISSERTAÇÃO DE MESTRADO

António Miguel Nunes Gomes

MESTRADO EM MATEMÁTICA

ORIENTADORA
Délia Canha Gouveia Reis

Júri:

Doutora Ana Maria Cortesão Pais Figueira da Silva Abreu

– Professora Auxiliar da Universidade da Madeira

Doutora Sandra Maria Freitas Mendonça

– Professor Auxiliar da Universidade da Madeira

Doutora Délia Canha Gouveia Reis

– Professora Auxiliar da Universidade da Madeira

Agradecimentos

A elaboração desta dissertação de Mestrado contou com o apoio e incentivos cruciais das pessoas que sempre acreditaram em mim desde a minha entrada no Mestrado em Matemática até aos dias de hoje. Assim, devo dizer que sem o vosso apoio não teria conseguido concretizar esta importante etapa da minha vida.

Desta forma, quero agradecer:

- Aos professores pelos conhecimentos transmitidos ao longo do Mestrado e que em muito contribuíram para o meu enriquecimento intelectual e pessoal;
- Aos meus familiares que sempre se mostraram orgulhosos com os feitos académicos por mim alcançados, nomeadamente os meus pais e irmãos pela transmissão de confiança e motivação para finalizar este curso;
- Aos meus grandes amigos e colegas de curso Ivo Ferreira, Duarte Sousa e Vítor Jesus que se mostraram de vital importância tanto pelo apoio curricular que me deram quer pelo companheirismo. Tenho a dizer que foi um enorme prazer ter feito este curso com vocês;
- À minha amiga Deise Faria que mesmo não tendo feito o curso comigo sempre se disponibilizou a ajudar, nomeadamente com a língua inglesa, e a quem muito devo agradecer pois a sua ajuda foi fundamental;
- À minha amiga e companheira de estudo na reta final desta dissertação Tânia Pinto pelo seu apoio e transmissão de motivação para os estudos após um dia de trabalho, muitas vezes cansativo;
- À Catarina Rebolo e Cleodice Fernandes, amigas de longa data, que sempre me motivaram e incentivaram ao longo deste curso;
- Aos meus verdadeiros amigos que de maneira alguma deixaram de se mostrar disponíveis para ajudar e sempre depositaram em mim muita confiança e orgulho.

Quanto a minha orientadora, a Professora Dr.^a Délia Canha Gouveia Reis, tenho a dar-lhe o meu maior agradecimento pois sem si nada disto seria possível. A Professora revelou-se um apoio fundamental para realização deste trabalho que me deixa cheio de orgulho. A sua forma de incentivar, motivar, ensinar e transmitir conhecimento, aliado sua simpatia e paciência fazem de si uma formidável Professora para além de uma ótima pessoa. Fico imensamente grato por tudo o que fez por mim e espero um dia voltar a trabalhar consigo.

A todas as pessoas aqui mencionadas o meu mais sincero obrigado.

Resumo

A necessidade de estudar e compreender os eventos extremos que surgem nas mais diversas áreas do nosso cotidiano levou ao aparecimento de diferentes metodologias de estudo de tais acontecimentos. Uma dessas metodologias é a denominada metodologia POT, na qual são analisados os valores que excedem um determinado nível elevado. Neste contexto, um dos problemas de grande importância na prática é a escolha da distribuição a utilizar na modelação desses valores extremos.

O principal objetivo desta dissertação é a elaboração de uma coletânea de métodos estatísticos existentes na literatura científica relativa ao problema da escolha entre a distribuição exponencial e a distribuição generalizada de Pareto não exponencial. A realização de uma breve análise da metodologia POT é também um dos objetivos deste estudo, assim como a exemplificação, com recurso ao *software* estatístico R, dos procedimentos necessários para a realização dos métodos estatísticos analisados.

Nesta dissertação é realizada uma breve introdução à Teoria dos Valores Extremos, com principal incidência nas metodologias de Gumbel e POT. A metodologia POT é estudada com mais pormenor, sendo dada a merecida importância à estimação dos parâmetros das distribuições exponencial e generalizada de Pareto não exponencial. A par da estimação de parâmetros, a estimação de quantis extremos é também abordada dada a sua relevância na gestão do risco. De forma a ser realizada uma escolha entre as distribuições exponencial e generalizada de Pareto não exponencial são sugeridos alguns métodos estatísticos que assentam em testes de hipóteses, em intervalos de confiança e em métodos gráficos. De modo a exemplificar a aplicação destes métodos estatísticos, assim como dos conceitos estudados, é realizada uma ilustração prática na qual é utilizada uma base de dados da área das Finanças.

Palavras-Chave: Metodologia POT, Distribuição Exponencial, Distribuição Generalizada de Pareto não Exponencial, Testes de Hipóteses, Intervalos de Confiança, Métodos Gráficos.

Abstract

The need to study and understand the extreme events that arise in the most diverse areas of our daily life led to the appearance of different approaches for the study of such events. One of these approaches is the POT approach, in which the values that exceed a high threshold are analyzed. In this context, one of the problems of great importance in practice is the choice of the distribution to be used in the modeling of these extreme values.

The main objective of this dissertation is the elaboration of a collection of statistical methods for the problem of statistical choice between exponential and non-exponential generalized Pareto distributions. A brief analysis of the POT approach is also one of the objectives of this study, as well as the exemplification, by means of the statistical software R, of the procedures required to perform the statistical methods analysed.

In this dissertation, a brief introduction to the Extreme Value Theory is carried out which focuses mainly on Gumbel's and POT approaches. The POT approach is studied in more detail, with the deserved importance being given to the estimation of exponential and generalized Pareto distributions parameters. In addition to parameter estimation, the estimation of extreme quantiles is also addressed given their relevance in risk management. In order to make a statistical choice between the exponential and the Generalized Pareto distributions, some statistical methods based on hypothesis tests, confidence intervals and graphical methods are suggested. In order to exemplify the application of these statistical methods, as well as the studied concepts, a practical illustration is performed using a database from the Finance field.

Keywords: POT Approach, Exponential Distribution, Non-exponential Generalized Pareto Distributions, Statistical Tests, Confidence Intervals, Graphical Methods.

Índice

Lista de Figuras	xi
Lista de Tabelas	xiii
Notação	xv
1 Introdução	1
2 Sobre a Teoria dos Valores Extremos	5
2.1 Máximos por blocos	5
2.2 Excessos acima de um nível elevado	10
3 Metodologia POT	17
3.1 Estimação de parâmetros	20
3.2 Estimação de quantis extremos	28
4 Escolha estatística de modelos extremos	31
4.1 Teste de hipóteses	31
4.1.1 Estatística de teste T_1	33
4.1.2 Estatística de teste T_2	35
4.1.3 Estatística de teste T_3	36
4.1.4 Estatística de teste T_4	37
4.1.5 Estatística de teste T_5	38
4.1.6 Estatística de teste T_6	38
4.2 Intervalos de confiança e métodos gráficos	40
5 Ilustração prática	45
5.1 Resultados	47
5.1.1 Testes de hipóteses	51
5.1.2 Intervalos de confiança e métodos gráficos	60
5.1.3 Estimação dos quantis extremos	64
6 Conclusão	67

Bibliografía	71
Índice Remissivo	76

Lista de Figuras

2.1	A função densidade de probabilidade λ e as funções densidade probabilidade ψ e φ para $\alpha = 1$	9
2.2	Função de distribuição F e função de distribuição condicional F_u	11
2.3	Função densidade de probabilidade $h_{\gamma,1}$ para $\gamma = 0.5, 0$ e -0.25	14
4.1	Gráficos de diagnóstico da biblioteca <code>ismev</code>	43
5.1	Gráficos das bibliotecas <code>evd</code> , <code>evir</code> e <code>ismev</code>	48
5.2	Gráfico MRL da biblioteca <code>evmix</code>	49
5.3	Gráficos MRL considerando 0.5%, 1% e 2.5% dos valores extremos.	50
5.4	Gráficos MRL considerando 2.5%, 5% e 10% dos valores extremos.	51
5.5	Estimativas e intervalos de confiança para γ	62
5.6	Gráficos de diagnóstico da biblioteca <code>ismev</code> para $u_{0.05} = 1.499139$	63
5.7	Gráficos de diagnóstico da biblioteca <code>ismev</code> para $u_{0.10} = 1.059779$	63

Lista de Tabelas

4.1	Estadísticas de teste - distribuições.	40
5.1	Descrição dos dados <i>S&P500</i>	46
5.2	Limiars e número de excedências.	49
5.3	Limiars e valores resultantes do procedimento 7.	52
5.4	Limiars e valores resultantes do procedimento 9.	53
5.5	Limiars e valores resultantes do procedimento 10.	54
5.6	Limiars e valores resultantes do procedimento 11.	55
5.7	Limiars e valores resultantes do procedimento 12.	56
5.8	Limiars e valores resultantes do procedimento 13.	57
5.9	Limiars e valores resultantes do procedimento 14.	57
5.10	Limiars e valores resultantes do procedimento 15.	58
5.11	Decisão de rejeitar ou não a distribuição exponencial ($\alpha = 0.05$).	59
5.12	Estimativas para γ e intervalos de 95% confiança resultantes do procedimento 16.	61
5.13	Estimação do VaR e CTE para os diferentes limiars.	65

Notação

EVT	Teoria dos Valores Extremos.
POT	<i>Peaks Over Threshold.</i>
GP	Generalizada de Pareto não exponencial.
GEV	Generalizada de valores extremos
u	Limiar.
Método MOM	Método dos momentos.
Método PWM	Método dos momentos ponderados de probabilidade.
Método ML	Método de máxima verosimilhança.
VaR	<i>Value-at-Risk.</i>
CTE	Valor esperado de cauda condicional.
$X_i \stackrel{d}{=} X$	Têm a mesma distribuição de probabilidades.
$N_u = m$	Número de observações que excedem um limiar u .
$\{\cdot\}$	Arredondamento para o número inteiro mais próximo.

Capítulo 1

Introdução

A Teoria dos Valores Extremos (EVT do inglês *Extreme Value Theory*) emergiu como uma das disciplinas estatísticas mais importantes para a ciência aplicada nos últimos 60 anos. Como o próprio nome sugere, o foco da EVT é o estudo dos valores extremos, valores esses que, apesar de não serem muito frequentes, têm consequências imprevisíveis quando surgem. Assim, a EVT contém ferramentas que são usadas para considerar probabilidades associadas a eventos raros mas de grande impacto nas mais diversas áreas de aplicação, tais como Hidrologia, Oceanografia, Meteorologia, Poluição, Seguros, Telecomunicações ou Finanças.

A área das Finanças tem sido alvo de inúmeros estudos por parte da EVT, uma vez que os acontecimentos extremos que por vezes ocorrem nos mercados financeiros são de grande importância para os investidores em particular e para toda a economia em geral. Esses acontecimentos ainda não são bem compreendidos pelos estudiosos da área, de tal forma que mesmo depois de passados vários anos desde o dia 19 de outubro de 1987, a “*Black Monday*”, ainda se está a tentar perceber o que causou o colapso do mercado de ações. O desenvolvimento da EVT ao longo do tempo fornece assim uma base teórica mais sólida sobre a qual são construídos modelos estatísticos com o intuito de melhor compreender esses eventos extremos.

A EVT tal como a conhecemos hoje, teve como um dos seus primeiros artigos o trabalho realizado por Fisher [19] nos anos 20 do século passado. Para além de Fisher, Tippet e Gnedenko muito fizeram para terem um lugar na história da EVT, chegando a ter os seus nomes atribuídos a um dos teoremas fundamentais da EVT, o Teorema de *Fisher-Tippet-Gnedenko*. Outras personalidades da área como Weibull, Gumbel, este tendo sido um dos pioneiros e mais sonantes nomes

da EVT, ou Fréchet também se revelaram de vital importância, de tal forma que temos três importantes distribuições da EVT com os seus nomes, a Distribuição de Weibull, a Distribuição de Gumbel e a Distribuição de Fréchet.

A metodologia de Gumbel (ou dos blocos ou dos máximos anuais) e a metodologia POT, designação que advém do inglês *Peaks Over Threshold*, são duas das metodologias da EVT. Na metodologia POT, que também é denominada de metodologia dos Excessos Acima de um Nível Elevado, são considerados relevantes os maiores valores observados que se encontram acima do nível ou limiar elevado u . Isto constitui uma vantagem relativamente à metodologia de Gumbel na qual é apenas considerado o valor máximo de cada período em estudo, como por exemplo o máximo por cada ano. Assim, ao contrário da metodologia de Gumbel, na metodologia POT não corremos o risco de perder observações elevadas ou de considerar observações baixas [17]. Uma breve análise destas metodologias é apresentada no Capítulo 2 desta dissertação, sendo a metodologia POT estudada com mais pormenor no Capítulo 3.

Davison e Smith, duas personalidades da área que em muito contribuíram para o desenvolvimento da metodologia POT, afirmaram em [15] que a metodologia POT começou a ser desenvolvida de forma sistemática nos trabalhos realizados por Todorovic e Rousselle em [42] e Todorovic e Zelenhasic em [43]. Por sua vez, Davison e Smith também elaboram, entre outros trabalhos de grande importância para esta metodologia, os estudos em [14], [15] e [40]. Outro marco importante para a análise dos excessos (diferenças entre os valores acima do limiar e este valor) foi estabelecido por Pickands em [36] e Balkema e de Haan em [3]. Estes autores estabeleceram que a função de distribuição condicional dos excessos acima do limiar u pode ser bem aproximada por uma função de distribuição generalizada de Pareto não exponencial (GP).

De modo a analisarmos os valores acima de um limiar temos em primeiro lugar de encontrar esse valor. No entanto, a escolha do limiar não é consensual no meio científico uma vez que não se consegue obter um valor ótimo que assegure os pressupostos básicos inerentes a essa escolha. A escolha de um limiar demasiado elevado leva a suprimirmos valores que são realmente relevantes. Por outro lado, a escolha de um valor demasiado baixo tem como consequência incluírmos valores que podem não ser úteis para o estudo dos valores extremos. Assim, a escolha de um limiar adequado é de grande importância. Tal como o gráfico da função de excesso médio empírica, a avaliação da estabilidade das estimativas

dos parâmetros da função de distribuição GP para uma variedade de diferentes limiares permite estabelecer um método gráfico a utilizar nessa escolha. Estas estimativas podem ser obtidas por inúmeros métodos de estimação, tais como o método dos momentos (método MOM), o método dos momentos ponderados pela probabilidade (método PWM) ou o método de máxima verosimilhança (método ML), sendo este último o mais utilizado na bibliografia existente. Para além da estimação de parâmetros temos também a estimação de quantis extremos, a qual se revela de vital importância para a gestão do risco. Uma medida de risco muito utilizada neste tipo de gestão é o valor denominado *Value-at-Risk* (VaR), valor que é simplesmente um quantil extremo. Outra medida de risco muito utilizada na área das Finanças é o denominado valor esperado de cauda condicional que denotaremos por CTE do inglês *Condicional Tail Expectation*. Levando em conta estas duas medidas de risco obtemos, respetivamente, através de uma probabilidade associada muito baixa, um valor a partir do qual se espera que haja um acontecimento extremo e quantificar em média que valor decorrerá desse acontecimento.

A obtenção de tais valores associados a um acontecimento extremo está associada à escolha de uma função de distribuição. Na metodologia POT, essa escolha pode recair entre a escolha da função de distribuição exponencial ou função de distribuição GP. A função de distribuição exponencial é a favorita muito por conta da sua simplicidade na estimação tanto dos quantis extremos quanto dos parâmetros. Assim, a escolha entre a função de distribuição exponencial e a função de distribuição GP é um problema estatístico importante, sendo esta temática a motivação principal desta dissertação.

Desta forma, no Capítulo 4 são sugeridos alguns métodos a aplicar de forma a ser realizada uma escolha entre a função de distribuição exponencial e a função de distribuição GP. Assim, nesse capítulo apresentamos métodos de escolha por testes de hipóteses, por intervalos de confiança e por métodos gráficos. Na Secção 4.1 são sugeridos seis testes de hipóteses que podem ser encontrados na literatura da área da EVT. A primeira estatística de teste sugerida foi abordada por Castillo *et al.* em [10] e por Coles em [12]. Esta tem por base a função de *profile log-likelihood*. A segunda estatística de teste foi abordada por Chaouche e Bacro em [11], a qual pode ser escrita em função do parâmetro de forma. Em [37] encontramos a terceira estatística de teste que se escreve em função do quadrado do coeficiente de variação. Outra das estatísticas abordadas na Secção 4.1 é

analisada por Gomes e van Montfort em [25]. Esta quarta estatística de teste foi proposta por van Montfort e Witter no trabalho realizado em [44]. A quinta estatística de teste sugerida pode ser encontrada em [23], a qual é dada pela razão entre a diferença do máximo e a mediana e a diferença desta e o mínimo. Baseada nesta estatística de teste, Brilhante em [8] apresentou a sexta estatística de teste sugerida nesta dissertação, definida pela razão entre a diferença do quarto superior e a mediana e a diferença desta e o quarto inferior. Tendo em conta as estimativas dos parâmetros de forma que podem ser obtidas pelos métodos de estimação de parâmetros anteriormente referidos, é de especial interesse construir os respetivos intervalos de confiança para o parâmetro de forma. A escolha entre a função distribuição exponencial e a função de distribuição GP mencionada na Secção 4.2 assenta na observação da inclusão ou não do valor zero nos intervalos de confiança obtidos, respetivamente. Na mesma secção, encontramos também um método informal para a validação de modelos através de métodos gráficos que podemos aplicar como método de escolha entre a função de distribuição exponencial e a função de distribuição GP.

Com o objetivo de colocar em prática os temas abordados, foi realizada uma ilustração prática no Capítulo 5. Dada a franca expansão da EVT nas últimas décadas, observamos que o *software* estatístico R, de entre alguns *softwares* estatísticos, tem acompanhado esse desenvolvimento. Este *software*, para além de ser de utilização gratuita, contém uma grande variedade de bibliotecas que podem ser encontradas na *task view* denominada *Extreme Value Analysis*.

Por fim, temos no Capítulo 6 desta dissertação as considerações finais resultantes do estudo realizado, assim como, alguns comentários efetuados aos resultados observados na ilustração prática realizada.

Capítulo 2

Sobre a Teoria dos Valores Extremos

A Teoria dos Valores Extremos tem como foco principal o estudo do comportamento de valores invulgarmente grandes ou pequenos. Diferentes formas de definir estes valores, denominados de extremos, originaram diferentes metodologias. Na metodologia de Gumbel é considerado um máximo por bloco, facto que pode limitar o estudo [12]. Na metodologia POT são integradas na análise todas as observações acima de um certo limiar, o que possibilita uma utilização mais eficiente dos dados [21]. Apesar da existência de outras metodologias, faremos nas secções seguintes apenas uma breve análise destas duas metodologias ditas mais clássicas.

2.1 Máximos por blocos

Seja $x = (x_1, \dots, x_n)$ uma realização do vetor aleatório (X_1, \dots, X_n) . Denotamos a ordenação ascendente dos elementos de x por $x_{1:n}, \dots, x_{n:n}$, onde $x_{1:n} \leq \dots \leq x_{n:n}$. O vetor que para a realização (x_1, \dots, x_n) de (X_1, \dots, X_n) tem como realização $x_{1:n}, \dots, x_{n:n}$ é denominado de vetor das estatísticas ordinais associado ao vetor aleatório (X_1, \dots, X_n) . Tomamos

$$X_{1:n} := \min_{1 \leq i \leq n} X_i$$

e

$$X_{n:n} := \max_{1 \leq i \leq n} X_i. \tag{2.1}$$

Cada componente univariada $X_{i:n}$ do vetor das estatísticas ordinais $(X_{1:n}, \dots, X_{n:n})$ é denominada i -ésima estatística ordinal ascendente do vetor aleatório (X_1, \dots, X_n) [45]. Consideremos ainda que as marginais do vetor (X_1, \dots, X_n) são variáveis aleatórias independentes e identicamente distribuídas a uma variável aleatória X ($X_i \stackrel{d}{=} X$) com função de distribuição F definida por

$$F(x) := P(X \leq x). \quad (2.2)$$

A função de distribuição empírica é dada por

$$\hat{F}_n(x) = \begin{cases} 0, & \text{se } x < x_{1:n}; \\ \frac{i}{n}, & \text{se } x \in [x_{1:n}, x_{i+1:n}[, i = 1 \dots n-1; \\ 1, & \text{se } x \geq x_{n:n}, \end{cases} \quad (2.3)$$

onde $x_{i:n}$ é o i -ésimo valor amostral [4].

Denotamos $X_{n:n}$ por M_n . A função de distribuição de M_n é dada por

$$\begin{aligned} F(x) = P[M_n \leq x] &= P\left[\max_{1 \leq i \leq n} X_i \leq x\right] \\ &= P[X_1 \leq x, X_2 \leq x, \dots, X_n \leq x] \\ &= \prod_{i=1}^n P[X_i \leq x] \\ &= (P[X \leq x])^n \\ &= F^n(x) \end{aligned} \quad (2.4)$$

No entanto, como a função de distribuição F é desconhecida, a igualdade (2.4) não tem grande utilidade na prática. Uma possível abordagem de resolução deste problema seria estimar a função de distribuição F a partir de dados observados, na qual a função de distribuição empírica, \hat{F}_n , teria um papel importante. No entanto, não é esta a abordagem mais aplicada porque pequenas discrepâncias na função estimativa para F podem originar grandes discrepâncias para F^n [12]. A função de distribuição F é assim considerada desconhecida e o estudo é focado no comportamento da função de distribuição F^n quando $n \rightarrow +\infty$. Assim, surge naturalmente a questão sobre quais distribuições podem surgir para a variável limite de M_n .

Antes de enunciar a resposta desta questão, relembremos as definições de convergência em distribuição e em probabilidade [45]:

Seja $\{X_n\}_{n \geq 1}$ uma sucessão de variáveis aleatórias.

- **Convergência em distribuição**

Dizemos que X_n converge em distribuição para a variável aleatória Y se e só se

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_Y(x),$$

para todo o x que seja ponto de continuidade de F_Y , e denotamos tal propriedade por $X_n \xrightarrow[n \rightarrow \infty]{d} Y$.

- **Convergência em probabilidade**

Dizemos que X_n converge em probabilidade para a variável aleatória Y (eventualmente degenerada) se $\forall \varepsilon > 0$,

$$\mathbb{P}[|X_n - Y| > \varepsilon] \xrightarrow[n \rightarrow \infty]{} 0$$

e escrevemos $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} Y$.

Ora, quando $n \rightarrow \infty$ obtemos:

$$\lim_{n \rightarrow \infty} F_{M_n}(x) = \begin{cases} 0, & \text{se } F(x) < 1; \\ 1, & \text{se } F(x) = 1, \end{cases} \quad (2.5)$$

pelo que $M_n \xrightarrow[n \rightarrow \infty]{d} x^F$, onde x^F é o limite superior de suporte de F ($x^F := \sup\{F(x) < 1\} \leq \infty$). Esta convergência em distribuição implica a convergência em probabilidade para a mesma constante, ou seja, temos $M_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} x^F$, mesmo que $x^F = \infty$ [24] e, assim observamos que a distribuição limite de M_n é degenerada. No entanto, a dificuldade de termos como distribuição limite uma distribuição degenerada é evitada se for possível encontrar sucessões de constantes reais $\{a_n\}_{n \geq 1}$ ($a_n > 0$) e $\{b_n\}_{n \geq 1}$ tais que

$$\frac{M_n - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{d} Y \quad (2.6)$$

com Y variável aleatória não degenerada, ou seja, tais que

$$F^n(a_n x + b_n) \xrightarrow[n \rightarrow \infty]{} G(x), \quad \forall x \in C(G),$$

com G função de distribuição de uma variável aleatória não degenerada e $C(G)$ o conjunto dos pontos de continuidade de G . Nestas condições, dizemos que F pertence ao max-domínio de atração da lei G ($F \in D_M(G)$) e denominamos as

sucessões de constantes $\{a_n\}_{n \geq 1}$ ($a > 0$) e $\{b_n\}_{n \geq 1}$ de coeficientes de atração de F para G . A resposta à questão realizada foi obtida por Gnedenko no trabalho realizado em [22], na sequência dos estudos de Fisher e Tippett em [18] e de von Mises em [47]. O enunciado do teorema de tipos extremais (cf. [16]), também denominado de teorema de Gnedenko ou teorema de Fisher-Tippett-Gnedenko, é apresentado em seguida:

Teorema 2.1 (Teorema de tipos extremais) *Seja $\{X_n\}_{n \geq 1}$ uma sucessão de variáveis aleatórias. Se existem sucessões de constantes $\{a_n\}_{n \geq 1}$ ($a_n \geq 0$), $\{b_n\}_{n \geq 1}$ e uma função de distribuição não degenerada G tais que*

$$\lim_{n \rightarrow \infty} P \left[\frac{M_n - b_n}{a_n} \leq x \right] = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = G(x), \quad \forall x \in C(G), \quad (2.7)$$

então, G pertence a um dos três tipos de distribuição de valores extremos:

$$\text{Fréchet :} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & x > 0 \end{cases} \quad \alpha > 0 \quad (2.8)$$

$$\text{Weibull :} \quad \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha), & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \alpha > 0 \quad (2.9)$$

$$\text{Gumbel :} \quad \Lambda(x) = \exp(-\exp(-x)), \quad x \in \mathbb{R}. \quad (2.10)$$

As funções de distribuição definidas em (2.8), (2.9) e (2.10) têm, respetivamente, como função de densidade de probabilidade as seguintes funções:

$$\psi(x, \alpha) = \alpha \cdot x^{-\alpha-1} \cdot \exp(-x^{-\alpha}), \quad x > 0 \quad (2.11)$$

$$\varphi(x, \alpha) = -(-1)^\alpha \cdot x^{\alpha-1} \cdot \alpha \cdot \exp[-(-1)^\alpha x^\alpha] \quad x < 0 \quad (2.12)$$

$$\lambda(x) = \exp(-x - \exp(-x)) \quad (2.13)$$

Na Figura 2.1 temos a representação da função densidade de probabilidade em (2.13) como também das funções densidade de probabilidade ψ e φ para $\alpha = 1$.

As distribuições de valores extremos podem ser representadas por meio de uma única família de funções denominada de distribuição generalizada de valores extremos (GEV, do inglês *generalized extreme value*), definida do seguinte modo

$$G_\gamma(x) = \begin{cases} \exp(-(1 + \gamma x)^{-1/\gamma}), & \text{para } 1 + \gamma x > 0, \text{ se } \gamma \neq 0, \\ \exp(-\exp(-x)), & \text{para } x \in \mathbb{R} \text{ se } \gamma = 0. \end{cases} \quad (2.14)$$

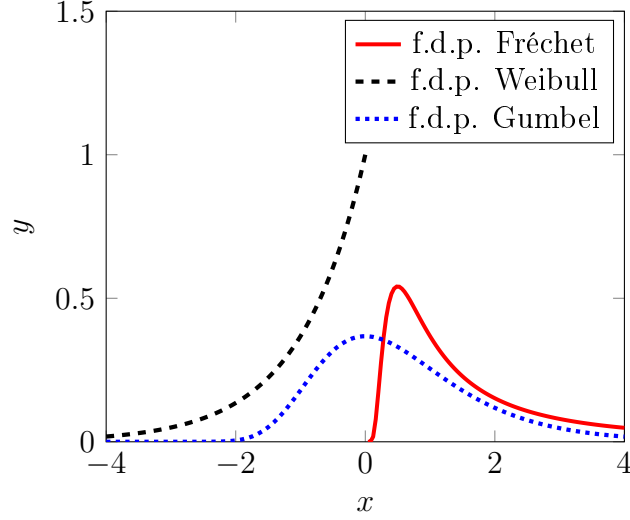


Figura 2.1: A função densidade de probabilidade λ e as funções densidade de probabilidade ψ e φ para $\alpha = 1$.

A função densidade de probabilidade correspondente a (2.14) é dada por (cf. [29])

$$g_\gamma(z) = \begin{cases} \exp \left[- (1 + \gamma x)^{-\frac{1}{\gamma}} \right] \cdot [1 + \gamma x]^{-\frac{1}{\gamma}-1} & , \text{ para } 1 + \gamma x > 0, \\ & \text{se } \gamma \neq 0, \\ \exp(-\exp(-x)) \cdot \exp(-x) & , \text{ para } x \in \mathbb{R} \text{ se } \gamma = 0. \end{cases} \quad (2.15)$$

O parâmetro de forma γ é denominado de índice de valores extremos (EVI). A unificação das três famílias, Fréchet, Weibull e Gumbel, pela distribuição GEV é atribuída a von Mises [47] e Jenkinson [28], sendo por tal denominada de parametrização de von Mises-Jenkinson. Esta parametrização pode ser escrita em função das funções de distribuição em (2.8), (2.9) e (2.10) do seguinte modo (cf. [39])

$$G_\gamma(x) = \begin{cases} \Phi_{\frac{1}{\gamma}}(1 + \gamma x) & \text{se } \gamma > 0, \\ \Psi_{-\frac{1}{\gamma}}(-(1 + \gamma x)) & \text{se } \gamma < 0, \\ \Lambda(x) & \text{se } \gamma = 0. \end{cases}$$

Notemos que $\lim_{\gamma \rightarrow 0} G_\gamma(x) = G_0(x)$ onde G_0 é a função de distribuição Gumbel. A introdução de parâmetros de escala, $\sigma > 0$, e de localização, $\mu \in \mathbb{R}$,

resulta na família completa de funções de distribuição GEV definida do seguinte modo

$$G_{\gamma,\sigma,\mu}(z) = \exp \left\{ - \left[1 + \gamma \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\gamma} \right\} \quad (2.16)$$

definida em $\{z : 1 + \gamma(z - \mu)/\sigma > 0\}$, onde $-\infty < \mu < +\infty$, $\sigma > 0$ e $-\infty < \gamma < +\infty$ [12].

Isto permite reescrever o Teorema 2.1 (cf., e.g. [24]) de outra forma:

Teorema 2.2 (Teorema Unificado dos Tipos Extremais) *Se existem as sucessões $\{a_n\}_{n \geq 1}$ ($a_n > 0$) e $\{b_n\}_{n \geq 1}$, tais que, quando $n \rightarrow \infty$*

$$P \left[\frac{M_n - b_n}{a_n} \leq z \right] \longrightarrow G(z)$$

para alguma função de distribuição G não-degenerada, então G é do mesmo tipo da distribuição GEV, definida em (2.16), para algum $\gamma \in \mathbb{R}$.

2.2 Excessos acima de um nível elevado

Seja X uma variável aleatória com função de distribuição F e x^F o limite superior de F . Para $u < x^F$ fixo, a função de distribuição definida por

$$F_u(y) = P[X - u \leq y | X > u], \quad 0 \leq y \leq x^F - u \quad (2.17)$$

é denominada de função de distribuição condicional dos excessos acima de u [cf., [21, 24]]. O valor de u é denominado de nível ou limiar elevado e os valores $y = x - u$ são designados de excessos.

Observemos que

$$\begin{aligned} F_u(y) &= P[X - u \leq y | X > u] \\ &= \frac{P[X - u \leq y, X > u]}{P[X > u]} \\ &= \frac{P[u < X \leq u + y]}{P[X > u]} \\ &= \frac{F(u + y) - F(u)}{1 - F(u)} \end{aligned} \quad (2.18)$$

Assim, as funções de distribuição F_u e F , representadas na Figura 2.2, estão relacionadas.

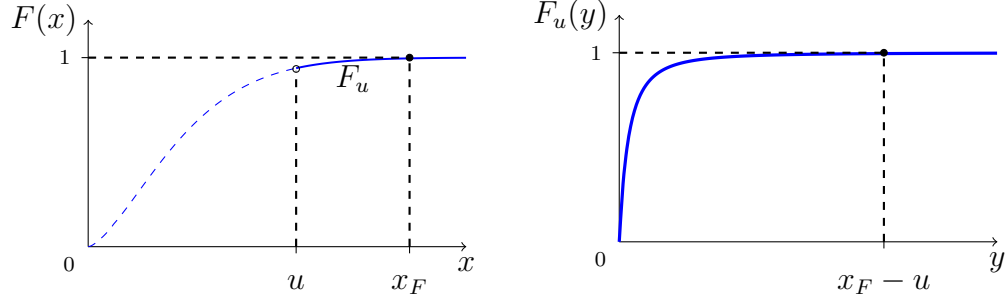


Figura 2.2: Função de distribuição F e função de distribuição condicional F_u .

Segundo Pickands em [36] e Balkema e de Haan em [3], a função de distribuição F_u pode ser bem aproximada por uma função de distribuição GP, i.e. $\exists y \in \mathbb{R}$, $\exists \sigma_u > 0$:

$$F_u(y) \approx W_{\gamma, \sigma_u}(y) := \begin{cases} 1 - \left(1 + \frac{\gamma y}{\sigma_u}\right)^{-1/\gamma}, & \text{para } y \geq 0, \text{ se } \gamma > 0, \\ 1 - \exp\left(-\frac{y}{\sigma_u}\right), & \text{para } y \geq 0, \text{ se } \gamma = 0, \\ 1 - \left(1 + \frac{\gamma y}{\sigma_u}\right)^{-1/\gamma}, & \text{para } 0 \leq y \leq -\frac{\sigma_u}{\gamma}, \text{ se } \gamma < 0. \end{cases} \quad (2.19)$$

Isto é, podemos considerar

$$F_u(y) \approx W_{\gamma, \sigma_u}(y)$$

para $y \in [0, x^F - u]$, se $\gamma \geq 0$, e para $y \in [0, -\frac{\sigma_u}{\gamma}]$ se $\gamma < 0$. O teorema demonstrado por Pickands, Balkema e de Haan é o seguinte:

Teorema 2.3 (Teorema Pickands-Balkema-de Haan)

$$F \in \mathcal{D}(G_\gamma), \gamma \in \mathbb{R} \iff \lim_{u \rightarrow x^F} \sup_{0 < y < x^F - u} |F_u(y) - W_{\gamma, \sigma_u}(y)| = 0.$$

É de salientar que existe uma forte dualidade entre a distribuição GP e a distribuição GEV. Analisemos com mais pormenor essa dualidade [12, 30]. Seja novamente X uma variável aleatória com função de distribuição F . Nas condições do Teorema 2.2, para n suficientemente grande, temos

$$F^n(z) \approx \exp \left\{ - \left[1 + \gamma \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\gamma} \right\}$$

para algum μ , $\gamma \in \mathbb{R}$ e $\sigma > 0$. Consequentemente

$$n \log F(z) \approx - \left[1 + \gamma \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\gamma}. \quad (2.20)$$

Para valores grandes de z , a expansão em série de Taylor implica que

$$\log F(z) \approx -\{1 - F(z)\}.$$

Substituindo em (2.20), obtemos

$$n(-\{1 - F(z)\}) \approx -\left[1 + \gamma\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\gamma},$$

da qual decorre

$$1 - F(u) \approx \frac{1}{n} \left[1 + \gamma\left(\frac{u - \mu}{\sigma}\right)\right]^{-1/\gamma},$$

se tomarmos para z um valor u elevado. Analogamente, tomando $y > 0$,

$$P(X > u + y) = 1 - F(u + y) \approx \frac{1}{n} \left[1 + \gamma\left(\frac{u + y - \mu}{\sigma}\right)\right]^{-1/\gamma}. \quad (2.21)$$

Consequentemente,

$$\begin{aligned} Pr(X > u + y | X > u) &\approx \frac{n^{-1}[1 + \gamma(u + y - \mu)/\sigma]^{-1/\gamma}}{n^{-1}[1 + \gamma(u - \mu)/\sigma]^{-1/\gamma}} \\ &= \left[1 + \frac{\gamma y / \sigma}{1 + \gamma(u - \mu)/\sigma}\right]^{-1/\gamma} \\ &= \left[1 + \frac{\gamma y}{\sigma_u}\right]^{-1/\gamma}, \end{aligned}$$

onde $\sigma_u = \sigma + \gamma(u - \mu)$.

Em resumo, temos o seguinte teorema (cf. [12]).

Teorema 2.4 *Seja X_i um termo arbitrário da sucessão de variáveis aleatórias independentes e identicamente distribuídas $\{X_n\}_{n \geq 1}$ com função de distribuição comum F . Suponhamos que F satisfaz as condições do Teorema 2.2, para um n grande,*

$$Pr\{M_n \leq z\} \approx G(z),$$

onde

$$G(z) = \exp \left\{ - \left[1 + \gamma\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\gamma} \right\}$$

para algum μ , $\gamma \in \mathbb{R}$ e $\sigma > 0$. Então, para um valor u suficientemente grande, a função distribuição de $(X - u)$ condicional a $X > u$, é aproximadamente

$$H(y) = 1 - \left(1 + \frac{\gamma y}{\sigma_u}\right)^{-1/\gamma} \quad (2.22)$$

com $\{y : y > 0 \text{ e } (1 + \gamma y/\sigma_u) > 0\}$

$$\sigma_u = \sigma + \gamma(u - \mu). \quad (2.23)$$

Observamos assim que os parâmetros da função de distribuição W_{γ,σ_u} podem ser obtidos exclusivamente dos parâmetros da função de distribuição $G_{\gamma,\sigma,\mu}$ considerada. É de salientar que as funções de distribuição $G_{\gamma,\sigma,\mu}$ e W_{γ,σ_u} partilham o mesmo parâmetro de forma γ , e que o parâmetro de escala de W_{γ,σ_u} , σ_u , é escrito à custa dos três parâmetros de $G_{\gamma,\sigma,\mu}$ e do limiar elevado u .

Observemos que, sendo $y = x - u$, a função de distribuição GP em (2.19) também pode ser expressa em termos de três parâmetros do seguinte modo

$$W_{\gamma,\sigma_u,u}(x|u) := \begin{cases} 1 - \left(1 + \frac{\gamma(x-u)}{\sigma_u}\right)^{-1/\gamma}, & \text{para } x \geq u, \text{ se } \gamma > 0, \\ 1 - \exp\left(-\frac{x-u}{\sigma_u}\right), & \text{para } x \geq u, \text{ se } \gamma = 0, \\ 1 - \left(1 + \frac{\gamma(x-u)}{\sigma_u}\right)^{-1/\gamma}, & \text{para } u \leq x \leq 1 - \frac{\sigma_u}{\gamma}, \text{ se } \gamma < 0, \end{cases} \quad (2.24)$$

com parâmetro de forma γ , parâmetro de escala σ_u e parâmetro de localização u .

A função de distribuição GP, H_γ , pode também ser escrita em função da parametrização de von Mises-Jenkinson do seguinte modo

$$H_\gamma(x) = 1 + \ln G_\gamma(x) = \begin{cases} 1 - (1 + \gamma x)^{-1/\gamma}, & 1 + \gamma x > 0, x > 0 \text{ se } \gamma \neq 0, \\ 1 - \exp(-x), & x > 0 \text{ se } \gamma = 0. \end{cases} \quad (2.25)$$

Além disso, pode ser escrita à custa das seguintes funções de distribuição (e.g., cf., [24, 39])

$$\mathbf{Exponencial} : F_1(x) = 1 - e^{-x}, \quad x > 0$$

$$\mathbf{Pareto} : F_{2,\alpha}(x) = 1 - x^{-\alpha}, \quad x > 1, \alpha > 0$$

$$\mathbf{Beta simétrica} : F_{3,\alpha}(x) = 1 - (-x)^{-\alpha}, \quad -1 < x < 0, \alpha > 0$$

da seguinte forma:

$$H_\gamma(x) = \begin{cases} F_{2,\frac{1}{\gamma}}(1 + \gamma x) & \text{se } \gamma > 0, \\ F_{3,-\frac{1}{\gamma}}(-(1 + \gamma x)) & \text{se } \gamma < 0, \\ F_1(x) & \text{se } \gamma = 0. \end{cases}$$

Notemos que $\lim_{\gamma \rightarrow 0} H_\gamma(x) = H_0(x)$, onde H_0 representa a função de distribuição exponencial, e que podemos também definir sucintamente a função de distribuição GP do seguinte modo

$$H_{\gamma,\beta}(x) = 1 - \left(1 + \gamma \frac{x}{\beta}\right)^{-1/\gamma}, \text{ para } x \in D(\gamma, \beta) \quad (2.26)$$

com $D(\gamma, \beta) = [0, +\infty[$ se $\gamma \geq 0$ e $D(\gamma, \beta) = [0, -\beta/\gamma]$ se $\gamma < 0$ [16]. A função densidade de probabilidade correspondente é dada por

$$h_{\gamma,\beta}(x) = \frac{1}{\beta} \left(1 + \gamma \frac{x}{\beta}\right)^{-\frac{1}{\gamma}-1}, \text{ } x \in D(\gamma, \beta). \quad (2.27)$$

Na Figura 2.3 temos as representações gráficas das funções densidade de probabilidade $h_{\gamma,\beta}$ para $\beta = 1$ e para γ igual a 0.5, 0 e -0.25 .

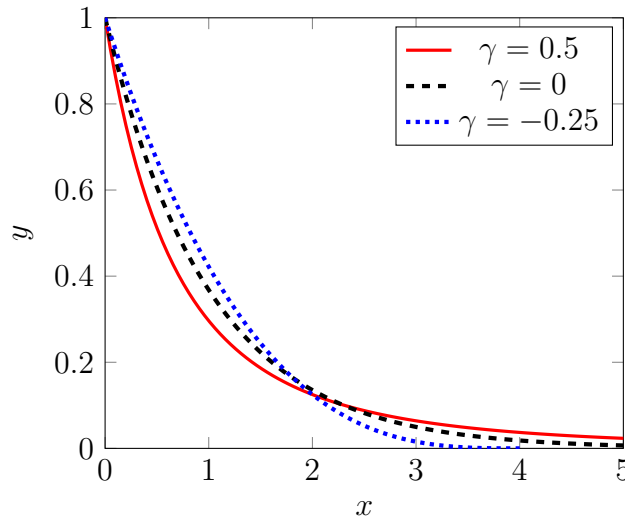


Figura 2.3: Função densidade de probabilidade $h_{\gamma,1}$ para $\gamma = 0.5$, 0 e -0.25 .

De entre as propriedades da função de distribuição GP enunciadas por Embrechets *et al.* em [16], salientamos a propriedade verificada em seguida.

Seja $x_i \in D(\gamma, \beta)$, $i \in \{1, 2\}$. Considerando $\bar{H}_{\gamma,\beta} = 1 - H_{\gamma,\beta}$ com a função $H_{\gamma,\beta}$ definida em (2.26), temos que

$$\begin{aligned}
\frac{\bar{H}_{\gamma,\beta}(x_1 + x_2)}{\bar{H}_{\gamma,\beta}(x_1)} &= \frac{1 - 1 + \left(1 + \gamma \frac{x_1 + x_2}{\beta}\right)^{-1/\gamma}}{1 - 1 + \left(1 + \gamma \frac{x_1}{\beta}\right)^{-1/\gamma}} \\
&= \left(\frac{\beta + \gamma(x_1 + x_2)}{\beta + \gamma x_1}\right)^{-1/\gamma} \\
&= \left(\frac{\beta + \gamma x_1 + \gamma x_2}{\beta + \gamma x_1}\right)^{-1/\gamma} \\
&= \left(1 + \frac{\gamma x_2}{\beta + \gamma x_1}\right)^{-1/\gamma} \\
&= 1 - \left[1 - \left(1 + \frac{\gamma x_2}{\beta + \gamma x_1}\right)^{-1/\gamma}\right] \\
&= 1 - H_{\gamma,\beta+\gamma x_1}(x_2)
\end{aligned}$$

De onde vem que

$$\frac{\bar{H}_{\gamma,\beta}(x_1 + x_2)}{\bar{H}_{\gamma,\beta}(x_1)} = 1 - H_{\gamma,\beta+\gamma x_1}(x_2) = \bar{H}_{\gamma,\beta+\gamma x_1}(x_2). \quad (2.28)$$

A igualdade (2.28) permite constatar que a probabilidade da variável X exceder o limiar $x_1 + x_2$ dado que excede x_1 pode ser obtida por uma função de distribuição GP. Se tomarmos a função de distribuição GP W_{γ,σ_u} e $x_i \in D(\gamma, \beta)$, $i \in \{1, 2\}$ com $\beta = \sigma_u$ temos

$$\begin{aligned}
\frac{\bar{W}_{\gamma,\sigma_u}(x_1 + x_2)}{\bar{W}_{\gamma,\sigma_u}(x_1)} &\approx \frac{1 - F_u(x_1 + x_2)}{1 - F_u(x_1)} \\
&= \frac{1 - P[X - u \leq x_1 + x_2 | X > u]}{1 - P[X - u \leq x_1 | X > u]} \\
&= \frac{P[X - u > x_1 + x_2 | X > u]}{P[X - u > x_1 | X > u]} \\
&= \frac{P[X > x_1 + x_2 + u | X > u]}{P[X > x_1 + u | X > u]} \\
&= \frac{P[X > x_1 + x_2]}{P[X > x_1]},
\end{aligned}$$

e se considerarmos $x_1 \leq x_2$, então obtemos

$$\frac{\bar{W}_{\gamma,\sigma_u}(x_1 + x_2)}{\bar{W}_{\gamma,\sigma_u}(x_1)} = P[X > x_1 + x_2 | X > x_1]. \quad (2.29)$$

De (2.28) e (2.29), vem que

$$P[X > x_1 + x_2 | X > x_1] = \bar{W}_{\gamma, \sigma_u + \gamma x_1}(x_2).$$

Dado isto, enunciarmos a propriedade (2.28) é o mesmo que indicar que a classe das funções de distribuição GP é fechada relativamente a alterações no limiar [16, 34].

Capítulo 3

Metodologia POT

Com a metodologia POT é possível estudar as observações que excedem um determinado limiar, ajustando o modelo paramétrico aos excessos decorrentes desse limiar, desde que seja assegurado um número suficiente de dados acima do limiar selecionado. Se tivermos então um número suficiente de dados temos de procurar a função de distribuição conveniente para esses excessos [46].

Consideremos x_1, \dots, x_n as concretizações das variáveis aleatórias independentes e identicamente distribuídas X_1, \dots, X_n . A identificação de acontecimentos extremos na metodologia POT é consequência da definição de um limiar elevado u . As observações que excedem este limiar são denominadas de excedências $\{x_i : x_i > u\}$, as quais serão aqui denotadas por $x_{(1)}, \dots, x_{(k)}$. Pelo Teorema 2.4, temos que as diferenças entre as excedências e o limiar u , ou seja, os excessos dados por

$$y_i = x_{(i)} - u, \text{ para } i = 1, \dots, k$$

são concretizações independentes de uma variável aleatória Y cuja distribuição pode ser aproximada por uma função de distribuição GP [12].

Assim, é de grande importância nesta metodologia a escolha do limiar u . A solução de um valor para u é um problema controverso [24], que implica um compromisso entre um menor viés ou uma menor variância para o estimador do parâmetro de forma γ consoante a escolha recaia num valor elevado ou não tão elevado para u [4].

Segundo Coles [12], é prática comum na seleção do limiar u considerar para este o menor valor possível que proporcione uma aproximação razoável nos termos do Teorema 2.4. Davison e Smith em [15] propuseram para a escolha

de um valor conveniente para o limiar u o estudo da função de excesso médio, a qual é dada por

$$e(t) = E[X - t | X > t] \quad (3.1)$$

para X variável aleatória tal que $E[X] < \infty$. Na prática, é utilizada a função de excesso médio empírica definida do seguinte modo

$$\hat{e}_n(t) = \frac{\sum_{i=1}^n x_i I_{(t,+\infty)}(x_i)}{\sum_{i=1}^n I_{(t,+\infty)}(x_i)} - t, \text{ com } I_{(t,+\infty)}(x_i) = \begin{cases} 1 & \text{se } x_i > t \\ 0 & \text{se } x_i \leq t \end{cases}$$

para uma dada amostra x_1, \dots, x_n [4]. Se X for uma variável aleatória com função de distribuição GP (2.26) com parâmetros $\gamma < 1$ e $\beta \in \mathbb{R}$, então temos

$$e(u) = E[X - u | X > u] = \frac{\beta + \gamma u}{1 - \gamma} = \frac{\gamma}{1 - \gamma} u + \frac{\beta}{1 - \gamma}, \quad (3.2)$$

para $u < x^F$ e $\beta + u\gamma > 0$ [16]. Notemos que, neste caso, a função e é uma função linear relativamente a u . Observemos também que

$$e(u) = E[X - u | X > u] = E[Y | Y > 0] = \frac{\sum_{j=1}^{N_u} Y_j}{N_u} \quad (3.3)$$

onde N_u denota o número de excedências e Y_j os excessos dados por

$$Y_j := X_i - u | X_i > u, \quad j = 1, 2, \dots, N_u. \quad (3.4)$$

Temos, assim a média dos excessos acima do limiar u , para a qual podemos determinar a seguinte estimativa pontual

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n_u} y_j}{n_u}.$$

A localização dos pontos

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\},$$

onde $x_{(1)}, \dots, x_{(n_u)}$ são as n_u observações que excedem u e x_{\max} é a maior observação de X_i , forma o gráfico da função de excesso médio empírica.

Assim, este gráfico constitui uma ferramenta exploratória no qual a escolha de um valor para o limiar u resulta da observação de linearidade à direita desse valor. Mas, o gráfico da função de excesso médio empírica pode ser de difícil

interpretação, apresentando além disso grande sensibilidade a alterações nos dados para valores grandes de u devida a uma maior escassez de dados [16]. De modo a ultrapassar esta segunda desvantagem têm sido sugeridas alternativas mais robustas tais como a baseada na função de excesso mediana [cf. Beirlant, Teugels e Vynckier [5], Mendes [34] e Rootzén e Tajvidi [38]] dada por

$$e^*(u) = \text{Mediana}\{Y_i(u), i = 1, \dots, N_u\}.$$

A dificuldade de interpretação do gráfico da função de excesso médio como método de escolha de limiar u pode ser compensada com uma avaliação da estabilidade das estimativas dos parâmetros da função de distribuição GP para uma variedade de diferentes valores do limiar.

Consideremos que a função de distribuição GP constitui uma aproximação válida para a função de distribuição dos excessos acima de um limiar u_0 resultantes de variáveis independentes X_i , tais que $X_i \stackrel{d}{=} X$. Como observamos na Secção 2.2, se uma função de distribuição GP constitui uma aproximação válida para a função de distribuição dos excessos acima de u_0 , então uma função de distribuição GP também o é para a função de distribuição dos excessos acima de um limiar $u > u_0$. Estas duas funções de distribuição GP terão o mesmo parâmetro de forma γ e os respetivos parâmetros de escala, σ_{u_0} e σ_u , satisfazem a seguinte igualdade

$$\sigma_u = \sigma_{u_0} + \gamma(u - u_0). \quad (3.5)$$

De facto, como $\sigma_u = \sigma + \gamma(u - \mu)$ temos

$$\sigma_{u_0} = \sigma + \gamma(u_0 - \mu) \quad (3.6)$$

e

$$\sigma_u = \sigma + \gamma(u - \mu). \quad (3.7)$$

Substituindo em (3.7) σ por $\sigma = \sigma_{u_0} - \gamma(u_0 - \mu)$ (equivalente a (3.6)) obtemos

$$\begin{aligned} \sigma_u &= \sigma_{u_0} - \gamma(u_0 - \mu) + \gamma(u - \mu) \\ &= \sigma_{u_0} + \gamma(u - u_0). \end{aligned}$$

Notemos que, ao contrário dos parâmetros de forma das duas funções de distribuição GP, os respetivos parâmetros de escala só são iguais se $\gamma = 0$. No entanto, se tomarmos

$$\sigma^* = \sigma_u - \gamma u$$

observamos que σ^* é constante relativamente a u , ou seja,

$$\sigma^* = \sigma_u - \gamma u = \sigma_{u_0} + \gamma(u - u_0) - \gamma u = \sigma_{u_0} - \gamma u_0.$$

Assim, as estimativas para σ^* e γ deverão ser constantes (ou quase constantes devido à variabilidade inerente à amostragem) para valores u acima do limiar u_0 . Os gráficos com as estimativas $\hat{\sigma}^*$ e $\hat{\gamma}$ num dos eixos e os correspondentes valores de u no outro eixo constituem outra ferramenta exploratória para a seleção de um limiar. Para tal, basta selecionar o limiar u_0 como o menor valor para o qual as estimativas correspondentes a valores à direita desse valor permanecem quase constantes.

3.1 Estimação de parâmetros

Vários métodos para estimar os parâmetros da função de distribuição GP têm sido propostos na literatura da área. Como ponto de partida, referimos uma revisão detalhada em dois artigos [6, 7] de Zea Bermudez e Kotz, focalizada naqueles métodos que revelam uma aplicabilidade relativamente simples a dados reais. Nesta secção, analisaremos apenas os três métodos habitualmente reconhecidos como os mais clássicos.

Segundo Beirlant *et al.* em [4], o método MOM e o método PWM foram sugeridos por Hosking e Wallis em [27]. Estes dois métodos partilham a ideia de que os estimadores para os parâmetros desconhecidos podem ser deduzidos das expressões dos momentos populacionais.

Seja $y = (y_1, \dots, y_{N_u})$ uma amostra de dimensão N_u , realização da amostra aleatória $Y = (Y_1, \dots, Y_{N_u})$ com Y_1, \dots, Y_{N_u} variáveis independentes e identicamente distribuídas à variável Y com função de distribuição GP.

Em [39], foi demonstrado que o momento de ordem r é dado por

$$E[Y^r] = \frac{r! \beta^r}{\prod_{i=1}^r (1 - i\gamma)}, \quad (3.8)$$

para $r = 1, 2, \dots$ e $\gamma < \frac{1}{r}$.

Observamos assim por (3.8) que, quando $\gamma < \frac{1}{2}$, o valor médio e a variância da função de distribuição GP são dados, respetivamente, por

$$E[Y] = \frac{\beta}{1 - \gamma} \quad (3.9)$$

e

$$\text{var}[Y] = E[Y^2] - E[Y]^2 = \frac{2\beta^2}{(1-\gamma)(1-2\gamma)} - \frac{\beta^2}{(1-\gamma)^2} = \frac{\beta^2}{(1-\gamma)^2(1-2\gamma)}. \quad (3.10)$$

O método MOM consiste em exprimir os parâmetros que pretendemos estimar em termos dos momentos do modelo e depois equacionar momentos populacionais com momentos empíricos. Os estimadores pelo método dos momentos de β e γ , $\hat{\beta}_{\text{MOM}}$ e $\hat{\gamma}_{\text{MOM}}$, resultam da resolução de (3.9) e (3.10) em função de γ e β seguida da substituição de $E[Y]$ por $\bar{Y} = \sum_{i=1}^{N_u} Y_i/N_u$ e de $\text{var}(Y)$ por $S_{\bar{Y}, N_u-1}^2 = \sum_{i=1}^{N_u} (Y_i - \bar{Y})^2 / (N_u - 1)$.

Resolvendo (3.9) em ordem a β temos que

$$\beta = E[Y](1 - \gamma). \quad (3.11)$$

Assim, substituindo o valor de β , calculado em (3.11), na equação (3.10) temos que, quando $\gamma < \frac{1}{2}$

$$\begin{aligned} \text{var}[Y] &= \frac{E[Y]^2(1-\gamma)^2}{(1-\gamma)^2(1-2\gamma)} \\ \Leftrightarrow \text{var}[Y] &= \frac{E[Y]^2}{1-2\gamma} \\ \Leftrightarrow 1-2\gamma &= \frac{E[Y]^2}{\text{var}[Y]} \\ \Leftrightarrow -2\gamma &= \frac{E[Y]^2}{\text{var}[Y]} - 1 \\ \Leftrightarrow \gamma &= \frac{1}{2} \left(1 - \frac{E[Y]^2}{\text{var}[Y]} \right). \end{aligned} \quad (3.12)$$

Por (3.11) e (3.12) obtemos

$$\begin{aligned} \beta &= E[Y] \left(1 - \frac{1}{2} + \frac{E[Y]^2}{2\text{var}[Y]} \right) \\ &= \frac{E[Y]}{2} \left(1 + \frac{E[Y]^2}{\text{var}[Y]} \right). \end{aligned} \quad (3.13)$$

Por fim, substituindo $E[Y]$ por \bar{Y} e $\text{var}[Y]$ por $S_{\bar{Y}, N_u-1}^2$ em (3.12) e (3.13) obtemos os estimadores

$$\hat{\gamma}_{\text{MOM}} = \frac{1}{2} \left(1 - \frac{\bar{Y}^2}{S_{\bar{Y}, N_u-1}^2} \right) \quad (3.14)$$

e

$$\hat{\beta}_{\text{MOM}} = \frac{\bar{Y}}{2} \left(1 + \frac{\bar{Y}^2}{S_{Y, N_u-1}^2} \right) \quad (3.15)$$

para os parâmetros γ e β , respetivamente. Se $\gamma = 0$ temos $E[Y] = \beta$. Assim, neste caso, o estimador de β pelo método MOM, $\hat{\beta}_{\text{MOM}}$, é dado pela média de Y ($\hat{\beta}_{\text{MOM}} = \bar{Y}$).

Os momentos ponderados de probabilidade (PWM) foram introduzidos por Greenwood *et al.* em [26]. Seguindo a ideia geral usada no método MOM, a abordagem pelo método PWM aparece como uma alternativa para estimar os parâmetros de uma função de distribuição.

Os momentos ponderados de probabilidade de uma variável aleatória Y com função distribuição F são dados por

$$M_{p,r,s} = E\{Y^p [F(Y)]^r [1 - F(Y)]^s\}, \quad (3.16)$$

com $p, r, s \in \mathbb{R}$ [4]. Na prática, o procedimento usual é fixar $p = 1$ e escolher para r e s os menores valores inteiros não negativos possíveis. Segundo Beirlant *et al.* em [4], é mais adequado no caso da função de distribuição GP utilizar

$$M_{1,0,s} = E[Y(1 - F(Y))^s] = \frac{\beta}{(s+1)(s+1-\gamma)} \quad \gamma < 1, \quad (3.17)$$

ou seja, $M_{p,r,s}$ com $p = 1$, $r = 0$ e $s = 0, 1, 2, \dots$

De (3.17) obtemos para $s = 0$ a seguinte expressão

$$M_{1,0,0} = E[Y] = \frac{\beta}{1-\gamma}, \quad (3.18)$$

enquanto que para $s = 1$ obtemos

$$M_{1,0,1} = E[Y(1 - F(Y))] = \frac{\beta}{4 - 2\gamma}. \quad (3.19)$$

Denotemos $M_{1,0,0}$ e $M_{1,0,1}$ por w_0 e w_1 , respetivamente. Assim, resolvendo (3.18) em ordem a β resulta

$$\beta = w_0(1 - \gamma). \quad (3.20)$$

Substituindo (3.19) o valor de β , calculado em (3.20), temos que

$$\begin{aligned}
w_1 &= \frac{w_0(1-\gamma)}{4-2\gamma} \\
\Leftrightarrow 4w_1 - 2w_1\gamma &= w_0 - w_0\gamma \\
\Leftrightarrow w_0\gamma - 2w_1\gamma &= w_0 - 4w_1 \\
\Leftrightarrow \gamma(w_0 - 2w_1) &= w_0 - 4w_1 \\
\Leftrightarrow \gamma &= \frac{w_0 - 4w_1}{w_0 - 2w_1} \\
\Leftrightarrow \gamma &= \frac{2w_0 - 4w_1 - w_0}{w_0 - 2w_1} \\
\Leftrightarrow \gamma &= 2\frac{w_0 - 2w_1}{w_0 - 2w_1} - \frac{w_0}{w_0 - 2w_1} \\
\Leftrightarrow \gamma &= 2 - \frac{w_0}{w_0 - 2w_1}. \tag{3.21}
\end{aligned}$$

Por fim, substituindo (3.21) em (3.20) temos que

$$\begin{aligned}
\beta &= w_0 \left(1 - 2 + \frac{w_0}{w_0 - 2w_1} \right) \\
&= -w_0 + \frac{w_0^2}{w_0 - 2w_1} \\
&= \frac{-w_0^2 + 2w_0w_1 + w_0^2}{w_0 - 2w_1} \\
&= \frac{2w_0w_1}{w_0 - 2w_1}. \tag{3.22}
\end{aligned}$$

Assim, de (3.21) e (3.22) obtemos os seguintes estimadores

$$\hat{\gamma}_{\text{PWM}} = 2 - \frac{\hat{M}_{1,0,0}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}} \tag{3.23}$$

e

$$\hat{\beta}_{\text{PWM}} = \frac{2\hat{M}_{1,0,0}\hat{M}_{1,0,1}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}}, \tag{3.24}$$

onde (cf. [4])

$$\hat{M}_{1,0,s} = \frac{1}{N_u} \sum_{j=1}^{N_u} \left(1 - \frac{j}{N_u + 1} \right)^s Y_{j:N_u}.$$

Consideremos agora $\gamma = 0$. De (3.19) temos $M_{1,0,1} = \frac{1}{4}\beta$ que resulta em

$$\beta = 4w_1. \tag{3.25}$$

Assim, de (3.25) obtemos

$$\hat{\beta}_{\text{PWM}} = 4\hat{M}_{1,0,1},$$

com $M_{1,0,1}$ substituído por

$$\hat{M}_{1,0,1} = \frac{1}{N_u} \sum_{j=1}^{N_u} \left(1 - \frac{j}{N_u + 1}\right) Y_{j:N_u}.$$

Hosking e Wallis em [27] efetuaram também uma comparação entre os estimadores obtidos pelos métodos MOM e PWM com os estimadores obtidos pelo método ML que apresentamos seguidamente. Os estimadores pelo método ML foram deduzidos por Davison em [14] e Smith em [40].

Consideremos, novamente, Y_1, \dots, Y_{N_u} variáveis independentes e identicamente distribuídas à variável Y com função de distribuição GP.

A função de verosimilhança é dada por

$$L_{\gamma,\beta}(y) = \prod_{i=1}^{N_u} h_{\gamma,\beta}(y_i) = \prod_{i=1}^{N_u} \left[\frac{1}{\beta} \left(1 + \gamma \frac{y_i}{\beta}\right)^{-\frac{1}{\gamma}-1} \right]$$

com $y = (y_1, y_2, \dots, y_{N_u})$, $\gamma \neq 0$ e $1 + \gamma \frac{y_i}{\beta} > 0$, e por

$$L_{0,\beta}(y) = \prod_{i=1}^{N_u} \left[\frac{1}{\beta} \exp\left(-\frac{y_i}{\beta}\right) \right]$$

para $\gamma = 0$ e $y_i > 0$.

O par de parâmetros desconhecidos $\theta = (\gamma, \beta)$ é então estimado pelo maximizante da função de verosimilhança, ou então pelo maximizante da função do logaritmo da função verosimilhança.

Assim, quando $\gamma \neq 0$ temos

$$\begin{aligned} \ln L_{\gamma,\beta}(y) &= \ln \prod_{i=1}^{N_u} \left[\frac{1}{\beta} \left(1 + \gamma \frac{y_i}{\beta}\right)^{-\frac{1}{\gamma}-1} \right] \\ &= \sum_{i=1}^{N_u} \ln \left[\frac{1}{\beta} \left(1 + \gamma \frac{y_i}{\beta}\right)^{-\frac{1}{\gamma}-1} \right] \\ &= \sum_{i=1}^{N_u} \ln \left(\frac{1}{\beta} \right) + \sum_{i=1}^{N_u} \ln \left[\left(1 + \gamma \frac{y_i}{\beta}\right)^{-\left(\frac{1}{\gamma}+1\right)} \right] \\ &= -N_u \ln \beta - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{N_u} \ln \left(1 + \gamma \frac{y_i}{\beta}\right). \end{aligned} \quad (3.26)$$

Quando $\gamma = 0$, vem que

$$\begin{aligned}
 \ln L_{0,\beta}(y) &= \ln \prod_{i=1}^{N_u} \left[\frac{1}{\beta} \exp \left(-\frac{y_i}{\beta} \right) \right] \\
 &= \sum_{i=1}^{N_u} \ln \left[\frac{1}{\beta} \exp \left(-\frac{y_i}{\beta} \right) \right] \\
 &= \sum_{i=1}^{N_u} \ln \left(\frac{1}{\beta} \right) + \sum_{i=1}^{N_u} \ln \left(\exp \left(-\frac{y_i}{\beta} \right) \right) \\
 &= -N_u \ln \beta - \frac{1}{\beta} \sum_{i=1}^{N_u} y_i.
 \end{aligned} \tag{3.27}$$

Observemos que se utilizarmos a seguinte reparametrização

$$(\gamma, \beta) \rightsquigarrow (\gamma, \tau) \text{ com } \tau := \frac{\gamma}{\beta},$$

então a expressão em (3.26) escreve-se do seguinte modo

$$\begin{aligned}
 \ln L_{\gamma,\tau}(y) &= -N_u \ln \left(\frac{\gamma}{\tau} \right) - \left(\frac{1}{\gamma} + 1 \right) \sum_{i=1}^{N_u} \ln(1 + \tau y_i) \\
 &= -N_u \ln \gamma + N_u \ln \tau - \left(\frac{1}{\gamma} + 1 \right) \sum_{i=1}^{N_u} \ln(1 + \tau y_i).
 \end{aligned} \tag{3.28}$$

Para encontrar o maximizante de (3.28) é então necessário encontrar a solução do sistema

$$\begin{cases} \frac{\partial}{\partial \gamma} \ln L_{\gamma,\tau}(y) = 0 \\ \frac{\partial}{\partial \tau} \ln L_{\gamma,\tau}(y) = 0. \end{cases}$$

De

$$\frac{\partial}{\partial \gamma} \ln L_{\gamma,\tau}(y) = 0$$

obtemos

$$-N_u \left(\frac{1}{\gamma} \right) + \frac{1}{\gamma^2} \sum_{i=1}^{N_u} \ln(1 + \tau y_i) = 0$$

de onde resulta

$$\gamma = \frac{\sum_{i=1}^{N_u} \ln(1 + \tau y_i)}{N_u}. \tag{3.29}$$

De

$$\frac{\partial}{\partial \tau} \ln L_{\gamma, \tau}(y) = 0$$

vem que

$$N_u \left(\frac{1}{\tau} \right) - \left(\frac{1}{\gamma} + 1 \right) \sum_{i=1}^{N_u} \left(\frac{y_i}{1 + \tau y_i} \right) = 0$$

que é o mesmo que ter

$$\frac{1}{\tau} - \left(\frac{1}{\gamma} + 1 \right) \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{y_i}{1 + \tau y_i} = 0.$$

Assim, os estimadores de máxima verosimilhança de γ e τ , $\hat{\gamma}_{ML}$ e $\hat{\tau}_{ML}$, resultam da resolução de

$$\frac{1}{\hat{\tau}_{ML}} - \left(\frac{1}{\hat{\gamma}_{ML}} + 1 \right) \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{Y_i}{1 + \hat{\tau}_{ML} Y_i} = 0,$$

com

$$\hat{\gamma}_{ML} = \frac{1}{N_u} \sum_{i=1}^{N_u} \ln(1 + \hat{\tau}_{ML} Y_i). \quad (3.30)$$

Notemos que, ao contrário do que é observado nos dois métodos anteriores, não conseguimos obter uma solução explícita para os estimadores dos parâmetros. Assim, a obtenção das estimativas para os parâmetros da função de distribuição GP pelo método ML resulta da aplicação de métodos numéricos.

Quando temos $\gamma = 0$, para encontrar o maximizante de (3.27), é necessário encontrar a solução da equação

$$\frac{\partial}{\partial \beta} \ln L_{\beta}(y) = 0. \quad (3.31)$$

de onde obtemos

$$-N_u \frac{1}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{N_u} y_i = 0$$

que resulta em

$$\begin{aligned} \beta &= \frac{1}{N_u} \sum_{i=1}^{N_u} y_i \\ &= \bar{y}. \end{aligned} \quad (3.32)$$

A derivada de segunda ordem da função de verosimilhança é, neste caso, dada por

$$N_u \frac{1}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{N_u} y_i.$$

Notemos que

$$\begin{aligned} \left[N_u \frac{1}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{N_u} y_i \right]_{\beta=\bar{y}} &= \frac{N_u}{(\bar{y})^2} - \frac{2}{(\bar{y})^3} \times N_u(\bar{y}) \\ &= \frac{N_u}{(\bar{y})^2} - \frac{2N_u}{(\bar{y})^2} \\ &= -\frac{N_u}{(\bar{y})^2} < 0 \end{aligned}$$

Assim, concluímos que o estimador de β pelo método da máxima verosimilhança, $\hat{\beta}_{ML}$, é dado pela média de (Y_1, \dots, Y_{N_u}) .

Zea Bermudez e Kotz [6] enumeram várias vantagens e desvantagens dos métodos abordados nesta secção. Uma das vantagens do método MOM é o facto de ser de fácil utilização. Por esta razão, este método é muitas vezes aplicado para fornecer estimativas dos parâmetros da função de distribuição GP a utilizar como valores iniciais em algoritmos numéricos relativos a outros métodos de estimação. Uma das desvantagens deste método consiste no facto de apenas existir estimadores determinados por este método quando $\gamma < 0.5$. Ao contrário do método MOM, o método PWM permite determinar estimadores quando temos $0.5 \leq \gamma < 1$. No entanto, as estimativas obtidas por este método podem ser inviáveis quando o $\gamma < 0$. Em particular, Castillo e Hadi em [9] constataram, através da realização de simulações, que o método PWM é mais adequado quando temos amostras pequenas e $0 \leq \gamma \leq 0.5$. Os métodos MOM e PWM apresentam a desvantagem comum de fornecerem estimativas para γ não pertencem ao domínio deste parâmetro [10]. Por outro lado, de acordo com Zea Bermudez e Kotz em [6], o método ML é mais eficiente na estimação dos parâmetros apesar de requerer a utilização de algoritmos numéricos. A necessidade de utilização destes algoritmos pode ser uma desvantagem na medida que podem existir problemas de convergência mesmo quando as amostras são grandes.

3.2 Estimação de quantis extremais

De acordo com Gilli e K ellezi em [21], a estima ao de quantis extremais tem um papel importante na gest ao quantitativa do risco. Uma das medidas mais utilizadas nessa gest ao   o *Value at Risk* (VaR), que n o   mais do que um quantil extremal. Assim, a estima ao de quantis extremais no contexto da metodologia POT ser  analisada em seguida.

A fun ao inversa generalizada de uma dada fun ao de distribui ao F dada por

$$F^{\leftarrow}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}, \quad 0 < t < 1$$

  tamb m denominada de fun ao quantil da fun ao de distribui ao F (cf., e.g., [16]).

O quantil extremal x_p , para p pequeno, tal que

$$x_p := F^{\leftarrow}(1 - p),$$

  denominado de quantil extremal [24].

Designamos a probabilidade de um n vel elevado c ser excedido

$$p^c := P[X > c]$$

por probabilidade de cauda. Uma vez que $P[X > x_p] = p$, o quantil extremal x_p   tamb m denominado de quantil de probabilidade de exced ncia p . Assim, tamb m   definido o VaR ($\text{VaR}_p = x_p$), valor que   ultrapassado para uma probabilidade p muito pequena:

$$\text{VaR}_p : P[X > \text{VaR}_p] = p.$$

Na pr tica, os valores de p mais usuais s o 0.01, 0.001 e 0.0001 [24]. O valor esperado de cauda condicional, CTE, definido por

$$\text{CTE}_p = \mu_p := E[X|X > x_p] = \frac{1}{p} \int_{x_p}^{\infty} xf(x)dx.$$

  o valor m dio condicional a 100 p % do topo da popula ao [24]. O CTE   outra medida de avalia ao do risco que, de acordo com Gilli e K ellezi em [21], n o   t o utilizada na pr tica como a medida VaR.

Sabemos que o valor x_p satisfaz a seguinte relação (cf. [24]) entre o valor esperado de cauda condicional e a função de excesso médio definida em (3.1)

$$\mathbb{E}[X|X > x_p] = c + e(x_p). \quad (3.33)$$

Observamos também que da seguinte igualdade obtida na Secção 2.2 (ver (2.18))

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}$$

resulta

$$F(u+y) = F_u(y)(1 - F(u)) + F(u). \quad (3.34)$$

Se substituirmos $x = u + y$ em (3.34) obtemos

$$F(x) = F_u(x-u)(1 - F(u)) + F(u). \quad (3.35)$$

De (3.35) vem que

$$\begin{aligned} 1 - F(x) &= 1 - [F_u(x-u)(1 - F(u)) + F(u)] \\ &= (1 - F(u)) - F_u(x-u)(1 - F(u)) \\ &= [1 - F(u)][1 - F_u(x-u)]. \end{aligned}$$

Assim, denotando $1 - F$ por \bar{F} , temos que

$$\bar{F}(x) = \bar{F}(u)(1 - F_u(x-u)).$$

Considerando que a distribuição condicional dos excessos acima de u é aproximada pela função de distribuição GP ($F_u(y) \approx H_{\alpha,\beta}(y)$) temos que (cf., e.g., (2.26))

$$\bar{F}(x) \approx \bar{F}(u) \left(1 + \gamma \frac{(x-u)}{\beta} \right)^{-1/\gamma}, \quad (3.36)$$

para algum γ, β , o que permite estimar a probabilidade de excedência do valor elevado x .

Estimemos $\bar{F}(u)$ por N_u/n com N_u o número de observações que excedem u na amostra de dimensão n . Se, além disso, substituirmos γ e β pelos respectivos estimadores obtidos por um dos métodos mencionados na Secção 3.1 vem que

$$\hat{\bar{F}}(x) = \frac{N_u}{n} \left(1 + \hat{\gamma} \frac{(x-u)}{\hat{\beta}} \right)^{-1/\hat{\gamma}}. \quad (3.37)$$

Assim, considerando $F(x) = 1 - p$ na expressão (3.37) obtemos

$$\begin{aligned}
p &= \frac{N_u}{n} \left(1 + \hat{\gamma} \frac{(x-u)}{\hat{\beta}} \right)^{-1/\hat{\gamma}} \\
\Leftrightarrow \left(\frac{np}{N_u} \right)^{-\hat{\gamma}} &= 1 + \hat{\gamma} \frac{(x-u)}{\hat{\beta}} \\
\Leftrightarrow \left(\frac{np}{N_u} \right)^{-\hat{\gamma}} - 1 &= \frac{\hat{\gamma}x - \hat{\gamma}u}{\hat{\beta}} \\
\Leftrightarrow \hat{\gamma}u + \hat{\beta} \left(\left(\frac{np}{N_u} \right)^{-\hat{\gamma}} - 1 \right) &= \hat{\gamma}x \\
\Leftrightarrow u + \frac{\hat{\beta}}{\hat{\gamma}} \left(\left(\frac{np}{N_u} \right)^{-\hat{\gamma}} - 1 \right) &= x.
\end{aligned}$$

Assim, o estimador de quantis elevados de $H_{\gamma,\beta}$ é então dado por

$$\hat{x}_p = u + \frac{\hat{\beta}}{\hat{\gamma}} \left(\left(\frac{np}{N_u} \right)^{-\hat{\gamma}} - 1 \right). \quad (3.38)$$

A igualdade (3.38) permite determinar uma estimativa para a medida de risco VaR_p a partir dos estimadores dos parâmetros da função de distribuição GP. Assim, temos

$$\widehat{\text{VaR}}_p = u + \frac{\hat{\beta}}{\hat{\gamma}} \left(\left(\frac{np}{N_u} \right)^{-\hat{\gamma}} - 1 \right). \quad (3.39)$$

Por outro lado, tendo em conta a igualdade (3.33), escrevemos

$$\widehat{\text{CTE}}_p = \widehat{\text{VaR}}_p + E(X - \widehat{\text{VaR}}_p | X > \widehat{\text{VaR}}_p). \quad (3.40)$$

Uma vez que, para um dado valor r , temos (ver (3.2))

$$e(r) = E[X - r | X > r] = \frac{\beta + \gamma r}{1 - \gamma}, \quad \beta + \gamma r > 0,$$

obtemos da igualdade (3.40)

$$\widehat{\text{CTE}}_p = \widehat{\text{VaR}}_p + \frac{\hat{\beta} + \hat{\gamma}(\widehat{\text{VaR}}_p - u)}{1 - \hat{\gamma}} = \frac{\widehat{\text{VaR}}_p}{1 - \hat{\gamma}} + \frac{\hat{\beta} - \hat{\gamma}u}{1 - \hat{\gamma}}, \quad (3.41)$$

com $r = \widehat{\text{VaR}}_p - u$.

Estas duas medidas de risco serão determinadas na ilustração prática realizada no Capítulo 5 após a aplicação dos métodos de escolha estatística de modelos extremais estudados no próximo capítulo.

Capítulo 4

Escolha estatística de modelos extremais

Neste capítulo, abordamos a temática da escolha estatística de modelos extremais. Em particular, queremos analisar a escolha entre a função de distribuição exponencial e a função de distribuição GP para os excessos acima de um dado limiar u . Assim sendo, apresentamos algumas estatísticas de testes presentes na literatura específica da área utilizadas para testar a nulidade do parâmetro de forma γ . Finalizamos o capítulo com a análise da escolha de modelos mencionada, quer através de intervalos de confiança, quer através de métodos gráficos.

4.1 Teste de hipóteses

Representemos uma população por um modelo X , variável aleatória com função de distribuição F . Denominamos de hipótese estatística qualquer asserção sobre aspetos desconhecidos de F . Quando essa asserção é relativa apenas ao parâmetro da função de distribuição F conhecida, seja esse parâmetro escalar ou vetorial, então temos uma hipótese paramétrica.

Suponhamos, então, que X é uma variável aleatória com uma função densidade de probabilidade f_X que pertence a uma família parametrizada por θ ,

$$\{f_\theta, \theta \in \Theta\}$$

onde Θ é o espaço de parâmetros. Qualquer hipótese paramétrica estabelece uma partição do espaço de parâmetros Θ ($\{\Theta_0, \Theta_1\}$ com $\Theta_0 \cup \Theta_1 = \Theta$ e $\Theta_0 \cap \Theta_1 = \emptyset$)

onde $H_0 : \theta \in \Theta_0$ é a hipótese a testar, e $H_1 : \theta \in \Theta_1$ é a hipótese que corresponde ao conjunto das alternativas [35]. Esta hipótese é, assim, designada por hipótese alternativa. A designação tradicional para H_0 é hipótese nula uma vez que geralmente corresponde a algo que se pretende manter.

Suponhamos que X_1, X_2, \dots, X_n são variáveis aleatórias independentes e identicamente distribuídas a X e designemos por estatística de teste a estatística $T = T(X_1, X_2, \dots, X_n)$. O espaço amostral T , relativo à estatística T , é o conjunto de todos os valores particulares $t = T(x_1, x_2, \dots, x_n)$ com (x_1, x_2, \dots, x_n) uma concretização de (X_1, X_2, \dots, X_n) . Um teste de hipóteses estabelece uma regra que permite determinar um conjunto $R_t \subset T$ tal que rejeitamos H_0 se $t \in R_t$ e não rejeitamos H_0 se $t \notin R_t$ [35]. Este conjunto R_T é denominado de região de rejeição ou região crítica.

Assim, rejeitar ou não rejeitar H_0 constitui o resultado do teste de hipóteses. A ideia subjacente é de apenas rejeitar H_0 se existir uma evidência substancial nos dados para tal. Queremos pois tomar uma decisão objetiva, mas essa decisão estará sujeita a erro. Ao testarmos a hipótese nula H_0 *versus* a hipótese alternativa H_1 podemos cometer dois tipos de erros. Denominamos de erro de primeira espécie ou de rejeição à decisão errada de rejeitar H_0 quando H_0 é verdadeira. O erro de segunda espécie ou não rejeição consiste em não rejeitar a hipótese nula quando esta é falsa.

Designamos por nível de significância α do teste (dimensão ou nível de teste) a probabilidade de cometer um erro de primeira espécie, ou seja,

$$\alpha = P[\text{rejeitar } H_0 | H_0 \text{ é verdadeira}].$$

Assim, dada uma amostra, ao adaptar a regra de rejeitar H_0 se o valor observado para a estatística de, $t_{\text{obs}} = t(x_1, x_2, \dots, x_n)$ pertencer à região de rejeição, estamos a tomar uma decisão objetiva na qual a probabilidade de erro de primeira espécie é α [45].

Existe, também, uma forma alternativa de realizar esta decisão. Denominamos de valor- p , (ou valor de prova p , ou nível de significância descritivo) a probabilidade, p_{obs} , de obter o valor t_{obs} ou outro mais desfavorável para a hipótese nula, admitindo que esta hipótese é verdadeira [35]. Por outras palavras, é a probabilidade de situações amostrais tão ou mais improváveis do que a observada, sob a validade de H_0 [45]. Assim esta outra forma de tomar a decisão consiste em rejeitar H_0 se o valor- p for menor ou igual a α .

Um teste de hipótese pode ser formulado com hipóteses nulas simples contra alternativas compostas. Uma hipótese estatística é designada de simples quando o correspondente subconjunto do espaço de parâmetros é constituído por apenas um elemento, sendo denominada de composta no caso contrário [35]. Em particular, as hipóteses alternativas compostas podem ser unilaterais, ou seja, temos um teste de hipóteses do tipo

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0 \quad (4.1)$$

ou

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0 \quad (4.2)$$

com $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \{\theta : \theta > \theta_0\}$ em (4.1) e $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \{\theta : \theta < \theta_0\}$ em (4.2). Quando a alternativa composta é bilateral temos um teste de hipóteses do tipo

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0. \quad (4.3)$$

No nosso caso, estamos interessados em testar um dos seguintes pares:

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma \neq 0 \quad (4.4)$$

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma < 0 \quad (4.5)$$

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma > 0, \quad (4.6)$$

onde γ representa o parâmetro de forma em (2.26).

Nas subsecções seguintes, vamos considerar Y_1, \dots, Y_m variáveis independentes e identicamente distribuídas à variável aleatória Y com função de distribuição GP, com agora m denotando o número de excedências.

4.1.1 Estatística de teste T_1

A primeira estatística de teste sugerida nesta dissertação aparece em alguns livros da área (cf [10], [12], [24]). Aqui denotamos esta estatística de teste por estatística de teste T_1 . Esta é dada por

$$T_1 = -2 \ln \left(\frac{L_p(\hat{\gamma})}{L_p(\gamma)} \right), \quad (4.7)$$

onde, para cada valor de γ , a função L_p , designada por *profile log-likelihood* é a log-verosimilhança maximizada relativamente ao parâmetro β , ou seja,

$$L_p(\gamma) = \max_{\beta|\gamma} L(\gamma, \beta).$$

Esta estatística de teste é aplicada para testar

$$H_0 : \gamma = \gamma_0 \text{ versus } H_1 : \gamma \neq \gamma_0.$$

Sob a validade da hipótese nula

$$T_1 \xrightarrow[m \rightarrow \infty]{d} V \quad (4.8)$$

onde V é uma variável aleatória qui-quadrado com um grau de liberdade e o m representa o número de excedências. Para um nível de significância α , a hipótese H_0 é rejeitada se $T_1 > v_{1-\alpha}$ onde $v_{1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da variável V . O valor de prova p é dado por

$$p(T_1) = 1 - F_V(T_1),$$

onde F_V é função de distribuição da variável aleatória V .

Reiss e Thomas em [37] recomendam ainda a aplicação da correção de Bartlett, resultando assim na seguinte estatística de teste

$$T_{1,b} = \frac{T_1}{1 + \frac{4}{m}}. \quad (4.9)$$

Sob a validade da hipótese nula temos também

$$T_{1,b} \xrightarrow[m \rightarrow \infty]{d} V \quad (4.10)$$

onde V é uma variável aleatória qui-quadrado com um grau de liberdade. Ao nível de significância α , a hipótese H_0 é rejeitada se $T_{1,b} > v_{1-\alpha}$ onde $v_{1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da variável V .

O valor p , agora com a aplicação da correção de Bartlett, é

$$p(T_{1,b}) = 1 - F_V(T_{1,b}),$$

onde F_V é função de distribuição da variável aleatória V .

4.1.2 Estatística de teste T_2

Na Secção 3.1, relembramos como obter os estimadores pelo método MOM para os parâmetros γ e β . Da expressão em (3.8), verificamos que

$$E[Y] = \frac{\beta}{1-\gamma} \text{ e } E[Y^2] = \frac{2\beta^2}{(1-2\gamma)(1-\gamma)}$$

para $\gamma < \frac{1}{2}$.

Chaouche e Bacro em [11] constataram que

$$\frac{E[Y^2]}{2E[Y]^2} - 1$$

apenas se escreve em função do parâmetro γ . De facto, temos que

$$\begin{aligned} \frac{E[Y^2]}{2E[Y]^2} - 1 &= \frac{2\beta^2}{(1-2\gamma)(1-\gamma)} - 1 \\ &= \frac{\beta^2}{2(1-\gamma)^2} \\ &= \frac{2\beta^2(1-\gamma)^2}{2\beta^2(1-2\gamma)(1-\gamma)} - 1 \\ &= \frac{1-\gamma}{1-2\gamma} - 1 \\ &= \frac{1-\gamma-1+2\gamma}{1-2\gamma} \\ &= \frac{\gamma}{1-2\gamma} \end{aligned}$$

Assim, Chaouche e Bacro [11] propuseram a seguinte estatística de teste

$$T_2 = \frac{E[Y^2]}{2E[Y]^2} - 1, \quad (4.11)$$

e demonstraram que, sob a validade da hipótese nula,

$$T_2^* = \sqrt{m} \left(T_2 - \frac{1-\gamma}{1-2\gamma} \right) \xrightarrow[m \rightarrow \infty]{d} V, \quad (4.12)$$

onde V é uma variável aleatória com distribuição normal com média zero e variância $\frac{(1-\gamma)^2(1-\gamma+6\gamma^2)}{(1-2\gamma)^3(1-3\gamma)(1-4\gamma)}$.

Ao nível de significância α , rejeitamos a hipótese H_0 em (4.4) se $|T_2^*| \geq v_{1-\alpha/2}$, onde $v_{1-\alpha/2}$ representa o quantil de probabilidade $1 - \alpha/2$ da variável V . O correspondente valor p é obtido por

$$p(T_2^*) = 2 - 2F_V(|T_2^*|),$$

onde F_V é função de distribuição da variável aleatória V .

O valor observado para T_2 , $t_{2,obs}$, é obtido pela substituição dos momentos de ordem um e dois, $E[Y]$ e $E[Y^2]$, pelos momentos empíricos de ordem um e dois dados, respetivamente, por

$$\frac{1}{m} \sum_{j=1}^m y_j \quad \text{e} \quad \frac{1}{m} \sum_{j=1}^m y_j^2.$$

4.1.3 Estatística de teste T_3

Além da estatística de teste T_1 , Reiss e Thomas em [37] referem outra estatística de teste, que também pode ser encontrada em [31]. Esta estatística de teste, que se escreve em função do quadrado do coeficiente de variação, é dada por

$$T_3 = \frac{1}{2} \left(\frac{S_Y^2}{\bar{Y}^2} - 1 \right) \quad (4.13)$$

com $\bar{Y} = \frac{1}{m} \sum_{j=1}^m Y_j$ e $S_{Y,m}^2 = \frac{1}{m} \sum_{j=1}^m (Y_j - \bar{Y})^2$.

Gomes e van Montfort em [25] demonstraram que, sob a validade da hipótese nula, e $\gamma < \frac{1}{2}$,

$$T_3^* = \sqrt{m} T_3 \xrightarrow[m \rightarrow \infty]{d} V \quad (4.14)$$

onde V é uma variável aleatória com distribuição normal padrão.

Consideremos os valores $v_{1-\alpha/2}$, v_α e $v_{1-\alpha}$ correspondentes, respetivamente, aos quantis de probabilidade $1 - \alpha/2$, α e $1 - \alpha$ da variável aleatória V . Ao nível de significância α , rejeitamos a hipótese H_0 em (4.4), (4.5) e (4.6) se $|T_3^*| \geq v_{1-\alpha/2}$, $T_3^* \leq v_\alpha$ e $T_3^* \leq v_{1-\alpha}$, respetivamente. O correspondente valor p é obtido, respetivamente, por

$$p(T_3^*) = 2 - 2F_V(|T_3^*|),$$

$$p(T_3^*) = F_V(T_3^*)$$

e

$$p(T_3^*) = 1 - F_V(T_3^*),$$

onde F_V representa a função distribuição da variável aleatória V .

O valor observado para T_3 , $t_{3,obs}$, é obtido pela substituição de \bar{Y} por

$$\bar{y} = \frac{1}{m} \sum_{j=1}^m y_j,$$

e de $S_{Y,m}^2$ por

$$s_{y,m}^2 = \frac{1}{m} \sum_{j=1}^m (y_j - \bar{y})^2.$$

4.1.4 Estatística de teste T_4

Além da estatística de teste T_3 , Gomes e van Montfort em [25] analisaram também, no contexto da metodologia POT, a estatística de teste proposta por van Montfort e Witter em [44], que aqui denominamos de estatística de teste T_4 .

A estatística de teste T_4 é igual a razão entre o máximo ($Y_{m:m}$) e mediana (M_m) das variáveis Y_1, \dots, Y_m , ou seja,

$$T_4 = \frac{Y_{m:m}}{M_m}, \quad (4.15)$$

com

$$M_m = \begin{cases} Y_{\frac{m+1}{2}:m} & \text{se } m \text{ é ímpar,} \\ \frac{1}{2} (Y_{\frac{m}{2}:m} + Y_{\frac{m}{2}+1:m}) & \text{se } m \text{ é par.} \end{cases} \quad (4.16)$$

Gomes e van Montfort [25] demonstraram que, sob a validade da hipótese nula, e $\gamma > -\frac{1}{2}$,

$$T_4^* = T_4 \times \ln 2 - \ln m \xrightarrow[m \rightarrow \infty]{d} V \quad (4.17)$$

onde V é uma variável aleatória com distribuição de Gumbel padrão.

Para um nível de significância α , rejeitamos a hipótese nula em (4.5) se $T_4^* \leq v_\alpha$, com v_α a designar o quantil de probabilidade α da variável aleatória V . Se tomarmos a hipótese H_0 em (4.6), rejeitamos esta se $T_4^* \geq v_{1-\alpha}$, onde $v_{1-\alpha}$ denota o quantil de probabilidade $1 - \alpha$ da variável aleatória V .

Os valores p correspondentes são calculados, respetivamente, por

$$p(T_4^*) = F_V(T_4^*)$$

e

$$p(T_4^*) = 1 - F_V(T_4^*).$$

onde F_V representa a função de distribuição da variável aleatória V .

4.1.5 Estatística de teste T_5

No artigo elaborado por Gomes em [23], a estatística de teste dada pela razão entre a diferença do máximo e a mediana e a diferença desta e o mínimo, foi utilizada na escolha de modelos extremais no contexto da metodologia dos máximos anuais. Brilhante em [8] refere que esta estatística de teste também pode ser utilizada para testar a função de distribuição exponencial *versus* a função de distribuição GP e que, sob a validade da hipótese nula

$$T_5^* = T_5 \times \ln 2 - \ln(m/2) \xrightarrow{m \rightarrow \infty} V \quad (4.18)$$

onde V é uma variável aleatória com distribuição de Gumbel padrão e

$$T_5 = \frac{Y_{m:m} - M_m}{M_m - Y_{1:m}}, \quad (4.19)$$

com M_m definida em (4.16).

Para um nível de significância α , rejeitamos a hipótese H_0 em (4.5) se $T_5^* \leq v_\alpha$, com v_α a designar o quantil de probabilidade α da variável aleatória V . Se tomarmos a hipótese H_0 em (4.6), rejeitamos esta se $T_5^* \geq v_{1-\alpha}$, onde $v_{1-\alpha}$ denota o quantil de probabilidade $1 - \alpha$ da variável aleatória V .

Os valores p correspondentes são calculados, respetivamente, por

$$p(T_5^*) = F_V(T_5^*)$$

e

$$p(T_5^*) = 1 - F_V(T_5^*),$$

onde F_V representa a função de distribuição da variável aleatória V .

4.1.6 Estatística de teste T_6

Brilhante em [8], tendo em conta a estatística de teste T_5 , apresentou a estatística de teste dada pela razão entre a diferença do quarto superior e a mediana e a diferença desta e o quarto inferior. A estatística de teste, aqui denotada por T_6 , é escrita do seguinte modo

$$T_6 = \frac{F_U - M_m}{M_m - F_L} \quad (4.20)$$

com M_m definida em (4.16), $F_U = Y_{m - \{\frac{m}{4}\} + 1:m}$ e $F_L = Y_{\{\frac{m}{4}\}:m}$ onde $\{\cdot\}$ representa o arredondamento para o número inteiro mais próximo.

Assim, a estatística de teste definida em (4.20) é dada por

$$T_6 = \frac{Y_{m-\{\frac{m}{4}\}+1:m} - Y_{\frac{m+1}{2}:m}}{Y_{\frac{m+1}{2}:m} - Y_{\{\frac{m}{4}\}:m}}, \quad (4.21)$$

se m é ímpar, e por

$$T_6 = \frac{Y_{m-\{\frac{m}{4}\}+1:m} - \frac{1}{2} \left(Y_{\frac{m}{2}:m} + Y_{\frac{m}{2}+1:m} \right)}{\frac{1}{2} \left(Y_{\frac{m}{2}:m} + Y_{\frac{m}{2}+1:m} \right) - Y_{\{\frac{m}{4}\}:m}}, \quad (4.22)$$

se m é par.

Brilhante em [8] demonstrou que, sob a validade da hipótese nula

$$T_6^* = \ln(3/2) \sqrt{\frac{m}{2}} \left(T_6 - \frac{\ln 2}{\ln(3/2)} \right) \xrightarrow{m \rightarrow \infty} V \quad (4.23)$$

onde V é uma variável aleatória com distribuição normal padrão.

Consideremos os valores $v_{1-\alpha/2}$, v_α e $v_{1-\alpha}$ correspondentes aos quantis de probabilidade $1 - \alpha/2$, α e $1 - \alpha$ da variável aleatória V . Ao nível de significância α , rejeitamos a hipótese H_0 em (4.4), (4.5) e (4.6) se $|T_6^*| \geq v_{1-\alpha/2}$, $T_6^* \leq v_\alpha$ e $T_6^* \geq v_{1-\alpha}$, respetivamente. Esta regra de decisão permite assim o cálculo dos valores p correspondentes por

$$p(T_6^*) = 2 - 2F_V(|T_6^*|),$$

$$p(T_6^*) = F_V(T_6^*)$$

e

$$p(T_6^*) = 1 - F_V(T_6^*),$$

onde F_V representa a função de distribuição da variável aleatória V .

Numa comparação entre as estatísticas de teste T_4 , T_5 e T_6 , Brilhante em [8] constatou que a estatística de teste T_6 tem um pior desempenho na deteção da distribuição exponencial. No entanto, segundo o mesmo artigo, Brilhante defende a importância da utilização desta estatística de teste dado esta ser resistente a existência de dados perturbadores na amostra.

Em resumo, na Tabela 4.1 temos as estatísticas de teste sugeridas, a indicação da distribuição associada quando $m \rightarrow \infty$ sob a validade da hipótese nula e uma correspondente referência para consulta.

Tabela 4.1: Estatísticas de teste - distribuições.

	V	
$T_1 \xrightarrow[m \rightarrow \infty]{d} V$	Qui Quadrado	[10]
$T_{1,b} \xrightarrow[m \rightarrow \infty]{d} V$		[37]
$T_2^* = \sqrt{m} \left(T_2 - \frac{1-\gamma}{1-2\gamma} \right) \xrightarrow[m \rightarrow \infty]{d} V$	Normal	[11]
$T_3^* = \sqrt{m} T_3 \xrightarrow[m \rightarrow \infty]{d} V$		[25]
$T_4^* = T_4 \times \ln 2 - \ln m \xrightarrow[m \rightarrow \infty]{d} V$	Gumbel	[25]
$T_5^* = T_5 \times \ln 2 - \ln(m/2) \xrightarrow[m \rightarrow \infty]{d} V$		[8]
$T_6^* = \ln(3/2) \sqrt{\frac{m}{2}} \left(T_6 - \frac{\ln 2}{\ln(3/2)} \right) \xrightarrow[m \rightarrow \infty]{d} V$	Normal	[8]

4.2 Intervalos de confiança e métodos gráficos

Um outro método de escolha entre as funções de distribuição exponencial e função de distribuição GP consiste em observar a presença ou não do valor zero num intervalo de confiança obtido e decidir de acordo com o verificado (e.g. [4, 10]). Se o valor zero pertencer ao intervalo obtido então os dados em estudo não fornecem evidência suficiente para considerarmos, com um grau de confiança de $(1 - \alpha) \times 100\%$, a função de distribuição GP. Por outro lado, teremos essa evidência estatística para escolher a função de distribuição GP se o valor zero não pertencer ao intervalo de confiança observado.

Os estimadores obtidos pelos métodos MOM, PWM e ML para os parâmetros da função de distribuição GP têm uma distribuição assintoticamente normal (cf. [4, 10]).

Assim, quando na presença de grande amostras, um intervalo a $(1 - \alpha) \times 100\%$ de confiança para γ é dado aproximadamente por

$$\hat{\gamma} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{v}}{N_u}} \quad (4.24)$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição normal padrão.

Para os métodos MOM, PWM e ML tomamos em (4.24)

$$\hat{\gamma} = \hat{\gamma}_{\text{MOM}} \quad \text{e} \quad \hat{v} = \frac{(1 - 2\hat{\gamma}_{\text{MOM}})(1 - \hat{\gamma}_{\text{MOM}} + 6\hat{\gamma}_{\text{MOM}}^2)(1 - \hat{\gamma}_{\text{MOM}})^2}{(1 - 3\hat{\gamma}_{\text{MOM}})(1 - 4\hat{\gamma}_{\text{MOM}})},$$

$$\hat{\gamma} = \hat{\gamma}_{\text{PWM}} \quad \text{e} \quad \hat{v} = \frac{(1 - \hat{\gamma}_{\text{PWM}})(2 - \hat{\gamma}_{\text{PWM}})^2(1 - \hat{\gamma}_{\text{PWM}} + 2\hat{\gamma}_{\text{PWM}}^2)}{(1 - 2\hat{\gamma}_{\text{PWM}})(3 - 2\hat{\gamma}_{\text{PWM}})},$$

e $\hat{\gamma} = \hat{\gamma}_{\text{ML}}$ e $\hat{v} = (1 + \hat{\gamma}_{\text{ML}})^2$, com γ_{MOM} , γ_{PWM} e γ_{ML} definidos em (3.14), (3.23) e (3.30), respetivamente [4].

A necessidade de avaliar se um qualquer modelo probabilístico apresenta um ajustamento adequado à distribuição subjacente a um determinado conjunto de dados motivou o aparecimento de procedimentos gráficos. Além de permitirem essa validação informal, os procedimentos referidos em seguida também podem ser perspetivados como métodos gráficos de seleção de modelos.

Seja $x = (x_1, \dots, x_n)$ uma realização do vetor aleatório (X_1, \dots, X_n) , cujas marginais são variáveis aleatórias independentes e identicamente distribuídas a X com função distribuição F . Seja $(x_{1:n}, \dots, x_{n:n})$ a realização do vetor das estatísticas ordinais $(X_{1:n}, \dots, X_{n:n})$. Observamos que para cada valor de $x_{i:n}$, com $i \in \{1, \dots, n\}$, tem-se i observações inferiores ou iguais a $x_{i:n}$. Assim, é natural considerar a proporção i/n ($= \hat{F}_n(x_{i:n})$ com \hat{F}_n definida em (2.3)) como uma estimativa pontual para $F(x_{i:n})$. No entanto, é usual considerar outra proporção alternativa tal como

$$\frac{i}{n+1}. \quad (4.25)$$

Tanto esta alternativa como as seguintes

$$\frac{i-0.5}{n} \quad \text{e} \quad \frac{i-0.35}{n}$$

constituem casos particulares de

$$p_i = \frac{i-c}{n+d}, \quad 1 \leq i \leq n \quad (4.26)$$

com $c \geq 0$, $d \leq 0$ e $1 \leq i \leq n$ [10]. Os valores p_i são denominados de posições de marcação (*plotting positions*).

Seja θ o vetor de parâmetros desconhecidos da função de distribuição F . O gráfico probabilidade - probabilidade (ou gráfico P-P) é o diagrama de dispersão definido pelos pontos

$$\left\{ (F_{\hat{\theta}}(x_{i:n}), p_i), i = 1, \dots, n \right\},$$

onde $\hat{\theta}$ é um estimador consistente de θ . Relembremos que um estimador T_n de um parâmetro θ diz-se consistente se converge em probabilidade para o verdadeiro valor do parâmetro, ou seja, $\mathbb{P}[|T_n - \theta| > \varepsilon] \xrightarrow[n \rightarrow \infty]{} 0$.

A representação dos seguintes pontos

$$\left\{ (F_{\hat{\theta}}^{\leftarrow}(p_i), x_{i:n}), i = 1, \dots, n \right\},$$

é denominada de gráfico quantil-quantil (ou gráfico Q-Q) [4, 24].

Se o modelo analisado for adequado então a nuvem de pontos obtida estará razoavelmente perto da diagonal em ambos os gráficos. Esta análise pode ser complementada pela comparação do histograma obtido dos dados com a função densidade de probabilidade do modelo a ajustar. A título exemplificativo, consideremos $H_{\gamma, \beta}$ a função de distribuição GP e $h_{\gamma, \beta}$ a respetiva função densidade de probabilidade com $\gamma = 0.25$ e $\beta = 1$. Para tal, usamos a função `gpdSim` de modo a gerar os dados para a função de distribuição GP assim como a função `findThreshold` para definirmos um limiar, ambas as funções da biblioteca `fExtremes` do R. A Figura 4.1 resulta então da utilização de funções da biblioteca `ismev` com uma modificação mínima do seu código (tradução para português):

```
##### Procedimento 1 #####
library(fExtremes)
Exemplo<-gpdSim(model=list(xi=0.25, mu=0, beta=1), n=10000, seed=1)
Percentagem=5
nf = floor(Percentagem/100*length(as.vector(Exemplo)))
threshold<-findThreshold(Exemplo, nf, doplot = FALSE)
library(ismev)
z<-gpd.fit(Exemplo,threshold)
  gpd.pp_alt(z$mle, z$threshold, z$data)
  gpd.qq_alt(z$mle, z$threshold, z$data)
  gpd.his_alt(z$mle, z$threshold, z$data)
```

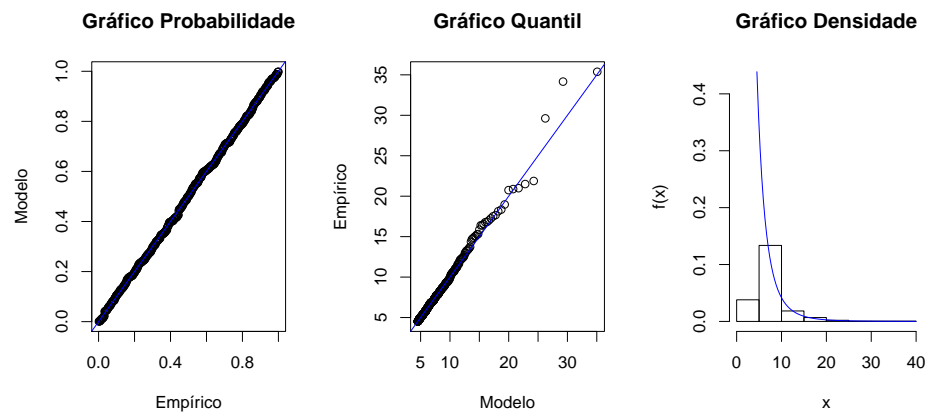


Figura 4.1: Gráficos de diagnóstico da biblioteca ismev.

Capítulo 5

Ilustração prática

Uma das principais preocupações dos agentes financeiros, sejam eles instituições financeiras ou investidores, é a gestão do risco de acontecimentos extremos que possam ocorrer nos mercados financeiros. Para estes agentes é crucial compreender a frequência e as consequências das descidas e subidas repentinas nos mercados financeiros [13]. A EVT é um importante ramo da estatística ao dispor da gestão do risco, onde os agentes financeiros visam obter respostas a questões tais como o que são os movimentos extremos que podem acontecer nos mercados financeiros ou se já aconteceram as maiores oscilações nos mercados financeiros.

A metodologia POT, um dos ramos mais importantes na EVT, será utilizada nesta ilustração prática. Esta metodologia, ao invés da metodologia de Gumbel, tem sido a mais utilizada em inúmeros estudos realizados nas mais diversas áreas de aplicação da EVT, nomeadamente em Finanças. Esta opção resulta de alguns autores terem observado que a metodologia POT permite uma utilização mais eficiente dos dados, os quais podem ser escassos na maior parte dos casos [32]. Os estudos realizados pelos autores Allen *et al.* em [2] e Gilli e Këllezli em [21] são exemplos de estudos na área das Finanças onde temos uma descrição e posterior comparação entre as duas metodologias da EVT acima mencionadas e que foram abordadas no Capítulo 2.

Outro trabalho realizado e de enorme interesse no âmbito desta ilustração prática foi elaborado por Curto e Rios em [13], no qual os autores testam se os valores acima de um limiar elevado seguem uma função de distribuição exponencial ou função de distribuição GP. Entre outros índices bolsistas mundiais, Curto e Rios utilizam o índice bolsista *S&P500*. Este índice bolsista

foi utilizado como base de dados para alguns dos estudos realizados na área das Finanças, incluindo os dois estudos já mencionados [2, 21]. Para colocarmos em prática as temáticas abordadas no Capítulo 4 escolhemos aqui analisar também com os retornos diários dos dados do índice bolsista *S&P500* (ver Tabela 5.1). Segundo a plataforma Plus500, que opera no mercado bolsista, o índice bolsista *S&P500* é constituído pelas 500 ações ou ativos americanos com as maiores capitalizações de mercado.

A base de dados escolhida foi retirada do site <http://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC> e nela estão incluídos os preços diários das ações. À semelhança dos artigos anteriormente referidos, trabalharemos com as taxas dos retornos diários, que segundo Curto e Rios em [13], são obtidas da seguinte forma:

$$\begin{aligned} r_t &= 100 \times [\log P_t - \log P_{t-1}] \\ &= 100 \times \log \left(\frac{P_t}{P_{t-1}} \right), \end{aligned} \quad (5.1)$$

onde P_t e P_{t-1} correspondem ao preço das ações no instante t e no instante $t-1$, respetivamente.

Tabela 5.1: Descrição dos dados *S&P500*.

	<i>S&P500</i>
Mínimo	-22.9
Mediana	0.04183
Máximo	10.96
Quantil 0.01	-2.709599
Quantil 0.05	-1.512276
Quantil 0.9	1.059779
Quantil 0.95	1.499139
Quantil 0.99	2.672995
Início	05/01/1960
Fim	05/01/2016
Dimensão	14097

5.1 Resultados

Nesta secção, são apresentados os resultados da aplicação aos dados em análise dos procedimentos de escolha estatística de modelos extremos estudados no capítulo anterior. O *software* R foi o *software* escolhido dada a quantidade de bibliotecas nele contida que abordam a teoria dos valores extremos. O R, *software* de utilização gratuita, era já em 2012 o *software* que continha a maior variedade de metodologias na área [20]. De modo a facilitar a procura das bibliotecas e das funções adequadas é aconselhável aceder à *task view* denominada *Extreme Value Analysis* <https://cran.r-project.org/web/views/ExtremeValue.html>. Até ao momento, nesta *task view* são mencionadas dezassete bibliotecas relativas à metodologia POT. Além da biblioteca com a mesma denominação que a metodologia em estudo (biblioteca POT), são mencionadas outras tais como as bibliotecas *evd*, *evir* e *ismev* que permitem, por exemplo, a estimação dos parâmetros da distribuição GP pelo método ML. As bibliotecas *evd*, *evir* e *ismev* são também indicadas na *task view* mencionada como bibliotecas a utilizar na obtenção do gráfico da função de excesso médio empírica. A Figura 5.1 exibe os três gráficos resultantes da aplicação aos dados *S&P500* considerados. Para tal, temos o seguinte procedimento:

```
##### Procedimento 2 #####
Graf.Limiar<-par(mfrow = c(1, 3))
library(evd) # Carrega a biblioteca que contém a função gráfica
             # “mrlplot” (Mean Residual Life Plot)
mrlplot(SP500)
library(evir) # Carrega a biblioteca que contém a função gráfica
             # “meplot” (Mean Excess Plot)
meplot(SP500)
library(ismev) # Carrega a biblioteca que contém a função gráfica
              # “mrl.plot” (Mean Residual Life Plot)
mrl.plot(SP500)
par(Graf.Limiar)
```

Além das bibliotecas já mencionadas, as bibliotecas *texmex*, *QRM* e *Relns* são referenciadas na *task view* “*Extreme Value Analysis*” como possíveis bibliotecas a aplicar na obtenção do mesmo gráfico. A biblioteca *evmix* possui também uma função *mrlplot*, a qual está baseada na correspondente função da biblioteca *evd* (<http://CRAN.R-project.org/doc/Rnews/> [41]).

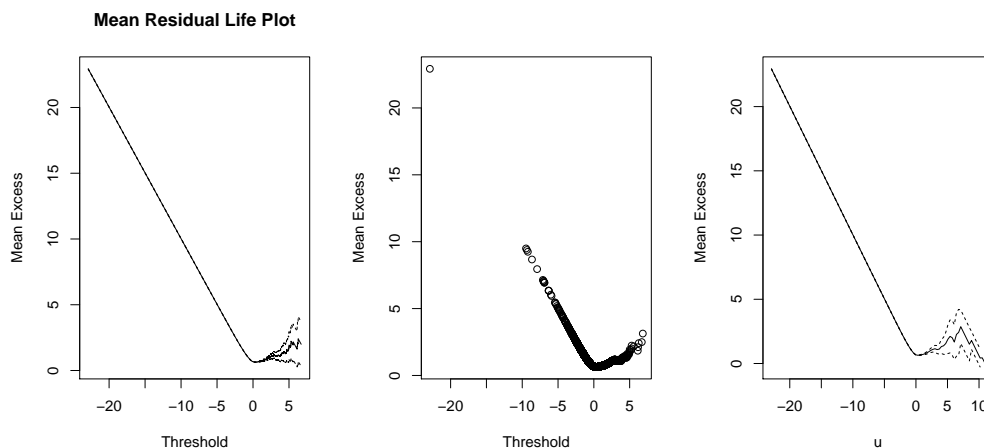


Figura 5.1: Gráficos das bibliotecas evd, evir e ismev.

No gráfico gerado por esta função há indicação das estimativas de máxima verossimilhança dos parâmetros da função de distribuição GP correspondente a um dado valor de referência para o limiar u . Como, na prática, a escolha de u resulta da observação no gráfico de uma linearidade à direita de u , a indicação gráfica (linha vertical) e numérica deste valor de referência é muito útil [1].

Um bom exemplo de um gráfico para a obtenção de um limiar plausível pode ser encontrado na biblioteca *evmix*. No gráfico há indicação de um valor para o limiar como também estimativas para os parâmetros de escala e de forma. Esse limiar é obtido considerando o valor do quantil 0.9 dos dados ordenados de forma ascendente (Figura 5.2). Com este gráfico ainda é possível comparar três estimativas para o limiar ao compararmos as três retas resultantes de modo a procurar a reta que melhor representa a tendência dos excessos médios (Figuras 5.3 e 5.4).

Uma nova função denominada de `mrplot_alt` foi criada de modo a alterar o texto da legenda (notação e tradução para português) no gráfico obtido da aplicação desta função `mrplot` da biblioteca *evmix*. A aplicação da função `mrplot_alt` aos dados S&P500 está descrita no seguinte procedimento gerado no R e resulta no gráfico exibido na Figura 5.2.

```
##### Procedimento 3 #####
library(evmix)

mrplot_alt(SP500, nt=min(100,length(SP500)), try.thresh=quantile (SP500,
0.975, na.rm=TRUE), main="Gráfico da vida média residual", xlab="Limiar u",
ylab="Excesso médio")
```

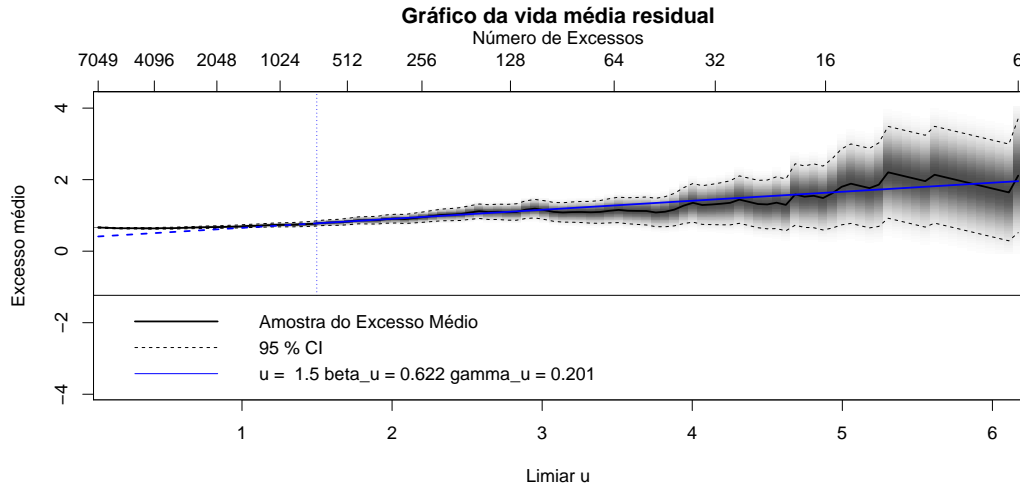


Figura 5.2: Gráfico MRL da biblioteca evmix.

A obtenção dos outros valores para o limiar u também considerados nesta ilustração prática é realizada por meio da aplicação da função `findthreshold()` da biblioteca `fExtremes`. A título exemplificativo, o procedimento efetuado no caso da utilização dos 10% maiores valores da amostra é apresentado de seguida:

```
##### Procedimento 4 - Limiar u #####
SP500<-(sp500$SP_ret) # Disponibiliza a amostra de valores utilizada
N<-length(SP500) # Determina a dimensão da amostra
library(fExtremes) # Carrega a biblioteca que contém a função aplicada
# na escolha de u
Percentagem=10 # Percentagem dos maiores valores
n=floor(Percentagem/100*length(as.vector(SP500)))
u<-findThreshold(SP500,n)
```

Na Tabela 5.2, estão indicados os valores obtidos para o limiar u e os respetivos números de excedências n_u considerando as 0.5%, 1%, 2.5%, 5% e 10% maiores observações da amostra.

Tabela 5.2: Limiares e número de excedências.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
Nº de exc. (n_u)	70	140	352	704	1409

Estes valores percentuais foram escolhidos por serem os mais usuais na literatura da área. Em [33], foram consideradas as 0.5%, 1% e 5% maiores observações da amostra, enquanto que no artigo de Allen *et al.* em [2] foram consideradas as 5% e 10% maiores observações da amostra. Por sua vez, Curto e Rios em [13] consideraram as 1%, 2.5% e 5% maiores observações da amostra.

É de realçar que a percentagem de 5% é aquela que reúne mais consensos na escolha do limiar u na aplicação da metodologia POT. Na prática, é considerado que a utilização dos 5% maiores valores da amostra permite obter um bom equilíbrio dos pressupostos inerentes à escolha do limiar. Isto é, o valor de u correspondente não é muito elevado de forma a não correremos o risco de termos poucos valores extremos para analisar, nem muito baixo para não correremos o risco de considerar valores que não são realmente extremos.

É de frisar que a função `mrlplot` da biblioteca `evmix` é útil para podermos comparar vários tipos de limiar. Desta forma, e tendo em conta os valores da Tabela 5.2, temos a representação gráfica para os diferentes limiares nas Figuras 5.3 e 5.4. A concretização destas figuras foi obtida através dos procedimentos 5.1 e 5.2 utilizando a mesma modificação usada na obtenção da Figura 5.2, respetivamente.

```
##### Procedimento 5.1 #####
library(evmix)

mrlplot_alt(SP500, nt=min(100,length(SP500)), try.thresh=quantile (SP500,
c(0.995, 0.99 ,0.975), na.rm=TRUE), main="Gráfico da vida média residual",
xlab="Limiar u", ylab="Excesso médio")
```

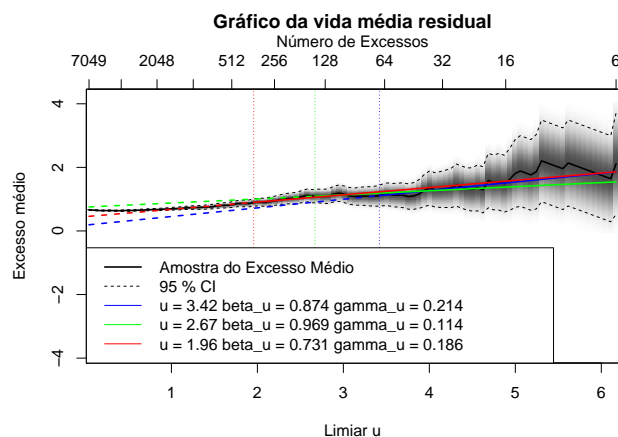


Figura 5.3: Gráficos MRL considerando 0.5%, 1% e 2.5% dos valores extremos.

```
##### Procedimento 5.2 #####
mrlplot_alt(SP500, nt=min(100,length(SP500)), try.thresh=quantile (SP500,
c(0.975, 0.95, 0.9), na.rm=TRUE), main="Gráfico da vida média residual",
xlab="Limiar u", ylab="Excesso médio")
```

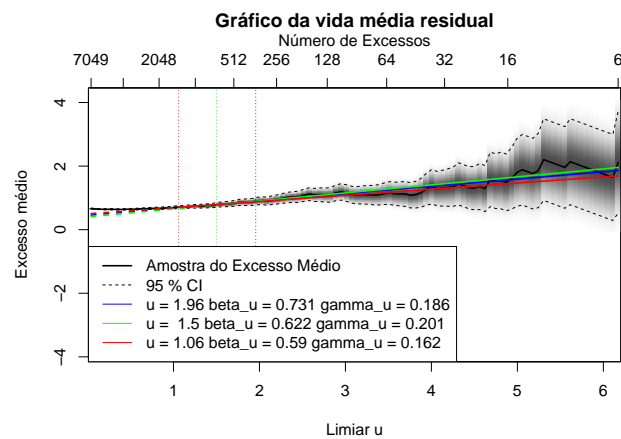


Figura 5.4: Gráficos MRL considerando 2.5%, 5% e 10% dos valores extremos.

Uma vez completada uma das etapas mais importantes na aplicação da metodologia POT, ou seja, a escolha de valores para o limiar u e que no nosso caso são vários (ver Tabela 5.2), veremos agora a aplicação das estatísticas de teste mencionadas na Secção 4.1.

5.1.1 Testes de hipóteses

Antes de aplicarmos as estatísticas de teste estudadas, apresentamos o procedimento aplicado na obtenção das excedências e dos excessos relativos a cada limiar considerado. Além disso, também apresentamos o código aplicado na determinação de algumas medidas de cada amostra a utilizar no cálculo dos valores observados das estatísticas de teste estudadas.

```
##### Procedimento 6 - Excedências e Excessos #####
SP500_s<-sort(SP500) # Ordena de forma ascendente a amostra
SP500_u<-SP500_s[which(SP500_s>u)] # Obtém as excedências amostrais
Y<-(SP500_u-u)# Obtém os excessos amostrais
Med_Y<-median(Y) # Determina a mediana dos excessos amostrais
M_Y<-mean(Y) # Determina a média dos excessos amostrais
m<-length(Y) # Determina o número total dos excessos amostrais
Y_mm<-max(Y) # Determina o máximo dos excessos amostrais
Y_1m<-min(Y) # Determina o mínimo dos excessos amostrais
```

Estatística de teste T_1

Começamos por aplicar a estatística de teste introduzida na Subsecção 4.1.1 e que podemos encontrar, por exemplo, em [10] e em [12]. Estamos interessados em testar as hipóteses em (4.4) e assim obtemos os valores da Tabela 5.3, que é construída através do seguinte procedimento:

```
##### Procedimento 7 - T_1 #####
library(Renext) # Biblioteca que contém a estatística T_1
T_1<-LRExp.test(Y,alternative= "GPD") # method = "asympt" quando m > 500
T_1$statistic
T_1b<-T_1$statistic/(1+4/m)
T_1b
pT1Bil<-1-pchisq(T_1$statistic,df=1)
pT1Bil # Valor de prova para T_1
pT1bBil<-1-pchisq(T_1b,df=1)
pT1bBil # Valor de prova para T_1b
```

Tabela 5.3: Limiares e valores resultantes do procedimento 7.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_1	3.221817	2.148515	12.7984	30.11005	43.9808
T_{1b}	3.047665	2.088834	12.6546	29.93994	43.85629
$p(T_1)$	0.07266292	0.1427079	0.00034692	4.082e-08	3.3161e-11
$p(T_{1b})$	0.08085331	0.1483793	0.00037464	4.456e-08	3.5339e-11

Observamos na Tabela 5.3 valores de prova p superiores ao nível de significância 0.05 para os dois maiores valores de u e para ambas as estatísticas de teste T_1 e T_{1b} . Nestes casos, ao nível de significância de 0.05, temos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo. Para os restantes valores de u considerados observamos, quer seja considerada a estatística de teste T_1 quer seja a estatística de teste T_{1b} , valores de prova p inferiores ao nível de significância 0.05. Logo, para estes casos, concluímos que os dados fornecem evidência suficiente para considerarmos não nulo o parâmetro de forma da função de distribuição GP para a modelação dos excessos populacionais em estudo.

Estatística de teste T_2

Aplicamos agora a estatística de teste introduzida na Subsecção 4.1.2 e que podemos encontrar, por exemplo, em [11]. À semelhança da aplicação da estatística anterior, estamos interessados em testar as hipóteses em (4.4). Assim obtemos os valores da Tabela 5.4, que é construída através do seguinte código:

```
##### Procedimento 8 #####
library(evir)# biblioteca para gerarmos o parâmetro de forma  $\gamma$ 
gamma<-gpd(SP500,u)$par.ests[1] # estimativa para o parâmetro  $\gamma$  pelo
                                # método ML

##### Procedimento 9 - T_2 #####
E1<-1/m*sum(Y)
E2<-1/m*sum(Y^2)
VAR<-((1-gamma)^2*(1-gamma+6*gamma^2))/((1-2*gamma)^3*(1-3*gamma)*
(1-4*gamma))
T_2<-E2/(2*(E1)^2)-1
T_2star<-sqrt(m)*(T_2-(1-gamma)/(1-2*gamma))
pT_2start<-(2-2*pnorm(abs(T_2star), mean=0, sd=sqrt(VAR)))
pT_2start
```

Tabela 5.4: Limiaries e valores resultantes do procedimento 9.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_2	0.2820213	0.1633753	0.2559924	0.2874925	0.2513468
T_2^*	-8.834884	-11.55749	-19.5115	-27.79946	-37.14262
$p(T_2^*)$	0.1674192	2.016e-08	7.027e-05	9.689e-06	0

Observamos na Tabela 5.4 valores de prova p não superiores ao nível de significância 0.05 para todos os valores de u , exceto quando $u_{0.005} = 3.421284$. Assim, ao nível de significância de 0.05, temos que os dados fornecem evidência suficiente para considerarmos não nulo o parâmetro de forma da função de distribuição GP para a modelação dos excessos populacionais em estudo, exceto quando $u_{0.005} = 3.421284$. Assim, para este caso, concluímos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo.

Estatística de teste T_3

Agora aplicamos a estatística de teste introduzida na Subsecção 4.1.3 e que podemos encontrar em [31], por exemplo. Em primeiro lugar vamos testar as hipóteses em (4.4). Para tal, geramos no R o seguinte código:

```
##### Procedimento 10 - T_3 #####
barY<-mean(Y)
S2<-1/m*sum((Y-M_Y)^2)
T_3<-1/2*(S2/barY^2-1)
T_3
T_3star<-sqrt(m)*T_3
T_3star

# Bilateral
pT_3startBil<-(2-2*pnorm(T_3star))
pT_3startBil
```

Os resultados decorrentes deste procedimento estão apresentados na Tabela 5.5.

Tabela 5.5: Limiares e valores resultantes do procedimento 10.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_3	0.2820213	0.1633753	0.2559924	0.2874925	0.2513468
T_3^*	2.35956	1.933083	4.802842	7.628039	9.434717
$p(T_3^*)$	0.01829663	0.05322596	1.564e-06	2.376e-14	0

Observamos na Tabela 5.5 valores de prova p não superiores ao nível de significância 0.05 para todos os valores de u , exceto quando $u_{0.01} = 2.672995$. Assim, ao nível de significância de 0.05, temos que os dados fornecem evidência suficiente para considerarmos não nulo o parâmetro de forma da função de distribuição GP para a modelação dos excessos populacionais em estudo, enquanto que para o valor $u_{0.01} = 2.672995$, observamos que o valor de prova p é superior ao nível de significância 0.05. Logo, para este caso, concluímos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo.

Ao analisarmos a Tabela 5.5 notamos que temos valores superiores a zero para T_3^* o que nos leva a testar as hipóteses em (4.6). Para tal, temos a Tabela 5.6 que é construída através do seguinte código gerado no R:

```
##### Procedimento 11 - T_3 #####
barY<-mean(Y)
S2<-1/m*sum((Y-M_Y)^2)
T_3<-1/2*(S2/barY^2-1)
T_3
T_3star<-sqrt(m)*T_3
T_3star
# Unilateral Maior
pT_3startMaior<-1-pnorm(T_3star)
pT_3startMaior
```

Tabela 5.6: Limiares e valores resultantes do procedimento 11.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_3	0.2820213	0.1633753	0.2559924	0.2874925	0.2513468
T_3^*	2.35956	1.933083	4.802842	7.628039	9.434717
$p(T_3^*)$	0.009148	0.02661	7.821e-07	1.188e-14	0

Ao analisarmos a Tabela 5.6 observamos que os valores de prova p não são superiores ao nível de significância 0.05 para qualquer que seja o valor de u em causa. Logo, para este caso, concluímos que os dados fornecem evidência suficiente para considerarmos o parâmetro de forma da função de distribuição GP maior do que zero para a modelação dos excessos populacionais em estudo.

Estatística de teste T_4

Pomos agora em prática a estatística de teste abordada na Subsecção 4.1.4 e que podemos encontrar no estudo realizado por Gomes e van Montfort em [25]. Os resultados decorrentes da aplicação desta estatística de teste estão apresentados na Tabela 5.7 e foram obtidos através do seguinte código gerado no R:

```
##### Procedimento 12 - T_4 #####
T_4<-Y_mm/Med_Y
T_4
T_4star<-T_4*log(2)-log(m)
T_4star
library(evd)
# Unilateral maior
pT_4startMaior<-1-pgumbel(T_4star)
pT_4startMaior
```

Tabela 5.7: Limiares e valores resultantes do procedimento 12.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_4	12.08568	10.95703	17.45866	20.53441	22.52686
T_4^*	4.128659	2.653195	6.237791	7.676587	8.363795
$p(T_4^*)$	0.01597547	0.0680032	0.0019523	0.0004635	0.000233

Na Tabela 5.7 observamos que os valores de prova p não são superiores ao nível de significância 0.05, exceto quando $u_{0.01} = 2.672995$. Logo, para estes casos, concluímos que os dados fornecem evidência suficiente para considerarmos o parâmetro de forma da função de distribuição GP maior do que zero para a modelação dos excessos populacionais em estudo, ao invés dos excessos populacionais que associados ao limiar $u_{0.01} = 2.672995$, onde obtemos um valor de prova p superior a 0.05. Assim, para $u_{0.01} = 2.672995$ os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo.

Estatística de teste T_5

Agora pomos em prática a estatística de teste abordada na Subsecção 4.1.5 e que podemos encontrar em [23], por exemplo. Os resultados decorrentes da aplicação desta estatística de teste estão apresentados na Tabela 5.8 e foram obtidos através do seguinte código gerado no R:

```
##### Procedimento 13 - T_5 #####
T_5<-(Y_mm-Med_Y)/(Med_Y-Y_1m)
T_5
T_5star<-T_5*log(2)-log(m/2)
T_5star

library(evd)
# Unilateral maior
pT_5startMaior<-1-pgumbel(T_5star)
pT_5startMaior
```

Quando analisamos a Tabela 5.8 observamos que os valores de prova p não são superiores ao nível de significância 0.05, exceto quando $u_{0.01} = 2.672995$. Logo, para estes casos, concluímos que os dados fornecem evidência suficiente para considerarmos o parâmetro de forma da função de distribuição GP maior

Tabela 5.8: Limiares e valores resultantes do procedimento 13.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_5	11.36914	9.995661	16.49246	19.54463	21.53057
T_5^*	4.325136	2.679969	6.261216	7.683675	8.366365
$p(T_5^*)$	0.01314459	0.0662675	0.0019071	0.0004602	0.0002325

do que zero para a modelação dos excessos populacionais em estudo ao invés dos excessos populacionais correspondentes ao limiar $u_{0.01} = 2.672995$. Assim, para $u_{0.01} = 2.672995$ temos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo.

Estadística de teste T_6

Por fim, vamos aplicar a estatística de teste abordada na Subsecção 4.1.6 e que podemos encontrar em [8]. Em primeiro lugar, vamos testar as hipóteses em (4.4). Para tal, geramos no R o seguinte código:

```
##### Procedimento 14 - T_6 #####
F_U<-Y[m-round(m/4)+1]
F_L<-Y[round(m/4)]
T_6<-(F_U-Med_Y)/(Med_Y-F_L)
T_6star<-log(3/2)*sqrt(m/2)*(T_6-log(2)/log(3/2))
# Bilateral
pT_6starBil<-(2-2*pnorm(abs(T_6star)))
pT_6starBil
```

Os resultados decorrentes deste procedimento estão apresentados na Tabela 5.9.

Tabela 5.9: Limiares e valores resultantes do procedimento 14.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_6	2.998013	1.264464	2.530439	1.844091	1.732886
T_6^*	3.09081	-1.509762	4.415856	1.023775	0.251558
$p(T_6^*)$	0.00199611	0.1311042	1.006e-05	0.3059415	0.8013827

Na Tabela 5.9 observamos que os valores de prova p são superiores ao nível de significância 0.05 para qualquer que seja o valor de u em causa, exceto quando $u_{0.005} = 3.421284$ e $u_{0.025} = 1.959207$. Assim, ao nível de significância de

0.05, temos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo.

Mas quando abordamos os excessos populacionais associados aos limiares $u_{0.005} = 3.421284$ e $u_{0.025} = 1.959207$ concluímos que os dados fornecem evidência suficiente para considerarmos não nulo o parâmetro de forma da função de distribuição GP para a modelação dos excessos populacionais em estudo.

Ao analisarmos a Tabela 5.9 notamos que temos valores superiores a zero para T_6^* o que nos leva a testar as hipóteses em (4.6). Para tal, temos a Tabela 5.10 que é construída através do seguinte código gerado no R:

```
##### Procedimento 15 - T_6 #####
F_U<-Y[m-round(m/4)+1]
F_L<-Y[round(m/4)]
T_6<-(F_U-Med_Y)/(Med_Y-F_L)
T_6star<-log(3/2)*sqrt(m/2)*(T_6-log(2)/log(3/2))
# Unilateral Maior
pT_6starMaior<-(1-pnorm(abs(T_6star)))
pT_6starMaior
```

Tabela 5.10: Limiares e valores resultantes do procedimento 15.

	0.5%	1%	2.5%	5%	10%
Limiar (u)	3.421284	2.672995	1.959207	1.499139	1.059779
T_6	2.998013	1.264464	2.530439	1.844091	1.732886
T_6^*	3.09081	-1.509762	4.415856	1.023775	0.251558
$p(T_6^*)$	0.0009981	0.0655521	5.031e-06	0.1529707	0.4006914

Já quando analisamos a Tabela 5.10, observamos que os valores de prova p são superiores ao nível de significância 0.05, exceto quando $u_{0.005} = 3.421284$ e $u_{0.025} = 1.959207$. Assim, concluímos que os dados não fornecem evidência suficiente para rejeitarmos a distribuição exponencial para os excessos populacionais em estudo. Já quando temos $u_{0.005} = 3.421284$ e $u_{0.025} = 1.959207$, concluímos que os dados fornecem evidência suficiente para considerarmos o parâmetro de forma da função de distribuição GP maior do que zero para a modelação dos excessos populacionais em estudo.

Resumo das estatísticas de teste

A decisão de rejeitar ou não a distribuição exponencial para cada uma das seis estatísticas de teste analisadas está apresentada na Tabela 5.11. Ao analisarmos a mesma, observamos que para todas as estatísticas de teste analisadas rejeitamos a distribuição exponencial para os excessos relacionados com o limiar $u_{0.025} = 1.959207$. De igual forma, podemos verificar que a função de distribuição exponencial é rejeitada quando a estatística de teste T_3 é aplicada, qualquer que seja o limiar considerado.

No caso do limiar $u_{0.005} = 3.421284$, notamos que a função de distribuição exponencial só não é rejeitada quando são realizados testes bilaterais, ou seja, quando são aplicadas as estatísticas de teste T_1 , T_{1b} e T_2 . Observamos assim para este limiar diferentes decisões para as estatísticas de teste T_2 e T_3 , ao contrário do que é observado para o limiar igual a 2.672995. Para este limiar, a função de distribuição exponencial não é rejeitada para todas as restantes estatísticas de teste. Observamos também que, para os limiares iguais a 1.499139 e 1.059779, a distribuição exponencial só não é rejeitada quando é aplicada a estatística de teste T_6 .

Relembremos que Brillhante em [8] constatou que a estatística de teste T_6 tem um pior desempenho na deteção da distribuição exponencial em comparação com as estatísticas de teste T_4 e T_5 . No entanto, observamos aqui que a função de distribuição exponencial é rejeitada para estas estatísticas de teste. Observamos também que as estatísticas de teste T_4 , T_5 e T_6 apenas indiciam decisões de rejeição diferentes para os dois valores de limiar mais baixos correspondente a uma maior quantidade de valores observados.

Tabela 5.11: Decisão de rejeitar ou não a distribuição exponencial ($\alpha = 0.05$).

	0.5%	1%	2.5%	5%	10%
Limiar u	3.421284	2.672995	1.959207	1.499139	1.059779
T_1 ($\gamma \neq 0$)	Ñ. Rej. H_0	Ñ. Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
$T_{1,b}$ ($\gamma \neq 0$)	Ñ. Rej. H_0	Ñ. Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
T_2 ($\gamma \neq 0$)	Ñ. Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
T_3 ($\gamma > 0$)	Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
T_4 ($\gamma > 0$)	Rej. H_0	Ñ. Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
T_5 ($\gamma > 0$)	Rej. H_0	Ñ. Rej. H_0	Rej. H_0	Rej. H_0	Rej. H_0
T_6 ($\gamma > 0$)	Rej. H_0	Ñ. Rej. H_0	Rej. H_0	Ñ. Rej. H_0	Ñ. Rej. H_0

5.1.2 Intervalos de confiança e métodos gráficos

No cálculo dos valores observados para as estatísticas de teste T_1 e T_{1b} foi utilizado o método de estimação ML. A obtenção da estimativa de γ por este método é parte integrante na construção dos intervalos de confiança referidos na Secção 4.2. Na biblioteca POT estão implementados os métodos MOM, PWM e ML, sendo que estes dois últimos métodos apresentam uma complexidade mais elevada do que o método MOM.

Outras bibliotecas onde podemos encontrar os métodos PWM e ML são as bibliotecas denominadas *eva*, *evir*, *fExtremes* e *QRM*. Nestas bibliotecas os estimadores obtidos pelo método PWM para os parâmetros de forma γ e escala β resultam da aplicação das equações (3.21) e (3.22), respetivamente. Por outro lado, as bibliotecas *evd*, *Renext* e *extRemes* apenas apresentam a possibilidade de aplicar o método ML na obtenção das estimativas para os parâmetros em estudo.

Nesta aplicação, optamos por utilizar a biblioteca POT porque engloba todos os métodos de estimação referidos (métodos ML, PWM e MOM). Assim sendo, para cada valor de u obtido, é aplicado o seguinte procedimento:

```
##### Procedimento 16 #####
library(POT) # Carrega a biblioteca que contém as funções aplicadas
MOM<- fitgpd(SP500, u,est = "moments") # Determina as estimativas dos
                                         # parâmetros pelo método MOM
PWM<- fitgpd(SP500, u, est = "pwmu") # Determina as estimativas dos
                                         # parâmetros pelo método PWM
ML <- fitgpd(SP500, u, est= "mle", std.err.type ="expected") # Determina
                                         # as estimativas dos parâmetros pelo método ML

gpd.fishape(MOM) # Determina o intervalo de confiança para o parâmetro
                  #  $\gamma$  pelo método MOM
gpd.fishape(PWM) # Determina o intervalo de confiança para o parâmetro
                  #  $\gamma$  pelo método PWM
gpd.fishape(ML) # Determina o intervalo de confiança para o parâmetro
                 #  $\gamma$  pelo método ML
```

Os resultados obtidos da aplicação deste procedimento estão apresentados na Tabela 5.12. Notemos que os intervalos de confiança obtidos possuem todos um grau de confiança de 95%.

Ao analisarmos a Tabela 5.12, observamos que os intervalos de confiança calculados quando consideramos 0.5% e 1% dos valores extremos contêm o valor

zero, qualquer que seja o método de estimação de parâmetros utilizado. Assim, para estes valores em particular, não temos evidência suficiente para rejeitar a distribuição exponencial.

Tabela 5.12: Estimativas para γ e intervalos de 95% confiança resultantes do procedimento 16.

	0.5%	1%	2.5%	5%	10%
Limiar u	3.421284	2.672995	1.959207	1.499139	1.059779
$\hat{\gamma}_{\text{MOM}}$	0.1849	0.1258	0.1702	0.1830	0.1675
I.C. _{MOM}	-0.26,0.63	-0.1, 0.35	-0.01,0.35	0.04,0.32	0.08,0.26
$\hat{\gamma}_{\text{PWM}}$	0.1916	0.0828	0.1898	0.2027	0.1579
I.C. _{PWM}	-0.09, 0.47	-0.11,0.27	0.06, 0.32	0.12, 0.29	0.1,0.22
$\hat{\gamma}_{\text{ML}}$	0.2016	0.1094	0.1859	0.2006	0.1626
I.C. _{ML}	-0.07,0.47	-0.06,0.28	0.06, 0.31	0.11,0.29	0.10, 0.22

Ao analisarmos a Tabela 5.12, observamos que os intervalos de confiança calculados quando consideramos 0.5% e 1% dos valores extremos contêm o valor zero, qualquer que seja o método de estimação de parâmetros utilizado. Assim, para estes valores em particular, não temos evidência suficiente para rejeitar a distribuição exponencial. Ao considerarmos 2.5% dos valores extremos, observamos que o intervalo de confiança calculado contém o valor zero apenas quando foi aplicado o método MOM na estimação do parâmetro γ . Assim, apenas neste caso, não temos evidência suficiente para rejeitar a distribuição exponencial. Por fim, observamos que os intervalos de confiança calculados quando consideramos 5% e 10% dos valores extremos não contêm o valor zero, qualquer que seja o método de estimação de parâmetros utilizado. Logo, para estes valores em particular, temos evidência suficiente para rejeitar a hipótese nula, ou seja, para considerar que os excessos em estudo seguem uma função de distribuição GP.

Temos também a função gráfica shape na biblioteca `evir` que nos auxilia no estudo dos intervalos de confiança, fornecendo uma representação gráfica para a estimativa do parâmetro de forma e os respectivos intervalos de confiança, para um dado conjunto de excedências. Na Figura 5.5 temos um exemplo dessa representação gráfica correspondente aos dados em estudo. O tracejado a representa os limites inferior e superior do intervalo de 95% de confiança, a linha contínua representa as estimativas para o parâmetro de forma, tendo em conta os diferentes limiares. A estimativa nula para o parâmetro de forma é salientada pela representação de uma reta.

O procedimento para a obtenção de Figura 5.5 é apresentado em seguida. Uma nova função denominada `shape_graf` foi criada a partir da função `shape` de modo a obtermos uma estimativa e respetivo intervalo de confiança utilizando o método ML (gráfico da esquerda) e o método PWM (gráfico da direita), sendo também realizadas ligeiras alterações no código no âmbito da tradução de textos de inglês para português.

```
##### Procedimento 17 #####
library(evir) # Carrega a biblioteca que contém a função aplicada
shape_graf(SP500,models=500,start=70,end=1409,reverse=F)
abline(0, 0, col = "green")
```

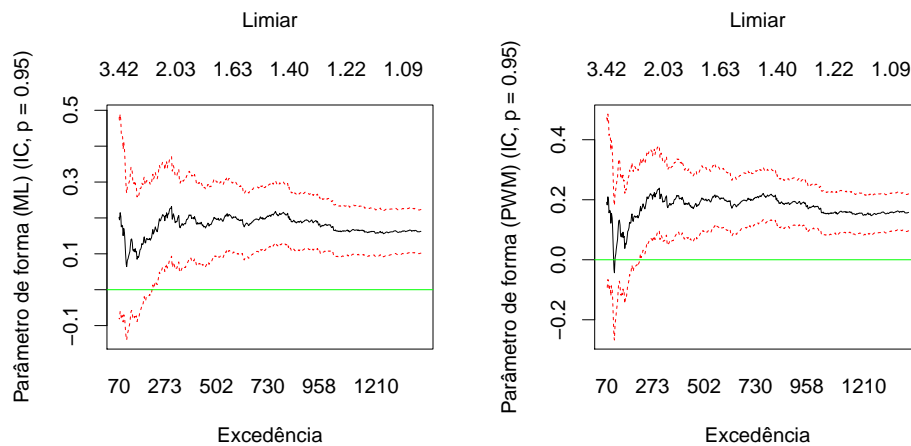


Figura 5.5: Estimativas e intervalos de confiança para γ .

Aplicamos em seguida os métodos de validação de modelos através dos métodos gráficos abordados na Secção 4.2 no caso de termos 5% e 10% dos valores extremos. Relembramos que o objetivo é percebermos efetivamente se temos a linearidade característica da boa adequação dos dados ao modelo da distribuição GP através da sua visualização. A título exemplificativo, apresentamos em seguida o procedimento a realizar no caso de termos 5% dos valores.

```
##### Procedimento 18 #####
library(fExtremes)
Porcentagem=5
nf = floor(Porcentagem/100*length(as.vector(SP500)))
threshold<-findThreshold(SP500, nf)
library(ismev)
z<-gpd.fit(SP500,threshold)
```

```

Repr <- par(mfrow = c(1, 3))
gpd.pp_alt(z$mle, z$threshold, z$data)
gpd.qq_alt(z$mle, z$threshold, z$data)
gpd.his_alt(z$mle, z$threshold, z$data)
par(Repr)

```

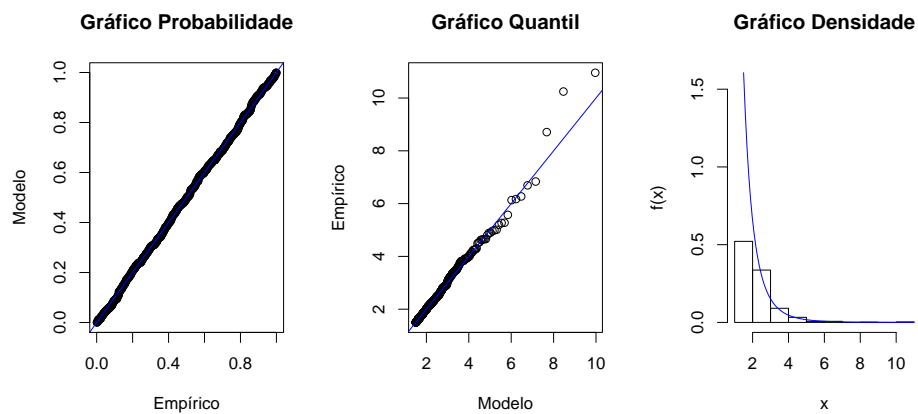


Figura 5.6: Gráficos de diagnóstico da biblioteca ismev para $u_{0.05} = 1.499139$.

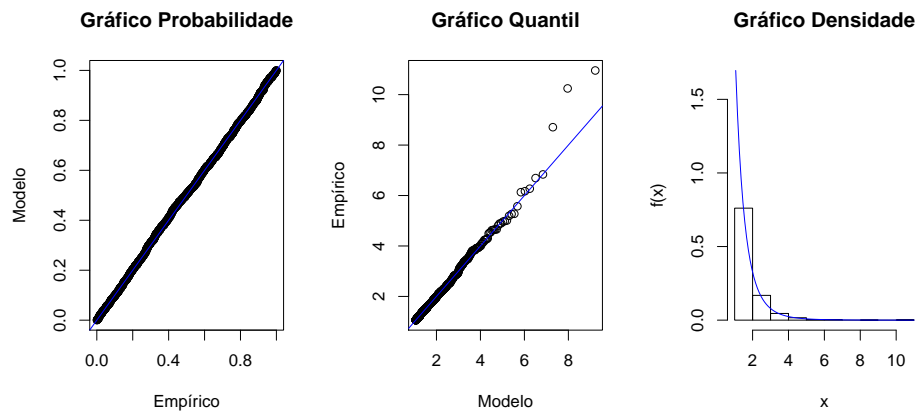


Figura 5.7: Gráficos de diagnóstico da biblioteca ismev para $u_{0.10} = 1.059779$.

Através da análise das Figuras 5.6 e 5.7, constatamos que os dados têm uma aproximação razoável à diagonal apresentada nos gráficos tanto para 5% como para 10% dos valores extremos considerados. De igual forma, podemos observar um bom ajustamento dos dados do histograma à curva no terceiro gráfico de cada um dos casos.

5.1.3 Estimação dos quantis extremos

Nesta subsecção, o pretendido é calcularmos as duas medidas de risco abordados na Secção 3.2, denotadas de VaR e CTE. Estas duas medidas são de vital importância para a gestão do risco, dado o risco de eventos extremos estar presente nas mais diversas áreas financeiras, como é o caso do mercado bolsista. Assim, a gestão do riscos tem como um dos seus maiores desafios a implementação de modelos que quantifiquem e qualifiquem o risco de eventos raros, como são o VaR e o CTE.

Para estimarmos os valores do VaR e CTE consideremos o seguinte procedimento gerado no R (particularizando no caso em que é utilizada a sub-amostra dos 2.5% maiores valores da amostra).

```
##### Procedimento 19 - VaR e CTE #####
library(fExtremes)
Percentagem=2.5
nf = floor(Percentagem/100*length(as.vector (SP500)))
u<-findThreshold(SP500, nf)

library(QRM)
z<-fit.GPD(SP500, threshold = u,information = "observed")
RiskMeasures(z,p)
gamma<-z$par.ests[1]
beta<-z$par.ests[2]
p<-c(0.01,0.001,0.0001) # Probabilidades consideradas
N_u=m
VaR_p=u+(beta/gamma)*((n*p/N_u)^(-gamma)-1)
ES_p=VaR_p/(1-gamma)+(beta-gamma*u)/(1-gamma)
```

Os valores decorrentes da aplicação do procedimento anterior para cada um dos limiares considerado estão compilados na Tabela 5.13.

Na Tabela 5.11 pudemos constatar que recusamos a hipótese nula (4.6) qualquer que seja a estatística de teste aplicada quando consideramos 2.5% dos dados extremos. Assim sendo, consideraremos apenas este caso na interpretação dos valores VaR e CTE apresentados na Tabela 5.13.

Desta forma, com uma probabilidade de 0.01, o que esperamos é que as taxas dos retornos diários obtidos sejam superiores a $VaR_{0.01}^{2.5\%} = 2.688481$, sendo o valor médio dos retornos acima de $VaR_{0.01}^{2.5\%}$ de 3.752985, que não é mais do que o valor de $CTE_{0.01}^{2.5\%}$.

Tabela 5.13: Estimação do VaR e CTE para os diferentes limiares.

	0.5%	1%	2.5%	5%	10%
Limiar u	3.421284	2.672995	1.959207	1.499139	1.059779
$\hat{\gamma}$	0.2018	0.1094	0.1859	0.2005	0.1627
$\hat{\beta}$	0.8966	0.9784	0.7311	0.6228	0.5897
$\text{VaR}_{0.01}$	2.83592	2.66624	2.68848	2.68106	2.70632
$\text{CTE}_{0.01}$	3.811214	3.764006	3.752985	3.756363	3.730387
$\hat{\gamma}$	0.2018	0.1094	0.1859	0.2005	0.1627
$\hat{\beta}$	0.8966	0.9784	0.7311	0.6228	0.5897
$\text{VaR}_{0.001}$	5.117647	5.226331	5.179065	5.196792	5.101357
$\text{CTE}_{0.001}$	6.669698	6.638571	6.812187	6.902840	6.590693
$\hat{\gamma}$	0.2018	0.1094	0.1859	0.2005	0.1627
$\hat{\beta}$	0.8966	0.9784	0.7311	0.6228	0.5897
$\text{VaR}_{0.0001}$	8.748708	8.519808	9.000026	9.188197	8.584536
$\text{CTE}_{0.0001}$	11.218585	10.336613	11.505502	11.894981	10.750525

Já quando consideramos uma probabilidade de 0.001, esperamos que as taxas dos retornos diários obtidos sejam superiores a $\text{VaR}_{0.001}^{2.5\%} = 5.179065$, sendo $\text{CTE}_{0.001}^{2.5\%} = 6.812187$. Por fim, quando consideramos uma probabilidade 0.0001, esperamos que as taxas dos retornos diários obtidos sejam superiores a $\text{VaR}_{0.0001}^{2.5\%} = 9.000026$, sendo $\text{CTE}_{0.0001}^{2.5\%} = 11.505502$.

Capítulo 6

Conclusão

Desde a sua introdução até a atualidade, a metodologia POT tem sido uma das metodologias que mais tem merecido a atenção dos estudiosos nas áreas de aplicação da EVT, dado permitir uma melhor utilização dos dados em estudo segundo a bibliografia consultada. Os valores em estudo na metodologia POT são aqueles que se encontram acima de um limiar selecionado, no entanto, essa seleção não é consensual. Apesar de não haver um consenso no meio científico quanto à escolha do melhor limiar a ser utilizado, é habitual o uso de métodos gráficos que auxiliam na escolha desse valor. Nesta dissertação foram referidos alguns desses métodos, nomeadamente o método de escolha baseado na função gráfica do excesso médio empírico, que não se revela de fácil interpretação.

No estudo da metodologia POT, foi de especial interesse estudar a estimação dos parâmetros das funções de distribuições exponencial e de distribuições GP. Para tal, foram abordados os métodos MOM, PWM e ML, os quais são considerados como os métodos de estimação mais usuais. Na aplicação destes três métodos, foi observado que o método MOM é, matematicamente, o mais acessível de entre os três métodos, de tal forma que este método é utilizado para fornecer valores iniciais em algoritmos numéricos relativos aos métodos de estimação PWM e ML numa grande variedade de bibliotecas consideradas nesta dissertação.

O objetivo do estudo aqui realizado foi sugerir métodos de escolha entre a função de distribuição exponencial e a função de distribuição GP. Além dos métodos estatísticos que assentam em testes de hipóteses foram também sugeridos métodos baseados em intervalos de confiança e em métodos gráficos. De modo a exemplificar a aplicação destes métodos estatísticos, foi realizada

uma ilustração prática no quinto capítulo. Nesta ilustração prática foi utilizado o *software* R, dada a grande variedade de bibliotecas já implementadas. Muitas destas bibliotecas, como por exemplo as bibliotecas *evir*, *evmix*, *evd*, *isnev* e *POT*, já incorporam as diversas funções usadas para a aplicação dos temas abordados ao longo desta dissertação.

A base de dados utilizada na ilustração prática realizada é formada pelos retornos diários dos dados do índice bolsista *S&P500*. Ao aplicarmos as estatísticas de teste T_1 e $T_{1,b}$ (cf. Subsecção 4.1.1) obtivemos as mesmas conclusões para qualquer limiar considerado. Para 0.5% e 1% dos valores extremos, os dados não forneceram evidência suficiente para rejeitar a hipótese nula, ou seja, não forneceram evidência suficiente para rejeitar que os excessos seguem uma distribuição exponencial. Já quanto aos restantes limiares considerados, os dados forneceram evidência suficiente para concluirmos que os excessos em estudo seguem uma distribuição GP. Conclusões análogas foram extraídas dos intervalos de confiança obtidos quando foi considerada a estimativa para o parâmetro de forma γ obtida tanto pelo método ML como pelo método PWM. De facto, foi observado que o valor zero apenas pertencia aos intervalos de confiança calculados quando foram considerados 0.5% e 1% dos valores extremos. Por outro lado, foi observado que o intervalo de confiança calculado quando foi considerada a estimativa para γ obtida pelo método MOM foi o único intervalo de confiança que continha o valor zero.

Aplicando a estatística de teste T_2 (cf. Subsecção 4.1.2), foi considerado que os dados forneceram evidência suficiente para considerar que os excessos em estudo seguem uma distribuição GP quando todos os limiares são considerados, exceto quando foi considerado o limiar correspondente a 0.5% dos valores extremos. Esta exceção não foi observada na aplicação da estatística de teste T_3 (cf. Subsecção 4.1.3), uma vez que qualquer que seja o limiar considerado, os dados forneceram sempre evidência suficiente para a rejeição do modelo exponencial para os excessos em estudo.

Por sua vez, ao aplicarmos as estatísticas de teste T_4 e T_5 (cf. Subsecções 4.1.4 e 4.1.5) foram obtidas as mesmas conclusões para quaisquer limiares utilizados, existindo apenas evidência para não rejeitar a distribuição exponencial quando foi considerado 1% dos valores extremos. Por fim, as estatísticas de teste T_4 , T_5 e T_6 apenas indicaram decisões de rejeição diferentes para os dois valores de limiar mais baixos correspondentes a uma maior quantidade de valores observados.

De modo a finalizar a nossa ilustração prática calculámos as estimativas para quantis extremos. Ao analisar os 2.5% valores extremos estimamos que, para probabilidades no máximo iguais a 1% as taxas dos retornos diários obtidos no instante $t + 1$ seriam, em média, no mínimo iguais a 3.75%. Assim, por cada 1.000.000,00€ é de esperar que no instante $t + 1$ os investidores obtenham um retorno superior a 37.500,00€.

Bibliografia

- [1] Abreu, A. e Reis, D. Análise de sobrevivência e valores extremos em R. *Congresso da Sociedade Portuguesa de Estatística* (2015), 1–14.
- [2] Allen, E., Singh, A. e Powell, R. Extreme market risk-an extreme value theory approach. *Edith Cowan University* (2011), 1–26
- [3] Balkema, A. e de Haan, L. Residual life time at great age. *The Annals of Probability* (1974), 792–804.
- [4] Beirlant, J., Goegebeur, Y., Teugels, J. e Segers J. *Statistics of Extremes: Theory and Applications*, John Wiley & Sons Ltd, (2004).
- [5] Beirlant, J., Teugels, J. e Vynckier, P. *Practical Analysis of Extreme Values*, Leuven University Press, (1996).
- [6] Bermudez, P. e Kotz, S. Parameter estimation of the generalized Pareto distribution - Parte I. *Journal of Statistical Planning and Inference* (2008), 1353–1373.
- [7] Bermudez, P. e Kotz, S. Parameter estimation of the generalized Pareto distribution - Parte II. *Journal of Statistical Planning and Inference* (2008), 1374–1388.
- [8] Brillhante, M. Exponentiality versus generalized Pareto - a resistant and robust test. *REVSTAT - Statistical Journal*, Vol. 2(1), (2004), 1–13.
- [9] Castillo, E. e Hadi, A. Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, Vol. 92(440) (1997), 1609–1620.

- [10] Castillo, E., Hadi, A., Balakrishnan, N. e Sarabia, J. *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley Hoboken, NJ, (2005).
- [11] Chaouche, A. e Bacro, J. Statistical inference for the generalized Pareto distribution: Maximum likelihood revisited. *Communications in Statistics - Theory and Methods*, Vol 35, (2006), 785–802.
- [12] Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer, (2001).
- [13] Curto, J. e Rios, A. Teoria de valores extremos: aplicação aos mercados de capitais. *Actas do XV Congresso Anual da SPE*, (2007), 175–184.
- [14] Davison, A. Modelling excesses over high thresholds, with an application. *Statistical Extremes and Application*, (1984), 461–482.
- [15] Davison, A. e Smith, R. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B*, Vol. 52(3), (1990), 393–442.
- [16] Embrechts, P., Klüppelberg, C. e Mikosch, T. *Modelling extremal events for Insurance and Finance*. Springer-Verlag, Vol. 33(1) (1997).
- [17] Ferreira, A. e de Haan, L. On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, Vol. 43(1), (2015), 276–298.
- [18] Fisher, A. e Tippett, L. Limiting forms of the frequency distributions of the largest or smallest member in a sample. *Proceedings Cambridge Philosophical Society*, 24, (1928), 180–190.
- [19] Fisher, M. Sur le loi de probabilité de l'écart maximum. *Ann. Société Polonaise de Mathématique*, 6, (1927), 93–116.
- [20] Gilleland, E., Ribatet, M. e Stephenson, A. A software review for extreme value analysis. *Extremes*, Vol. 26, (2013), 103–119
- [21] Gilli, M. e Këllezli, E. An application of extreme value theory for measuring financial risk. *Computational Economics*, Vol. 27(2-3) (2006), 207–228.

- [22] Gnedenko, J. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, (44), (1943), 423–453.
- [23] Gomes, M. I. A note on statistical choice of extremal models. *Actas de las IX Jornadas Matemáticas Hispano-Lusas: Salamanca 12-16 abril 1982*, Vol. 1 (1982), 653–655.
- [24] Gomes, M. I., Alves, M. e Neves, C. *Análise de Valores Extremos: Uma Introdução*. Sociedade Portuguesa de Estatística, (2013).
- [25] Gomes, M. I. e van Montfort, J. Exponentiality versus generalized Pareto-quick tests. *Third International Conference on Statistical Climatology Vienna*, (1986), 185–195.
- [26] Greenwood, J., Landwehr, J., Matalas, N. e Wallis, J. Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, Vol. 15(5) (1979).
- [27] Hosking, J. e Wallis, J. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, Vol. 29(3), (1987), 339–349.
- [28] Jenkinson, A. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, Vol. 81(348), (1955), 158–171.
- [29] Kotz, S. e Nadarajah, S. *Extreme Value Distributions - Theory and Applications*. Imperial College Press, 2002.
- [30] Leadbetter, M., Lindgren, G. e Rootzen, H. Extremes and Related Properties of Random Sequences and Series. *Springer Verlag, New York*, (1) (1983).
- [31] Marohn, F. Testing extreme value models. *Extremes*, Vol. 3(4), (2000), 363–384.
- [32] McNeil, A. Extreme value theory for risk managers. *Departement Mathematik ETH Zentrum*, (1999).
- [33] McNeil, A. e Frey, R. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, (2000), 271–300.

- [34] Mendes, B. *Introdução à Análise de Eventos Extremos*. E-papers Serviços Editoriais, Ltda, Vol. 1, (2004).
- [35] Murteira, B., Ribeiro, C., Silva, J., Pimenta, F. e Pimenta, C. *Introdução à Estatística*. Escolar Editora, 2015.
- [36] Pickands, J. Statistical inference using extreme order statistic. *The Annals of Statistics* (1975), 119–131.
- [37] Reiss, R. e Thomas, M. Statistical Analysis of Extreme Values - Applications to Insurance, Finance, Hydrology and Other Fields. *Springer*, Vol. 2, (2007).
- [38] Rootzén, H. e Tajvidi, N. Extreme value statistics and wind storm losses: a case study. *Scandinavian Actuarial Journal*, (1997), 70–94.
- [39] Santos, P. Testes não-paramétricos para validação de modelos extremos. *Dissertação na Universidade de Lisboa Faculdade de Ciências*, (2011).
- [40] Smith, R. Threshold methods for sample extremes. *Statistical Extremes and Application*, (1984), 621–638.
- [41] Stephenson, A. EVD: Extreme value distributions. *R News* , Vol. 2(2) (2002).
- [42] Todorovic, P. e Rousselle, J. Some problems of flood analysis. *Water Resources Research* (1971), 1641–1648.
- [43] Todorovic, P. e Zelenhasic, E. A stochastic model for flood analysis. *Water Resources Research* (1970), 1144–1150.
- [44] van Montfort, J. e Witter, J. Testing exponentiality against generalised Pareto distribution. *Journal of Hydrology* 78 (1985), 305–315.
- [45] Velosa, S. e Pestana, D. *Introdução à Probabilidade e à Estatística*. Fundação Calouste Gulbenkian, Vol. 1 (2008).
- [46] Vicente, S. Extreme value theory: an application to sports. *Dissertação na Universidade de Lisboa*, (2012).

- [47] von Mises, R. La distribution de la plus grande de n valeurs. *Revue Math. Union Interbalcanique* (1936), 141–160. Reprinted in Selected Papers of Richard von Mises Amer. Math. Soc. (1954), 271-294.

Índice Remissivo

- CTE ou valor esperado de cauda
 - condicional, 3
 - aplicação, 64
 - definição, 28
 - estimação, 30
- Estimação de parâmetros
 - ML, 26, 60
 - MOM, 21, 60
 - PWM, 23, 60
- Função
 - de excesso médio, 18
 - de excesso mediana, 19
 - profile log-likelihood, 34
- Função de distribuição
 - Beta simétrica, 13
 - Exponencial, 13
 - Fréchet, 8
 - GEV, 8
 - GP, 14
 - Gumbel, 8
 - Pareto, 13
 - Weibull, 8
- Função densidade de probabilidade
 - Fréchet, 8
 - GEV, 9
 - GP, 14
 - Gumbel, 8
- Weibull, 8
- Intervalo de confiança
 - aplicação, 60
 - estimação, 40
- Métodos gráficos
 - aplicação, 62
 - estimação, 41
 - gráfico P-P, 42
 - gráfico Q-Q, 42
- Metodologia
 - Gumbel, 5
 - POT, 10, 17
- Taxas dos retornos diários, 46
- Teoria dos Valores Extremos, 5
- Teste de hipóteses
 - Estatística de Teste T_1 , 33, 52
 - Estatística de Teste T_2 , 35, 53
 - Estatística de Teste T_3 , 36, 54
 - Estatística de Teste T_4 , 37, 55
 - Estatística de Teste T_5 , 38, 56
 - Estatística de Teste T_6 , 38, 57
- Value-at-Risk, 3
 - aplicação, 64
 - definição, 28
 - estimação, 30